

TFM – MII: Emprendimiento (Servicio de análisis de datos en recién nacidos para detección de patrones y desarrollo de tratamientos preventivos)

Autor: Álvaro Torán Jiménez – 201601866@alu.comillas.edu

Director: Ana Laguna Pradas

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título

Servicio de análisis de datos en recién nacidos para detección de patrones y desarrollo de tratamientos preventivos

en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el

curso académico 2023/24 es de mi autoría, original e inédito y

no ha sido presentado con anterioridad a otros efectos. El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido tomada

de otros documentos está debidamente referenciada.

Fdo.: Álvaro Torán Jiménez Fecha: 08/ 07/ 2024

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO



Fdo.: Ana Laguna Pradas Fecha: 10/07/2024

AUTORIZACIÓN PARA LA DIGITALIZACIÓN, DEPÓSITO Y DIVULGACIÓN EN RED DE PROYECTOS FIN DE GRADO, FIN DE MÁSTER, TESINAS O MEMORIAS DE BACHILLERATO

1º. Declaración de la autoría y acreditación de la misma.

El autor D. ALVARO TORAN JIMENEZ

DECLARA ser el titular de los derechos de propiedad intelectual de la obra: *Servicio de análisis de datos de lactancia en recién nacidos para detección de patrones y desarrollo de tratamientos preventivos*, que ésta es una obra original, y que ostenta la condición de autor en el sentido que otorga la Ley de Propiedad Intelectual.

2º. Objeto y fines de la cesión.

Con el fin de dar la máxima difusión a la obra citada a través del Repositorio institucional de la Universidad, el autor **CEDE** a la Universidad Pontificia Comillas, de forma gratuita y no exclusiva, por el máximo plazo legal y con ámbito universal, los derechos de digitalización, de archivo, de reproducción, de distribución y de comunicación pública, incluido el derecho de puesta a disposición electrónica, tal y como se describen en la Ley de Propiedad Intelectual. El derecho de transformación se cede a los únicos efectos de lo dispuesto en la letra a) del apartado siguiente.

3º. Condiciones de la cesión y acceso

Sin perjuicio de la titularidad de la obra, que sigue correspondiendo a su autor, la cesión de derechos contemplada en esta licencia habilita para:

a) Transformarla con el fin de adaptarla a cualquier tecnología que permita incorporarla a internet y hacerla accesible; incorporar metadatos para realizar el registro de la obra e incorporar "marcas de agua" o cualquier otro sistema de seguridad o de protección.
b) Reproducirla en un soporte digital para su incorporación a una base de datos electrónica, incluyendo el derecho de reproducir y almacenar la obra en servidores, a los efectos de garantizar su seguridad, conservación y preservar el formato.
c) Comunicarla, por defecto, a través de un archivo institucional abierto, accesible de modo libre y gratuito a través de internet.
d) Cualquier otra forma de acceso (restringido, embargado, cerrado) deberá solicitarse expresamente y obedecer a causas justificadas.
e) Asignar por defecto a estos trabajos una licencia Creative Commons.
f) Asignar por defecto a estos trabajos un HANDLE (URL *persistente)*.

4º. Derechos del autor.

El autor, en tanto que titular de una obra tiene derecho a:

a) Que la Universidad identifique claramente su nombre como autor de la misma
b) Comunicar y dar publicidad a la obra en la versión que ceda y en otras posteriores a través de cualquier medio.
c) Solicitar la retirada de la obra del repositorio por causa justificada.
d) Recibir notificación fehaciente de cualquier reclamación que puedan formular terceras personas en relación con la obra y, en particular, de reclamaciones relativas a los derechos de propiedad intelectual sobre ella.

5º. Deberes del autor.

- El autor se compromete a:

a) Garantizar que el compromiso que adquiere mediante el presente escrito no infringe ningún derecho de terceros, ya sean de propiedad industrial, intelectual o cualquier otro.
b) Garantizar que el contenido de las obras no atenta contra los derechos al honor, a la intimidad y a la imagen de terceros.
c) Asumir toda reclamación o responsabilidad, incluyendo las indemnizaciones por daños, que pudieran ejercitarse contra la Universidad por terceros que vieran infringidos sus derechos e intereses a causa de la cesión.
d) Asumir la responsabilidad en el caso de que las instituciones fueran condenadas por infracción de derechos derivada de las obras objeto de la cesión.

6º. Fines y funcionamiento del Repositorio Institucional.

La obra se pondrá a disposición de los usuarios para que hagan de ella un uso justo y respetuoso con los derechos del autor, según lo permitido por la legislación aplicable, y con fines de estudio, investigación, o cualquier otro fin lícito. Con dicha finalidad, la Universidad asume los siguientes deberes y se reserva las siguientes facultades:

- La Universidad informará a los usuarios del archivo sobre los usos permitidos, y no garantiza ni asume responsabilidad alguna por otras formas en que los usuarios hagan un uso posterior de las obras no conforme con la legislación vigente. El uso posterior, más allá de la copia privada, requerirá que se cite la fuente y se reconozca la autoría, que no se obtenga beneficio comercial, y que no se realicen obras derivadas.
- La Universidad no revisará el contenido de las obras, que en todo caso permanecerá bajo la responsabilidad exclusive del autor y no estará obligada a ejercitar acciones legales en nombre del autor en el supuesto de infracciones a derechos de propiedad intelectual derivados del depósito y archivo de las obras. El autor renuncia a cualquier reclamación frente a la Universidad por las formas no ajustadas a la legislación vigente en que los usuarios hagan uso de las obras.
- La Universidad adoptará las medidas necesarias para la preservación de la obra en un futuro.
- La Universidad se reserva la facultad de retirar la obra, previa notificación al autor, en supuestos suficientemente justificados, o en caso de reclamaciones de terceros.

Madrid, a 8 de JULIO de 2024

ACEPTA

Fdo……

Motivos para solicitar el acceso restringido, cerrado o embargado del trabajo en el Repositorio Institucional:

Contenido

Alcance del Proyecto ***14***

MÓDULO I – FACTIBILIDAD COMERCIAL ***16***

El Problema que se busca resolver ***17***

Stakeholders 17

Nivel 1 ..... *18*

Nivel 2 ..... *19*

Nivel 3 ..... *19*

Problemas 20

La Idea 21

Cambios *21*

Business Model Canvas 23

Segmentos de Clientes (Customer Segments) *24*

Propuesta de Valor (Value Propositions) *25*

Canales (Channels) *25*

Relaciones con Clientes (Customer Relationships) *25*

Fuentes de Ingresos (Revenue Streams) *26*

Recursos Clave (Key Resources) *26*

Actividades Clave (Key Activities) *26*

Socios Clave (Key Partnerships) *27*

Estructura de Costes (Cost Structure) *27*

Presentación de la Empresa ***28***

Misión 28

Visión 28

Valores 29

Macroentorno ***30***

PESTEL 32

FACTORES POLÍTICOS: *33*

FACTORES ECONOMICOS: *34*

FACTORES SOCIALES: *34*

FACTORES TECNOLÓGICOS: *35*

FACTORES ECOLÓGICOS *35*
FACTORES LEGALES *36*

Microentorno ***37***
5 Fuerzas de Porter 37
AMENAZA DE ENTRADA DE NUEVOS COMPETIDORES: *37*
RIVALIDAD ENTRE COMPETIDORES EXISTENTES: *38*
AMENAZA DE PRODUCTOS Y SERVICIOS SUSTITUTOS: *41*
PODER DE NEGOCIACIÓN DEL CLIENTE: *42*
PODER DE NEGOCIACIÓN DEL PROVEEDOR: *43*

Presentación del Servicio ***44***
Descripción del Servicio 44
Análisis DAFO 47

Dimensionamiento del Mercado ***48***
Mercado Total (TAM) 48
Mercado Disponible (SAM) 49
Mercado Objetivo (SOM) 50

Costes Asociados directos ***53***
Datos *53*
Tecnología y software *53*

Proveedores 55

Precio y Política ***58***
Precio 58
Política 59

Cálculo del P*Q ***60***

Plan de Ventas: Evolución en 5 Años ***61***

MÓDULO II – FACTIBILIDAD TÉCNICA ***62***

Introducción ***63***

Servicio 64

Descripción y proceso 64

Producto Mínimo Viable (proyecto Lactapp) 68

Inicio del proyecto 68

Recopilación Y procesamiento de datos 73

Recopilación y entendimiento 73

Verificación de los datos, análisis preliminar 75

Estructuración, formateo de los datos y EDA 77

Análisis Externo 87

Conclusiones del análisis 92

Planificación del servicio (proyectos) 94

Personal 96

Localización 97

Detección de Riesgos 101

Detección de Riesgos 101

Matriz de Riesgo 102

RSE 108

Objetivos de Desarrollo Sostenible 108

Otras Actividades 110

MÓDULO III – FACTIBILIDAD ADMINISTRATIVA LEGAL 111

Determinación de la Forma Jurídica 112

Cobertura de Responsabilidades 112

Registrar la Empresa 112

Reglamento y protección de datos 114

Reglas para el tratamiento de datos 114

GDPR (Reglamento General de Protección de Datos) 114

LOPD (Ley Orgánica de Protección de Datos y Garantía de los Derechos Digitales) 115

ISO 27001 115

Cloud Control Matrix (CCM) *116*
AI Act *117*

Organigrama. Cantidad de Personal ***118***

Perfiles de Puesto ***121***

Salarios de Cada Posición ***124***

Sistema Administrativo de Gestión (Indicadores, KPIs) ***126***

MÓDULO IV – FACTIBILIDAD ECONÓMICA FINANCIERA ***128***

Recursos 130

Cuenta de Pérdidas y Ganancias ***132***

Fondos Propios 2
Deuda *3*
KPIs DEL INVERSOR *2*

CONCLUSIONES ***5***

Conclusión Sobre el Proyecto ***6***

ANEXOS 8

Anexo A – Encuesta Sobre las tendencias en los cuidados infantiles y uso de tecnología 9

Anexo B – Documentación del archivo de datos de Lactapp ***17***

Anexo C - código del análisis ***3***

Anexo D - Análisis de datos externo (Natalia Gorrín, Cristina Salazar, Beatriz Martín) 70

Análisis de Datos ***71***

1. PREPROCESADO DE LOS DATOS ***74***

2. ANÁLISIS DE RESULTADOS ***111***

3. CONCLUSIONES Y PRÓXIMOS PASOS.....1

REFERENCIAS.....3

TABLAS 4

ILUSTRACIONES.....5

BIBLIOGRAFÍA.....6

ALCANCE DEL PROYECTO

Este Trabajo de Fin de Máster pretende llevar a cabo el desarrollo de un negocio de servicios de análisis de datos (y empleo de modelado de *ML o ML*), que permitan la detección preventiva de problemas relacionados con la salud, a empresas en el sector sanitario relacionadas con el cuidado infantil.

Para ello será importante realizar un análisis del mercado y el entorno en el que se sitúa esta empresa, para detectar desde los clientes potenciales, hasta los canales de venta, competidores y recursos necesarios.

Además, se realizará un análisis de la viabilidad técnica, siendo esta el servicio de análisis de datos, se estudiará la solución y metodología, así como el resto de los activos principales necesarios para el mismo. Dado que, para el análisis de datos, se necesita una muestra significativa de datos para poder tomar los resultados obtenidos como fiables, se justifica la necesidad de realizar *partnerships* con las empresas a las que se les ofrezca el servicio, como en el caso del MVP, con Lactapp. Lactapp Medical es la app que acompaña a los profesionales que atienden a madres lactantes. Ofrece soluciones autónomas y resuelve dudas de lactancia y maternidad de manera personalizada (Lactapp, s.f.). Ofrecen formación, información, canales de consulta y *trackers* para el monitoreo de lactantes a profesionales de la salud en todo el mundo. Los datos que han sido objeto de análisis en este proyecto provienen de Lactapp, la viabilidad del proyecto como creación de empresa de consultoría se basa en la presunción de que no será la única empresa que requiere de este tipo de análisis y que existen otras empresas predispuestas a contratar este servicio, tanto en el entorno público como en el privado.

El servicio consiste, por tanto, en un análisis extensivo y detallado de los datos que aporte el cliente (Lactapp en este caso). Lo que se busca con esto es poder ofrecer una solución ad-hoc, que se ajuste completamente a las necesidades del cliente apoyándose en el uso de la tecnología. De esta forma, si lo que el cliente requiere va más allá de los *insights* y recomendaciones que se pueden aportar desde el punto de vista del análisis, podemos apoyarnos en el uso de nuevas tecnologías como el ML, para ofrecer una solución que emplee modelos algorítmicos para analizar patrones en datos y predecir resultados.

Una vez estudiada la viabilidad técnica, se estudiará la viabilidad administrativa. La formación de la empresa, su personal, el tratamiento de los datos y su legislación y los marcadores internos que se estudiarán durante su desarrollo.

Por último, empleando todos los hallazgos anteriormente expuestos, se propone el plan de negocio a 5 años, estudiando un caso hipotético que se ajustaría más a la realidad de un proyecto de estas características.

MÓDULO I – FACTIBILIDAD COMERCIAL

EL PROBLEMA QUE SE BUSCA RESOLVER

La idea de este servicio, en un primer lugar, surge del entorno familiar. Dos familiares cercanos tienen recién nacidos en fechas muy cercanas, y con el nacimiento de los bebés surge un problema inicial.

Dentro de la rutina de los neonatos, las tomas en los primeros meses eran susceptibles de mucha incertidumbre: ¿Cada cuánto deben ser las tomas?, ¿Cuánto han tomado en la toma?, ¿Les ha quedado hambre en la toma?, etc.

Con estas dudas en mente, a la hora de atender al recién nacido cuando lloraba después o antes de una toma, era difícil saber a qué se debía el llanto. Podía ser por hambre o por otro problema relacionado con la salud del neonato.

Es por esto por lo que se comienza a investigar si hay una posible oportunidad de negocio.

STAKEHOLDERS

En primer lugar se realiza un análisis de los *stakeholders* presentes en el cuidado de los recién nacidos, clasificando cada *stakeholder* por nivel, siendo los niveles:

- **Nivel 1**: Involucrados directamente y de manera constante en el cuidado y decisiones sobre la salud del infante.
- **Nivel 2**: Profesionales y familiares que interactúan frecuentemente con el infante y sus padres, utilizando los datos y análisis para tomar decisiones diarias sobre el cuidado y la salud.
- **Nivel 3**: Instituciones y empresas que apoyan, regulan y complementan los servicios de salud, asegurando que se cumplan los estándares y que los recursos necesarios estén disponibles.

Una vez definidos los niveles situamos a los *stakeholders* obtenemos el siguiente mapa

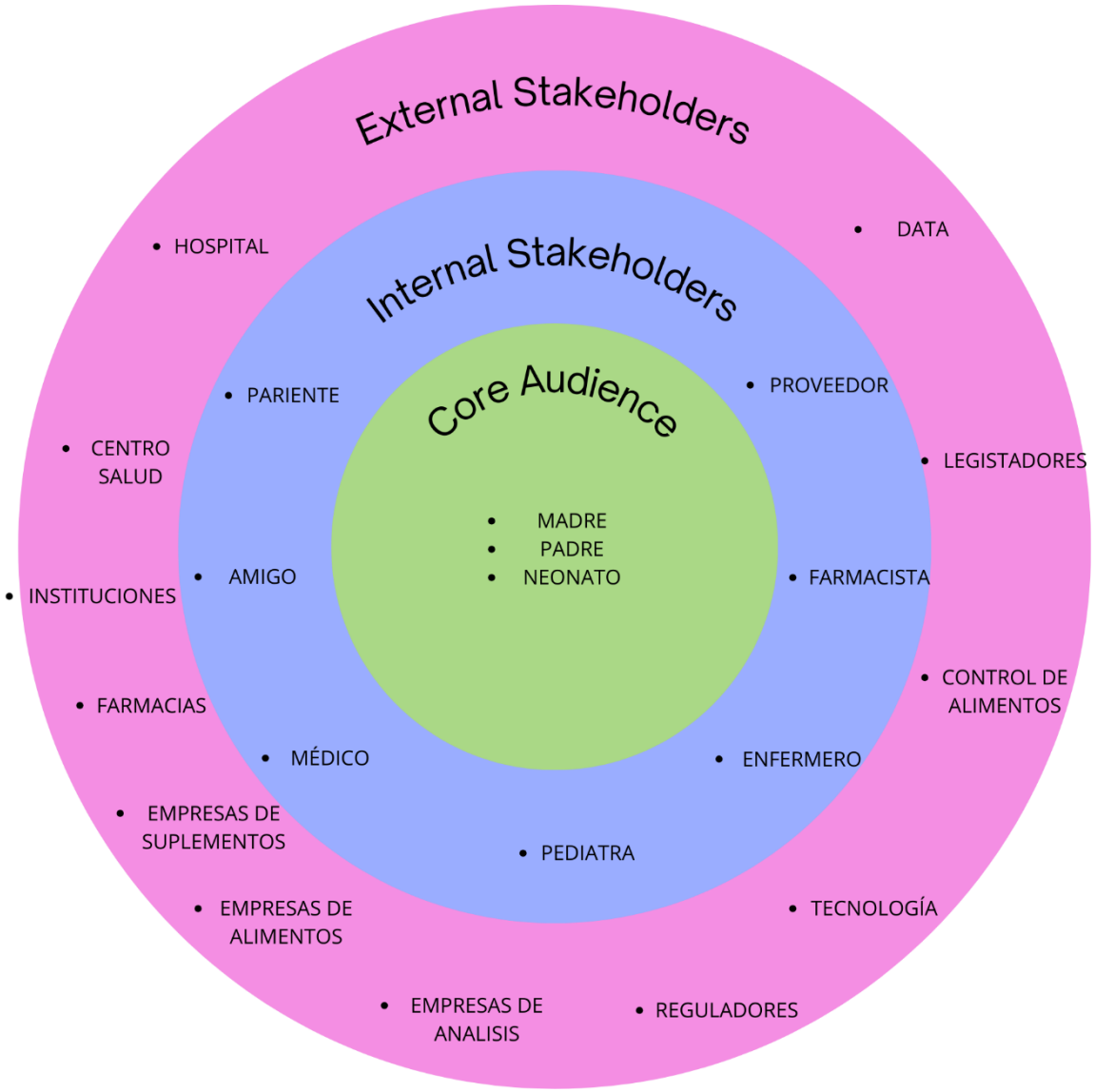


Ilustración 1:mapa de stakeholders. Fuente: propia

De acuerdo con el mapa, encontramos los siguientes *stakeholders* por nivel:

NIVEL 1

- **Madre:** La madre es un *stakeholder* esencial debido a su papel principal en el embarazo, el parto y el cuidado bebé, especialmente en los primeros meses.
- **Padre:** El padre también juega un papel fundamental en el cuidado del recién nacido, apoyando tanto a la madre como al bebé, y participando activamente en las decisiones de salud y bienestar.

- **Recién Nacido:** El bebé es el foco central del cuidado neonatal, siendo el receptor directo de todos los servicios y atenciones proporcionados por los demás *stakeholders*.

NIVEL 2

- **Parientes:** Abuelos, tíos y otros familiares cercanos suelen estar involucrados en el apoyo emocional y práctico a la madre y al recién nacido.
- **Pediatra:** El pediatra es el principal profesional de la salud que monitorea y trata al recién nacido, asegurando su desarrollo y bienestar.
- **Médico:** Otros médicos, como obstetras y especialistas en neonatología, juegan roles cruciales en el manejo de la salud de la madre y el bebé durante y después del parto.
- **Enfermero/a:** Las enfermeras y enfermeros son vitales en el cuidado diario del recién nacido, proporcionando atención continua y educando a los padres sobre el cuidado del bebé.
- **Farmaceuta:** Los farmacéuticos aseguran la disponibilidad de medicamentos y suplementos necesarios para la salud del recién nacido.

NIVEL 3

- **Hospital:** Los hospitales proporcionan los servicios médicos esenciales para el parto y el cuidado neonatal, incluyendo cuidados intensivos cuando es necesario.
- **Centro de Salud:** Proveen servicios continuos de salud y seguimiento para el recién nacido.
- **Instituciones:** Instituciones educativas y de investigación que contribuyen al conocimiento y desarrollo de mejores prácticas en el cuidado neonatal.
- **Farmacias:** Proveen acceso a medicamentos y productos de cuidado para el recién nacido y la madre.
- **Empresas de suplementos:** Desarrollan y suministran suplementos nutricionales para apoyar la salud del bebé.
- **Empresas de Comida de Bebé:** Producen alimentos diseñados específicamente para satisfacer las necesidades nutricionales de los recién nacidos.

- **Empresas de Análisis:** Contribuyen con análisis de datos e investigaciones que mejoran las prácticas de cuidado neonatal.
- **Reguladores:** Organismos que establecen y vigilan el cumplimiento de las normas y estándares de salud y seguridad.
- **Tecnología:** Proveedores de equipos y software que facilitan el cuidado y monitoreo de la salud neonatal.
- **Control de Alimentos:** Autoridades que garantizan la calidad y seguridad de los alimentos para bebés.
- **Legislación/Legisladores:** Desarrollan políticas y leyes que afectan el cuidado y los derechos de las madres y los recién nacidos.

PROBLEMAS

Una vez analizados los *stakeholders*, nos centramos en el análisis de una madre como *stakeholder* principal. Existen herramientas para hallar posibles problemas en la vida y en la forma que socializa un *stakeholder* como el *empathy map*, pero en este caso, dado que ya se ha encontrado posibles *pain-points* en el cuidado de los recién nacidos, carece de sentido emplear dichas herramientas para encontrar problemas adicionales. De esta forma, haciendo referencia a las incertidumbres mencionadas con anterioridad, buscamos dar solución a problemas como:

¿Cómo podría ayudar a una madre/padre a tener certidumbre con el estado de su bebé?

Como se ha mencionado antes, cuando el bebé llora, ¿existe alguna forma de saber a qué se debe el llanto de forma sencilla y sin acudir a algún profesional?

¿Cómo podría saber cuál es el estado de salud de un bebé?

¿Existen herramientas o información veraz que indique de forma sencilla el estado de salud de un bebé?, y al mismo tiempo, ¿hay forma de prevenir problemas de salud que puedan surgir?

¿Cómo podría saber la cantidad de leche que toma un bebé por toma?

Encontrando la solución a este problema se podría al menos descartar que la causa del llanto de los recién nacidos fuera por hambre, si es que la nutrición es correcta.

Centrándonos en el problema: **¿Cómo podría ayudar a una madre a tener certidumbre con el estado de su bebé?**, se busca emplear las afinidades del autor, así como las tendencias tecnológicas para encontrar una solución viable para el problema. De este modo, se busca

emplear la tecnología, el uso de la ingeniería para la resolución de problemas, el estudio, análisis y la visualización de datos para extraer información y la búsqueda de soluciones innovadoras, en conjunto con tendencias como el data análisis, la IA, el ML, la tecnología *wearable*, la programación y otras tendencias para buscar una solución compatible.

LA IDEA

Con los nuevos y continuos avances tecnológicos y médicos, así como con el fácil acceso a la información, la concienciación y el comportamiento de los padres primerizos se han visto muy influenciados y por ello se ven nuevas tendencias en el monitoreo de los recién nacidos.
La disponibilidad de información más precisa y accesible ha educado a los padres en la importancia de la nutrición en las primeras etapas de la vida. Aplicaciones móviles, plataformas online y herramientas de monitoreo proporcionan recursos educativos que informan de cómo ciertas prácticas en la alimentación de sus hijos pueden afectar a su desarrollo en el futuro. Además, a medida que más padres adoptan este tipo de prácticas basadas en información con base tecnológica y evidencia científica, los paradigmas sociales y culturales alrededor del cuidado de los recién nacidos se van modificando, llevando a una mayor valoración en las prácticas del cuidado informado y consciente de una comunidad más amplia. (Álvarez, 2021)
Teniendo esto en cuenta, la idea original consistía en diseñar e implementar un dispositivo de tipo *wearable*, que fuera poco invasivo y fácil de usar, que se asemejara a una pezonera, ya que son accesorios que ya se emplean en el cuidado de los recién nacidos y ayudaría a la adopción de la tecnología, y que permitiera tomar medidas relacionadas con las tomas de los recién nacidos, con algún elemento electrónico como caudalímetro u otro. Estas podían ser de tiempo, de caudal y/o de contenido, proporcionando información directa y relevante en cuanto a la nutrición del recién nacido. Además, con la monitorización de las tomas, se podrían identificar patrones en los hábitos alimenticios, así como en el contenido nutricional de la leche, que influenciaran la manera en la que el recién nacido se desarrollara.

CAMBIOS

Una vez planteada la idea, quedaba someterla a estrés, para identificar posibles cambios que ayudaran al desarrollo e implementación de esta. Planteando modificaciones como

sustituciones, combinaciones, adaptaciones, eliminaciones y otras, de los distintos elementos de la idea, se encuentran puntos de mejora.

Además, analizando en detalle la electrónica necesaria para desarrollar dicho producto, encontramos barreras significativas, como el tamaño o el coste. Es por esto que se evalúa y se decide:

Sustituir la electrónica de medición: en lugar de emplear un caudalímetro para la medición de caudal y de tiempo de flujo de leche, con el objetivo de hacer que el dispositivo sea menos intrusivo, y teniendo en cuenta otras tecnologías ya empleadas en dispositivos de tipo *wearable,* se decide emplear un fotodiodo. Este es el método que se emplea para calcular el pulso en relojes Smart entre otros. Empleando fotodiodos se podría averiguar la cantidad de líquido que fluye. (myAnalog, 2019)

Sin embargo, el problema de coste y tamaño persiste, además el desarrollo de una idea como esta, seria de una elevada complejidad y excedería las limitaciones que presenta la elaboración de un trabajo como este. Por ello se decide adaptar la idea. En lugar de hacer un dispositivo para medición, se podría desarrollar un software para la monitorización. Un software que emplee tecnologías como ML e IA, para detectar patrones en datos y elaborar análisis predictivos en las tomas de los neonatos. De tal modo que con datos como la duración de las tomas, la periodicidad de las tomas, el estado de salud de la madre y otros, se pudiera extraer información relevante y precisa. (Tang, 2024)

Esta solución propone un nuevo obstáculo, ya que para que las conclusiones extraídas de los datos y mediciones sean fiables, sería necesaria una muestra de datos muy elevada, lo que de nuevo excedería las limitaciones asociadas a un trabajo de este tipo.

Por ello, la idea final propuesta es: **ofrecer un servicio de análisis de datos de lactancia en recién nacidos para detección de patrones y desarrollo de tratamientos preventivos.**

De este modo, el peso de adquisición de datos recae sobre la empresa a la que se le realiza el servicio.

Cabe mencionar, que, a lo largo de la instigación de problemas y propuesta de soluciones, se pudo poner en contacto con una empresa que estaba dispuesta a proporcionar sus datos para realizar dicho análisis. Esta empresa es Lactapp Medical, Lactapp Medical es la app que acompaña a los profesionales que atienden a madres lactantes. Ofrece soluciones autónomas y

resuelve dudas de lactancia y maternidad de manera personalizada (Lactapp, s.f.). Ofrecen formación, información, canales de consulta y *trackers* para el monitoreo de lactantes a profesionales de la salud en todo el mundo. Estos datos son los que se emplearán para crear el modelo de análisis y predicción del servicio.

BUSINESS MODEL CANVAS

Una vez elegida la solución que se propone como servicio, se analiza y se realiza el Business Model Canvas (BMC) como primer acercamiento al modelo de negocio e inicio del proceso Lean Startup. (JACKSON, 2023) El Business Model Canvas permite a los emprendedores visualizar todos los aspectos cruciales de su modelo de negocio en un solo documento. Esto ayuda a mantener el enfoque en lo esencial y evita distracciones con detalles innecesarios.
En este caso, los datos son aportados por Lactapp en este servicio, pero se asume que se dará servicio a más empresas o entidades con suficientes datos para extraer conclusiones relevantes. Con esto en mente el BMC resultante se presenta a continuación:

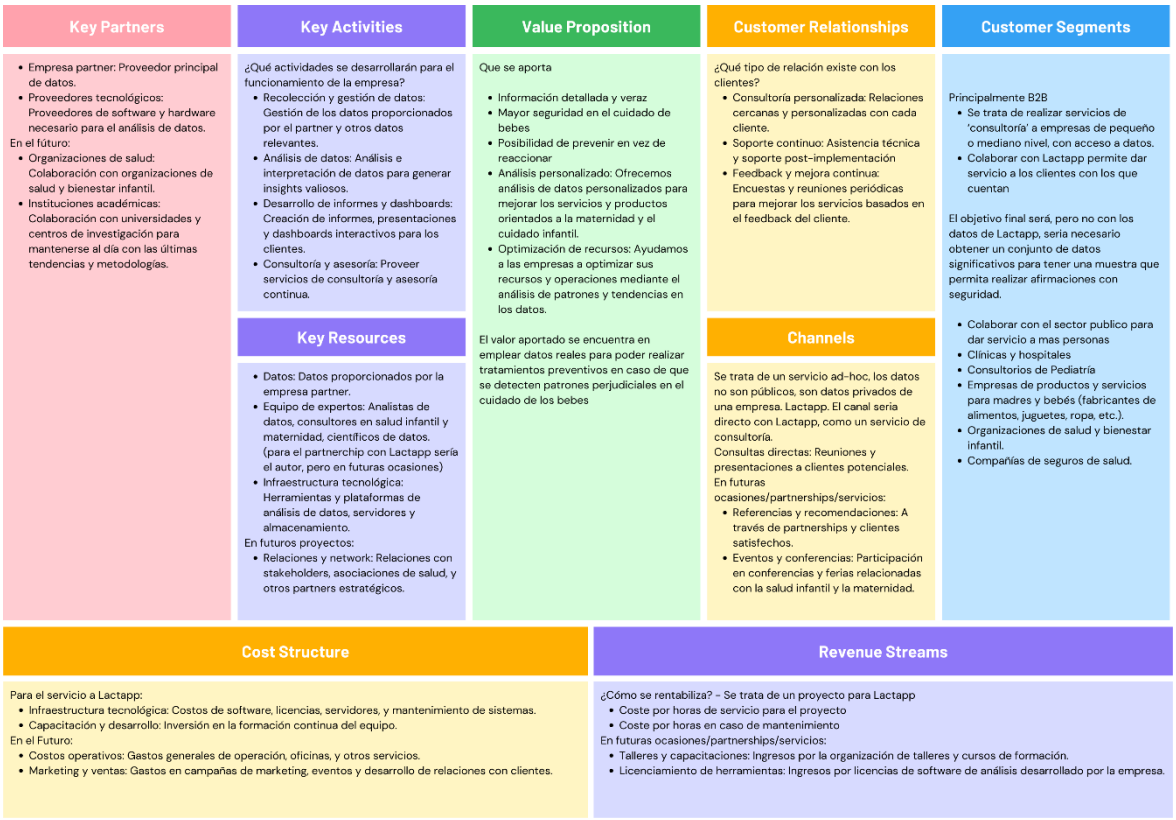


Ilustración 2: Business Model Canvas. Fuente: propia

SEGMENTOS DE CLIENTES (CUSTOMER SEGMENTS)

Se propone un servicio de análisis semejante al de consultoría, por ello principalmente se trataría de segmentos B2B.

- Clínicas y hospitales especializados en maternidad y cuidado infantil.
- Consultorios de pediatría.
- Empresas de productos y servicios para madres y bebés (fabricantes de alimentos, juguetes, ropa, etc.).
- Organizaciones de salud y bienestar infantil.
- Compañías de seguros de salud.
- Compañías privadas del sector de salud

PROPUESTA DE VALOR (VALUE PROPOSITIONS)

- **Información detallada y veraz:** Ofrecemos análisis precisos basados en datos reales y objetivos.
- **Análisis personalizado**: Ofrecemos análisis de datos personalizados para mejorar los servicios y productos orientados a la maternidad y el cuidado infantil.
- **Optimización de recursos**: Ayudamos a las empresas a optimizar sus recursos y operaciones mediante el análisis de patrones y tendencias en los datos.

El valor aportado se encuentra en emplear datos reales para poder realizar tratamientos preventivos en caso de que se detecten patrones perjudiciales en el cuidado de los bebés

CANALES (CHANNELS)

Se trata de un servicio ad-hoc, los datos no son públicos, son datos privados de una empresa. Lactapp. El canal sería directo con Lactapp, como un servicio de consultoría.

Consultas directas: Reuniones y presentaciones a clientes potenciales.

En futuras ocasiones/partnerships/servicios:

- **Marketing digital**: Sitio web, redes sociales, webinars, y newsletters.
- **Eventos y conferencias**: Participación en conferencias y ferias relacionadas con la salud infantil y la maternidad.
- **Consultas directas**: Reuniones y presentaciones a clientes potenciales.
- **Referencias y recomendaciones**: A través de partnerships y clientes satisfechos.

RELACIONES CON CLIENTES (CUSTOMER RELATIONSHIPS)

- **Consultoría personalizada**: Relaciones cercanas y personalizadas con cada cliente.
- **Soporte continuo**: Asistencia técnica y soporte post-implementación.
- **Feedback y mejora continua**: Encuestas y reuniones periódicas para mejorar los servicios basados en el *feedback* del cliente.

FUENTES DE INGRESOS (REVENUE STREAMS)

Principalmente:

- **Consultoría y análisis**: Cobro por proyecto o por horas de consultoría.

En el futuro:

- **Suscripciones**: Modelos de suscripción para acceso a análisis continuos y reportes periódicos.
- **Talleres y capacitaciones**: Ingresos por la organización de talleres y cursos de formación.
- **Licenciamiento de herramientas**: Ingresos por licencias de software de análisis desarrollado por la empresa.

RECURSOS CLAVE (KEY RESOURCES)

- **Datos**: Datos proporcionados por la empresa *partner*.
- **Equipo de expertos**: Analistas de datos, consultores en salud infantil y maternidad, científicos de datos.
- **Infraestructura tecnológica**: Herramientas y plataformas de análisis de datos, servidores y almacenamiento. Algunos ejemplos son Amazon Web Services, Azure, Google Cloud Platform, etc.
- **Relaciones y *network***: Relaciones con *stakeholders*, asociaciones de salud, y otros *partners* estratégicos.

ACTIVIDADES CLAVE (KEY ACTIVITIES)

- **Recolección y gestión de datos**: Gestión de los datos proporcionados por el *partner* y otros datos relevantes.
- **Análisis de datos**: Análisis e interpretación de datos para generar *insights* valiosos.
- **Desarrollo de informes y *dashboards***: Creación de informes, presentaciones y *dashboards* interactivos para los clientes.

- **Consultoría y asesoría**: Proveer servicios de consultoría y asesoría continua.

SOCIOS CLAVE (KEY PARTNERSHIPS)

- **Empresa *partner***: Proveedor principal de datos.
- **Proveedores tecnológicos**: Proveedores de software y hardware necesario para el análisis de datos.
- **Organizaciones de salud**: Colaboración con organizaciones de salud y bienestar infantil. ONGs como 'Save the Children', con programas de salud infantil a nivel global o la 'Fundación española del Corazón', que se enfoca en enfermedades cardiovasculares desde la infancia.
- **Instituciones académicas**: Colaboración con universidades y centros de investigación para mantenerse al día con las últimas tendencias y metodologías.

ESTRUCTURA DE COSTES (COST STRUCTURE)

- **Costes de personal**: Salarios y beneficios para analistas de datos, consultores, y personal técnico.
- **Infraestructura tecnológica**: Costos de software, licencias, servidores, y mantenimiento de sistemas.
- **Marketing y ventas**: Gastos en campañas de marketing, eventos y desarrollo de relaciones con clientes.
- **Capacitación y desarrollo**: Inversión en la formación continua del equipo.
- **Costes operativos**: Gastos generales de operación, oficinas, y otros servicios.

PRESENTACIÓN DE LA EMPRESA

La empresa que se constituye en este proyecto de Fin de Máster de Ingeniería Industrial recibe el nombre de "*Pedidata Consulting*". Surge de la idea de combinar los conocimientos técnicos aprendidos a lo largo del grado y el máster junto con los conocimientos de empresa y negocio, así como el deseo de realizar un proyecto de emprendimiento del autor.
Se pretende desarrollar un plan de negocio y un servicio de carácter técnico para dar solución a aquellas empresas que desean ofrecer a sus clientes información de carácter predictivo y preventivo, así como segura.
Este servicio pretende dar solución a varios problemas principales: En primer lugar, la falta de información a la hora de hallar problemas relacionados con la neonatalidad. En segundo lugar, emplear modelos predictivos para identificar riesgos de salud en neonatos y en tercer lugar facilitar una mejor comunicación entre profesionales de la salud y padres sobre el estado de salud y las necesidades de los neonatos.
Se desarrollará un modelo de análisis de datos y de ML para realizar predicciones.

MISIÓN

Proporcionar soluciones innovadoras y basadas en datos reales para mejorar la salud y el bienestar de los neonatos y sus familias. Utilizando técnicas avanzadas de análisis de datos y ML, buscamos ofrecer información predictiva y preventiva a las empresas del sector de la salud infantil y los profesionales de salud, ayudando a detectar y gestionar riesgos de salud de manera efectiva y segura, así como mejorando la comunicación entre profesionales y pacientes.

VISIÓN

Nuestra visión es convertirnos en líderes mundiales en el análisis de datos desde no estructurados hasta su monitorización para la salud neonatal, transformando la manera en que se aborda el cuidado de los neonatos a través de la tecnología y la innovación.

VALORES

1. **Innovación**: Nos comprometemos a estar a la vanguardia de la tecnología, ofreciendo soluciones innovadoras que transformen el cuidado neonatal.
2. **Precisión**: Garantizamos la exactitud y fiabilidad de nuestros análisis y predicciones, proporcionando información crucial para la toma de decisiones en el cuidado de los neonatos.
3. **Seguridad**: Valoramos la privacidad y seguridad de los datos, asegurando que la información de nuestros clientes y pacientes sea manejada con el máximo cuidado y protección.
4. **Compromiso**: Estamos dedicados a mejorar la salud infantil y a apoyar a los profesionales de la salud y a los padres en cada paso del camino.
5. **Colaboración**: Fomentamos una estrecha colaboración entre todos los actores involucrados en el cuidado de los neonatos, trabajando juntos para mejorar los resultados.
6. **Ética**: Actuamos con integridad y responsabilidad, asegurando que nuestras prácticas empresariales y nuestros servicios cumplan con los más altos estándares éticos.

MACROENTORNO

Como se ha mencionado con anterioridad, los nuevos y continuos avances en el ámbito de la tecnología y la medicina, así como el fácil acceso a la información, ha cambiado el comportamiento y la concienciación de los padres primerizos, quienes influenciados por estos cambios muestran mayor tendencia al cuidado con sumo detalle y el continuo monitoreo de sus hijos.

La tecnología hace más accesible una vasta cantidad de recursos y conocimientos que antes no estaban disponibles fácilmente. Esto incluye desde artículos y libros hasta servicios de asesoramiento psicológico y médico en línea. Los padres tienen fácil acceso a información sobre el desarrollo infantil y mejores prácticas de cuidado, lo que influye en su comportamiento y decisiones diarias. Según el Pew Research Center, muchos padres reconocen que la disponibilidad de información en internet ha hecho que la crianza sea más manejable y les ha proporcionado herramientas para educarse sobre el desarrollo y cuidado de sus hijos. (BROOKE AUXIER, 2020).

La encuesta de la National Survey of Children's Health muestra que los padres están cada vez más involucrados en el cuidado de la salud de sus hijos, utilizando tecnologías que permiten un seguimiento continuo y detallado. Esta tendencia refleja un cambio en el comportamiento parental hacia un enfoque más proactivo y preventivo en la gestión de la salud infantil. Según los datos de la encuesta, entre 2021-2022, más de un 85% de los niños entre 0 y 17 años reciben cuidados centrados en el entorno familiar (Health, 2021-2022). Estas tendencias también se han validado mediante una encuesta realizada a una muestra de más de 50 personas, en la que encontramos resultados que validan la idea de que el comportamiento de los padres está cambiando con los últimos avances tecnológicos. Cada vez los padres monitorean más el cuidado de la salud y nutrición de sus hijos, y buscan obtener respuestas basadas en datos fiables. Los resultados de la encuesta se encuentran reflejados en Anexo A – Encuesta Sobre las tendencias en los cuidados infantiles y uso de tecnología.

Es común encontrarse con datos que sugieren que los padres cada vez prestan más atención al detalle en el cuidado de sus hijos, no solo gracias al fácil acceso a la información, sino también

a los avances en medicina. Esta tendencia sugiere que existe un mercado creciente en cuanto a la necesidad de acceso a la información precisa y detallada por parte de los padres.

Por otro lado, es posible ver el auge de tecnologías nuevas como el ML y el análisis de datos para obtener ventajas competitivas en todos los mercados.

Según IBM, "el ML es una forma de inteligencia artificial que permite que un sistema aprenda y mejore automáticamente a partir de la experiencia sin ser programado explícitamente" (STATES, s.f.). Similarmente, Coursera define el ML como "el uso de algoritmos y datos para crear modelos que pueden realizar tareas sin intervención humana directa, mejorando su precisión a medida que se les proporciona más datos" (STAFF, 2024)

Las empresas emplean algoritmos de ML para analizar grandes cantidades de datos y a su vez hacer predicciones precisas sobre el comportamiento de los clientes. Por ejemplo, los minoristas pueden predecir qué productos comprarán los clientes a continuación, lo que mejora significativamente la experiencia del cliente y aumenta las ventas. Amazon es un claro ejemplo de cómo las recomendaciones personalizadas basadas en ML pueden impulsar las ventas (avcontentteam, 2024).

Otras empresas emplean esta tecnología para crear los sistemas de recomendación personalizados basados en ML mejorando la satisfacción del cliente al proporcionar sugerencias de productos y servicios que se alinean con sus preferencias (media, 2023).

También encontramos casos en los que se emplea ML para optimizar cadenas de suministro, predecir fallos de equipos y mejorar la asignación de recursos. Mejoras que pueden resultar en ahorros de costes significativos y en aumentos de la productividad. Por ejemplo, las compañías mineras están empleando ML para el mantenimiento predictivo de sus flotas, reduciendo costos y aumentando la producción (Digital, 2020).

En este artículo se muestran distintos casos de estudio en el que se emplean estas nuevas tecnologías para transformar productos y servicios, mejorando la experiencia de clientes, reduciendo costes y mejorando productividad. Casos como el de Goldman Sachs con su plataforma Marcus, ha demostrado cómo una filosofía digital primero puede transformar rápidamente productos y servicios para satisfacer las necesidades cambiantes de los clientes.

Otros casos como el de Petrosea, ha demostrado que el uso de una combinación de IA, sensores inteligentes y ML puede transformar su operación, reducir costes y aumentar la producción,

demostrando el poder de la digitalización para mantener la relevancia en un mercado competitivo.

Esto muestra que la adopción de estas tecnologías está en crecimiento y que existe un mercado en el empleo de las mismas para obtener ventajas competitivas, dejando hueco para que una solución como la de *Pedidata* sea capaz de abrirse en este mercado en el sector de salud.

Según un informe de McKinsey, el 60% de las empresas del sector financiero ya han implementado al menos una capacidad de IA en sus operaciones. Además, se proyecta que el mercado global de ML crezca de 21.17 mil millones de dólares en 2022 a 209.91 mil millones de dólares para 2029, con una tasa de crecimiento anual compuesta (CAGR) del 38.8% (Company, 2023) (Korobeyko, 2023)

PESTEL

A través del análisis PESTEL se busca identificar las tendencias del mercado que pueden afectar al negocio. Este enfoque examina seis áreas: Política, Económica, Social, Tecnológica, Ecológica y Legal. Inicialmente, se evaluaron las tendencias en cada una de estas áreas, asignándoles una puntuación de 0 a 5. Con estas puntuaciones, se determinó una puntuación global para cada uno de los seis temas.

A continuación, se presenta una imagen con las tendencias consideradas y las puntuaciones asignadas así como las puntuaciones globales para cada tema.

ANALISIS PESTEL	FACTORES	TOTAL
P	Regulaciones de salud y seguridad: 3/5 Iniciativas gubernamentales y subsidios: 2/5	2,5/5
E	Crecimiento del sector de la salud: 3/5 Inversión en tecnología: 5/5 Capacidad de gasto de los consumidores: 2/5	3,3/5
S	Envejecimiento de la población y bajas tasas de natalidad: 2/5 Mayor concienciación sobre la salud: 5/5	3,5/5
T	Avances en análisis de datos y big data: 5/5 Telemedicina y e-health: 4/5 Automatización e inteligencia artificial: 5/5	4,6/5
E	Sostenibilidad: 3/5 Regulaciones ambientales: 1/5	2/5
L	Regulaciones de privacidad y protección de datos: 4/5 Legislación en salud: 4/5 Propiedad intelectual: 5/5	4,3/5

Ilustración 3: Análisis PESTEL. Fuente: propia

FACTORES POLÍTICOS:

- **Regulaciones de salud y seguridad:** En muchos casos, las políticas gubernamentales enfatizan mucho el cuidado de la salud infantil y la maternidad, promoviendo la demanda de servicios que los mejoren. Según un informe de la HHS, hay iniciativas significativas en políticas de salud que apoyan y mejoran los resultados en la salud materna y neonatal, reflejando una prioridad gubernamental en este ámbito ((ASPA), 2023)
- **Iniciativas gubernamentales y subsidios:** Los programas gubernamentales a menudo proporcionan subsidios y fondos para apoyar la innovación en el sector de la salud. Por ejemplo, la administración Biden-Harris en los Estados Unidos ha implementado modelos de pago y atención que mejoran los resultados de salud materna, lo cual es paralelo a

iniciativas similares en Europa que podrían influir en las políticas en España (Services, 2022).

- **Conflictos geopolíticos:** Existe un entorno de incertidumbre, en el que la recuperación económica se encuentra amenazada por diversos factores macroeconómicos como son los altos niveles de inflación, las potenciales subidas de tipos, los actuales conflictos geopolíticos entre Rusia y Ucrania o el incremento de los costes energéticos. (Tamayo, 2022)

FACTORES ECONOMICOS:

- **Crecimiento del sector de la salud:** El sector de la salud en España, especialmente en el área de maternidad y cuidado infantil, está en constante crecimiento, generando una demanda continua de servicios de mejora y optimización. (Tamayo, 2022) Según un informe de McKinsey, las empresas están invirtiendo significativamente en tecnologías de salud, lo que incluye el análisis de datos para mejorar la operatividad y los servicios (Company, 2023)
- **Inversión en tecnología:** Las organizaciones de salud en España están dispuestas a invertir en tecnologías avanzadas, incluyendo el análisis de datos, para mejorar sus servicios. Esta es una tendencia global que se evidencia especialmente en sectores como la banca y la manufactura, donde la adopción de ML y AI está aumentando rápidamente (Korobeyko, 2023)

FACTORES SOCIALES:

- **Envejecimiento de la población y bajas tasas de natalidad:** España se enfrenta a tasas de natalidad bajas (Arranz, 2022), pero hay un enfoque creciente en la calidad del cuidado para los niños que nacen.
- **Mayor concienciación sobre la salud:** Los padres actuales están más informados y son más exigentes respecto a la calidad de los servicios de salud para sus hijos, buscando constantemente mejorar el cuidado y bienestar familiar.

FACTORES TECNOLÓGICOS:

- **Avances en análisis de datos y Big Data:** Las tecnologías de análisis de datos están evolucionando constantemente, permitiendo un análisis más profundo y eficiente de grandes volúmenes de datos. Esto es crucial para el éxito de empresas como *Pedidata* que se basan en análisis de datos para ofrecer sus servicios (Company, 2023)
- **Telemedicina y e-Health:** La adopción creciente de tecnologías de telemedicina y salud digital facilita la recolección y análisis de datos, creando más oportunidades para empresas de análisis de datos en el sector de la salud
- **Automatización e inteligencia artificial:** El uso de IA y automatización en el análisis de datos mejora la precisión y eficiencia, ofreciendo ventajas competitivas significativas a las empresas que implementan estas tecnologías, de cara a producción, reducción de costes y mejora de la satisfacción de los clientes entre otros

FACTORES ECOLÓGICOS

- **Sostenibilidad:** Las prácticas sostenibles son cada vez más importantes tanto para empresas como para consumidores. Cumplir con las regulaciones ambientales y promover prácticas ecológicas puede mejorar la reputación de una empresa y atraer a clientes conscientes del medio ambiente. En este sentido, cabe mencionar que esta propuesta se alinea con los siguientes objetivos de desarrollo sostenible:
    - 2-Hambre 0: El servicio contribuye a la mejora de la nutrición de los neonatos, asegurando que los problemas relacionados con la alimentación puedan ser identificados y tratados.
    - 3-Salud y Bienestar: Al proporcionar herramientas para detectar de forma temprana y facilitar el tratamiento de problemas de nutrición y posibles problemas de salud, este proyecto tiene un impacto directo en mejorar la salud y el bienestar de los recién nacidos.
    - 10-Reducción de las desigualdades: El objetivo sería implementar esta solución de forma pública en el futuro, o facilitar este paso mediante el trabajo realizado. El proyecto puede ayudar a reducir las desigualdades al proporcionar acceso a

tecnologías de salud avanzadas tanto en comunidades desarrolladas como en desarrollo.

- 4-Educación de calidad: A través de la educación y capacitación de los padres y profesionales de la salud sobre nutrición neonatal y el uso de tecnologías de análisis de datos, el proyecto promueve la educación en salud y bienestar, que es fundamental para el desarrollo de comunidades saludables y bien informadas.

FACTORES LEGALES

- **Regulaciones de privacidad y protección de datos:** Las leyes de protección de datos, como el GDPR en Europa, requieren que las empresas manejen la información de manera segura y ética. Cumplir con estas regulaciones puede aumentar la confianza de los clientes y es crucial para operar en el sector de la salud. Esto a su vez puede complicar su implementación de forma masiva, impidiendo el escalado del proyecto.
- **Legislación en salud:** Las regulaciones específicas del sector salud deben ser observadas y cumplidas, lo que puede implicar costes, pero también garantiza que la empresa opere en un entorno seguro y regulado.

MICROENTORNO

5 FUERZAS DE PORTER

El método de las 5 fuerzas de Porter es un marco de análisis estratégico desarrollado por Michael E. Porter, que se utiliza para evaluar la competitividad de una industria y determinar su atractivo general.

AMENAZA DE ENTRADA DE NUEVOS COMPETIDORES:

Barreras de entrada:

1. Acceso a datos de alta calidad:

Acceder a datos relevantes y de alta calidad es crucial para ofrecer servicios precisos de análisis de datos. Los nuevos entrantes podrían encontrar dificultades para obtener estos datos debido a acuerdos exclusivos y colaboraciones existentes entre proveedores de datos y empresas consolidadas. Además, al tratarse de ML, las empresas consolidadas podrían tener acceso a datos significativamente más pulidos que los nuevos entrantes, ya que es un proceso de aprendizaje continuo.

2. Regulaciones y cumplimiento normativo

Cumplir con las regulaciones de privacidad y protección de datos (como el GDPR en Europa) es una barrera significativa. Los nuevos entrantes deben establecer sistemas robustos para asegurar el cumplimiento legal, lo cual requiere tiempo y recursos.
Existe la posibilidad de que empresas que ya tienen un sistema de privacidad y protección de datos bien establecido, traten de incorporarse al sector de salud

3. Tendencias en la tecnología

Las tendencias de adopción de tecnologías avanzadas como IA, ML y Big Data están acelerando, lo que favorece a las empresas que ya están establecidas y tienen infraestructura tecnológica avanzada.
Además, al tratarse de una nueva empresa, no se contará con capital para hacer inversiones importantes, mientras que las empresas ya establecidas que no empleen estas tecnologías y que

ven que cada vez existe mayor ventaja al emplear estas tecnologías, tendrán mayor capital para realizar inversiones en I+D

4. Competidores

Existe una gran cantidad de competidores en el sector del análisis de datos, IA y ML. Empresas como IBM, Oracle y Philips ya cuentan con soluciones en el ámbito de la salud IBM Watson Health, Philips Healthcare y Oracle Health respectivamente.
No sólo esto, si no que, algunos datos indican que 9 de cada 10 empresas están invirtiendo en alguna forma de ML (hosteltur, 2019). Indicando que ya existe un mercado consolidado para este tipo de soluciones y de gran competencia.

5. Tiempo de desarrollo

El tiempo necesario para desarrollar una solución de análisis de datos de alta calidad puede ser una desventaja significativa para nuevos entrantes. Las empresas establecidas ya tienen productos en el mercado y pueden iterar y mejorar continuamente.
Asimismo, durante el tiempo invertido en el desarrollo de una solución que emplee esta tecnología, pueden surgir nuevos competidores que quieran entrar en el mercado. Como se ha mencionado con anterioridad, otras empresas con los recursos suficientes para invertir en I + D podrían desarrollar una solución más rápido que nuevas empresas.

RIVALIDAD ENTRE COMPETIDORES EXISTENTES:

El nivel de competitividad en el sector de análisis de datos y ML para la salud infantil y maternal es alto. Esto es principalmente debido a la presencia de importantes empresas tecnológicas y de salud que ya están consolidadas y ofrecen soluciones fiables. Estas empresas tienen muchos recursos para invertir en I+D, marketing y establecimiento de relaciones estratégicas, lo que les da una ventaja competitiva considerable. Por otro lado, la demanda creciente de soluciones de análisis de datos en salud también intensifica la competencia, ya que más empresas intentan ingresar al mercado y captar la atención de los clientes (technavio, 2023).
Estas son algunas de las empresas existentes en el mercado a día de hoy:

- IBM Watson: IBM Watson es una división de IBM que utiliza inteligencia artificial y análisis avanzado para ayudar a profesionales de la salud a mejorar el cuidado del paciente. Ofrecen soluciones que van desde la gestión de datos hasta la predicción de resultados de salud. Con más de un siglo de experiencia, IBM es pionera en innovación tecnológica y transformación digital. Sus áreas de especialización incluyen la inteligencia artificial, la computación en la nube, la seguridad de la información, y el análisis de datos. IBM Watson, su plataforma de inteligencia artificial, es líder en la industria, ofreciendo soluciones avanzadas para diversos sectores, incluyendo salud, finanzas y tecnología. La empresa también destaca por su compromiso con la investigación y desarrollo, impulsando continuamente los límites de la tecnología. (IBM, s.f.). IBM ofrece una variedad de planes de precios para sus soluciones de ML, adaptándose a diferentes necesidades empresariales. Los planes van desde opciones gratuitas para evaluación hasta soluciones premium para grandes empresas. La flexibilidad en el despliegue, la integración con herramientas avanzadas y el soporte para múltiples marcos de ML hacen que las soluciones de IBM sean una opción robusta para empresas que buscan implementar y escalar sus capacidades de IA. (Studio, s.f.) (Learning, s.f.)
- Oracle y Cerner: Cerner Corporation fue adquirida por Oracle y ahora se conoce como Oracle Health. Según Oracle Health, con la conexión de tecnologías y la unificación de datos, se le da poder a individuos permitiendo al ecosistema de salud acelerar su innovación e influencia en los resultados sanitarios. Están creando una plataforma abierta con herramientas centradas en el dato para conectar a los consumidores con proveedores y otros involucrados en el sistema sanitario (Oracle, s.f.). Tienen una gran integración con sistemas de registros electrónicos de salud, y en sus palabras el *Market Share* más amplio a nivel global. Esto significa que ya tienen una gran base de clientes, con presencia en hospitales y centros de salud. Oracle fue fundada en 1977, es decir, tiene aproximadamente 50 años de experiencia en el ámbito de sistemas de gestión de bases de datos relacionales. Está especializada en ámbitos como las bases de datos, los softwares

empresariales (CRM), la computación en la nube y el análisis de datos y el ML. En este último, ofrece soluciones específicas integradas en su plataforma Oracle Cloud Infrastructure, diseñadas para ofrecer información relevante de grandes volúmenes de datos. Entre sus productos se encuentran **Oracle Analytics Cloud (OAC)** que ofrece capacidades avanzadas de análisis, incluyendo visualización de datos, preparación de datos y análisis predictivo, y **Oracle ML (OML)** que permite la creación y despliegue de modelos de ML directamente en la base de datos. Los precios de OAC varían entre 16 y 80 dólares por usuario por mes, pero también ofrecen servicios de pago por uso CPU. De un modo similar, los precios por el uso de sus soluciones de ML varían desde 0,0125 a 2 dólares por hora de uso.

- Philips Healthcare: Philips Healthcare está centrado principalmente en dar asistencia a profesionales médicos. Entre sus soluciones encontramos una diversa gama de productos tecnológicos de última generación, pero también ofrecen lo que ellos llaman “informática clínica”. Sus sistemas son compatibles con IA y ofrecen monitorización avanzada, análisis predictivos y diagnósticos precisos entre otros. Philips Healthsuite es su solución en el ámbito del ML. Es una plataforma de tecnología de la salud que conecta dispositivos, datos y sistemas para mejorar la atención y los resultados de los pacientes. HealthSuite utiliza análisis de datos y ML para proporcionar insights accionables. ‘Al mismo tiempo, un amplio repositorio de datos recopilados a lo largo del tiempo y de miles de puntos de datos individuales permite a las organizaciones encontrar patrones y puntos de vista para asumir desafíos más grandes en la prestación de cuidados en poblaciones enteras.’ (Philips, s.f.). Philips es una empresa con más de 100 años de experiencia y aunque no están especializados en el análisis de datos, tienen especialización en otras áreas, incluyendo la iluminación, la electrónica de consumo y la tecnología de salud. Philips tiene una gran reputación global y es conocida por sus productos de alta calidad y tecnología avanzada. Esto indica que tiene una buena base de clientes y buen capital con un fuerte enfoque en I + D para desarrollar nuevas tecnologías

y soluciones. Los precios para su solución HeathSuite son muy personalizados y varían según el alcance del proyecto, es difícil extraer una estimación.

AMENAZA DE PRODUCTOS Y SERVICIOS SUSTITUTOS:

Existen otros competidores que aunque no están involucrados directamente en el ámbito de la salud, proporcionan soluciones avanzadas en el ámbito de la inteligencia artificial, ML, análisis predictivo y gestión de datos. Estas soluciones son fácilmente adoptables por aquellas empresas con las que *Pedidata* podría alcanzar alianzas estratégicas en un futuro, y por tanto, suponen potenciales competidores e incluso pueden llegar a tener una propuesta de valor similar a la que tiene *Pedidata*.

Entre ellos encontramos servicios como:

- SAS Institute Inc.: SAS es una empresa líder en software de análisis avanzado, fundada en 1976. Ofrece soluciones de inteligencia de negocios, análisis predictivo y gestión de datos, y es ampliamente utilizada en diversas industrias. Tiene una larga trayectoria en el análisis de datos y es conocida por su robustez y precisión en el manejo de grandes volúmenes de datos. (SAS, s.f.)
- Microsoft Azure: Microsoft Azure es una plataforma de servicios en la nube que proporciona una amplia gama de servicios, incluyendo análisis de datos, inteligencia artificial y ML. Fundada en 2010, Azure ha crecido para convertirse en una de las principales plataformas de nube pública. Ofrece servicios como **Azure ML**, servicio en la nube que permite construir, entrenar e implementar modelos de ML con una gran estructura escalable. Una de las claras ventajas de esta solución es que tiene el respaldo de ser parte de Microsoft, uno de los grandes gigantes tecnológicos, y por supuesto su integración con el resto de sus servicios (Microsoft, s.f.)
- Google Cloud Platform: Google Cloud Platform ofrece una variedad de servicios en la nube, incluyendo análisis de datos y ML. Google ha sido pionero en tecnologías de búsqueda y análisis de datos desde su fundación en 1998. Ofrece una suite de herramientas de IA y ML que incluyen TensorFlow y AutoML para el desarrollo de modelos personalizados. De nuevo una de las grandes ventajas de esta solución es que Google es

conocido por su innovación continua y su liderazgo en el desarrollo de tecnologías de IA y ML.

- Amazon Web Services (AWS): AWS es la plataforma de servicios en la nube de Amazon, lanzada en 2006. Ofrece una amplia gama de servicios, incluyendo computación, almacenamiento y análisis de datos. Ofrece servicios específicos como **Amazon SageMaker**, un servicio completamente gestionado que permite a los desarrolladores y científicos de datos construir, entrenar y desplegar modelos de ML a escala. No solo tiene la ventaja de ofrecer una amplia gama de servicios de fácil escalado, sino también el nombre y la reputación que existen detrás de la empresa Amazon.

PODER DE NEGOCIACIÓN DEL CLIENTE:

Teniendo en cuenta todas las opciones que se han ido mencionando con anterioridad, podríamos concluir que el poder de negociación del cliente es muy alto.

Como estrategia para penetrar en el mercado y evitar el abuso de poder del cliente, por un lado, se deben establecer buenas alianzas estratégicas con empresas en el sector de salud, tanto de ámbito privado como público. Esto nos permitirá diferenciarnos de otros competidores que abarca soluciones más amplias en más industrias. También nos permitirá tener un flujo de datos constante y seguro, que afianzará la fiabilidad de nuestro producto.

Por otro lado, debemos enfocarnos en la personalización de nuestros análisis, sería un error crear un software y ofrecerlo como Saas teniendo en cuenta todas las alternativas que existen en el mercado y su reputación. Hay que centrarse en el aspecto de consultoría, ofreciendo servicios de consultoría personalizada, informes detallados y recomendaciones personalizadas que aumenten la lealtad de nuestros clientes y aporten una buena reputación a la marca.

Adicionalmente, se ha de buscar un elemento diferenciador para poder penetrar en el mercado. Se podría plantear un sistema de monitorización, integrando los distintos datos no estructurados de las empresas con las que trabajemos, y cuyos datos especializados (Lactapp – lactancia, Zoundream – sueño, EEEG – Neurologicos, etc) se puedan integrar en un sistema de telemedicina para tomar acciones en *near/real time.*

Finalmente habría que estudiar con gran atención al detalle la estrategia de inversión. Debemos continuar invirtiendo en innovación continua para mejorar nuestro servicio, pero también

debemos invertir en marketing digital para dar a conocer nuestra marca, especialmente mediante las redes sociales y los propios *partners* con los que tratemos, que den a conocer nuestro servicio y sus resultados, así como en talleres y formaciones en el ámbito de la salud infantil concretamente, que nos permitan establecer otro canal para la adquisición de clientes así como para establecer nuevas relaciones estratégicas al mismo tiempo que nos diferencie del resto de industrias.

PODER DE NEGOCIACIÓN DEL PROVEEDOR:

El poder de negociación del proveedor reside en la necesidad de datos de alta calidad y tecnología avanzada para el análisis.

En cuanto a los datos de alta calidad, depende principalmente de las alianzas que se establezcan con otras empresas y los servicios que se proporcionen en el tiempo. El acceso a datos de gran calidad es complejo y está muy controlado. Será importante alcanzar buenos acuerdos con el resto de las empresas de forma que no impidan nuestro trabajo con restricciones de uso y distribución de estos.

Por otro lado, las empresas sustitutas también pueden ser proveedoras de servicios para *Pedidata*, permitiendo emplear sus servicios para el análisis de datos y ML. Los costes asociados pueden llegar a ser altos dependiendo de la solución escogida, y también puede afectar a decisiones futuras ya que cada uno de estos proveedores ofrece su propio *framework* e infraestructuras que puede afectar a la toma de decisiones.

Por último, será necesario emplear equipos electrónicos potentes y con cierta capacidad de computación. Afortunadamente existen numerosos proveedores de estos productos, sin embargo, será importante buscar contratos de servicio gestionado que ofrezcan soporte continuo y garantías. A pesar de ello, la computación en la nube puede permitirnos escalar las operaciones en función de la demanda de forma más flexible.

PRESENTACIÓN DEL SERVICIO

DESCRIPCIÓN DEL SERVICIO

Una vez estudiado tanto el entorno macro, como el micro que rodea al servicio que se pretende prestar, podemos dar una descripción más en detalle del servicio. Como ya se ha ido mencionando, el servicio que se pretende prestar es un servicio de consultoría/análisis de datos. Es difícil definir el servicio en sí, ya que cada cliente deberá aportar sus datos y esos serán diferentes por lo que definiremos una serie de entregables comunes a todos los proyectos. Podremos dividir los entregables en 2 bloques principales en función de la solución que solicite el cliente, el volumen de datos y la duración del proyecto.

- **Bloque Análisis**

Se trata de un servicio de análisis de datos en el que se buscará entregar un informe detallado, con visualizaciones y análisis de los datos de los que se extraerán patrones, tendencias y anomalías. El objetivo es que estos hallazgos permitan a nuestros clientes proporcionar recomendaciones e información más detallada a sus clientes en cuanto al estado de salud, nutrición, etc.
En primer lugar, se dará inicio al proyecto: en esta fase lo que se hará será reunirse con el cliente para definir los objetivos del proyecto, las expectativas, el alcance los entregables finales y un cronograma. El objetivo de esta reunión es definir de forma precisa el tipo de información que se busca extraer de los datos que se vayan a proporcionar, así como los plazos esperados para el resto de las fases y la definición de fechas estratégicas para reuniones donde se revise el progreso y se resuelvan dudas. Después de esta fase, se deberán tener claros los objetivos y se podrá elaborar un plan de proyecto.
En segundo lugar, se realizará la recopilación y procesamiento de datos: en esta fase, se deberá analizar la calidad de los datos del cliente, ya que para lograr alcanzar los objetivos deseados es necesario que los datos que se empleen en el análisis sean limpios y claros, de forma que las conclusiones extraídas sean fiables. Para facilitar esta fase, se deberá establecer otra reunión, el

objetivo de esta es resolver cualquier duda que surja durante el proceso de tratamiento de datos. De esta forma se entenderá correctamente el tipo de patrones que puede existir entre los distintos datos. Tras realizar la evaluación de la integridad, exactitud y consistencia de los datos, se procederá a su limpieza y tratamiento de valores anómalos en función de lo establecido. Finalmente, se realizará el análisis exploratorio de los datos (EDA): Una vez establecidos los KPI's a los que se debe prestar especial atención en busca de patrones, y teniendo la seguridad de que los datos empleados tienen la correcta integridad, consistencia y limpieza, se analizarán los datos y se realizará un informe detallado con visualizaciones y análisis de los datos, con posibles patrones identificados, así como tendencias y anomalías. Este informe se presentará en otra reunión, con recomendaciones de actuación sobre las principales variables en las que se ha identificado alguna tendencia o patrón.
Este será el entregable que concluirá el proyecto en caso de que la solución que solicita el cliente sea un análisis a alto nivel.

- **Bloque ML**

En caso de que la solución solicitada por el cliente sea más avanzada y requiera del empleo de modelos de ML que deban ser desplegados, las siguientes fases serán:
Posterior a la elaboración del informe EDA, se desarrollarán los modelos de ML: en esta fase en función de los patrones hallados será necesario seleccionar los algoritmos que se vayan a utilizar, empleando algoritmos como regresiones logísticas, árboles de decisión u otras. Estos algoritmos se evaluarán entrenándolos con un conjunto de datos seleccionados como entrenamiento. En base a la precisión del algoritmo, se seleccionará finalmente uno u otro. Se validará la precisión y rendimiento del algoritmo con métricas como F1-score o las curvas ROC, y se comparará con el resto de los algoritmos seleccionados anteriormente. Al final de esta fase se deberá tener un modelo entrenado con los datos del cliente, así como un informe con las validaciones y KPI's empleados para su selección.
Con el modelo seleccionado, el siguiente paso será el despliegue para el cliente: será necesario integrar la solución con la infraestructura del cliente, creando las APIs necesarias para que se pueda interactuar con el modelo. Esta solución deberá desplegarse en la nube, ya que no se contará con los recursos para alojarlos 'in-site' en *Pedidata*.

Por último, con el modelo desplegado, se capacitará al cliente: Será necesario documentar de forma detallada la solución desplegada para poder crear un manual de uso para el cliente. Es posible que sean necesarias reuniones posteriores a este entregable para asegurarse que el cliente esté familiarizado con la solución

Adicionalmente, se estudiará la posibilidad de desarrollar un sistema de telemedicina en un futuro. Una vez realizado el servicio a distintas empresas especializadas en datos concretos no estructurados, como por ejemplo Lactapp, que se especializa en lactancia, o Zoundream, que estudia el sueño, tendremos una base de datos no estructurados integral con la que se podría desarrollar un sistema de monitorización en *near/real time* para poder dar accionables valiosos a los profesionales relacionados con cada una de las empresas con las que se haya trabajado. Para ello sería necesaria la colaboración de las distintas empresas *partner*, de modo que acepten proporcionar sus datos para el desarrollo de este sistema.

ANÁLISIS DAFO

A continuación, se ha realizado un análisis DAFO de la empresa.



Ilustración 4: Análisis DAFO. Fuente: Propia

DIMENSIONAMIENTO DEL MERCADO

MERCADO TOTAL (TAM)

Para estimar el tamaño del mercado total (TAM – Total Addressable Market), se comienza por el número de empresas en el sector de salud de España. Al tratarse de una nueva empresa no parece razonable comenzar con una estrategia global.

Según datos del INE, en 2023 existían 158.057 empresas dedicadas a la industria sanitaria. Esta es una cifra que ha seguido una tendencia lineal creciente en los últimos 10 años, con 131.740 empresas en 2014 como se puede observar en el siguiente gráfico (INE, s.f.):

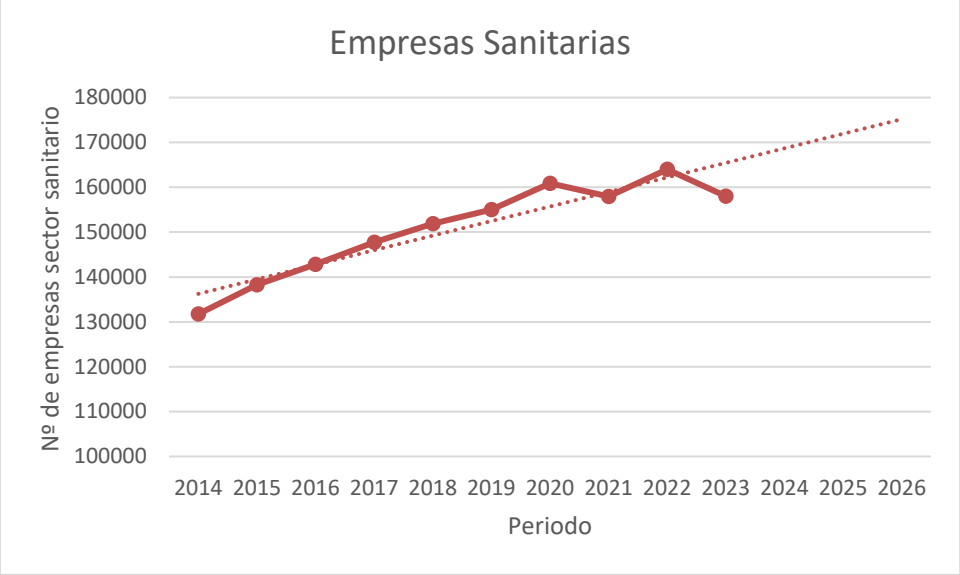


Ilustración 5: Número de empresas en el sector sanitario 2014-2023. Fuente: propia. Datos: INE

Tal y como se puede observar en el gráfico superior, la evolución del número de empresas en el sector sanitario ha sido creciente a excepción de los años 2021 y 2023, posiblemente a causa del contexto político y social, debido a que dichos años coinciden con la llegada de la COVID y los efectos de una recesión económica. Sin embargo, extrapolando esta tendencia llegamos a la conclusión de que sería posible esperar un crecimiento en el número de empresas en este sector hasta una cifra superior a 180.000 para el año 2028.

Sabiendo además que los datos indican que 9 de cada 10 empresas están invirtiendo en alguna forma de ML (hosteltur, 2019), asumiremos de forma conservadora que al menos el 75% de estas empresas buscan una solución similar o están invirtiendo en esta tecnología.
Con estos datos, se deduce que el mercado total en 2023 es **de 118.543 empresas** en el sector sanitario.

MERCADO DISPONIBLE (SAM)

No es posible acceder a todo el mercado (TAM) debido a las limitaciones del modelo de negocio y al hecho de ser una empresa nueva en un mercado con competidores consolidados. Por lo tanto, es importante calcular la porción del mercado que se puede captar con el modelo de negocio actual, conocida como el Mercado Disponible (Serviceable Available Market, SAM). Parece poco razonable tratar de penetrar el mercado de España al completo siendo que *Pedidata* es una *start-up* y posiblemente no cuente con los recursos en un inicio para un mercado tan amplio. Dado que se constituirá en Madrid, consideraremos este como el mercado disponible ya que facilitará el establecimiento de las alianzas estratégicas.
Según datos del INE, el número total de empresas del sector sanitario en Madrid es de 29.902, si se aplica la misma lógica que al TAM y se asume de nuevo que 75% están dispuestas a adoptar este tipo de servicios, se deduce que 22,427 empresas son el número de empresas que podrían contratar nuestros servicios.
Resulta complicado estimar el porcentaje de estas empresas que tiene carácter público y el que tiene carácter privado, sin embargo, en base a estadísticas de empleo de 2023 sabemos que de 21.246.900 personas empleadas, 3.593.300 son en empleo público, lo que supone un 17% (INE, s.f.). Asumiremos por tanto que al menos el 10% de estas empresas son públicas.
Sabemos además que, dentro de la atención primaria, parte de los recursos se destinan al cuidado infantil (pediatría y otros) por tanto del total de empresas al menos el 10% tienen relación con los servicios específicos que se ofrecen. Parece razonable asumir que al menos el 50% de las empresas privadas dedicadas a la salud, ofrecen servicios de cuidado infantil de un modo u otro, por tanto, nuestro **SAM** será de **13.456 empresas.**

MERCADO OBJETIVO (SOM)

No sería razonable pensar que con todas las alternativas que existen en el mercado, todo el mercado disponible (SAM) elegiría nuestro servicio, particularmente sabiendo la reputación, nombre, base de clientes y especialización de algunas de las alternativas. Teniendo en cuenta que es una nueva empresa, se reduciría considerablemente nuestro mercado disponible objetivo (SOM – Service Obtainable Market).

Por este motivo, se pretende alcanzar el 3% del *market share* en los primeros 5 años, pero teniendo en cuenta la tendencia del crecimiento en el número de empresas en el sector sanitario, parece razonable asumir que este mercado aumentará en el tiempo proyectado.

Dado que la cifra de *market share* irá creciendo con el tiempo según establezcamos un mayor número de alianzas estratégicas con estas empresas, y a su vez nuestra reputación mejore facilitando otras futuras, esperamos una captación de mercado como la reflejada en la siguiente tabla, que supone que para el año 5 tendremos proyectos con **404 empresas.**

Año	% de MS	% acc de MS
1	0.2	0.2
2	0.3	0.5
3	0.5	1
4	1	2
5	1	3

Tabla 1: Evolución Market Share. Fuente: Propia

A continuación, se presenta un esquema del TAM, SAM y SOM para visualizar qué se pretende conseguir y cuál es el mercado total.

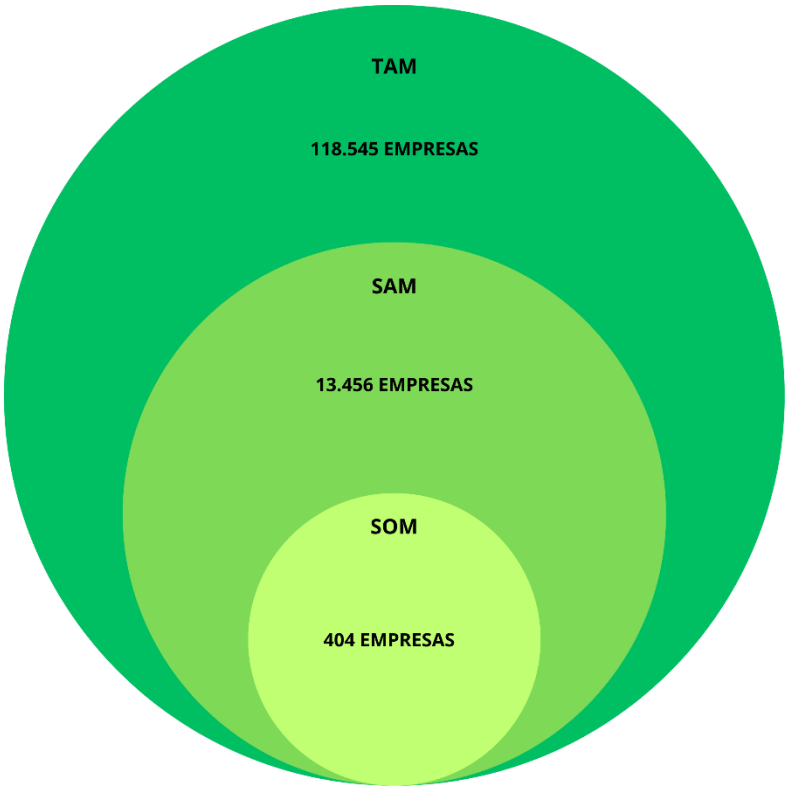


Ilustración 6: TAM, SAM, SOM. Fuente: propia

Asumiendo que la tendencia de crecimiento en el número de empresas del sector sanitario se mantiene como los datos del INE indican, el número de empresas estimado anteriormente será mayor también. Aplicando este ajuste a la estimación del SOM obtenemos los siguientes resultados.

Año	Proyectos mensuales
2024	3
2025	4
2026	7
2027	14
2028	14

Tabla 2: Proyectos por mes en los primeros 5 años. Fuente: Propia

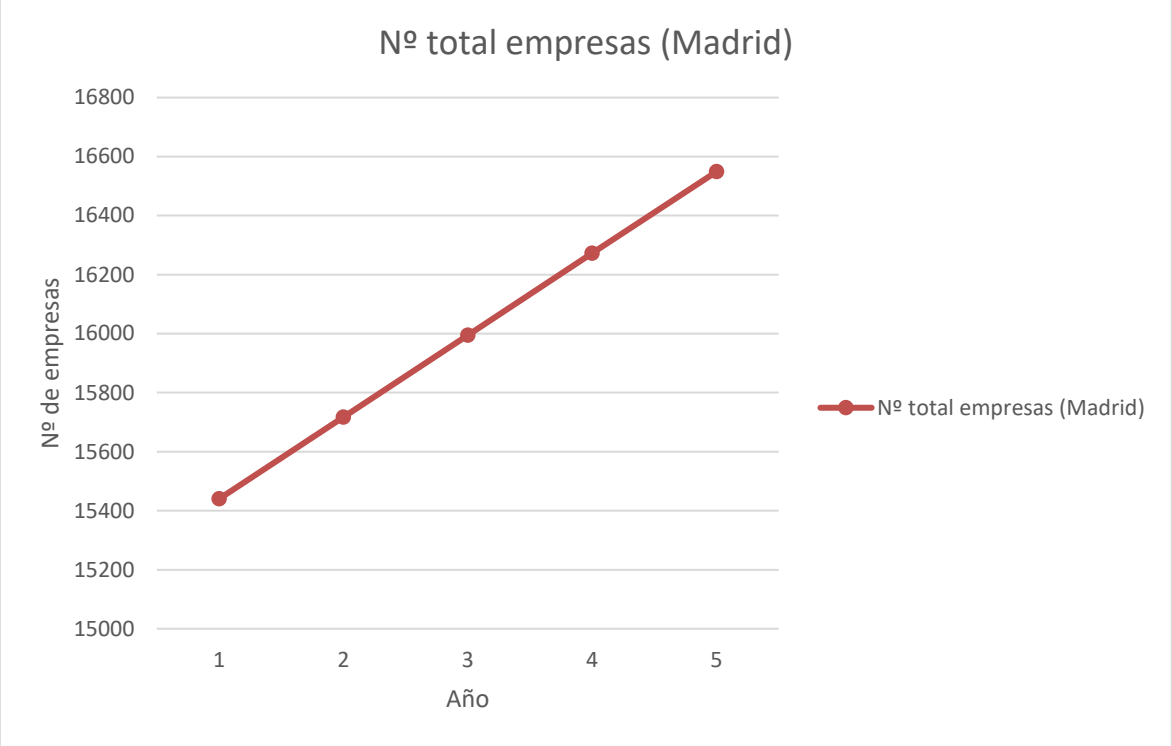


Ilustración 7: Crecimiento de Empresas sector sanitario. Fuente: Propia

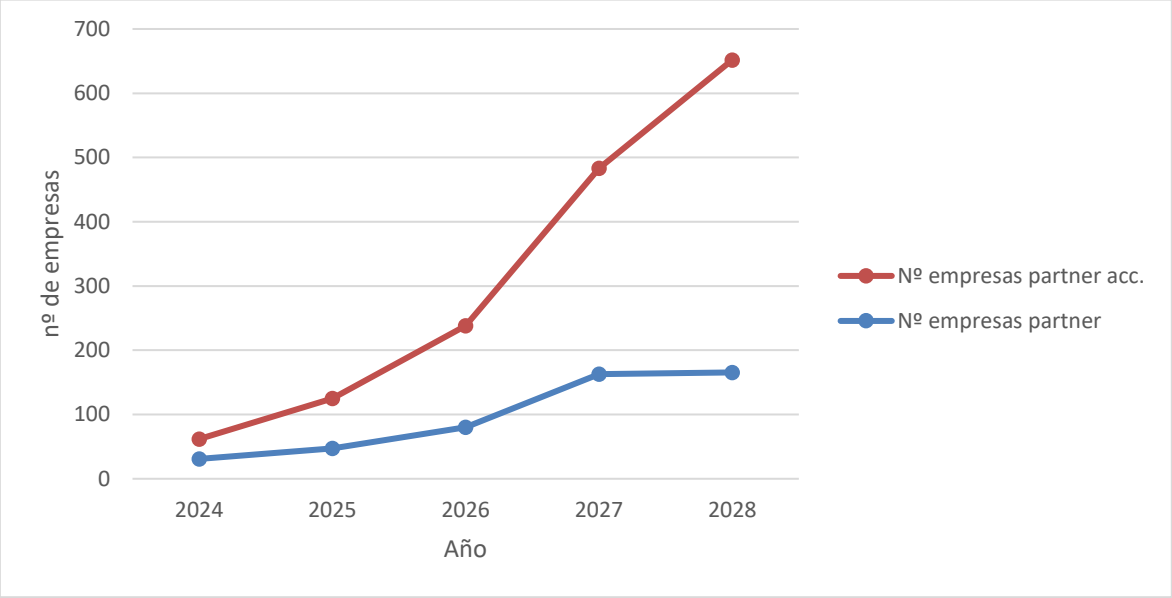


Ilustración 8: Número de empresas partner/proyectos por año. Fuente: Propia

Esto daría lugar a los siguientes proyectos por mes según el año:

COSTES ASOCIADOS DIRECTOS

Dado que se trata de un servicio, *Pedidata* no incurre en costes asociados a materias primas. Sin embargo, para poder realizar el servicio, sí que se incurren en costes asociados a la adquisición de datos y tecnología y software principalmente.
Además, en base al análisis del entorno en el que se sitúa *Pedidata*, sería necesario mantener inversión constante en innovación e investigación de la tecnología empleada para así poder mantener una ventaja competitiva frente a nuevas alternativas o empresas existentes. Por esta razón se deberá elegir una tecnología que permita el escalado fácil. De este modo lograremos mantener el número de servicios alto a lo largo del tiempo.

DATOS

Para el servicio que ofrecemos es necesario tener una fuente de datos continua y fiable. El objetivo es lograr partnerships con las empresas con las que trabajemos, al igual que con Lactapp. Sin embargo, en caso de que se necesitara una base de datos propia para poder elaborar nuestros análisis y sacar las conclusiones pertinentes en un futuro, sería posible obtener los mismos de hospitales, clínicas, centros de salud, bases de datos gubernamentales, bases de datos públicas y organizaciones de salud.

TECNOLOGÍA Y SOFTWARE

Dependiendo del proyecto y de la solución específica que requieran las empresas con las que trabajemos, es probable que no se requieran emplear softwares con licencias de pago.
Dependiendo de la cantidad de datos aportada por el cliente, se podrá emplear los dispositivos locales como almacenamiento para su análisis, pero en caso de ser una cantidad de datos inasumible se pueden emplear plataformas de almacenamiento en la nube como:

- AWS
- Azure
- Google Cloud Platform

Por otro lado, existen numerosas bibliotecas gratuitas y de *open source* para el análisis de datos, sin embargo, si en el futuro queremos desarrollar soluciones que integren ML, pudiendo emplear soluciones de los mismos proveedores de almacenamiento de datos.
Finalmente, si requerimos desplegar nuestras soluciones a los clientes, podemos emplear de nuevo los mismos proveedores de almacenamiento, que también tienen soluciones para el despliegue de herramientas de ML.

PROVEEDORES

Una vez identificados los recursos necesarios, se estudiarán los posibles proveedores de estos recursos a nivel industrial.
Los proveedores mencionados trabajan a nivel global por tanto la localización no se tendrá en cuenta a la hora de seleccionar al proveedor. Sí se tendrá en cuenta, sin embargo, la huella de carbono estimada que generan estas empresas a la hora de seleccionar al proveedor, ya que muchos de estos sistemas requieren de mucha refrigeración para su mantenimiento. Algunos de los proveedores cuentan con sus propios estándares que indican el nivel de CO2 emitido a la atmósfera. En función de la opción se dará una puntuación de 1-5, siendo 1 poca emisión y 5 mucha. Consideramos que es importante que nuestros clientes perciban nuestra marca como una marca con concienciación medioambiental.

PROVEEDORES:

Uno de los problemas al que nos enfrentamos a la hora de seleccionar al proveedor, es que los precios varían mucho en función de la solución que se necesite en cualquiera de los aspectos. En función del crecimiento de la demanda de nuestros servicios se tendrá una estimación de costes de los proveedores u otra, de manera que tomaremos el año en el que tengamos la mayor demanda para seleccionar nuestro proveedor.
No solo eso, sino que influye también el *framework* que propone cada uno de estos proveedores. Estos tienen sus propias soluciones con sus costes asociados para tanto el almacenamiento como para el ML. En función de la escalabilidad que proponen estos *frameworks*, la interacción entre las soluciones y la integración con otras herramientas, se seleccionará un proveedor u otro.
Teniendo en cuenta que el año con mayor demanda será 2028, con unos 14 proyectos mensuales. Asumiendo que el volumen de datos por proyecto es similar al aportado por Lactapp: 1.5GB, se considera que puede ser mayor y rondar los 2.5GB por proyecto:
2.5 GB/proyecto * 14 proyectos/mes = 35GB/mes.
137.5GB/mes * 12 meses = 420GB/año

Considerando este volumen de datos, parece razonable emplear alguna de las soluciones de los proveedores mencionados.

Emplear el almacenamiento en la nube permitiría escalabilidad, accesibilidad, seguridad tanto contra pérdida de datos como contra accesos denegados y recuperación de datos.

Estimaremos que para estos proyectos no será necesaria gran capacidad computacional a la hora calcular los precios (selección computacional). Y que los modelos generados de ML no tendrán un peso superior a 0.1GB suponiendo unas 10h de entrenamiento por proyecto, estos valores se pueden modificar a medida que se escala, pero para tomar una decisión es necesario estipular condiciones. Adicionalmente, se dará una puntuación del 1 al 5 en función de lo buena o mala que sea la integración con el *framework* respectivamente.

Empleando las calculadoras de precios de los servicios obtenemos los siguientes resultados:

Proveedor	C/Almacenamiento de datos	C/Computación y Entrenamiento	C/Almacenamiento de modelos	C/Implementación	Medioambiente	Framework	Columna1
	$/mes	$/mes	$/mes	$/mes	-	-	
AWS	0.84	496.32	0.02	9.08	2	3	
Azure	0.66	596.2	0.02	1.17	1	4	
GCP	0.61	495.49	0.03	6.15	2	1	504.72

Tabla 3: Coste por servicio y proveedor. Fuente: Propia

Una vez estimados los costes, se normalizan para poder seleccionar el mejor proveedor de acuerdo con los factores decididos. De este modo, el proveedor con el valor total normalizado menor será el seleccionado:

Proveedor	C/Almacenamiento de datos	C/Computación y	C/Almacenamiento de modelos	C/Implementación	Medioambiente	Framework	total

		Entrenamiento					
AWS	1.0	0.8	1	1	2	3	8.8
Azure	0.8	1.0	1.0	0.1	1	4	7.9
GCP	0.7	0.8	1.5	0.7	2	1	6.7

Tabla 4: Coste por servicio normalizado. Fuente: Propia

PRECIO Y POLÍTICA

PRECIO

Es bastante complicado establecer un precio para un servicio como este, al fin y al cabo, buena parte de los costes serán gastos asociados al resto de consultores/analistas que formen parte de los proyectos.

Los competidores como IBM, Oracle y Philips ofrecen servicios especializados, cuyo valor es difícil de estimar, ya que depende exclusivamente de la solución que se necesite para cada caso particular. Por tanto, el precio se estimará como servicio de consultoría/análisis de datos.

Algunas fuentes indican que las consultoras estratégicas de entre 10 y 40 empleados, establecen su precio por hora en torno a los 100-250eur/h (Torres, 2022). Suponiendo que, como se ha mencionado antes los proyectos a realizar son de reducido tamaño, de entre 3 a 6 semanas con jornadas laborales de 8h/día y 5 días laborables, podemos establecer que el precio por proyecto de este tipo sería de entre 5000 y 10.000eur. Otras fuentes indican que el precio por un servicio de consultoría Big Data está más próximo a los 200-350eur/h. Haciendo el mismo cálculo, el precio por proyecto de este tipo sería de entre 8000 y 14.000eur. (webfx, s.f.)

Una buena manera de estimar el precio por hora para un servicio de consultoría comienza con una estimación de los gastos fijos. Incluyendo licencias, cuotas de autónomo o mutualidad, gestoría, colegiación, responsabilidad civil, amortización de equipos, etc. A esto se le suman los costes variables, que podemos estimar a priori como un 30%. También hay que tener en cuenta sueldos de trabajadores y cuotas de seguridad social, así como los impuestos. Si esto lo partimos de las horas que estimamos trabajar anuales, sabiendo que el año habitual tiene 365 días; si le restas 52 fines de semana, 14 festivos y 22 días de vacaciones te quedan 225 días. Si consideras jornadas de 8 horas diarias, obtienes 1.800 horas/año, obtenemos un umbral mínimo de precio/h. Asumiendo actualmente unos costes fijos de 505eur/mes por licencias, que para el primer año no se necesitarían más trabajadores, que el salario será de unos 30.000eur que la cuota de autónomos es de 1.148eur, sumándole a esto un 20% por el resto de costes y otro 30% por los costes variables llegamos a una cifra de unos 60.000eur anuales, o 85.000 después de impuestos. Estos cálculos son cualitativos y poco representativos y lo más probable es que sea necesario

contratar a al menos un equipo de más de 10px para la carga de trabajo del 4º y 5º año (14 proyectos * 12 meses * 4 semanas * 5 días*8h=26.880h de trabajo, 26880/1800=15px).
De este modo podemos estimar que la cantidad de consultores necesarios es:
(15*12*4*5*8)/(1800)=15px
Con este número de empleados, asumiendo un coste de salarios anuales de unos 30.000eur, un alquiler de unos 750eur/mes, costes de licencias de 505eur/mes, costes de seguridad social de 10.000eur/px, un 30% de costes variables y un 40% de impuestos llegamos a una cifra de aproximadamente 1.120.410 eur. Asumiendo que el número de horas disponibles será de 15*1800, obtenemos un umbral mínimo de unos **42eur.**

Estos valores se analizarán en detalle más adelante, de este modo se obtendrán cifras de carácter cuantitativo, pero nos permite establecer unos precios base con los que valorar nuestra situación en el mercado.

POLÍTICA

En cuanto a la política de precio, se ha decidido marcar un precio en función principalmente de los costes en los que se puedan incurrir, especialmente en el año de mayor demanda. Sin embargo, para hacer esto adecuadamente se ha tenido también en cuenta el precio que marca el mercado, definiendo un margen viable para obtener beneficios sin perder ventajas competitivas

CÁLCULO DEL P*Q

Teniendo en cuenta la previsión de ventas de proyectos calculada con anterioridad y que se muestra en - *Tabla 2: Proyectos por mes en los primeros 5 años. Fuente: Propia* -, así como el precio estimado por hora por proyecto, asumiendo que las cifras de proyectos que podríamos asumir no pueden superar los 15 proyectos, con un precio por hora de 84eur/h y un tiempo estimado de 4 semanas por proyecto se pueden calcular los ingresos anuales como:

Año	proyectos anuales	beneficios anuales
2024	36	483.840
2025	48	645.120
2026	84	1.128.960
2027	168	2.257.920
2028	168	2.257.920

Tabla 5: Beneficios anuales proyectados. Fuente: propia

Cabe destacar que se ha utilizado una moneda constante sin incluir valoración por inflación en este cálculo.

PLAN DE VENTAS: EVOLUCIÓN EN 5 AÑOS

Una vez realizado el análisis de mercado externo e interno y el potencial del producto, se recopila el análisis con la información relativa a las ventas que se ha obtenido en este módulo de factibilidad comercial.

En primer lugar, se concluye que se pretende obtener un 3% de *market share* en los primeros 5 años,

Año	% de MS	% acc de MS	Nº empresas anual partner	Nº de empresas acc partner
1	0.2	0.2	36	36
2	0.3	0.3	48	84
3	0.5	0.5	84	168
4	1	1	168	336
5	1	1	168	504

Tabla 6: Porcentaje de MS y Número de partners por año de proyecto. Fuente: Propia

Por otro lado, teniendo en cuenta los costes asociados a proveedores, alquileres, salarios y cuotas, así como un 30% de costes variable, concluimos que para un equipo de 15px, el umbral mínimo de precio por hora que se requeriría para no entrar en pérdidas sería de 42Eur/h. Asumiendo un beneficio de 200%, el precio por hora se establecerá en 84Eur/h.

Con la selección de precios y la estimación de ventas se pueden calcular los ingresos esperados por los proyectos realizados, reflejado en *Tabla 5: Beneficios anuales proyectados. Fuente: propia.*

MÓDULO II –

FACTIBILIDAD TÉCNICA

INTRODUCCIÓN

Una vez analizado el entorno comercial, se concluye que existe una demanda por un servicio como el propuesto, con una tendencia creciente en la adopción de las tecnologías propuestas, y la posibilidad de ofrecer una solución ad-hoc que se ajuste a las necesidades específicas de cada proyecto.

Sin embargo, para definir el servicio de manera más precisa, se estudiará la viabilidad técnica del mismo.

En esta sección se definirán con detalle los procesos y entregables comunes a todos los proyectos, estos definirán el proceso que se siga para el servicio proporcionado. También se hará referencia a los materiales, licencias, y útiles necesarios para el correcto desempeño laboral de la *start-up*.

Es en esta sección en la que se emplearán los datos obtenidos de Lactapp para realizar un análisis. Como se ha mencionado anteriormente, Lactapp Medical es la app que acompaña a los profesionales que atienden a madres lactantes. Ofrece soluciones autónomas y resuelve dudas de lactancia y maternidad de manera personalizada (Lactapp, s.f.). Ofrecen formación, información, canales de consulta y *trackers* para el monitoreo de lactantes a profesionales de la salud en todo el mundo.

El objetivo será ilustrar el tipo de servicio que se realizará en *Pedidata*, incluyendo el contacto con la empresa y definición de objetivos, la adquisición de los datos, el estudio de la estructura y contenido de los datos, así como su integridad y validez, y su análisis.

Para el análisis de viabilidad técnica, también se estudiará la mano de obra necesaria para la estimación de proyectos que se realizarán a lo largo de los primeros 5 años

SERVICIO

DESCRIPCIÓN Y PROCESO

Como ya se ha ido mencionando, el servicio que se pretende prestar es un servicio de consultoría/análisis de datos. Con el objetivo de estandarizar el servicio, para lograr los objetivos establecidos por el cliente con el mejor rendimiento posible, se definen las distintas fases en las que se dividirán los proyectos. Dado que, dependiendo del cliente, la solución escogida será un análisis más profundo, con resultados empleando ML, o un estricto análisis de los datos, sobre el que tomar futuras decisiones, las fases de cada proyecto se dividirán en 2 bloques.

.

- **Bloque Análisis**

Servicio de análisis de datos en el que se buscará entregar un informe detallado, con visualizaciones y análisis de los datos de los que se extraerán patrones, tendencias y anomalías. Además, se realizará una definición de objetivos para futuras fases. El objetivo es que los hallazgos permitan a nuestros clientes proporcionar recomendaciones e información más detallada a sus clientes en cuanto al estado de salud, nutrición u otros.

<u>En primer lugar, se iniciará el proyecto, definirá su alcance y elaborará el plan de proyecto:</u>

1. El primer paso será el inicio del proyecto y la definición de objetivos. En este paso será necesario realizar una reunión con el cliente de manera que se puedan comprender las necesidades y expectativas del cliente. Se deben identificar los problemas del cliente, definir los objetivos del proyecto y los resultados esperados y recopilar la información relevante sobre los datos disponibles y su formato.
2. Posteriormente se deberá definir el alcance del proyecto. Se deben establecer claramente el alcance del trabajo para evitar malentendidos y crear expectativas realistas. Habrá que definir las tareas del proyecto e identificar los recursos necesarios y los entregables del proyecto.

3. Finalmente se elaborará el plan de proyecto: En este se documentarán los aspectos clave del proyecto de forma detallada. Esto permitirá proporcionar un marco de trabajo con una estructura clara tanto para el equipo como para el cliente. Se creará un documento que incluya los objetivos, alcance, recursos y entregables. Además, servirá para establecer un posible cronograma y agendar reuniones de seguimiento de proyecto.

En segundo lugar, se realizará la recopilación y procesamiento de datos:

1. La primera acción será recopilar los datos del cliente en el formato acordado. Permitiendo que los datos requeridos estén disponibles y accesibles. En esta fase, será importante verificar la integridad de los datos aportados.
2. Una vez adquiridos los datos, se estudiarán para identificar posibles problemas de calidad, de este modo se garantizará que los datos son adecuados para su análisis y modelado (si es que se requiere). Será necesario realizar un análisis exploratorio inicial para evaluar y detectar valores faltantes, duplicados, errores y otros. Este análisis también permitirá constatar la consistencia de estos. Es en este momento en el que se podrá agendar alguna reunión adicional para comentar hallazgos en los datos con el cliente, así como aclarar cualquier duda que surja sobre la estructura de los mismos.
3. Finalmente, se limpiarán los datos para poder procesarlos en el análisis. En función de los objetivos y la estructura de los datos, se deberá tratar valores nulos como imputación o eliminación, codificar variables de categoría, normalizar o escalar variables numéricas, crear nuevas variables relevantes, etc.

Finalmente, se realizará el Análisis exploratorio de los datos (EDA):

1. Se comenzará por un análisis descriptivo, analizando características básicas de los datos. Esto permitirá tener una comprensión inicial de los datos y su distribución.
2. Después se procederá con la visualización de los datos, creando visualizaciones que ayuden a identificar patrones y relaciones en los datos, así como visualizar la distribución de las variables iniciales. Es aquí donde se comenzará la relacionar las variables con

gráficos de dispersión, diagramas de correlación o gráficos temporales para comprobar si existen relaciones significativas entre las variables.

Este será el entregable que concluirá el proyecto en caso de que la solución que solicita el cliente sea un análisis a alto nivel.

- **Bloque ML**

En caso de que la solución solicitada por el cliente sea más avanzada y requiera el empleo de modelos de *ML* que deban ser desplegados, las siguientes fases serán:

Posterior a la elaboración del informe EDA, se desarrollarán los modelos de ML:

1. Lo primero será seleccionar los algoritmos de ML más adecuados según los objetivos del proyecto y las características de los datos. Entre los más típicos se encuentran árboles de decisión, regresiones lineales, redes neuronales, etc. (Na8, 2017)
2. Posteriormente se entrenarán los modelos seleccionados y se valorará su rendimiento, empleando técnicas como la valoración cruzada para evaluar la robustez de los modelos seleccionados.
3. Finalmente se evaluará el rendimiento de los modelos empleando las métricas adecuadas en cada caso. Métricas como el F1-score, ROC-AUC u otras. (Díaz, s.f.). Este proceso se documentará y se seleccionará el modelo que ofrezca el mejor equilibrio entre rendimiento y eficiencia.

Con el modelo seleccionado, el siguiente paso será el despliegue para el cliente:

1. El primer paso será integrar los modelos de ML en las plataformas de los clientes. Para ello será necesario crear APIs con una arquitectura de microservicios para poder desplegarlo y escalarlo. En este paso será importante testear las APIs para asegurar su correcto funcionamiento.
2. Una vez creada la integración, lo siguiente será desplegar la solución. Esto supone desplegar los modelos en un entorno de producción. En nuestro caso, el entorno seleccionado será el de la nube de Google, ya que, en comparación con los otros servicios principales de alojamiento en la nube y ML, éste es el más competitivo en cuanto a precios

y tiene el *framework* más integrable. Además, este servicio permite escoger las localizaciones en las que se alojan los servidores con indicadores de co2 bajos (Google, s.f.)

Por último, con el modelo desplegado, se capacitará al cliente:

Será necesario tener el proceso bien documentado, de este modo se podrá capacitar al cliente para el uso de la solución.

PRODUCTO MÍNIMO VIABLE (PROYECTO LACTAPP)

En esta sección se va a realizar el análisis de datos a Lactapp medical. Para ello se seguirá el proceso establecido con anterioridad. De este modo se planteará el análisis como un servicio.

INICIO DEL PROYECTO

En primer lugar, se da inicio al proyecto. Se contacta con Lactapp Medical, una App que acompaña a los profesionales cuya labor es atender a las madres lactantes, ofreciendo información relevante, formación, y monitorización con la ayuda de una aplicación que permite *trackear* aspectos relevantes a la hora de las tomas.
En las conversaciones con Lactapp, se estudian sus necesidades. Consideran que tienen información relevante en cuanto a las tomas de las madres, pero esperan poder aprovechar los *trackers* más a fondo, estudiando posibles relaciones entre estos que permitan establecer tendencias que indiquen problemas que puede llegar a tener la madre, el/los bebés o ambos.
Entre los objetivos se definen:

- Estudiar la relación que existen entre las tomas y los problemas que puede tener la madre
- Explorar la relación entre tomas y la edad del bebé
- Explorar la relación entre tomas y si el bebé es primer, segundo o tercer hijo
- Investigar la relación entre ciertos *trackers* y la Región
- Investigar la relación entre los *trackers* y la situación laboral

Los resultados esperados serán las relaciones o tendencias con las que se pueda llegar a predecir los problemas que se encuentran reflejados en los *trackers* y que afectan a las madres. Además, se estima que, por el volumen de datos y la calidad de los mismos, con un análisis exploratorio de los datos, sea suficiente para encontrar tendencias y patrones claros que permitan predecir posibles resultados futuros, por lo que el empleo de ML podría ser redundante.
En cuanto al alcance del proyecto, una vez establecidos los objetivos, se pueden definir las tareas y subtareas que serán necesarias para poder alcanzarlos. Además, se identificarán las

herramientas a emplear para la correcta realización de estas. Finalmente se podrá establecer una hoja de ruta para llegar al objetivo final.

En relación con los objetivos, podemos establecer las tareas a realizar desde el momento en que comience el proyecto como:

1. Recolección y entendimiento de los datos: El primer paso será conseguir que Lactapp proporcione los datos necesarios para realizar el estudio. Será importante que proporcionen toda la información necesaria para la comprensión de los datos y de su relación y estructura. Esto es clave para poder llegar a entender los datos y así establecer las hipótesis de relaciones que pueden llegar a existir entre ellos y que confirmen las tendencias esperadas por el cliente.
2. Verificación y limpieza de los datos: Una vez tengamos los datos en posesión, será necesario realizar un análisis preliminar de los datos, de esta forma sabremos la integridad y consistencia de estos, y nos permitirá limpiar/filtrar los datos de acuerdo con nuestras necesidades.
3. Una vez limpios, será necesario estructurarlos para darles el formato adecuado para el análisis, recogiendo en muestras significativas los datos necesarios para estudiar las estadísticas relevantes que nos den la información necesaria para establecer las relaciones que buscamos.
4. Con los datos formateados, se comenzará el análisis exploratorio (EDA), lo primero será estudiar los *trackers* que consideramos más relevantes para el objeto de estudio, describiendo estadísticas básicas de estos y creando visualizaciones iniciales de los datos
5. Una vez comprendidos más a fondo los datos, se puede comenzar a buscar tendencias, patrones y anomalías entre los datos que tengan relación con los objetivos establecidos inicialmente por el cliente.
6. Se realizarán hipótesis preliminares sobre las relaciones, y se verificarán con el estudio y relación de los datos. Se emplearán visuales como gráficos de dispersión, de calor, estudio de las estadísticas básicas, etc y se modelarán con algoritmos como regresiones lineales para identificar posibles tendencias.

7. Una vez concluido el EDA, se realizará un informe para el cliente, incluyendo la documentación de los pasos seguidos para la obtención de los resultados, y las relaciones y tendencias detectadas en los datos, que concuerden con los objetivos iniciales, empleando los visuales realizados durante el análisis.

Con la correcta realización de estas tareas, se dará por terminado el proyecto. El modelado de modelos algorítmicos de ML quedaría por tanto fuera del alcance de este proyecto en este caso. Con esta definición del alcance se espera poder evitar malentendidos durante y tras el desarrollo del proyecto con el cliente. Será importante mantener una línea de comunicación constante con el cliente para asegurarse de su satisfacción en todo momento, de esta manera demostraremos que lo que nos diferencia de las soluciones existentes en el mercado, es la personalización de nuestras soluciones.

Con el alcance del proyecto definido, se procede a realizar el Plan de Proyecto. Esta es una herramienta que nos ayudará tanto al cliente como a nosotros la gestión y realización del proyecto, es una **"alineación estratégica de todos los elementos que requieres para lograr un proyecto"** (Santos, 2024). En este se incluirá:

1. Visión general
2. Alcance
3. Desarrollo de objetivos
4. Roles y responsabilidades
5. Entregables
6. Calendario
7. Presupuesto

No se incluirá el Plan de Proyecto completo, ya que, en la definición del proceso, ya se incluyen algunos de los puntos del Plan, como la visión y alcance, que ya se han detallado.

En cuanto al desarrollo de los objetivos, estos ya se mencionan con anterioridad, y cumplen con la clasificación SMART.

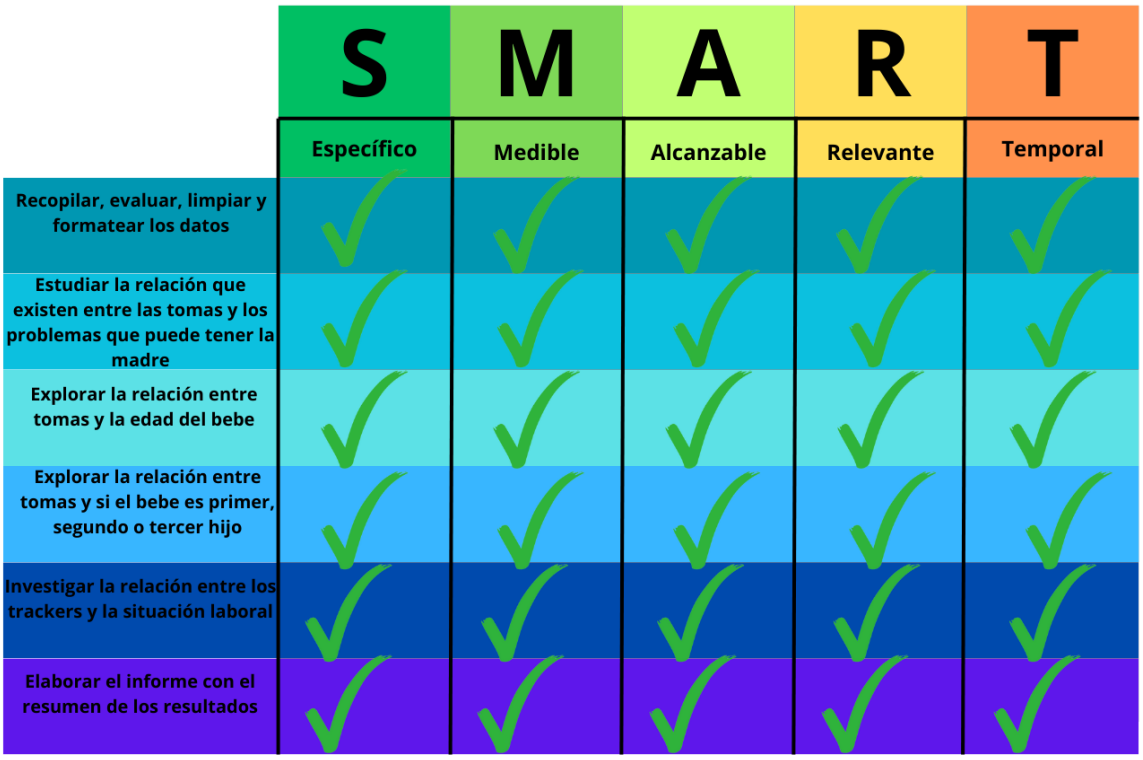


Ilustración 9: Objetivos del proyecto. Fuente: Propia

Podemos decir que todos estos objetivos cumplen con la metodología SMART, ya que son específicos, medibles, alcanzables, relevantes y limitados en el tiempo.

Roles y responsabilidades:

Dado que, para este proyecto, solo el autor del TFM está encargado de realizar el análisis, consideraremos que tanto el mando como el flujo del proyecto dependen exclusivamente de este, por tanto, no es necesario el establecimiento de roles y responsabilidades, ya que recaen únicamente sobre el autor.

Entregables:

Los entregables irán asociados a las tareas establecidas en el alcance del proyecto. De esta manera se establecen los entregables como:

1. Informe preliminar de los datos: mediante este informe se pretende verificar con el cliente la validez, integridad y consistencia de los datos

2. Documentación del proceso de formateo: de esta forma ofreceremos total transparencia con el cliente en cuanto al tratamiento que se le ha dado a los datos a la hora de prepararlos para el análisis
3. Informe final del EDA: este será el último entregable, con las conclusiones del análisis.

Calendario:

Para establecer el calendario del proyecto será necesario establecer la duración de éste. Consideraremos de acuerdo a lo mencionado en otros módulos que el proyecto se realizará en unas 4 semanas, desde el inicio del proyecto hasta la entrega del informe y fin del proyecto, de esta forma establecemos un diagrama de Gantt con los hitos, tareas y entregables establecidos anteriormente.

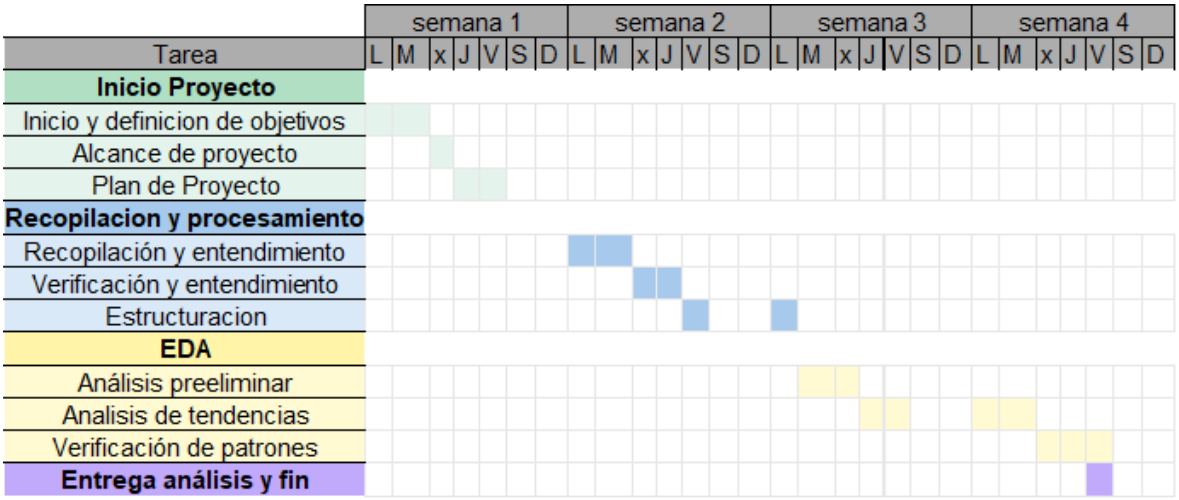


Ilustración 10: Diagrama de Gantt del proyecto. Fuente: Propia

Presupuesto:

Dado el volumen y la calidad de los datos que son aportados, así como el hecho de que no será necesario el empleo de ML para la solución, no será necesario el pago de licencias para la realización del proyecto. El presupuesto se podrá considerar como la cantidad de tiempo invertida en el proyecto en este caso. Para el proyecto se estimará en 3-4 semanas, que con un precio de 84 euros la hora, deja el presupuesto en unos 13.500 euros (4 semanas).
Con estos aspectos estudiados, daríamos por finalizado el Plan de Proyecto, y por tanto se podría dar inicio al proyecto.

RECOPILACIÓN Y PROCESAMIENTO DE DATOS

RECOPILACIÓN Y ENTENDIMIENTO

Con las bases del proyecto establecidas, y una vez realizado el Plan, se puede proceder con la recopilación y entendimiento de los datos.

Los datos son proporcionados por Lactapp, el formato de los datos entregados es un archivo csv con casi 7 millones de entradas y 35 columnas.

Adicionalmente, Lactapp entregó un documento interno con la explicación del significado de los datos, así como con el proceso de formación del archivo. A simple vista podemos comprender el significado de algunos de estos trackers, como el trackID, userID, city, bornDate y otros, sabiendo esto y con la ayuda del documento proporcionado podemos entender cada uno de estos trackers, y estimar el contenido que puede haber en cada uno (tipo de dato). *Anexo B – Documentación del archivo de datos de Lactapp*

Además, dado que el documento es un csv, podremos abrirlo para ver el tipo de datos que nos podemos encontrar, así como su formato.

1. trackId: nº identificador de la entrada, entenderemos cada entrada como una toma de leche. [int]
2. userID: nº identificador único para cada usuario (madre) [int]
3. baby1ID: nº identificador único para cada bebé (si se dan tomas a más de un bebé y no se da en tándem, el segundo bebé tendrá otro identificador único para el mismo usuario). [int]
4. baby2ID: nº idententificador único para el segundo bebé en caso de que la toma se de en tándem, solo soporta 2 bebés. [int | null | 0]
5. userName: nombre del usuario (madre) (censurado por privacidad)
6. bornDate: fecha de nacimiento del usuario [date (mm/dd/yyy) | null]
7. city: ciudad de alojamiento del usuario [string | null]
8. numSons: nº de hijos del usuario [int | null]
9. dueDate: fecha de salida de cuentas [date (mm/dd/yyy) | null)]
10. babyName: nombre del bebé (censurado)
11. genre: genero del bebé, puede ser femenino ‘F’ o masculino ‘M’ [string] (asociado a baby1)

12. bornDate.1: fecha de nacimiento del bebé [date (mm/dd/yyy)] (asociado a baby1)
13. premature: Si el bebé nació de forma prematura. Si nace de forma prematura = '1' si no nace de forma prematura = '0' [int | bool] (asociado a baby1)
14. gestationWeek: semana de gestación en la que nació el bebé [int | null | '0'] (asociado a baby1)
15. bornHeight: longitud / altura del bebé al nacer [float | null] (asociado a baby1)
16. babyTandemName: nombre del segundo bebé al dar la toma en tándem (censurado)
17. babyTandemGenre: genero del segundo bebé, puede ser femenino 'F' o masculino 'M' [string] (asociado a baby2)
18. babyTandemBornDate: fecha de nacimiento del segundo bebé [date (mm/dd/yyy)] (asociado a baby2)
19. babyTandemPremature: Si el bebé nació de forma prematura. Si nace de forma prematura = '1' si no nace de forma prematura = '0' [int | bool] (asociado a baby2)
20. babyTandemGestationWeek: (censurado)
21. babyTandemBornHeight: longitud / altura del segundo bebé al nacer [float | null] (asociado a baby2)
22. startDate: fecha comienzo de la toma [datetime (mm/dd/yyyy hh:mm:ss)]
23. endDate: fecha fin de la toma [datetime (mm/dd/yyyy hh:mm:ss) | null]
24. breast: Pezón con el que se da la toma, puede ser el izquierdo, derecho ambos o desconocido [string]
25. suplBreastmilk: cantidad de leche materna suplementaria, puede ser desconocida [float | string]
26. suplFormula: cantidad de leche de formula suplementaria, puede ser desconocida [float | string]
27. painLevel: nivel de dolor en la toma, puede ser del 0-10 o desconocida [int | string]
28. nipleShape: forma del pezón de la toma. Puede ser tipo 'lipstic', redonda o desconocida [string]
29. nipleColor: color del pezón en la toma. Puede ser normal, morado, blanco o desconocido [string]

30. milkPearl: indica la presencia de ampollas de leche, estas pequeñas y dolorosas manchas blancas que se producen por la obstrucción de un poro de leche (clinic, 2024). Si existen ampollas = ‘1’ si no existen = ‘0’, también puede ser desconocida [int | bool | string]
31. cracks: indica la presencia de grietas en el pezón. Si existen grietas = ‘1’si no existen = ‘0’, también puede ser desconocida [int | bool | string]
32. emotionalStatus: indica el status emocional de la madre. Puede ser calmado, contento, triste, cansado, frustrado o una combinación de ellas o desconocida [string]
33. notes: notas de la toma [string | null]
34. creationDate: fecha de creación de la entrada [datetime (mm/dd/yyyy hh:mm:ss) | null]
35. modificationDate: fecha de modificación de la entrada [datetime (mm/dd/yyyy hh:mm:ss)]

Con toda esta información, así como con la app de la que se extraen los datos, podemos tener una comprensión exhaustiva de tanto la procedencia de los datos como de sus características y relaciones. Con esto en mente podemos pasar a la verificación de los datos.

VERIFICACIÓN DE LOS DATOS, ANÁLISIS PRELIMINAR

El objetivo del análisis preliminar es verificar la integridad y consistencia de los datos con respecto a la fuente, así como verificar que no hay datos faltantes, duplicados u otros. Durante este análisis también podremos extraer información adicional de los datos que queremos emplear y que están relacionados con las tendencias que buscamos identificar. Sabiendo el tipo de datos que podemos encontrarnos en cada columna y la relación que debería existir con el resto de los datos, podremos limpiarlos, ya sea por proceso de eliminación, filtrado o normalización.
En este análisis también podremos estudiar los datos en más detalle identificando valores relevantes (máximos, mínimos, medias, medianas, etc) que nos puedan ayudar de cara a su estructuración.

Para el análisis preliminar y posteriormente el EDA, será necesario instalar un entorno IDE, Python como lenguaje de programación, jupyter notebook por la interfaz que presenta y la facilidad con la que se puede analizar el código, y las librerías de pandas, numpy, matplotlib,

seaborn. Estas librerías facilitarán la manipulación y análisis de datos y permitirán realizar gráficos con interfaces visuales potentes.

Con las librerías instaladas en el entorno IDE se procede a verificar los datos.

Entre los datos, nos encontramos con columnas censuradas por protección de los clientes de Lactapp, estos datos se consideran irrelevantes para el estudio en cuestión y por tanto se eliminan de la base. Adicionalmente, encontramos numerosas inconsistencias en los datos, pero especialmente en campos como ‘city’ o ‘notes’, estos campos pueden ser relevantes para el estudio en cuestión pero dadas las inconsistencias (casi 4000 valores únicos en city, entre los que encontramos erratas, cambios de capitalización, etc.), debidas a que la madre es la que pone el valor en estos campos, se considera que no se pueden emplear en el análisis. Además, se analizan campos que se consideran de crucial importancia en el estudio, como los identificadores únicos así como las fechas de inicio y fin de las entradas, ya que serán importantes para estudiar las tomas. Encontramos que existen filas con valores nulos en alguno de estos campos y se eliminan estas filas.

Una vez realizados estos cambios, se analizan el resto de las columnas, tratando de evitar la eliminación para conservar el mayor número de datos posible en la muestra, pero realizando los cambios necesarios para poder filtrar los valores en pasos futuros.

ESTRUCTURACIÓN, FORMATEO DE LOS DATOS Y EDA

Una vez verificados los datos, se procede a su estructuración. El objetivo es conseguir una estructura y formato de los datos que nos pueda ayudar a identificar las tendencias propuestas en los objetivos. Adicionalmente, se deberán crear nuevas variables para datos que sean necesarios para el análisis, así como de muestras específicas de los datos donde podamos identificar estas relaciones y tendencias. Con los datos estructurados se procederá a crear los visuales que nos permitan comprender más en profundidad la relación entre los datos.
Durante la verificación se detectan inconsistencias en la calidad de los datos, especialmente en algunas columnas críticas de cara a los objetivos establecidos anteriormente.
Concretamente, en la columna 'city', encontramos unos 3984 valores únicos, entre los que se encuentran múltiples erratas y caracteres desconocidos. Esta columna es necesaria tanto para establecer la relación entre las tomas y el país del usuario, como para establecer una relación entre las tomas y la situación laboral de la madre. Esto es porque, al no tener un tracker de esta, la manera más razonable de establecer la relación entre estos valores sería estudiando la diferencia entre la primera entrada de un usuario único y las consecuentes para saber si hay cambios a partir del 4º mes, pero eso solo aplicaría en el caso de España, y de nuevo, para saber la región sería necesaria la columna 'city'. La única manera de limpiar los datos sería establecer variables concretas (ciudades) y asignar el valor de la variable concreta a aquellos valores únicos que se presuman que coinciden con la variable así creando la variable 'BCN' podríamos agrupar los valores únicos que consideramos representan Barcelona entre los valores de la columna city que sean erratas.

Teniendo en cuenta los objetivos:
Identificar la relación que existe entre las tomas y los problemas que tiene la madre
Según los *trackers*, los principales problemas que puede tener la madre son:
Nivel de dolor en la toma.
Forma del pezón al terminar la toma: el pezón en forma de pintalabios puede indicar un agarre incorrecto del bebé al pecho lo que puede causar molestias y afectar a la eficiencia de la lactancia. (RUIZ, 2023).

El color del pezón en la toma o al terminar: un color morado puede ser símbolo de isquemia, que es una reducción del flujo sanguíneo en el pezón. Esto puede causarse por una mala técnica de agarre del bebé. También puede estar relacionado con el síndrome de Raynaud, que puede causar espasmos vasculares debido al frío o estrés. Por otro lado, un color blanquecino, puede indicar una condición conocida como 'perla de lactancia' ya mencionada con anterioridad, y que se provoca por la obstrucción de los conductos galactóforos. Otra posible causa es la isquemia, que puede deberse a un mal agarre o compresión del pezón, y también causa dolor. (Mónica Machicado Sotoca, 2018).

Perlas de lactancia: provocadas por la obstrucción de los contactos galactóforos, puede causar dolor en las tomas debido a una succión incorrecta del bebé.

Será necesario aislar las muestras de los datos que sí son conocidas para cada uno de estos problemas.

Para este caso en primer lugar se estudia la posible relación entre la duración de las tomas y el nivel de dolor. En primer lugar, se crea la variable duración, como la resta entre startDate y endDate, como se observan resultados anómalos, a la hora de realizar los visuales se filtran los datos para evitar datos anómalos. Se encontraron valores de duración que eran negativos, así como valores positivos superiores a los 3.000 minutos. Se podría asumir que los valores positivos altos son debidos a que la madre inicia la toma en la aplicación, pero se le olvida pararla, por tanto se filtraran las tomas por valores que superen las 2h como umbral superior. Por otro lado, se encuentran valores negativos de duración. Esto es una imposibilidad, y posiblemente se deba a un error en el registro del dato, por ello se filtra por valores de duración positivos.

Una vez creada la variable y filtrado el conjunto de datos, se relaciona la variable duración con el nivel de dolor. Tras estudiar los resultados, concluimos que existen una gran cantidad de valores atípicos. En base a esto se decide filtrar aún más el conjunto, eliminando los valores desconocidos ya que son irrelevantes para el estudio. Una vez filtrados, se crea un histograma con cada nivel de dolor, para tratar de aislar una tendencia entre el nivel de dolor y la duración de las tomas.

De nuevo, al analizar los resultados, se observa que la tendencia de dolor a lo largo del tiempo de toma no varía significativamente entre los distintos niveles de dolor, lo que sugiere que contrario a lo que se esperaba observar originalmente, las tomas no varían mucho en duración dependiendo del nivel de dolor.

Una vez analizadas estas variables, se procede a analizar si existe relación entre alguna de las variables que indican problemas en la toma y estas dos previamente analizadas.
Para estudiar estas variables, en primer lugar, se filtran los valores desconocidos, al ser irrelevantes para este estudio, y se emplea el mismo filtro en la duración de las tomas que el empleado anteriormente (0 a 120 minutos).
Adicionalmente se ordenan los valores de la variable de dolor para poder observar con mayor claridad los resultados.
Se obtienen los siguientes resultados relevantes:

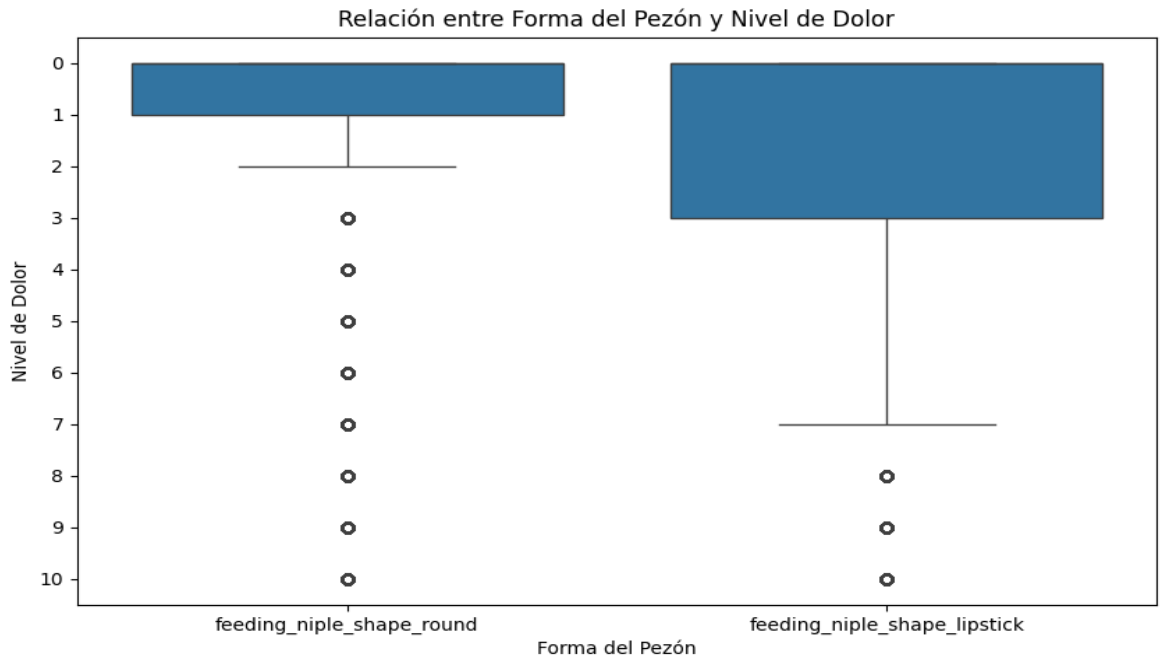


Ilustración 11: Relación entre forma de pezón y nivel de dolor. Fuente: Propia

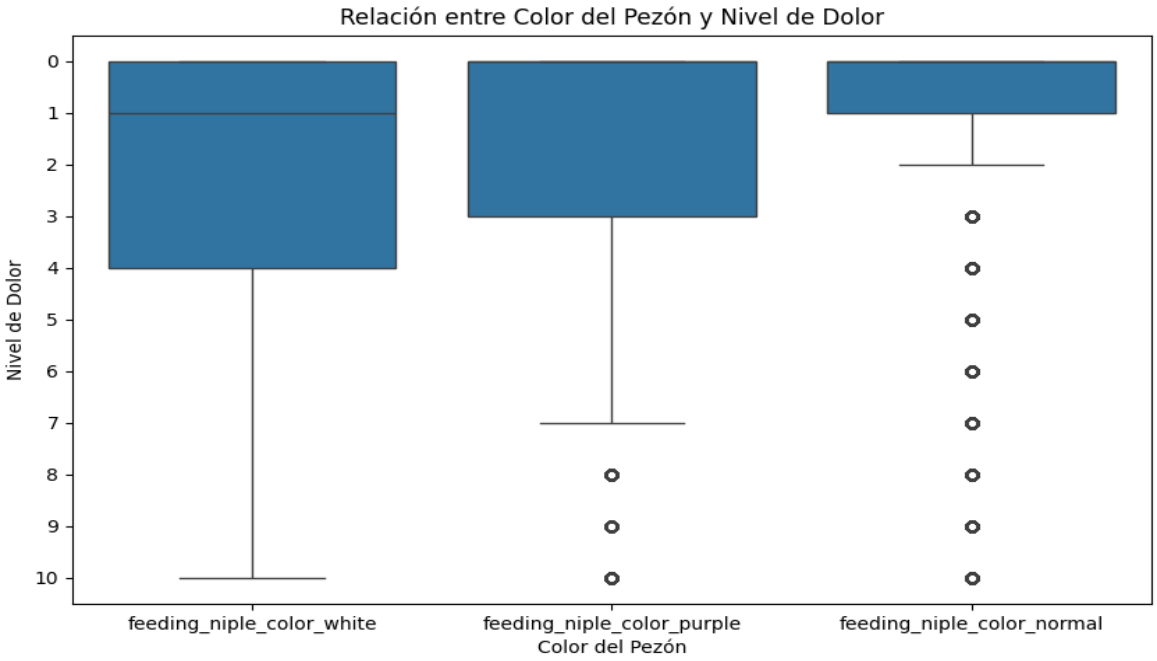


Ilustración 12: Relación entre color de pezón y nivel de dolor. Fuente: Propia

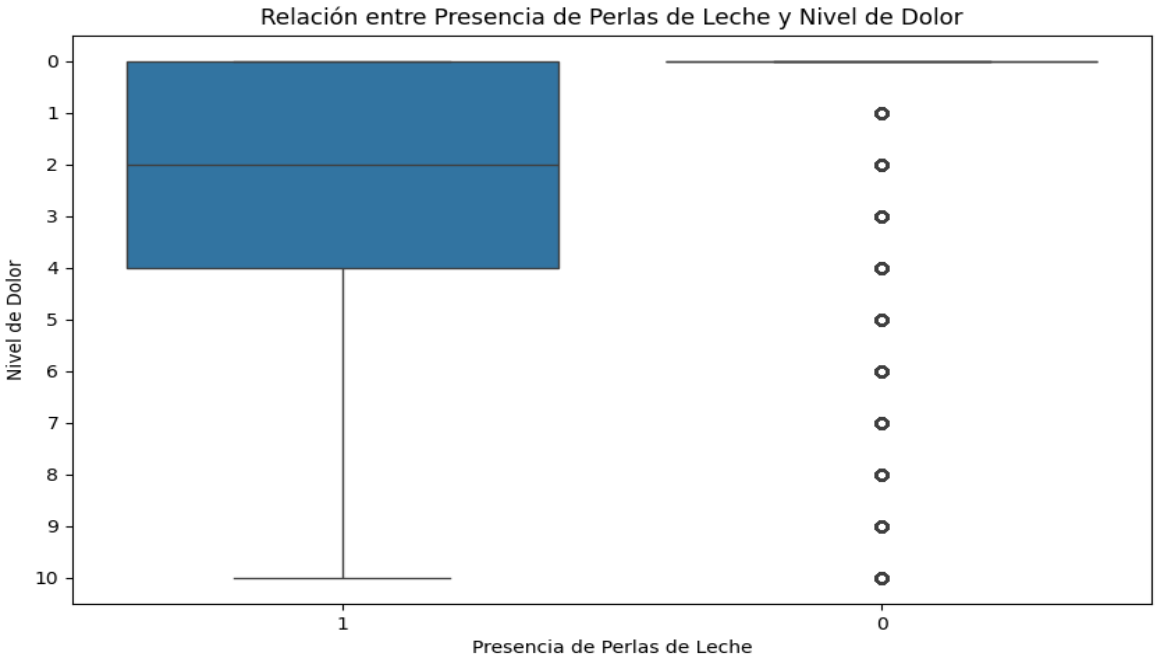


Ilustración 13: Relación entre existencia de perlas de lactancia y nivel de dolor. Fuente: Propia

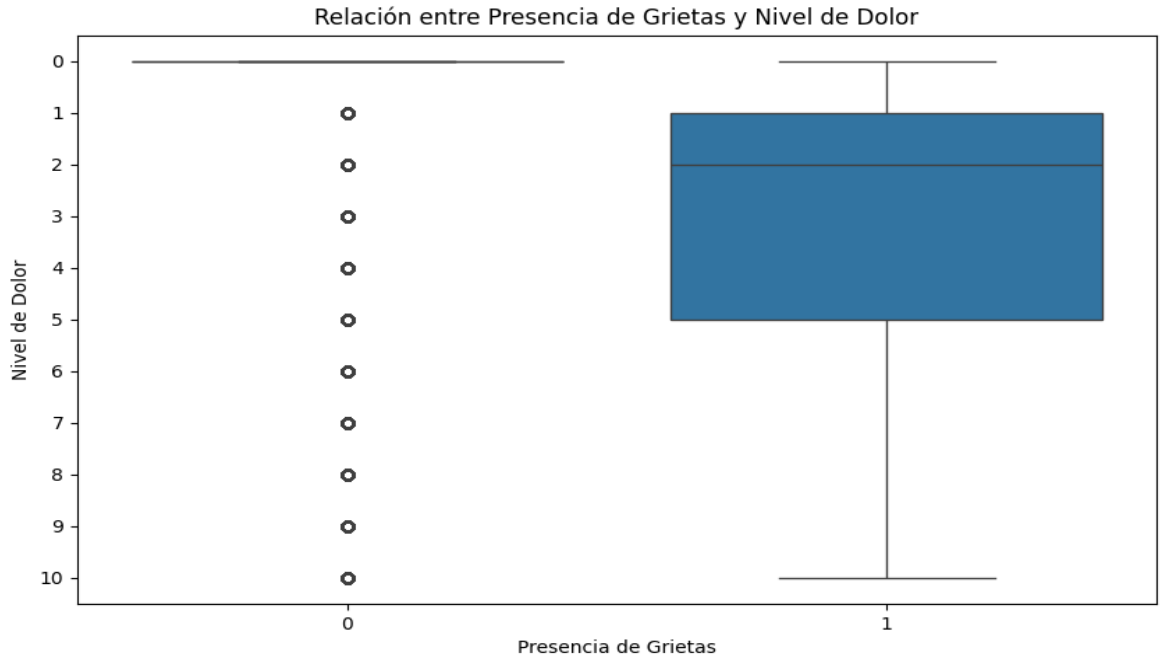


Ilustración 14: Relación entre existencia de grietas en pezón y nivel de dolor. Fuente: Propia

Como se puede observar en los gráficos, al observar la relación entre el nivel de dolor y la forma del pezón, se observa que claramente hay un aumento en el nivel de dolor percibido por la madre cuando el pezón tiene forma de pintalabios. Esto concuerda con lo esperado tras la investigación de los distintos problemas de la madre. Es posible que la forma se deba a una mala sujeción del bebé al pecho, y cause dolor a la madre en la toma.

De forma similar, vemos como dependiendo del color del pezón, las madres padecen más o menos dolor, siendo aparente que el color de pezón que más dolor causa en las tomas, el color blanco, seguido del morado. De nuevo esto corrobora los hallazgos relacionados con el color del pezón anteriormente.

En cuanto a la existencia de perlas de leche, se observa como claramente el dolor experimentado es mayor con la existencia de perlas que sin ella, y en este caso, la duración de las tomas aumentó ligeramente. Esto puede estar relacionado con la dificultad del bebé de tomar la leche con la existencia de poros bloqueados.

Algo muy similar ocurre con la existencia de grietas, el nivel de dolor aumenta significativamente, siendo esta la causa aparente de mayores niveles de dolor de la madre.

Tras estudiar la relación entre el dolor y los problemas de la madre, se estudia la relación entre el género y los problemas de la madre, la hipótesis, es que exista una relación entre el genero y los problemas de la madre. Podría ser que los niños, tuvieran más fuerza y agarrasen el pezón con mas fuerza causando dolor a la madre o similares. Sin embargo, al analizar los resultados, concluimos que no existe tal relación entre género y los problemas.

Tras concluir que no existe relación entre estas dos variables, se procede a estudiar la relación entre los problemas de la madre y la necesidad de dar leche suplementaria materna o fórmula. Para ello será necesario crear nuevas variables. Se crearán 3 variables booleanas.

- La primera indicará verdadero en caso de que exista cualquier valor distinto de ‘0’ o ‘null’ en la columna de suplBreastmilk.
- La segunda indicará verdadero en caso de que exista cualquier valor distinto de ‘0’ o ‘null’ en la columna de suplFormula
- La tercera será verdadera si cualquiera de las dos anteriores lo es.

Con el objetivo de poder visualizar de forma más clara los resultados, se normalizan las variables, de este modo se observa con más claridad si la frecuencia en el uso de suplementos aumenta con los problemas de las tomas.

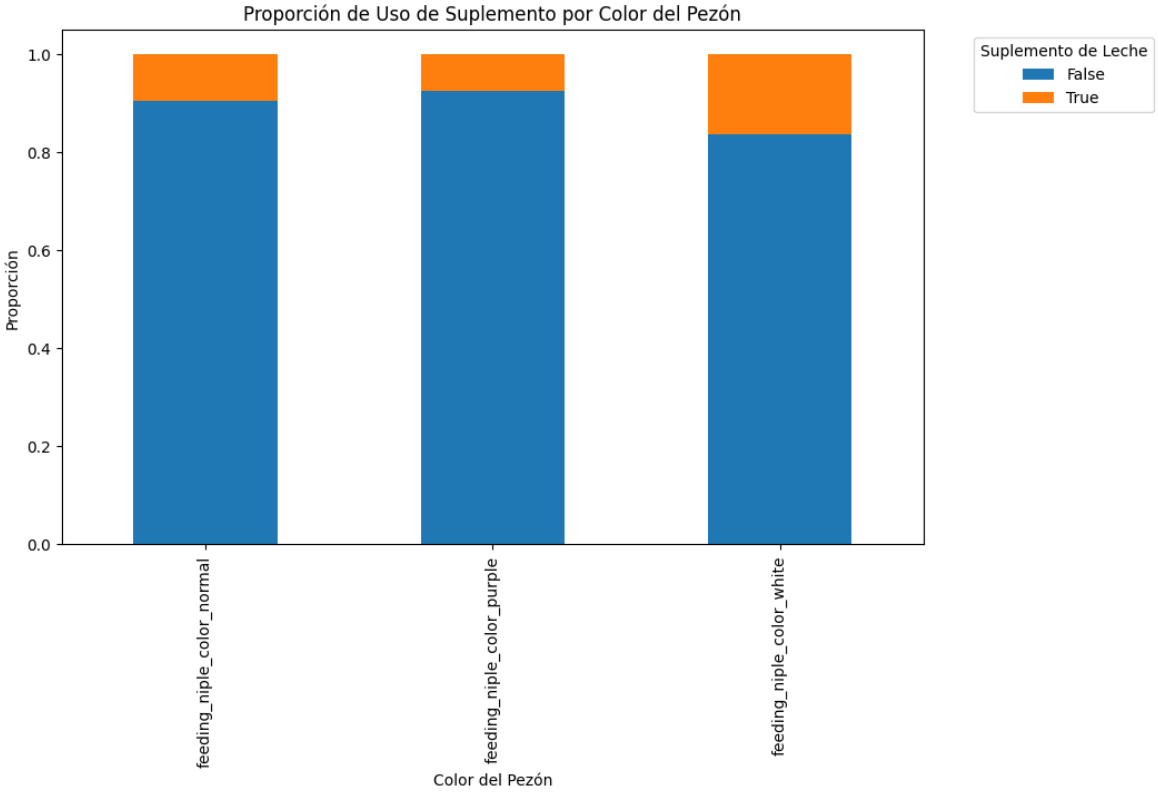


Ilustración 15:Proporción de uso de suplemento de alimentación y color del pezón. Fuente: propia

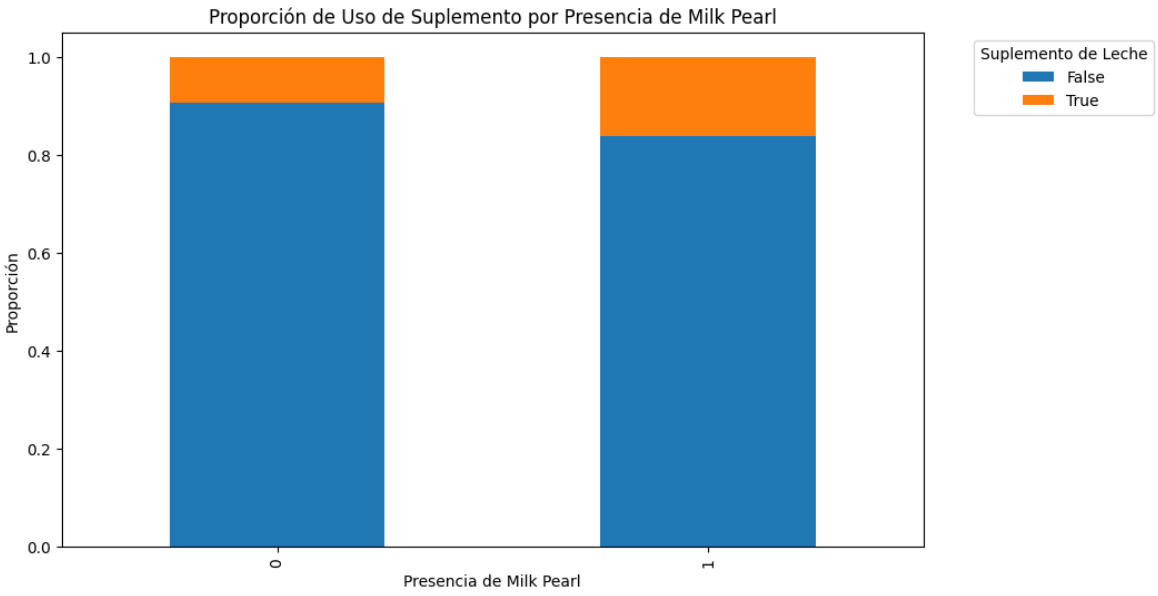


Ilustración 16: Proporción de uso de suplemento de alimentación y existencia de perlas de lactancia. Fuente: propia

Observando los gráficos, encontramos una mayor frecuencia en el uso de suplementos nutricionales (en mayor o menor medida) cuando la madre padecía alguno de los otros problemas. Siendo el color de pezón blanco y las perlas de lactancia (problemas que pueden además estar relacionados entre sí) los problemas que causan un mayor empleo de suplementos nutricionales.

Por último, se quiere averiguar si existe alguna relación entre la hora a la que se da la toma y el nivel de dolor.
En primer lugar, se crea la variable hora_toma como las horas y minutos de la variable startDate. Posteriormente se evalúa el nivel de dolor en relación a la hora. Sin embargo, al analizar los resultados, nos encontramos con que la densidad de frecuencia con la que se registra un nivel de dolor igual a '0' dificulta significativamente el analizar los datos. Se contempla estudiar la media del dolor en relación a la hora, y con los resultados, se concluye que no hay relación aparente entre estas variables.

Explorar la relación entre tomas y la edad del bebé:

Para esto, en primer lugar, será necesario averiguar la edad del bebé, para ello se creará una nueva variable edad, que será la diferencia entre la fecha de nacimiento y la fecha de inicio de la toma. Una vez calculado, podemos averiguar su relación con el resto de las variables.

Una vez calculada nuestra variable, la analizamos para detectar posibles anomalías y nos encontramos con valores anómalos negativos y otros positivos muy elevados, así que se filtra por positivos y menores de 10 años.

Una vez filtrados, sacamos sus estadísticas básicas:

- (6,781,418): El número total de registros considerados para calcular estas estadísticas es aproximadamente 6.78 millones de registros en el conjunto de datos.
- (83.21): La edad promedio del bebé en días es de aproximadamente 83 días.

- (97.95): La desviación estándar de la edad del bebé es de aproximadamente 97.95 días. Esto indica la variabilidad de las edades alrededor de la media.
- (0): La edad mínima del bebé es 0 días, lo que significa que algunos registros corresponden a tomas hechas el mismo día del nacimiento.
- (23): El 25% de los bebés tienen una edad de 23 días o menos.
- (55): El 50% de los bebés tienen una edad de 55 días o menos. Esta es la edad mediana, indicando que la mitad de los bebés tienen menos de 55 días y la otra mitad tienen más de 55 días.
- (111): El 75% de los bebés tienen una edad de 111 días o menos.
- (3,647): La edad máxima del bebé en los datos es de 3,647 días, lo que equivale a aproximadamente 10 años.

Ante estos resultados podemos concluir que siguen existiendo valores anómalos, ya que es extraño encontrarnos con valores superiores a los 2 años en las tomas. El resto de los valores proporcionan información interesante acerca de los datos.

Con esto en mente, se procede a comprobar si existe una relación entre la edad del bebé y el dolor en la toma.

De nuevo nos encontramos con una frecuencia en los valores de 0 tanto para las edades como para el dolor que dificultan la extracción de tendencias relevantes, por tanto, se filtran para el estudio.

En primer lugar, se representan los datos de las edades (días) junto con el nivel de dolor, pero el gráfico resultante resulta muy complicado extraer conclusiones ya que hay una gran dispersión en los datos. Ante los resultados, se decide calcular la media del dolor en función de la edad para poder reflejarla en un gráfico de línea para comprobar si existe alguna relación entre estos datos.

Observando los resultados se aprecia una tendencia descendente a pesar de la gran cantidad de ruido en la muestra. Adicionalmente, parecen existir picos de dolor en las edades del niño cercanas a los 180 días (6 meses) y 300 días (10 meses).

Para saber si la tendencia descendente, es relevante, se decide hacer una regresión lineal:

$$y = \beta_0 + \beta_1 x$$

Dónde:

- y es la variable dependiente (en este caso, el nivel de dolor medio).
- x es la variable independiente (en este caso, la edad del bebé en días).
- β0 es la intersección (el valor de y cuando x es 0).

β1 es el coeficiente de regresión (la pendiente de la línea).

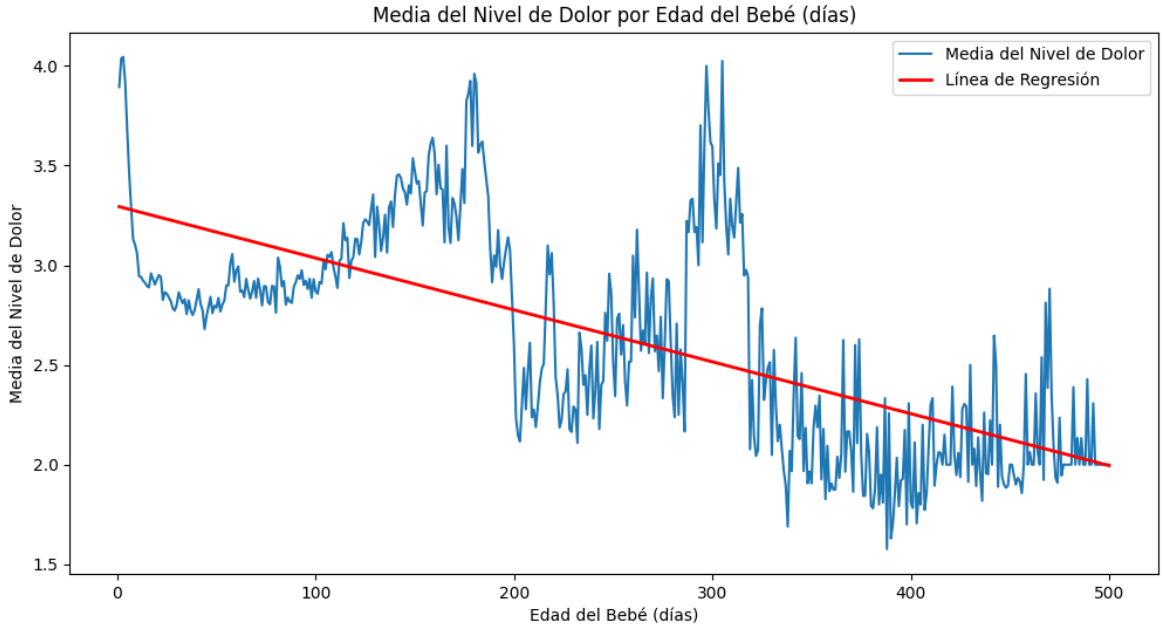


Ilustración 17. Regresión lineal aplicada a la media del nivel de dolor por la edad del bebé. Fuente: Propia

Como se observa por el resultado existe una tendencia de descenso de dolor según aumenta la edad. Esto se puede relacionar con el hecho de que los recién nacidos mejoran su capacidad de coordinación, desarrollan sus capacidades motrices, se desarrollan y ganan fuerza a medida que van creciendo, facilitando la lactancia y reduciendo el dolor que las madres perciben durante las tomas.

En conclusión, dada la inconsistencia de los datos, se ha dificultado sustancialmente la labor del análisis en relación con los objetivos del proyecto. Sin embargo, sí que se han encontrado relaciones y tendencias entre las tomas y los problemas de la madre, así como con la edad del bebé y las tomas.

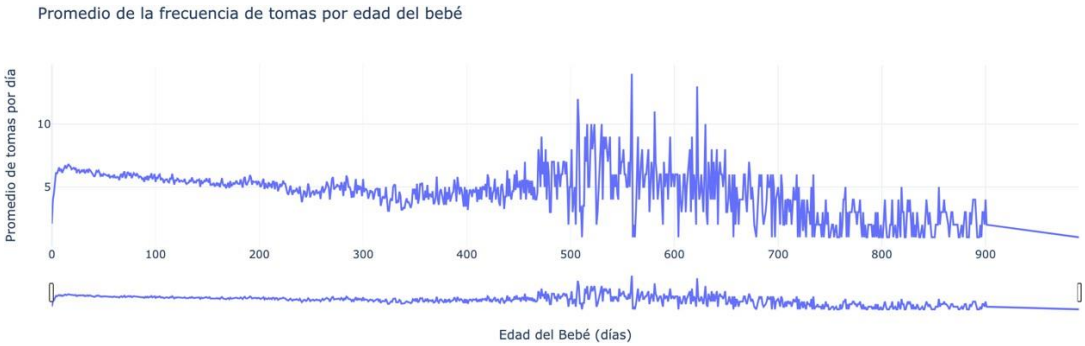
ANÁLISIS EXTERNO

Además del análisis propio, se obtiene un análisis de los mismos datos realizado por Natalia Gorrín Moreno, Cristina Salazar Diego y Beatriz Martín Cinos. Este es un análisis más extenso y detallado en el que se obtienen conclusiones adicionales relacionadas con los objetivos del proyecto. Cabe destacar este análisis, fue totalmente guiado por expertos y matronas así como por Lactapp, con hipótesis válidas en todos los casos, de las que se pueden extraer conclusiones relevantes.

Una vez procesados los datos, con una muestra final de aproximadamente 300.000 datos, proceden a analizar y extraer conclusiones como:

<u>En relación a las frecuencias:</u>

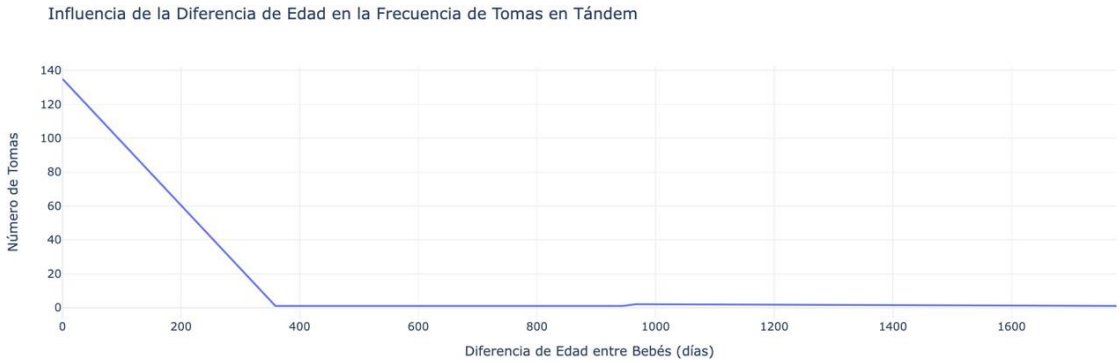
Promedio de la frecuencia de tomas por edad del bebé:



En sus resultados se observa una clara tendencia hacia una disminución de frecuencia de las tomas por edad del bebé, con 3 secciones principales, una con poca variabilidad y claro descenso entre los primeros 300 días de vida, una con gran variabilidad entre los 300 y los 700 días de vida, periodo en el que comienza a variar la dieta del bebé, y finalmente otra posterior a los 700 días de vida. Ante estos resultados se podría concluir que en los primeros 10 meses de vida, la frecuencia de las tomas es elevada debido a las necesidades nutricionales del bebé, pero con poca variabilidad, lo que se podría deber a patrones alimenticios uniformes. A partir del décimo mes, se observa una mayor variabilidad, lo que se puede deber a la introducción de suplementos nutricionales y dietas más variadas (cereales, sólidos, etc). Además, estas variaciones se pueden deber a diferencias individuales en la duración del periodo de lactancia así como a la situación laboral individual. A partir de los 700 dias, se observa una nueva

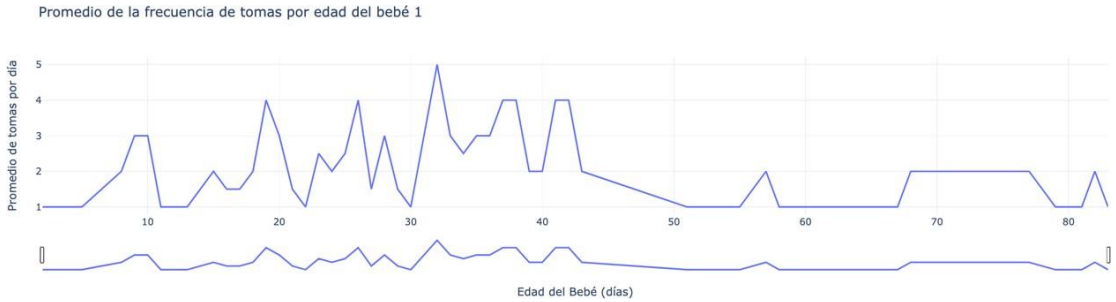
estabilización que se puede deber a una normalización de los patrones alimenticios, cuyas dietas cada vez son más predominantemente sólidas.

Influencia de la diferencia de edad en las tomas en tandem:



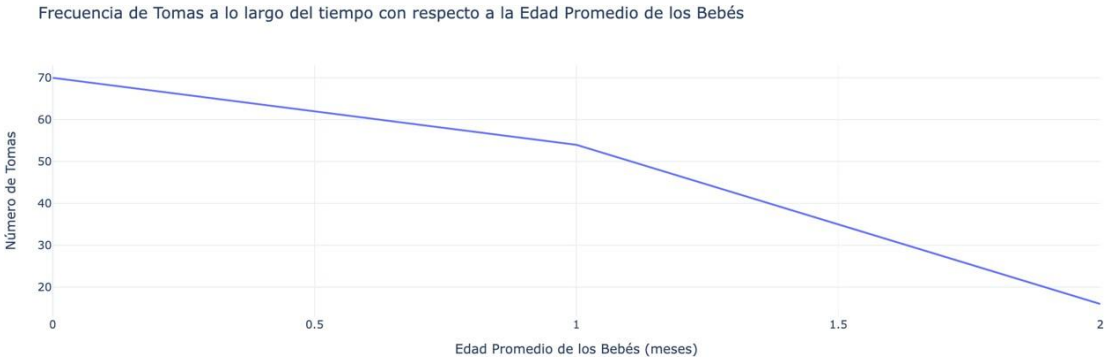
Relacionando la frecuencia de las tomas en tándem con la edad del bebé, las autoras observan una disminución a medida que aumenta la diferencia de edades. Esto se puede deber a que los bebes se encuentren en etapas muy distintas de desarrollo, con introducción de alimentos sólidos y cambios en la nutrición.

Promedio de la frecuencia de tomas por edad del primer bebé:



En sus resultados, encuentran una fluctuación en la frecuencia de tomas en los primeros 50 días, con un rápido descenso y estabilización a partir del día 50. Estas diferenciaciones se pueden relacionar con la capacidad de alimentación del bebe, que a medida que crece, es capaz de comer más, permitiendo intervalos más amplios entre tomas.

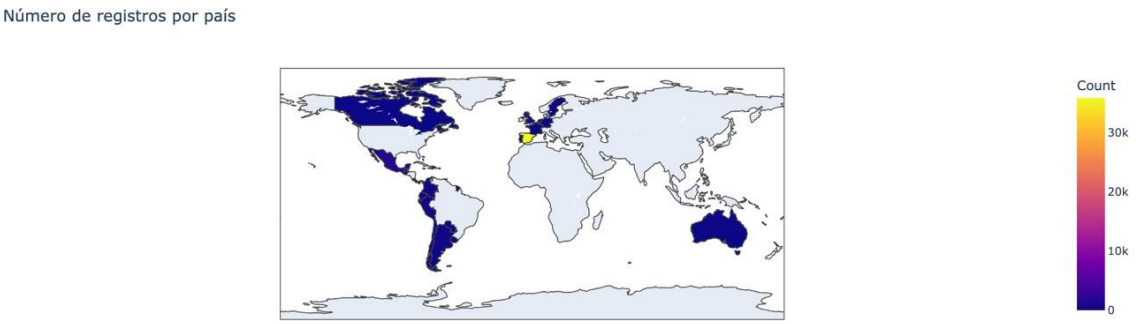
Frecuencia de Tomas a lo largo del tiempo:



En sus resultados, las autoras observan una clara y continua tendencia descendente en la frecuencia de las tomas, con un cambio en la pendiente descendiente a partir del primer año, que podría coincidir con los cambios en la alimentación como se ha ido mencionando.

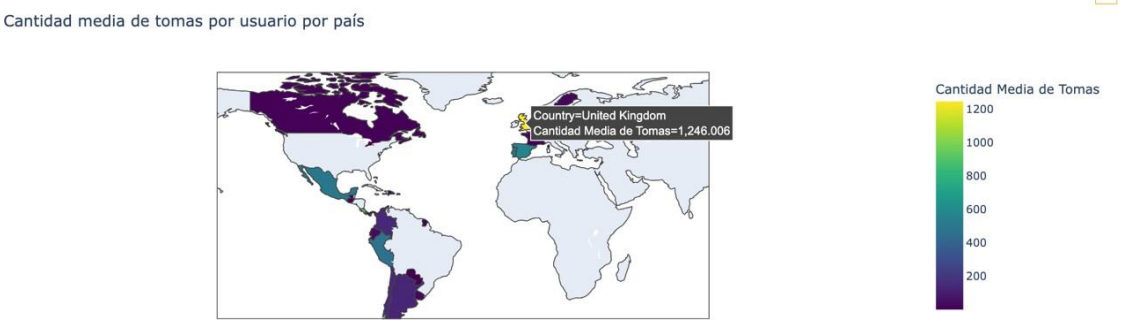
<u>En relación a los países:</u>

Número de Países:



Se obtienen registros para México, USA, Colombia, Costa Rica, Panamá, Perú, Ecuador, Francia, Alemania, UK, Noruega, Países Bajos y Australia. Siendo España el país con mayor número de registros.

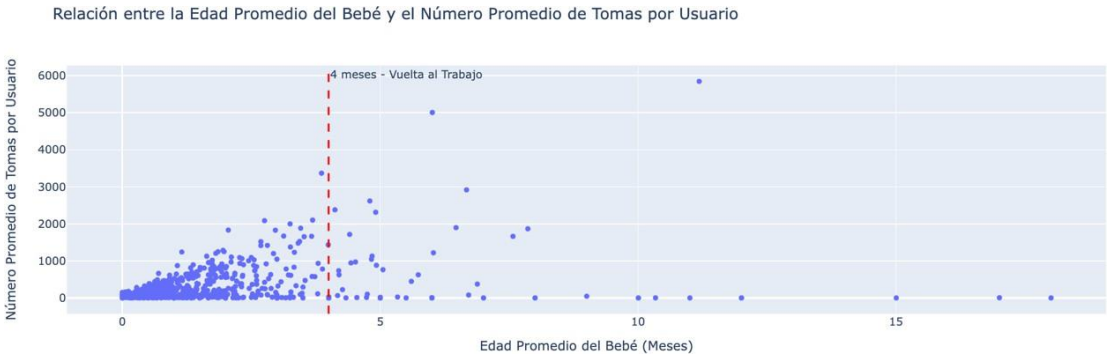
Cantidad de tomas por usuario en cada país:



En sus resultados, las autoras concluyen que Reino Unido es el país con mayor promedio de tomas por usuario con 1246, seguido por España con aproximadamente 800. En su análisis, concluyen que estos resultados pueden ser indicativos de variaciones en las prácticas de lactancia o en la adopción de la tecnología en diferentes contextos culturales y socioeconómicos, así como la diferencia de uso de la aplicación.

Relaciones con la Vuelta al Trabajo:

Relación entre la Edad promedio del Bebé y Número Promedio de Tomas por Usuario añadiendo la vuelta al trabajo:



En su estudio, representan la edad promedio del bebé (en meses) junto con el número promedio de tomas por usuario. Al mismo tiempo, representan el punto de 4 meses como el punto de 'Vuelta al trabajo de las madres' en países como España. En sus resultados, concluyen que hay una clara tendencia de disminución en el promedio de tomas a medida que aumenta la edad promedio del bebé, y que no se ve un claro punto de corte en el promedio de tomas asociado a la 'Vuelta al trabajo'.

Otras conclusiones:

Adicionalmente a los aspectos cubiertos anteriormente, realizan un análisis de las emociones de la madre en relación a distintos aspectos como la edad del bebé o alguno de los problemas relacionados con la lactancia. En sus conclusiones, admiten haber identificado patrones emocionales predominantes y su variación en el tiempo, así como la influencia de factores como la presencia de grietas en el pezón y la suplementación de la alimentación con leche de fórmula.

Con los resultados obtenidos, concluyen que existen medidas y mejoras en la recogida de los datos como futuros pasos. Recomiendan el empleo de desplegables para mejorar la precisión y consistencia de los datos.

CONCLUSIONES DEL ANÁLISIS

En base a los resultados obtenidos tanto por el análisis propio como por el externo, podemos llegar a las siguientes conclusiones y recomendaciones para Lactapp con relación a los objetivos establecidos para el proyecto.

Estudiar la relación que existen entre las tomas y los problemas que puede tener la madre

En base a los resultados propios obtenidos, podríamos recomendar que las madres con alguno de los problemas mencionados busquen ayuda profesional para tratar los problemas en cuestión ya que pueden ser la causa del dolor experimentado en las tomas. Adicionalmente, estos problemas parecen estar relacionados con la alimentación del bebé, ya que las madres que tienen algunos de estos problemas parecen emplear en mayor medida suplementos de nutrición, desfavoreciendo la estabilidad en los patrones alimenticios de las tomas.

Explorar la relación entre tomas y la edad del bebé

En base a los resultados propios obtenidos, podemos observar como el nivel de dolor disminuye a medida que aumenta la edad del bebe, lo que sugiere que a medida que el bebe se desarrolla, mejora su capacidad de agarre así como de succión y su coordinación, facilitando las tomas. Como sugerencia, se podría establecer una mayor monitorización en los primeros meses de los problemas de la madre, ya que es en este periodo en el que mayor dolor experimenta la madre.

En base a los resultados obtenidos en el análisis externo, parece que existen patrones en cuanto a la tendencia de la frecuencia de las tomas en función de la edad del bebe, que sugieren posibles cambios en los hábitos alimenticios de los bebés relacionados con la etapa de desarrollo en la que se encuentra, pero también posiblemente con la extensión del periodo de lactancia así como con la reincorporación al trabajo.

Investigar la relación entre ciertos *trackers* y la Región

En base al análisis externo, observamos que existe una relación entre el número de tomas por usuario y la región. Esta relación en primer lugar puede deberse a que distintas regiones tienen distintos niveles de utilización de la aplicación, pero también puede deberse a que existan diferencias en las tendencias alimenticias y de lactancia entre distintas regiones.

Investigar la relación entre los *trackers* y la situación laboral

En base a los resultados obtenidos en el análisis externo, no se puede observar un verdadero punto de corte a los 4 meses en cuanto al número promedio de tomas por usuario y la edad del bebé, lo que puede sugerir que las nuevas modalidades de trabajo como el teletrabajo han ayudado a facilitar la lactancia de las madres.

Adicionalmente a estas conclusiones, uno de los problemas comunes que se encuentra en ambos análisis es la inconsistencia en algunos de los datos, y como sugerencia, para evitar la variabilidad, se podría cambiar el tipo de input a los datos para acotarlo, de manera que en lugar de que la madre escriba la ciudad, exista una lista predeterminada de la que elegir, por ejemplo.

PLANIFICACIÓN DEL SERVICIO (PROYECTOS)

Teniendo en cuenta la estimación de proyectos realizados anualmente, calculada con anterioridad. ***¡Error! No se encuentra el origen de la referencia.***
Podremos calcular la cantidad de personal dedicado al análisis/consultoría que se necesitará para poder realizar los proyectos. Para poder realizar los cálculos, se asumirá que los proyectos asumidos durante este periodo serán pequeños, debido a la falta de reputación y nombre que tendrá la empresa al tratarse de una nueva *start-up* que ha entrado en el mercado. Asumiremos que los proyectos son de máximo 4 semanas.
Por otro lado, se estimará la cantidad de horas disponibles de trabajo anualmente sabiendo que el año habitual tiene 365 días; si le restas 52 fines de semana, 14 festivos y 22 días de vacaciones te quedan 225 días. Si consideras jornadas de 8 horas diarias, obtienes 1.800 horas/año.

Año 1:

En el primer año se estiman vender 36 proyectos, a una media de 4 semanas (asumimos laborales) significaría que sería necesario trabajar 36 proyectos, 4 semanas con 5 días por semana y 8h laborables por día, serían unas 5760h. Teniendo en cuenta estas previsiones, asumiendo que 1 consultor tiene disponibles 1800h anuales, llegamos a la conclusión que para poder asumir esa cantidad de trabajo serían necesarios al menos 4 consultores el primer año, asumiendo que pueden trabajar de forma simultánea en distintos proyectos.
Cabe indicar que de las 7200h que tendríamos disponibles sólo se emplearían 5760, por lo que se podría plantear contratar a 3 personas fijas y a una externa con contrato temporal. Otra posibilidad sería emplear las horas del cuarto consultor en optimizar procesos internos o apoyar alguna otra de las áreas en las que se necesitaría trabajar.

Año 2:

El segundo año, la cantidad de proyectos asumidos aumenta. En total se asumirán 48 proyectos. Siguiendo la misma lógica que el primer año esto supondrá unas 7680h de trabajo, que con los 4 consultores que tendríamos contratados del primer año no se podrían asumir. Por tanto, el segundo año sería necesario contratar a un quinto consultor. De nuevo, con esta organización,

tendríamos disponibles 9000h de trabajo, por tanto, se podría plantear contratar a una persona externa para cubrir las 480h de trabajo adicionales en caso de que solo contratásemos a 4 consultores y a alguien externo o con contrato parcial.

Año 3:

El tercer año se produce un gran crecimiento de las ventas de proyecto, hasta los 84 anuales. Con estas proyecciones serán necesarias 13440h de trabajo. Siguiendo los cálculos empleados los años anteriores llegamos a la conclusión de que serían necesarias 8 personas, por tanto, será necesario contratar a 3 personas adicionales (asumiendo que el segundo año se contratan a 5). Una vez más se podría contratar a 2 personas y una persona a tiempo parcial o temporalmente.

Año 4:

El año 4 de nuevo hay un enorme incremento en la proyección de las ventas, siendo necesarias 26880h para la realización de estos proyectos, según la lógica empleada anteriormente. Con estas horas en mente, la cantidad de consultores necesarios será de 15. En este caso, las horas disponibles estarán prácticamente cubiertas al completo por lo que no será razonable emplear un contrato temporal/parcial.

Año 5:

Dado que la cantidad de proyectos el quinto año se presume igual que el cuarto, los cálculos son iguales, y, por tanto, no se consideraría necesario contratar a una persona adicional.

Teniendo estas previsiones en cuenta, llegamos a las siguientes conclusiones:

Año	Proyectos Anuales	Proyectos mensuales	Personal	Horas de trabajo necesarias	Horas de trabajo disponibles
2024	36	3	4	5760	7200
2025	48	4	5	7680	9000
2026	84	7	8	13440	14400
2027	168	14	15	26880	27000
2028	168	14	15	26880	27000

Tabla 7: Personal contratado sin tiempo parcial. Fuente: Propia

Año	Proyectos Anuales	Proyectos mensuales	Personal	Personal a media jornada	Horas de trabajo necesarias	Horas de trabajo disponibles
2024	36	3	3	1	5760	6300
2025	48	4	4	1	7680	8100
2026	84	7	7	1	13440	13500
2027	168	14	15	0	26880	27000
2028	168	14	15	0	26880	27000

Tabla 8: Personal contratado con tiempo parcial. Fuente: Propia

PERSONAL

Una vez estimados la cantidad de personas necesarias por año para la labor de análisis de datos, y teniendo en cuenta otras áreas de negocio necesarias para el correcto desarrollo del servicio, concluimos

Año 1:

- 3 consultores a tiempo completo y 1 a tiempo parcial
- 1 empleado para el desarrollo software: cloud provider, ML, etc
- 1 empleado para reclutamientos y campañas de marketing.

Año 2:

- 4 consultores a tiempo completo y 1 a tiempo parcial
- 1 empleado para el desarrollo software: cloud provider, ML, etc
- 1 empleado para reclutamientos y campañas de marketing.

Año 3:

- 7 consultores a tiempo completo y 1 a tiempo parcial
- 1 empleado para el desarrollo software: cloud provider, ML, etc

- 1 empleado para reclutamientos y campañas de marketing.

Año 4:

- 15 consultores a tiempo completo
- 1 empleado para el desarrollo software: cloud provider, ML, etc
- 1 empleado para reclutamientos y campañas de marketing.

Año 5:

- 15 consultores a tiempo completo
- 1 empleado para el desarrollo software: cloud provider, ML, IT, etc
- 1 empleado para reclutamientos y campañas de marketing.

Cabe menciona que la labor de Project Manager (PM) aunque no mencionada de forma explícita, se desarrollará por el propio desarrollador de negocio en los primeros años, hasta que los consultores tengan suficiente experiencia y senioridad como para poder realizar la labor por su cuenta.

LOCALIZACIÓN

Para la selección del local se realiza un estudio en Madrid de las posibilidades existentes. Para la toma de decisión del local final se realiza una ponderación en base a los siguientes factores:

Ubicación	Superficie (m^2)	Precio/m^2	Espacio para reuniones
30%	20%	30%	20%

Tabla 9: Ponderaciones de factores de localización. Fuente: Propia

La elección será en base principalmente al espacio, el precio y su ubicación, pero dado que trataremos mucho con cliente y además se buscará que haya mucha colaboración en el equipo para favorecer rendimiento, se prestará atención específicamente a aquellas que dispongan de espacio previo para reuniones sin necesidad de obra el primer día.

Teniendo en cuenta que el espacio mínimo de trabajo para una persona es de al menos 2m^2 (dobuss, 2024), o 4m^2 según la norma NPT 242, y estudiando desde el punto de vista del

quinto año, se necesitaran mínimo unos 80m^2 de superficie, cualquier oficina por debajo de este umbral se descarta.

Las propiedades a las que se les presta atención son:

- Ubicación: que se puntúa entre 0-10 en base a la cercanía a la zona de nuevos ministerios, dado que es una zona tanto de ocio, como empresarial, con muchas otras empresas *start-up* localizadas en *coworkings* de la zona.
- Precio: dado que *Pedidata* es una *start-up* que cuenta con capital limitado, se buscara el precio más bajo posible. Por eso pondera como un 30% frente a los otros factores.
- Superficie: se tiene en cuenta si los espacios permiten una correcta distribución del mobiliario para la ergonomía de los puestos de trabajo
- Espacio para reuniones: se valora de forma positiva aquellos que tengan espacios especialmente dedicados a reuniones.

Se encuentran 4 locales a través de idealista que, por una combinación de los factores, se decide estudiar más en detalle. Por ello se decide realizar una matriz de ponderación para decidir cuál es el local adecuado.

Estas son las distintas cualidades de los locales:

Local	Ubicación	precio/mes (€/mes)	superficie (m^2)	precio/m2 (€/m^2)	Espacio para Reuniones	Amueblada	Reformada
María de Molina	Prosperidad	3031	120	25.26	SI	SI	SI
Colombia	Bernabeu-Hispanoamérica	2250	200	11.25	SI	NO	NO
Santa Engracia	Nuevos Ministerios-Rios Rosas	1700	100	17	NO	NO	NO

José Abascal	Nuevos Ministerios-Rios Rosas	2303	120	19.19	SI	SI	SI

Tabla 10: Locales elegidos. Fuente: propia

Teniendo estos factores en cuenta, se normalizan aquellos para los que se ponderará la matriz:

Local	Ubicación	precio/mes (€/mes)	superficie (m^2)	precio/m2 (€/m^2)	Espacio para Reuniones
María de Molina	1	0	0.6	0.445368171	1
Colombia	0.666666667	0.257670736	1	1	0
Santa Engracia	0.888888889	0.439129	0.5	0.661764706	0
José Abascal	1	0.240184758	0.6	0.586242835	1

Tabla 11: Propiedades de los locales normalizadas. Fuente: Propia

Por último, se pondera en base a las proporciones establecidas al comienzo:

Local	Ubicación	superficie (m^2)	precio/m 2 (€/m^2)	Espacio para Reuniones	TOTAL
María de Molina	0.35	0.12	0.155878 86	0.15	7.8
Colombia	0.233333333	0.2	0.35	0	7.8
Santa Engracia	0.311111111	0.1	0.231617 647	0	6.4
José Abascal	0.35	0.12	0.205184 992	0.15	8.3

Tabla 12: Matriz ponderación para selección de local. Fuente: Propia

El local donde se desarrollará la actividad, por tanto, será el local situado en José Abascal 41. Se trata de un local bien situado, con la parada de metro Alonso Cano a menos de 5 minutos y fácil

acceso al Paseo de la Castellana. Este local tiene 120m^2 con un diseño abierto. Tiene wifi de alta velocidad, e incluso una terraza con acceso a la azotea para los momentos de descanso. Cuentan con espacio adicional de escritorio disponible por h/días/meses, así como espacio de trabajo ampliable, lo que se ajustaría a las necesidades crecientes de nuestra plantilla. A pesar de considerar nuevas opciones de trabajo como el teletrabajo, o la hibridación, al tratarse de una *start-up,* con el objetivo de facilitar y favorecer la colaboración de todo el equipo para promover una cultura de mejora continua, se realizará el trabajo de forma presencial los primeros años.

DETECCIÓN DE RIESGOS

DETECCIÓN DE RIESGOS

La evaluación de los riesgos del proyecto se puede dividir en distintas áreas.

- Riesgos relacionados con la plantilla
- Riesgos relacionados con los proveedores de servicios *cloud*
- Riesgos relacionados con el mercado y los clientes

A continuación, se presenta una tabla que identifica diversos riesgos asociados con los temas mencionados. El objetivo es prevenir la aparición de estos riesgos, evaluarlos y establecer una jerarquía de importancia para minimizar su impacto en caso de que no se puedan evitar.

Evaluación de riesgos			
Riesgos relacionados con la plantilla			
ID	**Riesgo**	**Causa**	**Efecto**
RP1	Falta de mano de obra.	El personal para el desarrollo del servicio no es suficiente, no hay suficientes personas interesadas en el puesto de trabajo o existe demasiada rotación en los puestos.	No poder asumir el volumen de proyectos previsto. Retrasos en los proyectos
RP2	Falta de capacitación adecuada	Puede ser que los empleados seleccionados no tengan la suficiente capacidad en las últimas tecnologías	Peores resultados de los análisis y disminución de la satisfacción del cliente
RP3	Carga de trabajo excesiva	Sobrecarga de trabajo en picos determinados	Errores en proyectos Baja moral
Riesgos relacionados con los proveedores de servicios *cloud*			

RPSC1	Dependencia del proveedor	Alta dependencia de un único proveedor	Problemas de continuidad en caso de que el proveedor sufra interrupciones, cambie precios, políticas, etc
RPSC2	Seguridad y privacidad de los datos	Riesgos de seguridad de los datos en caso de almacenamiento en la nube	Brechas de seguridad pueden resultar en pérdidas de datos y de reputación
RPSC3	Escalabilidad y rendimiento	Problemas de escalabilidad asociados a picos de demanda con el proveedor seleccionado	Reducción de la calidad del servicio. Tiempos de respuesta más largos Retrasos en proyectos
Riesgos relacionados con el mercado y los clientes			
RMC1	Cambios en las necesidades del cliente	Cambios rápidos en la necesidades o expectativas del cliente	Puede resultar en la irrelevancia del servicio
RMC2	Competencia intensa	Aumento significativo en el sector de análisis de datos y ML	Presión sobre los precios y la necesidad de innovar

Tabla 13: Tabla de riesgos. Fuente: Propia

MATRIZ DE RIESGO

Una vez se han determinado los posibles riesgos, sus causas y sus efectos, se puede desarrollar una matriz de riesgo. Con esto se pretende evaluar la probabilidad y el impacto de cada riesgo para así poder determinar su importancia.

Para ello, se dará una puntuación del 1 al 5 tanto a la probabilidad como al impacto de que ocurra un riesgo, siendo 1 lo mínimo y 5 lo máximo.

Para evaluar la importancia se multiplican estos dos factores (probabilidad e impacto) sacando así un ranking que permite ordenar los diferentes riesgos.

Importancia de Riesgos				
Riesgos relacionados con la plantilla				
ID	**Riesgo**	**Probabilidad**	**Impacto**	**Importancia**
RP1	Falta de mano de obra	4	4	**16**
RP2	Falta de capacitación adecuada	2	4	**8**
RP3	Carga de trabajo excesiva	3	3	**9**
Riesgos relacionados con los proveedores y las materias primas				
ID	**Riesgo**	**Probabilidad**	**Impacto**	**Importancia**
RPSC1	Dependencia del proveedor	3	4	**12**
RPSC2	Seguridad y privacidad de los datos	2	5	**10**
RPSC3	Escalabilidad y rendimiento	2	3	**6**
Riesgos relacionados con el mercado y los clientes				
ID	**Riesgo**	**Probabilidad**	**Impacto**	**Importancia**
RMC1	Cambios en las necesidades del cliente	4	2	**8**
RMC2	Competencia intensa	3	2	**8**

Tabla 14: Probabilidad e Importancia de los Riesgos

Se presenta una matriz en la que se ordenan estos riesgos pudiendo evaluarlos de una manera más visual. Las zonas rojas son las de mayor importancia, seguido de las amarillas y por último las verdes son aquellas de menor importancia.

			IMPACTO				
			Muy Bajo	Bajo	Medio	Alto	Muy Alto
			1	**2**	**3**	**4**	**5**
PROBABILIDAD	Muy Alta	**5**					
	Alta	**4**		RMC1		RP1	
	Media	**3**			RP3	RPSC1	
	Baja	**2**		RMC2	RPSC3	RP2	RPSC2
	Muy Baja	**1**					

Tabla 15: Matriz de Riesgos (impacto y probabilidad)

Se pueden establecer ciertas medidas a tener en cuenta para prevenir o disminuir el impacto del riesgo en caso de que este ocurra.

Riesgos de alta importancia:

- RP1: Falta de mano de obra:

Dado que el trabajo a realizar es altamente demandado y existe una gran competencia, podría surgir una escasez de mano de obra. Este riesgo es significativo, ya que sin suficiente personal no se podría llevar a cabo la actividad ni cumplir con el volumen de proyectos previsto.

  - Para evitarlo:

- Publicar ofertas de trabajo con descripciones claras y detalladas de las tareas a realizar, evitando así que los empleados tengan expectativas erróneas.
- Ofrecer beneficios sociales atractivos para captar empleados: salarios justos, buenas condiciones laborales, ubicación de oficina accesible, oportunidades de asumir más responsabilidades en el futuro, vacaciones, horarios flexibles, entre otros.

Riesgos de importancia media:

- RP2: Falta de capacitación adecuada:

Este riesgo se considera de importancia media dado que al tratarse de una start-up nueva, lo que se desea es el crecimiento del empleado junto con la empresa.

- Para evitar este riesgo:
  - Inversión en formación constante.
  - Inversión en certificaciones de nuevas tecnologías
  - Establecer un proceso de selección detallado para evitar errores de contratación

- PR3: Carga de trabajo excesiva:

Puede ser que las estimaciones de proyectos no concuerden con la realidad, y que para poder mantener el negocio en marcha sea necesario aceptar más proyectos. También cabe la posibilidad de que los proyectos sean más grandes de lo estimado.

- Para evitar este riesgo:
  - Gestión adecuada de la carga de trabajo
  - Contratación de personal temporal o externo

- RPSC1: Dependencia del proveedor;

Cabe la posibilidad de que al emplear un único proveedor para los servicios de almacenamiento cloud y ML, suframos a causa de posibles interrupciones en el servicio de este.

- Para evitar este riesgo:
  - Diversificar en la selección de proveedores

- Ajustar correctamente los acuerdos de nivel de servicio para establecer correctamente las responsabilidades en caso de interrupciones o similares

- RPSC2: Seguridad y privacidad de los datos:

Es posible que, durante el servicio, se sufran brechas en la seguridad que causen corrupción de datos o similares. La probabilidad de que esto ocurra con las medidas de seguridad establecidas por los proveedores de servicios de cloud son bajas, por eso se considera de importancia media.

- Para evitar el riesgo:
  - Emplear medidas robustas de seguridad con encriptación de datos y controles de acceso restringido además será necesario cumplir con los reglamentos establecidos.

- RMC1: Cambios en las necesidades del cliente:

Es posible que los clientes en el mercado cambien repentinamente de necesidades y/o expectativas. Se considera de importancia media dada la baja probabilidad de que algo similar suceda.

- Para evitar el riesgo
  - Se debe mantener una comunicación constante con los clientes para entender sus necesidades y adaptar los servicios rápidamente.

Riesgos de baja importancia:

- RPSC3: Escalabilidad y rendimiento:

Dado que dependemos de servicios de *cloud* para los momentos de alta demanda, dependemos de estos para que los servicios sean escalables de forma sencilla para evitar problemas de rendimiento en los proyectos. Se considera de importancia baja dado que no se considera que vaya a existir un pico de demanda tan elevado que cause problemas en el modelo de escalado, y con el estudio de los proveedores realizado, se sabe que la transición y escalabilidad de este es muy elevada.

- Para evitarlo:
  - Elegir proveedores que ofrezcan escalabilidad automática
  - Realizar pruebas de carga regularmente para asegurar el rendimiento bajo condiciones de alta demanda.

- RMC2: Competencia intensa:

Si bien es cierto que se sabe que la tendencia por emplear estas tecnologías está aumentando, y existe una gran competencia en el sector, también consideramos que hay una gran oferta y demanda por tanto la importancia del riesgo se considera baja

  - Para evitarlo:
    - Invertir en innovación
    - Diferenciación de servicios
    - Establecer alianzas estratégicas.

RSE

OBJETIVOS DE DESARROLLO SOSTENIBLE

Pedidata es una empresa con distintas acciones relacionadas con la Responsabilidad Social Empresarial. A continuación, se enuncian las acciones que generan un impacto social con la intención de mejorar la vida de las personas, así como con la conciencia por el impacto medioambiental.

En primer lugar, los objetivos de *Pedidata* se alinean con algunos de los Objetivos de Desarrollo Sostenible destinados a poner fin a la pobreza extrema, garantizar los derechos humanos y proteger el medioambiente (UPND, s.f.)

De los 17 ODS, *Pedidata* está alineado con los siguientes:



Este proyecto está alineado con los siguientes Objetivos de Desarrollo Sostenible presentados.

2 – Hambre cero:

Aunque el objetivo de desarrollo sostenible está más enfocado en la agricultura, este objetivo también incluye metas relacionadas con la nutrición. El proyecto contribuye a mejorar la nutrición de los neonatos, asegurando que los problemas relacionados con la alimentación sean identificados y tratados, lo cual es esencial para prevenir el hambre y sus efectos adversos desde una edad temprana.

Además, al proporcionar datos y análisis que pueden ayudar a optimizar la alimentación de los neonatos, *Pedidata* puede apoyar iniciativas de salud pública destinadas a mejorar la nutrición infantil en diversas comunidades.

3 – Salud y bienestar:

Al proporcionar herramientas para detectar de forma temprana y facilitar el tratamiento de problemas de nutrición y posibles problemas de salud, este proyecto tiene un impacto directo en mejorar la salud y el bienestar de los recién nacidos. La prevención y la intervención precoz son claves para reducir las tasas de mortalidad neonatal y para asegurar un desarrollo saludable durante la infancia.

Al promover prácticas de lactancia saludable y proporcionar datos que pueden informar políticas de salud, *Pedidata* también puede contribuir a reducir enfermedades y mejorar la calidad de vida tanto de los neonatos como de sus madres.

- **4 – Educación de calidad:**

A través de la educación y capacitación de los padres y profesionales de la salud sobre nutrición neonatal y el uso de tecnologías de análisis de datos, el proyecto promueve la educación en salud y bienestar, que es fundamental para el desarrollo de comunidades saludables y bien informadas. Esto permitirá a los padres sentirse más seguros y capacitados para el cuidado de sus hijos

Suponiendo que en un futuro *Pedidata* pueda colaborar con organizaciones gubernamentales. Al colaborar con organizaciones gubernamentales y no gubernamentales, *Pedidata* puede extender su alcance a comunidades desfavorecidas, asegurando que todos los niños tengan acceso a una nutrición adecuada y cuidado de salud, independientemente de su situación económica o ubicación geográfica.

10 – Reducción de las desigualdades

El objetivo sería implementar esta solución de forma pública en el futuro, o facilitar este paso mediante el trabajo realizado. El proyecto puede ayudar a reducir las desigualdades al proporcionar acceso a tecnologías de salud avanzadas tanto en comunidades desarrolladas como en desarrollo. Al hacer la tecnología accesible a bajo costo o mediante programas de salud pública, se puede ayudar a nivelar el campo de juego en términos de acceso a cuidados de salud de calidad para todos los neonatos.

No solo eso, si no que, *Pedidata* puede desarrollar programas educativos y materiales de capacitación en conjunción con los clientes con los que establezca alianzas estratégicas para mejorar el conocimiento sobre la lactancia y la nutrición neonatal, beneficiando a una amplia gama de profesionales de la salud y padres.

OTRAS ACTIVIDADES

Aparte de ser una empresa alineada con los Objetivos de Desarrollo Sostenible, existen otras actividades llevadas a cabo por *Pedidata* que crean impacto social.

En primer lugar, la empresa estará creando puestos de trabajo, lo cual ayudará a varias personas a desarrollarse profesionalmente.

Por otro lado, gracias a las relaciones comerciales y alianzas estratégicas que se pretende desarrollar con otras empresas se pretende ayudar a que tanto *Pedidata* como otros negocios que comparten los mismos valores obtengan beneficios y prosperen mientras se contribuye a mejorar el bienestar social.

MÓDULO III –

FACTIBILIDAD

ADMINISTRATIVA LEGAL

DETERMINACIÓN DE LA FORMA JURÍDICA

COBERTURA DE RESPONSABILIDADES

La estructura legal de *Pedidata*, que determina la responsabilidad de los socios según la legislación, es una Sociedad Limitada. Esto se debe a que el mínimo legal de capital aportado es de 1€ (España, 2022) pudiendo ser en efectivo o en especie (propiedades materiales, inmuebles, etc.). En caso de deudas o problemas financieros, los socios sólo son responsables hasta el monto del capital invertido, sin comprometer su patrimonio personal. Además, la empresa tendrá varios socios con derecho a voto, cuya responsabilidad también estará limitada a su inversión. Estos socios podrán dividirse en dos clases: trabajadores y capitalistas. Serán ellos los que decidan sobre el camino y la estrategia a seguir de la empresa, asegurando una gestión democrática y participativa.

REGISTRAR LA EMPRESA

A la hora de registrar la empresa será necesario consultar en el Registro Mercantil si el nombre seleccionado ha sido registrado antes puesto que debe ser un nombre original que no haya sido

registrado. Se pretende utilizar el nombre de *Pedidata Consulting.*



Ilustración 18: Logo Empresa. Fuente: Propia

Con el nombre de la empresa establecido, y la estructura de la sociedad definida, lo único que seria necesario antes de poder registrarla seria la obtención de un NIF.

REGLAMENTO Y PROTECCIÓN DE DATOS

REGLAS PARA EL TRATAMIENTO DE DATOS

Para garantizar que *Pedidata* cumpla con las normativas y estándares de protección de datos, es esencial prestar atención a varios aspectos legales y de seguridad. A continuación, se detalla el análisis de los aspectos pertinentes según el Reglamento General de Protección de Datos (GDPR), la Ley Orgánica de Protección de Datos (LOPD), la norma ISO 27001, y los controles de seguridad de Cloud Control Matrix (CCM). Además de los reglamentos mencionados anteriormente (GDPR, LOPD, ISO 27001, y CCM), existe una nueva propuesta de reglamento a nivel europeo que es relevante para las empresas que desarrollan o utilizan inteligencia artificial (IA). Esta propuesta es conocida como **Reglamento de la Inteligencia Artificial (AI Act)** de la Unión Europea.

GDPR (REGLAMENTO GENERAL DE PROTECCIÓN DE DATOS)

El GDPR es la normativa europea que regula el tratamiento de datos personales de los ciudadanos de la UE. Dado que en nuestros proyectos nos encargaremos de dar servicio a empresas dedicadas a la mejora del bienestar de las madres y sus hijos, es crucial para *Pedidata* atenerse al cumplimiento de esta norma. Implementada el 25 de mayo de 2018, esta regulación establece directrices estrictas sobre cómo se deben recopilar, almacenar, procesar y eliminar los datos personales. (Comission, s.f.)

Sus aspectos clave son:

- **Consentimiento**: Los usuarios deben dar su consentimiento explícito antes de procesar sus datos. Debemos asegurar con nuestros clientes el correcto tratamiento de los datos y el consentimiento con los usuarios finales.
- **Derechos de los Sujetos de Datos**: Incluye derechos como el acceso, rectificación, eliminación y oposición al tratamiento de sus datos.

- **Transferencia de Datos**: Las transferencias de datos fuera de la UE deben realizarse con las garantías adecuadas.
- **Notificación de Brechas**: Se debe informar a las autoridades y a los sujetos afectados dentro de las 72 horas en caso de una brecha de seguridad.

LOPD (LEY ORGÁNICA DE PROTECCIÓN DE DATOS Y GARANTÍA DE LOS DERECHOS DIGITALES)

La LOPD es la legislación española que adapta el GDPR al contexto nacional, proporcionando detalles adicionales y especificidades relevantes para España. Entró en vigor el 6 de diciembre de 2018. Dado que el servicio de *Pedidata* es en el entorno nacional, concretamente en Madrid, se deberá atener a lo estipulado por este reglamento (BOE, 2018).

Sus aspectos clave son:

- **Registro de Actividades de Tratamiento**: Las empresas deben mantener un registro detallado de las actividades de tratamiento de datos.
- **Evaluaciones de Impacto:** Se requieren evaluaciones de impacto sobre la protección de datos cuando se introducen nuevas tecnologías o se realizan tratamientos de datos sensibles.
- **Delegado de Protección de Datos (DPO):** Las empresas deben designar un DPO si realizan tratamientos a gran escala de datos sensibles o si su actividad principal implica el tratamiento de datos que requieren un seguimiento sistemático y regular.

ISO 27001

ISO 27001 es un estándar internacional para la gestión de la seguridad de la información, que proporciona un marco para gestionar la seguridad de los activos de información de una organización. Está diseñado para proteger la confidencialidad, integridad y disponibilidad de la información (IEC, 2022). Es clave seguir los estándares internacionales para garantizar a los clientes la fiabilidad de nuestro trabajo.

Sus aspectos clave son:

- **Sistema de Gestión de Seguridad de la Información (SGSI):** Implementación de un SGSI que incluye políticas y procedimientos para la gestión de la seguridad de la información.
- **Evaluación de Riesgos:** Identificación y evaluación de riesgos para la seguridad de la información y la implementación de controles adecuados para mitigarlos.
- **Auditorías Internas y Externas:** Realización de auditorías periódicas para asegurar el cumplimiento con los requisitos del estándar y la mejora continua del SGSI.

CLOUD CONTROL MATRIX (CCM)

El Cloud Control Matrix (CCM) es un marco de controles de seguridad diseñado específicamente para entornos de computación en la nube. Desarrollado por la Cloud Security Alliance (CSA), el CCM proporciona una estructura detallada de controles de seguridad. A la hora de emplear los servicios de nuestro proveedor seleccionado, debemos asegurarnos de que sigue la estructura del marco CCM. (aliance, s.f.).

Sus aspectos clave son:

- **Control de Acceso:** Implementación de controles robustos para gestionar y restringir el acceso a los datos en la nube.
- **Seguridad de Aplicaciones e Infraestructura:** Protección de aplicaciones e infraestructuras en la nube contra amenazas y vulnerabilidades.
- **Gestión de Incidentes:** Procedimientos establecidos para la gestión y respuesta a incidentes de seguridad.
- **Conformidad y Auditoría:** Aseguramiento del cumplimiento con normativas aplicables y realización de auditorías regulares para verificar la efectividad de los controles de seguridad.

En este sentido, cabe mencionar que el proveedor seleccionado con anterioridad y cumple con las normativas enumeradas y es compatible con el modelo de *framework* establecido por CCM. (cloud, s.f.) (cloud G. , s.f.) (compliance, s.f.)

AI ACT

Esta es una propuesta legislativa de la Comisión Europea cuyo objetivo es regular el uso de la inteligencia artificial en la UE, asegurando que las aplicaciones de IA sean seguras y respeten los derechos fundamentales de las personas. (Parliment, 2023)

Sus aspectos clave son:

- **Clasificación del Sistema de IA**: Determinar si los sistemas de IA desarrollados son de alto riesgo y aplicar los requisitos pertinentes.
- **Gestión de Riesgos y Gobernanza de Datos**: Implementar procedimientos sólidos para la gestión de riesgos y la gobernanza de datos.
- **Transparencia y Supervisión Humana**: Asegurar que los usuarios sean informados adecuadamente y que exista una supervisión humana sobre los sistemas de IA.
- **Conformidad y Evaluaciones**: Realizar evaluaciones de conformidad y estar preparados para auditorías y supervisiones por parte de las autoridades competentes.

ORGANIGRAMA. CANTIDAD DE PERSONAL

En cuanto al tamaño de la empresa durante los primeros 5 años (para los que se ha hecho una predicción y un dimensionamiento de ventas), la empresa será una PYME. Los primeros 3 años será clasificada como 'microempresa' dado que contará con menos de 10 empleados, y los siguientes 2 como 'Pequeña empresa' dado que se tendrá más de 15.

Más concretamente, el personal se podrá dividir en 4 equipos principales dependiendo del área del negocio que estén desarrollando:

- Desarrollo de negocio: Este equipo lo compondrá un trabajador, concretamente el fundador (Álvaro Torán). Este se encargará de que la empresa siga una dirección concreta, así como de gestionar las alianzas empresariales y coordinar los equipos. De este depende que se puedan abrir nuevos canales de ventas con más empresas y que la reputación de la empresa aumente con el tiempo
- Desarrollo de software e IT: este equipo se encargará de desarrollar los sistemas alojados en los proveedores de almacenamiento *cloud*, así como del control, despliegue y mantenimiento de los modelos de ML que sean necesarios en los proyectos. También se encargará del mantenimiento de los equipos de los trabajadores.
- Recruiting: este equipo se encargará de buscar los perfiles más adecuados para el desarrollo del negocio. Será necesario que se encargue de mantener expectativas realistas, buscar perfiles con deseo de desarrollo profesional y con la experiencia necesaria para que estén capacitados para llevar a cabo las tareas que desarrollen.
- Consultores: estos se encargarán del correcto desarrollo de los proyectos de la empresa. Dependiendo del año aumentará el número de personas en este equipo, llegando a ser hasta 15 el último año. Es por esto que será necesario que exista una organización entre el equipo de consultores en función de la senioridad y experiencia, de forma que los consultores que entraron el primer año se encargarán de dirigir cada uno un equipo de 4 consultores junior el quinto año.

Debido al número tan bajo de empleados, es posible establecer relaciones horizontales dentro de la empresa sin la necesidad de tener muchos gerentes de por medio. Es por esto por lo que se considera que un organigrama horizontal o plano es la mejor opción. De esta manera, habrá comunicación directa entre los dos niveles (“superiores” y “empleados”) permitiendo un flujo de información mucho más directo y eficaz que agilizará la toma de decisiones y la gestión de problemas o imprevistos.

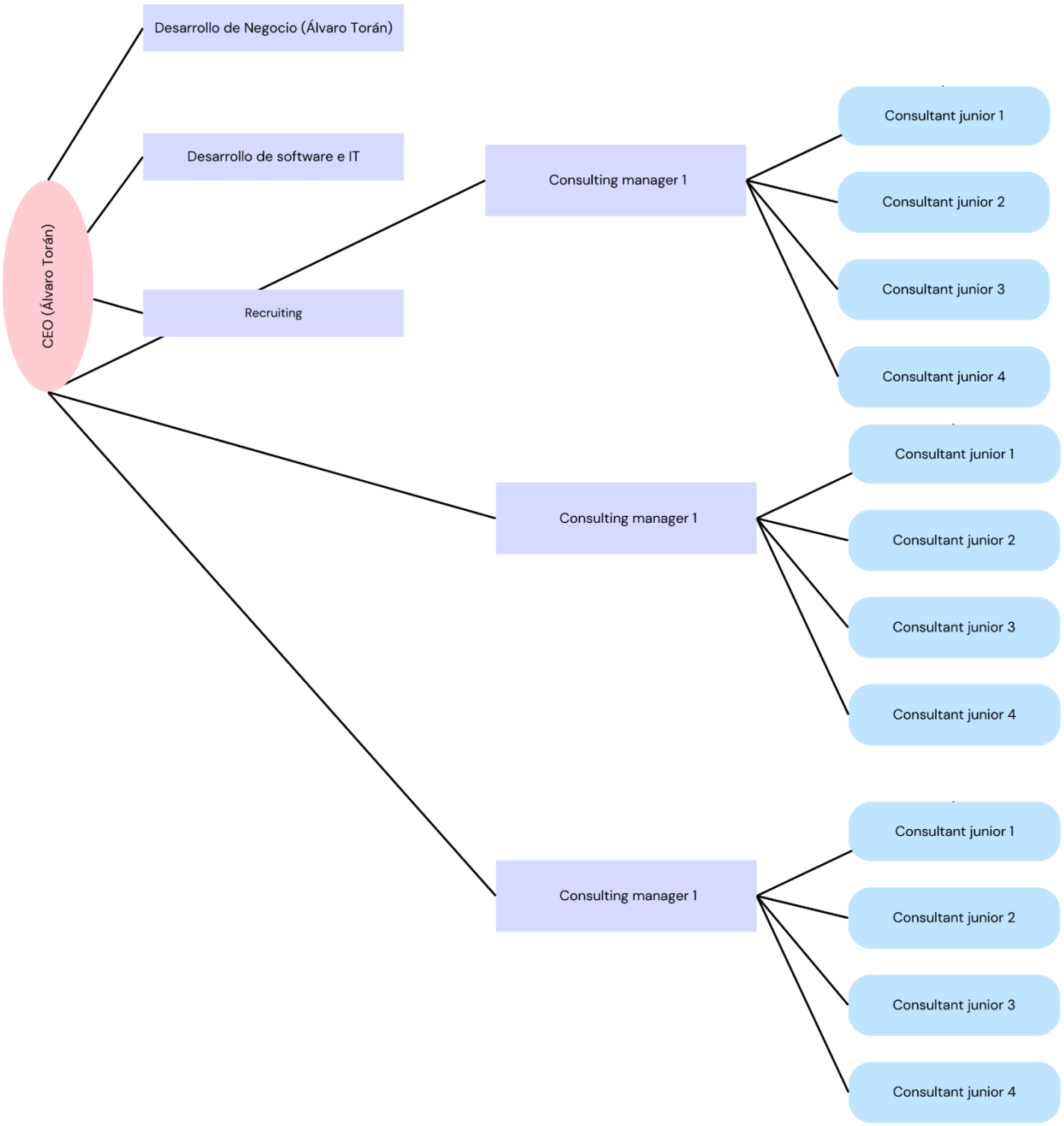


Ilustración 19: Organigrama de la empresa. Fuente: Propia

PERFILES DE PUESTO

A continuación, se desarrollan más específicamente los distintos perfiles que se necesitan para cada puesto de trabajo durante los primeros cinco años de vida de *Pedidata*.

En cuanto a requerimientos más específicos de cada puesto estos se detallan a continuación:

Desarrollo de negocio:

- Sobre la posición: Este rol suele conllevar mucha responsabilidad para desarrollar el negocio de acuerdo a relaciones con clientes, alianzas, etc. Además, se requiere una gran capacidad de análisis para evaluar las distintas oportunidades, tendencias de mercado, etc. En este caso, además, este asumirá la labor de PM para los primeros años.
- Sobre las habilidades necesarias:
    1. La mayoría de las ofertas de trabajo de este tipo de puestos buscan personas con conocimientos empresariales o de ingeniería que tengan esa capacidad de análisis y puedan tener una visión global de lo que está sucediendo.
    2. Experiencia o conocimiento en gestión y ejecución de proyectos.
    3. Conocimientos de idiomas son muy positivos para poder establecer relaciones comerciales o alianzas con empresas internacionales.
    4. Conocimiento de la parte técnica de la empresa para un mejor análisis de las oportunidades de negocio.
    5. Capacidad para dirigir y motivar equipos, y tomar decisiones estratégicas.
    6. Habilidad para negociar acuerdos y alianzas con otras empresas.
    7. Excelentes habilidades de comunicación verbal y escrita para interactuar con socios, clientes y el equipo interno.
    8. Capacidad para desarrollar e implementar planes de negocio efectivos.

Desarrollo de software:

- Sobre la posición: Este rol requiere amplios conocimientos técnicos a nivel de programación y de las distintas tecnologías disponibles. Por otro lado, será necesario que tenga conocimientos suficientes para llevar a cabo desarrollo y mantenimiento de sistemas alojados en proveedores *cloud*. También deberá tener control del despliegue y mantenimiento de estos sistemas.

- Sobre el rol:
    1. Licenciatura en informática, programación o experiencia técnica en este campo.
    2. Trabajo en equipo y capacidad de comunicación.
    3. Conocimiento y competencias técnicas en el lenguaje de programación pertinente y conocimiento sobre el funcionamiento de los sistemas *cloud*.
    4. Gestión de pantallas, *layouts* e integración con los distintos tipos de sistemas operativos de los smartphones.
    5. Conocimiento en desarrollo y mantenimiento de modelos de ML.
    6. Habilidad para gestionar sistemas operativos y software de mantenimiento.
    7. Capacidad para identificar y solucionar problemas técnicos rápidamente.

Recruiting:

- Sobre la posición: Este puesto está enfocado a la búsqueda y selección de los perfiles más adecuados para el desarrollo del negocio. Deben conocer bien el negocio y el mercado para mantener expectativas realistas y buscar candidatos con deseos de desarrollo profesional y experiencia necesaria
- Sobre el rol:
    1. Experiencia en procesos de reclutamiento y selección de personal.
    2. Excelentes habilidades de comunicación para interactuar con candidatos y equipos internos.
    3. Capacidad para evaluar las habilidades y competencias de los candidatos.
    4. Habilidad para construir y mantener una red de contactos profesionales.
    5. Capacidad para gestionar múltiples procesos de selección simultáneamente.

Consultoría:

- Sobre la posición: deben tener gran conocimiento en el desarrollo y ejecución de proyectos de análisis de datos, así como la capacidad de comprender las necesidades específicas de los clientes. Deberá ser capaz, con la experiencia adquirida a lo largo del tiempo de trabajo en *Pedidata*, de asumir las labores de PM para poder estar a cargo de su propio equipo como consultor sr.
- Sobre el rol:

1. Experiencia en técnicas de análisis de datos y herramientas como SQL, Excel, y software de análisis estadístico.
2. Conocimiento en el desarrollo y aplicación de modelos de ML.
3. Habilidades para presentar datos y resultados a clientes de manera clara y comprensible.
4. Capacidad para gestionar proyectos desde la conceptualización hasta la entrega final.
5. Habilidad para trabajar en equipos multidisciplinarios y colaborar efectivamente.

Adicionalmente existirán dos rangos en esta posición:

- Junior: Deberá apoyar en el desarrollo de proyectos, realizar análisis de datos bajo la supervisión de un consultor senior, y contribuir con ideas y soluciones innovadoras. También debe tener motivación para aprender y desarrollarse profesionalmente, atención al detalle y capacidad para trabajar bajo supervisión.
- Senior: debe ser capaz de dirigir proyectos complejos, supervisar equipos de consultores junior, y asegurar la calidad del trabajo entregado. Adicionalmente, tendrá experiencia en liderazgo de equipos, gestión de clientes y toma de decisiones estratégicas.

SALARIOS DE CADA POSICIÓN

A través de las comparativas y estudios realizados se concluye lo siguiente acerca de los sueldos medios que se perciben en España para los puestos de trabajo mencionados anteriormente.

Desarrollador de negocio: De acuerdo a fuentes como Salary.Com, el salario promedio de un desarrollador de negocio se encuentre entre los 40.000 y los 80.000 euros (Salary.com, s.f.), sin embargo, dado que este trabajo sería llevado a cabo por el fundador, se pretende percibir una distribución que comience sin beneficios el primer año, y escale con el tiempo hasta los 20.000euros el quinto año de la siguiente manera:

Año	2024	2025	2026	2027	2028
Sueldo (€/año)	0	5.000	10.000	15.000	20.000

Tabla 16: Distribución del salario de desarrollo de negocio. Fuente: Propia

Desarrollador de software: Según un estudio realizado a partir de 10.322 sueldos aportados a Glassdoor, el sueldo de desarrollador de Software es de entre 35.000 y 50.000 €/año considerándose un sueldo de 21.000 €/año como un sueldo bajo (Glassdoor, s.f.). Cabe destacar, que dadas las responsabilidades del puesto, incluyendo conocimiento de *cloud* y ML, parece razonable ofrecer **30.000 €/año.**

Recruiting: Nuevamente, de acuerdo con los estudios realizados por GlassDoor con muestras de casi 300 sueldos, se puede saber que el sueldo medio para el puesto recruiter es de 20.000 a 35.000 euros (Glassdoor, Glassdoor, 2024). Teniendo en cuenta el tamaño de la *start-up* y la escasez de recursos además de los sueldos medios, se considera que el sueldo medio óptimo para esta posición son **23.000 €/año.**

Consultoría: al tratarse de la posición más esencial en la empresa se deberá estudiar con especial atención los salarios que perciban nuestros consultores. Según un estudio de Glassdoor, el salario promedio de un 'consultant' es de 29.000 a 39.000 euros (Glassdoor, Glassdoor, 2025). Dado que tendremos niveles de senioridad en la empresa será importante diferenciar entre los consultores con mayor experiencia y los consultores junior. Teniendo en cuenta la

situación de la empresa como *start-up*, se decide establecer el salario de un consultor junior en **28.000 euros** y el de un consultor senior en **35.000 euros.**

SISTEMA ADMINISTRATIVO DE GESTIÓN (INDICADORES, KPIS)

En cuanto al sistema administrativo de la empresa, es necesario definir una serie de indicadores para ir controlando y midiendo el correcto desarrollo de las distintas áreas en *Pedidata*.
De este modo distribuimos los indicadores para las distintas áreas de la siguiente manera.

Desarrollo de negocio:

Será importante estimar la salud financiera de la empresa para presentar con transparencia a los inversores y empleados la situación en todo momento. Para ello revisaremos:
Ingresos Mensuales: Medir los ingresos generados por nuevos clientes y proyectos.
Tasa de Retención de Clientes: Porcentaje de clientes que continúan utilizando los servicios de *Pedidata*.
Número de Alianzas Estratégicas: Cantidad de nuevas alianzas o asociaciones establecidas.

Desarrollo de software y IT:

En este sentido será importante tener un registro de la eficiencia y fiabilidad de nuestros sistemas. En este caso será necesario revisar:
Tiempo de Despliegue de Modelos: Tiempo promedio desde el desarrollo hasta el despliegue de un modelo de ML.
Disponibilidad del Sistema: Porcentaje de tiempo que los sistemas están operativos y disponibles para los clientes.
Costes relacionados con el almacenamiento y procesado de datos en proveedores.
Tasa de Incidencias Técnicas: Número de problemas técnicos reportados y resueltos en un período de tiempo.

Recruiting:

Es importante mantenerse por delante de los competidores del mercado en todo momento, para ello es necesario prestar atención a los perfiles que reclutamos, así como de la retención de talento que tenemos como empresa. Se revisará:
Tasa de Retención de Empleados: Porcentaje de empleados que permanecen en la empresa a lo largo del tiempo.

Tiempo de Contratación: Tiempo promedio desde la apertura de una vacante hasta la contratación de un candidato.

Satisfacción del Empleado: Medida a través de encuestas de satisfacción periódicas.

Consultoría:

Por último, nuestra prioridad es el cliente, es importante tener buenos indicadores de nuestro servicio que nos permitan adaptarnos y cambiar de estrategia en caso de que el mercado lo indique. Por ello se revisará:

Satisfacción del Cliente: Medida a través de encuestas de *feedback* y Net Promoter Score (NPS).

Cumplimiento de Plazos: Porcentaje de proyectos entregados dentro del plazo acordado.

Eficiencia del Proyecto: Relación entre el tiempo estimado y el tiempo real invertido en los proyectos.

MÓDULO IV – FACTIBILIDAD ECONÓMICA FINANCIERA

La siguiente sección tiene por objeto analizar la viabilidad económica y financiera del proyecto. En concreto, se analizará el modelo de ingresos de la empresa y, en consecuencia, se estudiará la financiación del proyecto.

Además, esta sección permite realizar el análisis de rentabilidad y el análisis de sensibilidad. De este análisis económico se extraerán conclusiones adecuadas para evaluar la estrategia correcta en términos de precio, fijación de precios, financiación, inversión, etc.

RECURSOS

En primer lugar, es necesario analizar cuáles son los recursos necesarios para el desarrollo de *Pedidata*, teniendo en cuenta tanto el servicio como las instalaciones dedicadas al mismo. Estos recursos se pueden clasificar de la siguiente manera.

1. **Recursos tecnológicos:** El desarrollo del servicio de *Pedidata* no solo incluye los proveedores de servicios necesarios para el almacenamiento de datos en la nube y el ML. También es necesario contar con una página Web, Redes Sociales, software de programación, publicidad y campañas digitales, trabajo diario, etc.

    a. Desarrollo de software:

        - Recursos necesarios para establecer bases de datos en la nube
        - Recursos necesarios para desplegar y mantener modelos de ML en la nube

    c. Licencias: Otras licencias de programas básicos como Microsoft, programas de diseño, etc. Estos serán útiles para el desarrollo de publicidad y campañas digitales tanto para las publicaciones en Redes Sociales como para el diseño de los eventos, etc.

    d. Comunicaciones (teléfono y WiFi)

2. **Mano de obra:** En cuanto a la mano de obra, se pueden distinguir varios subgrupos dependiendo de las funciones a realizar.

    a. Consultoría: Debido a la previsión de proyectos a 5 años realizada anteriormente, se prevé un volumen de proyectos para el que es necesario contar varios equipos únicamente dedicados al servicio de consultoría.

    b. Recruiting: Cabe destacar que para poder contratar a los consultores que cumplan con expectativas de capacidad y cualificación será necesario contar con personal que dedique su tiempo a filtrar los potenciales candidatos para encontrar la mejor selección para Pedidata.

    c. Desarrollo del negocio: Mano de obra responsable de crear y aplicar estrategias que aumenten los ingresos, la rentabilidad y la cuota de mercado. Durante los primeros 5 años, el responsable será Álvaro Torán, con unos honorarios al directorio que aumentaría progresivamente a medida que lo hicieran los proyectos, según el dimensionamiento del mercado y la previsión a 5 años que se ha presentado anteriormente.

d. Desarrollo de software e IT: Para hacer frente a todas las funciones tecnológicas tanto del servicio, como del trabajo de oficina, como el manejo de los sistemas implementados en los proveedores de servicios *cloud*, el despliegue de soluciones de ML y otros, se requiere un responsable del diseño, desarrollo, pruebas y mantenimiento de las soluciones y sistemas.

4. **Infraestructura:** Para poder desarrollar el trabajo de forma eficiente, así como para ofrecer un espacio a los clientes en el que se pueda mantener una comunicación directa y cercana, se alquila una oficina en Madrid (Calle José Abascal). Los motivos por los que se ha elegido este local frente a otras opciones quedan reflejados en secciones anteriores del proyecto: *Localización*

CUENTA DE PÉRDIDAS Y GANANCIAS

Después de haber analizado el plan y la previsión de ventas a 5 años y una vez analizados los proveedores, la mano de obra, el alquiler del local, etc., se puede construir la cuenta de pérdidas y ganancias que se estima obtener a 5 años.

Cabe destacar que debido a la volatilidad del mercado y puesto que se trata de una start-up que aún no está implantada y no es conocida, estos números no son 100% representativos.

De cara al plan de negocio, se plantea:

- *Advertising*: Se plantea un presupuesto en campañas de anuncios por diferentes medios de 55.500eur iniciales, que crece año a año en función de los ingresos.
- Licencias de Microsoft para los empleados: 6.99eur por licencia/mes
- Licencia de Zoho para recruiting: 50eur mensual
- Gastos de mantenimiento: asociados a licencias y alquiler
- Servicios profesionales independientes: incluyen gastos de gestoría, abogados, etc, así como gastos de marketing. Se decide externalizar el marketing, con un presupuesto inicial en el año 1 de 77000eur y un 10% de incremento anual.
- Unos gastos en mobiliario inicial (equipos informáticos, mesas, sillas) de 50000 con una depreciación al 10% anual

De acuerdo con todo esto, así como los recursos mencionados anteriormente, la cuenta de Pérdidas y Ganancias resulta:

PyG		2024	2025	2026	2027	2028
Ingresos		483,840.00 €	645,120.00 €	1,128,960.00 €	2,257,920.00 €	2,257,920.00 €
Ventas de proyectos						
Ventas anuales		483,840.00 €	645,120.00 €	1,128,960.00 €	2,257,920.00 €	2,257,920.00 €
Ingresos Totales		483,840.00 €	645,120.00 €	1,128,960.00 €	2,257,920.00 €	2,257,920.00 €
Costes directos		55,500.00 €	74,000.00 €	129,500.00 €	259,000.00 €	259,000.00 €
Advertising		55,500.00 €	74,000.00 €	129,500.00 €	259,000.00 €	259,000.00 €
Advertising		55,500.00 €	74,000.00 €	129,500.00 €	259,000.00 €	259,000.00 €
resultado comercial		428,340.00 €	571,120.00 €	999,460.00 €	1,998,920.00 €	1,998,920.00 €
Costes indirectos		254,156.00 €	297,296.00 €	421,716.00 €	720,196.00 €	725,196.00 €
Costes de licencias		21,020.00 €	21,160.00 €	21,580.00 €	22,560.00 €	22,560.00 €
Licencias GCP		6,060.00 €	6,060.00 €	6,060.00 €	6,060.00 €	6,060.00 €
Microsoft Office		560.00 €	700.00 €	1,120.00 €	2,100.00 €	2,100.00 €
Hubspot		14,400.00 €	14,400.00 €	14,400.00 €	14,400.00 €	14,400.00 €
Gastos de estructura		27,636.00 €	27,636.00 €	27,636.00 €	27,636.00 €	27,636.00 €
Alquileres		27,636.00 €	27,636.00 €	27,636.00 €	27,636.00 €	27,636.00 €
Gastos de mantenimiento	10	9,731.20 €	9,759.20 €	9,843.20 €	10,039.20 €	10,039.20 €
servicios profesionales independientes		83,000.00 €	91,300.00 €	100,430.00 €	110,473.00 €	121,520.30 €
Sueldos y salarios (sgs incluida)		205,500.00 €	248,500.00 €	372,500.00 €	670,000.00 €	675,000.00 €
desarrollo de negocio						
sueldo anual		- €	5,000.00 €	10,000.00 €	15,000.00 €	20,000.00 €
desarrollo de sofware						
sueldo anual		40,000.00 €	40,000.00 €	40,000.00 €	40,000.00 €	40,000.00 €
recruiting						
sueldo anual		30,000.00 €	30,000.00 €	30,000.00 €	30,000.00 €	30,000.00 €
consultoria						
sueldo anual jr		114,000.00 €	152,000.00 €	228,000.00 €	456,000.00 €	456,000.00 €
sueldo anual sr		21,500.00 €	21,500.00 €	64,500.00 €	129,000.00 €	129,000.00 €
EBITDA		174,184.00 €	273,824.00 €	577,744.00 €	1,278,724.00 €	1,273,724.00 €
Intereses		- €	- €	- €	- €	- €
Depreciación/amortizacion	10	5,000.00 €	5,000.00 €	5,000.00 €	5,000.00 €	5,000.00 €
amortizacion de mobiliario		5,000.00 €	5,000.00 €	5,000.00 €	5,000.00 €	5,000.00 €
EBT		169,184.00 €	268,824.00 €	572,744.00 €	1,273,724.00 €	1,268,724.00 €
impuestos	30					
Impuestos		€ 50,755.20	€ 80,647.20	€ 171,823.20	€ 382,117.20	€ 380,617.20
Resultado		118,428.80 €	188,176.80 €	400,920.80 €	891,606.80 €	888,106.80 €

Tabla 17: Plan de negocio - hipótesis 1. Fuente: Propia

Ante estos resultados, resulta inverosímil pensar que sería posible generar beneficios desde el mes 1, es por esto, que se revisan los beneficios del año 1 mes a mes, estimando una venta de proyectos con esta distribución:

Mes 1	Mes 2	Mes 3	Mes 4	Mes 5	Mes 6	Mes 7	Mes 8	Mes 9	Mes 10	Mes 11	Mes 12
0	0	0	0	1	1	2	2	3	3	3	3

Tabla 18: Distribución alternativa de ventas año 1. Fuente: Propia

Adicionalmente a este ajuste, se calcula el CAPEX de *Pedidata* como el coste total del mobiliario necesario para el desarrollo del negocio proyectado en el tiempo. El primer año se estudiará a nivel mensual. De cara a la compra de mobiliario (mesa, silla, ordenador, licencias, etc) se considerará los servicios a tiempo parcial como persona adicional completa. Estos son los resultados:

	mes 1	mes 2	mes 3	mes 4	mes 5	mes 6	mes 7	mes 8	mes 9	mes 10	mes 11	mes 12	año 2	año 3	año 4	año 5
Desarrollo de negocio	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Recruiting									1	1	1	1	1	1	1	1
Consulting					**2**	**2**	**3**	**3**	**4**	**4**	**4**	**4**	**5**	**8**	**15**	**15**
Desarrollo de software	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Tabla 19: Ajuste de personal para estudio del plan de negocio. Fuente: Propia

El mobiliario que se ha considerado necesario para el correcto desarrollo del servicio es:

- Silla: Se estudian las medias de precios para sillas de gama baja y media, considerando el precio como la media de las 2. El precio de una silla ergonómica de gama baja ronda los 93eur y el de una silla de gama media los 200 (hinomi, s.f.) (Ergosistema, s.f.) por tanto se establece el precio en 146,5eur.
- Mesa: se estudian las medias de precios de gama baja, media y alta, siendo estos 200, 450 y 800 respectivamente. Por tanto, se establece el precio de la mesa en 483eur.

- Ordenador: para poder realizar las labores necesarias para el correcto desarrollo del negocio, se considera necesario el empleo de ordenadores de gama alta, se estudian los precios de 4 modelos distintos y se establece el precio de los ordenadores como la media de estos, siendo: 3000, 1400, 1500 y 1050. El precio se fija en 1889,8eur
- Teclado y ratón: se estudian los precios de distintos combos de teclado y ratón, en concreto 3 modelos distintos con precios: 27, 51 y 75. El precio de estos se establece en 51eur.
- Pantalla: de nuevo se estudian 3 modelos distintos de pantalla, con precios que varían desde 150 hasta 500. El precio se establece en 316,67eur.
- Licencia GCP: sabemos por estudios anteriores que el precio de la licencia GCP será de 505eur/mes
- Microsoft office: para poder realizar los informes y el trabajo diario será necesaria también la licencia de Microsoft, siendo esta, 6,99 euros mensuales por persona.
- Zoho recruit: para dar apoyo al área de recruit, se comprará la licencia de Zoho recruit, por 50eur mensuales por persona.

Una vez conocidos los mobiliarios, se establecen los que cada una de las personas de cada área necesita para el desarrollo de su trabajo:

Desarrollo de negocio	Silla	Mesa	Ordenador	Pantalla	Teclado	Ratón		Microsoft Office	
Recruiting	Silla	Mesa	Ordenador	Pantalla	Teclado	Ratón		Microsoft Office	Zoho
Consulting	Silla	Mesa	Ordenador	Pantalla	Teclado	Ratón		Microsoft Office	
Desarrollo de software	Silla	Mesa	Ordenador	Pantalla	Teclado	Ratón	Licencias GCP	Microsoft Office	

Tabla 20: Mobiliario necesario para cada empleado por área. Fuente: Propia

Conocidos el número de empleados, el mobiliario necesario y el precio de este, se proyecta el CAPEX, entendiendo CAPEX como el gasto en bienes de capital de una empresa, que incluye la adquisición, mejora o mantenimiento de activos físicos como edificios, maquinaria, equipos y

tecnología, necesarios para sostener y expandir sus operaciones a largo plazo, sabiendo que el mobiliario es un pago único asociado a cada persona adicional que se contrate, y las licencias serán pagos mensuales que podrán o no ser asociados a las personas contratadas dependiendo de si la licencia es por persona o de pago único mensual. Además, se tiene en cuenta el efecto del IPCA, que según Bankinter rondará el 2,3% el 2025 y de 2026 en adelante se establecerá en un 2,2% (Bankinter, 2024).

	mes 1	mes 2	mes 3	mes 4	mes 5	mes 6	mes 7	mes 8	mes 9	mes 10	mes 11	mes 12	año 2	año 3	año 4	año 5
Desarrollo de negocio	€ 2,894.29	€ 6.99	€ 6.99	€ 6.99	€ 6.99	€ 6.99	€ 6.99	€ 6.99	€ 6.99	€ 6.99	€ 6.99	€ 6.99	€ 85.81	€ 85.73	€ 85.73	€ 85.73
Recruiting									€ 2,944.29	€ 56.99	€ 56.99	€ 56.99	€ 699.61	€ 698.93	€ 698.93	€ 698.93
Consulting					€ 5,788.58	€ 13.98	€ 2,908.27	€ 20.97	€ 2,915.26	€ 27.96	€ 27.96	€ 27.96	€ 3,382.75	€ 9,538.26	€ 21,941.62	€ 1,285.88
Desarrollo de software	€ 2,894.29	€ 6.99	€ 6.99	€ 6.99	€ 6.99	€ 6.99	€ 6.99	€ 6.99	€ 6.99	€ 6.99	€ 6.99	€ 6.99	€ 85.81	€ 85.73	€ 85.73	€ 85.73

Tabla 21: Cálculo del CAPEX. Fuente: Propia

El total del CAPEX de la empresa para los primeros 5 años será de **60.924,4 euros**. Cabe mencionar, que se ha calculado el CAPEX incluyendo licencias, ya que se consideran esenciales para poder desarrollar el servicio.

En lo que se refiere a las inversiones en advertising y a los gastos de servicios profesionales independientes (marketing externo), se asignan unos presupuestos de 55.500 eur y 77.000 eur respectivamente. Estos gastos se distribuyen el primer año como inversión en los primeros meses de marketing, y según se comienzan a ver los resultados (nuevos proyectos) se realizan inversiones en advertising mensuales. Estas inversiones se irán aumentando, ya que se pueden considerar medidas efectivas dado el crecimiento en el número de proyectos. La inversión en estas áreas aumentará con la proyección en el tiempo de forma proporcional al aumento de ventas.

Por último, cabe mencionar, que también se ha tenido en cuenta la misma variación del IPCA en las cuentas de Pérdidas y Ganancias que en el cálculo del CAPEX.

Con estos ajustes, se calcula la cuenta de Pérdidas y ganancias como:

PyG	Mes 1	Mes 2	Mes 3	Mes 4	Mes 5	Mes 6	Mes 7	Mes 8	Mes 9	Mes 10	Mes 11	Mes 12	Año 2	Año 3	Año 4	Año 5
Ingresos	**-**	**-**	**-**	**-**	**13,440.00**	**13,440.00**	**26,880.00**	**26,880.00**	**40,320.00**	**40,320.00**	**40,320.00**	**40,320.00**	**659,957.76**	**1,153,797.12**	**2,307,594.24**	**2,307,594.24**
Ventas de proyectos																
Ventas	-	-	-	-	13,440.00	13,440.00	26,880.00	26,880.00	40,320.00	40,320.00	40,320.00	40,320.00	659,957.76	1,153,797.12	2,307,594.24	2,307,594.24
Ingresos Totales	-	-	-	-	13,440.00	13,440.00	26,880.00	26,880.00	40,320.00	40,320.00	40,320.00	40,320.00	659,957.76	1,153,797.12	2,307,594.24	2,307,594.24
Costes directos	**-**	**-**	**-**	**-**	**5,000.00**	**5,000.00**	**6,363.64**	**6,363.64**	**8,181.82**	**8,181.82**	**8,181.82**	**8,181.82**	**75,702.00**	**132,349.00**	**264,698.00**	**264,698.00**
Advertising	**-**	**-**	**-**	**-**	**5,000.00**	**5,000.00**	**6,363.64**	**6,363.64**	**8,181.82**	**8,181.82**	**8,181.82**	**8,181.82**	**75,702.00**	**132,349.00**	**264,698.00**	**264,698.00**
Advertising	-	-	-	-	5,000.00	5,000.00	6,363.64	6,363.64	8,181.82	8,181.82	8,181.82	8,181.82	75,702.00	132,349.00	264,698.00	264,698.00

resultado comercial	**-**	**-**	**-**	**-**	**8,440.00**	**8,440.00**	**20,516.36**	**20,516.36**	**32,138.18**	**32,138.18**	**32,138.18**	**32,138.18**	**584,255.76**	**1,021,448.12**	**2,042,896.24**	**2,042,896.24**
Costes indirectos	**10,947.27**	**10,947.27**	**10,947.27**	**10,947.27**	**17,197.27**	**17,197.27**	**21,863.93**	**21,863.93**	**27,030.60**	**29,530.60**	**29,530.60**	**29,530.60**	**379,265.19**	**502,483.20**	**807,947.46**	**819,237.80**
Gastos de estructura	**7,613.93**	**7,613.93**	**7,613.93**	**7,613.93**	**9,113.93**	**9,113.93**	**10,613.93**	**10,613.93**	**12,613.93**	**12,613.93**	**12,613.93**	**12,613.93**	**131,655.19**	**140,943.20**	**151,407.46**	**162,697.80**
Alquileres	2,303.00	2,303.00	2,303.00	2,303.00	2,303.00	2,303.00	2,303.00	2,303.00	2,303.00	2,303.00	2,303.00	2,303.00	28,271.63	28,243.99	28,243.99	28,243.99
Gastos de mantenimiento	810.93	810.93	810.93	810.93	810.93	810.93	810.93	810.93	810.93	810.93	810.93	810.93	9,983.66	10,059.75	10,260.06	10,260.06
servicios profesionales independientes	4,500.00	4,500.00	4,500.00	4,500.00	6,000.00	6,000.00	7,500.00	7,500.00	9,500.00	9,500.00	9,500.00	9,500.00	93,399.90	102,639.46	112,903.41	124,193.75
Sueldos y salarios (sgs incluida)	**3,333.33**	**3,333.33**	**3,333.33**	**3,333.33**	**8,083.33**	**8,083.33**	**11,250.00**	**11,250.00**	**14,416.67**	**16,916.67**	**16,916.67**	**16,916.67**	**247,610.00**	**361,540.00**	**656,540.00**	**656,540.00**
desarrollo de negocio																
sueldo	-	-	-	-	-	-	-	-	-	-	-	-	5,115.00	10,220.00	15,330.00	20,440.00
desarrollo de software	*3,333.33*	*3,333.33*	*3,333.33*	*3,333.33*	*3,333.33*	*3,333.33*	*3,333.33*	*3,333.33*	*3,333.33*	*3,333.33*	*3,333.33*	*3,333.33*	*40,920.00*	*40,880.00*	*40,880.00*	*40,880.00*
sueldo	3,333.33	3,333.33	3,333.33	3,333.33	3,333.33	3,333.33	3,333.33	3,333.33	3,333.33	3,333.33	3,333.33	3,333.33	40,920.00	40,880.00	40,880.00	40,880.00

recruiting		-	-	-	-	-	-	-	-	-	*2,500.00*	*2,500.00*	*2,500.00*	*30,690.00*	*30,660.00*	*30,660.00*	*30,660.00*
sueldo		-	-	-	-	-	-	-	-	-	2,500.00	2,500.00	2,500.00	30,690.00	30,660.00	30,660.00	30,660.00
consultoría		-	-	-	-	*4,750.00*	*4,750.00*	*7,916.67*	*7,916.67*	*11,083.33*	*11,083.33*	*11,083.33*	*11,083.33*	*176,000.00*	*290,000.00*	*585,000.00*	*585,000.00*
sueldo jr		-	-	-	-	4,750.00	4,750.00	7,916.67	7,916.67	11,083.33	11,083.33	11,083.33	11,083.33	133,000.00	247,000.00	456,000.00	456,000.00
sueldo sr		-	-	-	-	-	-	-	-	-	-	-	-	43,000.00	43,000.00	129,000.00	129,000.00
EBITDA		**(10,947.27)**	**(10,947.27)**	**(10,947.27)**	**(10,947.27)**	**(8,757.27)**	**(8,757.27)**	**(1,347.57)**	**(1,347.57)**	**5,107.58**	**2,607.58**	**2,607.58**	**2,607.58**	**204,990.57**	**518,964.92**	**1,234,948.78**	**1,223,658.44**
Intereses	*x*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Depreciación/amortización	*10*	*629.36*	*1.40*	*1.40*	*1.40*	*580.26*	*2.80*	*292.23*	*3.50*	*587.35*	*9.89*	*9.89*	*9.89*	*425.40*	*1,040.86*	*2,281.20*	*215.63*
amortización de mobiliario		629.36	1.40	1.40	1.40	580.26	2.80	292.23	3.50	587.35	9.89	9.89	9.89	425.40	1,040.86	2,281.20	215.63

EBT		**(11,576.62)**	**(10,948.66)**	**(10,948.66)**	**(10,948.66)**	**(9,337.52)**	**(8,760.06)**	**(1,639.79)**	**(1,351.06)**	**4,520.23**	**2,597.69**	**2,597.69**	**2,597.69**	**204,565.17**	**517,924.05**	**1,232,667.58**	**1,223,442.81**
impuestos	*30*	*-*	*-*	*-*	*-*	*-*	*-*	*-*	*-*	*1,356.07*	*779.31*	*779.31*	*779.31*	*61,369.55*	*155,377.22*	*369,800.27*	*367,032.84*
Impuestos		-	-	-	-	-	-	-	-	1,356.07	779.31	779.31	779.31	61,369.55	155,377.22	369,800.27	367,032.84
Resultado		(11,576.62)	(10,948.66)	(10,948.66)	(10,948.66)	(9,337.52)	(8,760.06)	(1,639.79)	(1,351.06)	3,164.16	1,818.38	1,818.38	1,818.38	143,195.62	362,546.84	862,867.31	856,409.97

Tabla 22: Cuenta de pérdidas y ganancias tras revisión de ventas año 1. Fuente: propia

Como se puede observar, se incurren en unas pérdidas en los primeros 8 meses que suman 63.998,74 eur. Financiación

Una vez calculada la cantidad de dinero que se necesita financiar desde el inicio como la suma del CAPEX y las pérdidas de los primeros meses, siendo esta **124.923.14eur,** se procede a establecer un plan para la financiación de la misma.

Con el objetivo de distribuir el riesgo entre deuda y fondos propios, se decide financiarlo con una estructura de 50% fondos propios y 50% deuda. Con esta estructura, reducimos la dependencia de la deuda y los riesgos asociados a los intereses y a los pagos fijos. Además, conseguiríamos evitar la dilución del control de la empresa al no tener que ceder tanta participación a los inversores.

Por otro lado, emplear una estructura con deuda permite conservar parte del capital de fundación para otras necesidades y los intereses de la deuda son deducibles de impuestos, pudiendo reducir la carga fiscal.

Con esta estructura se financiaría como **fondos propios 62.461,57eur**, y como **deuda** la cantidad restante **62.461,57eur.**

FONDOS PROPIOS

El total de los fondos propios será de 62.461,57eur. Para financiar esta cantidad, en primer lugar, se decide que el fundador de *Pedidata* (Álvaro Torán) aporte una inversión inicial de **40.000eur**, demostrando su compromiso y confianza con el proyecto.

Además, esta aportación le permitirá tener una participación de:

Porcentaje de Participación = (Aporte del Fundador/Total de Fondos Propios) × 100 = **64.04%**

La opción ideal para la financiación del monto restante de fondos propios sería un socio cofundador, alguien que permita complementarse y facilite el desarrollo de la empresa, pero en

caso de que no se pueda dar esta opción, se justifica la necesidad de emplear un Inversor Angel. También se podría tratar de financiar la cantidad total, en parte con la ayuda de premios como 'Everis/NTT Data' que ofrecen capital a *start-ups* e innovaciones tecnológicas o 'Open innovation Challenges'. También se podría buscar financiación a través de becas como 'ENISA' o 'NEOTEC'. Pero estas formas de financiación no son garantizadas y se estudia la posibilidad de necesitar inversores.

Los 22.461,57 se financiarán a través de un 'inversor ángel' y/o 'friends and families' ,entendiendo que es más sencilla esta segunda vía ya que no se cuentan con resultados reales, a cambio de una participación accionaria.

DEUDA

Para financiar la cantidad de deuda total de 62.461,57eur se pedirá un préstamo bancario.

Se considerará un préstamo bancario de 8% de interés con madurez de 5 años.

Conociendo esto, sabemos que la tasa de interés mensual será de 0.67%. Calculando los pagos mensuales para el primer año como:

$$M = P * \frac{r * (1+r)^n}{(1+r)^n - 1}$$

Donde:
M es el pago mensual.
P es el principal del préstamo (62,461.57 euros).
r es la tasa de interés mensual (0,08 / 12).
n es el número total de pagos (60meses).

Una vez calculado el pago mensual, podemos calcular el interés mensual como

$$It = R_t * r$$

Estos cálculos se repiten para todos los meses, todos los años, obteniendo los siguientes resultados:

Mes 1	Mes 2	Mes 3	Mes 4	Mes 5	Mes 6	Mes 7	Mes 8	Mes 9	Mes 10	Mes 11	Mes 12	Año 2	Año 3	Año 4	Año 5
416.41	410.74	405.04	399.3	393.51	387.69	381.84	375.94	370	364.02	358.01	351.95	3736.03	2784.69	1754.4	638.59

Tabla 23: Pagos de interés por el préstamo bancario. Fuente: Propia

Conocido el interés podemos calcular el resultado del ejercicio como:

PyG		Mes 1	Mes 2	Mes 3	Mes 4	Mes 5	Mes 6	Mes 7	Mes 8	Mes 9	Mes 10	Mes 11	Mes 12	Año 2	Año 3	Año 4	Año 5
EBITDA		**(10,947.27)**	**(10,947.27)**	**(10,947.27)**	**(10,947.27)**	**(8,757.27)**	**(8,757.27)**	**(1,347.57)**	**(1,347.57)**	**5,107.58**	**2,607.58**	**2,607.58**	**2,607.58**	**204,990.57**	**518,964.92**	**1,234,948.78**	**1,223,658.44**
Intereses	*8*	*416.41*	*410.74*	*405.04*	*399.30*	*393.51*	*387.69*	*381.84*	*375.94*	*370.00*	*364.02*	*358.01*	*351.95*	*3,736.03*	*2,784.69*	*1,754.40*	*638.59*
		416.41	410.74	405.04	399.30	393.51	387.69	381.84	375.94	370.00	364.02	358.01	351.95	3,736.03	2,784.69	1,754.40	638.59
Depreciación/amortización	*10*	*629.36*	*1.40*	*1.40*	*1.40*	*580.26*	*2.80*	*292.23*	*3.50*	*587.35*	*9.89*	*9.89*	*9.89*	*425.40*	*1,040.86*	*2,281.20*	*215.63*
amortización de mobiliario		629.36	1.40	1.40	1.40	580.26	2.80	292.23	3.50	587.35	9.89	9.89	9.89	425.40	1,040.86	2,281.20	215.63
EBT		**(11,993.03)**	**(11,359.40)**	**(11,353.70)**	**(11,347.96)**	**(9,731.03)**	**(9,147.75)**	**(2,021.63)**	**(1,727.00)**	**4,150.23**	**2,233.67**	**2,239.68**	**2,245.74**	**200,829.14**	**515,139.36**	**1,230,913.18**	**1,222,804.22**
impuestos	*30*	*-*	*-*	*-*	*-*	*-*	*-*	*-*	*-*	*1,245.07*	*670.10*	*671.90*	*673.72*	*60,248.74*	*154,541.81*	*369,273.95*	*366,841.27*

Impuestos	-	-	-	-	-	-	-	-	1,245.07	670.10	671.90	673.72	60,248.74	154,541.81	369,273.95	366,841.27
Resultado	(11,993.03)	(11,359.40)	(11,353.70)	(11,347.96)	(9,731.03)	(9,147.75)	(2,021.63)	(1,727.00)	2,905.16	1,563.57	1,567.78	1,572.02	140,580.40	360,597.55	861,639.23	855,962.96

Tabla 24: Resultado del ejercicio. Fuente: Propia

KPIs DEL INVERSOR

De cara a justificar la decisión del inversor, se calculan los siguientes KPIs que muestran una imagen transparente del estado de salud financiera de la empresa.

1. ROI

Teniendo en cuenta que, con la inversión inicial, la participación del inversor seria de un 35.96%, se calcula el Return on Investment como el beneficio neto a 5 años relativo a su participación/su inversión:

$$ROI = \frac{35.96\% * (Beneficio\ Neto\ a\ 5\ años)}{Inversión\ del\ IA} = \frac{(778.113{,}60)}{22.461{,}47} = 3464.21\%$$

Cabe mencionar que a pesar de lo sorprendente que pueda parecer el resultado, las hipótesis realizadas para el cálculo del plan de negocio serían realistas. Esta sería una cifra muy positiva de cara a la factibilidad comercial del negocio.

2. TIR

Calculando la tasa interna de retorno como:

$$0 = \sum_{t=0}^{n} \frac{CF_t}{(1+TIR)^t}$$

Obtenemos un TIR de la inversión de 140%. Este resultado es sorprendente, y aparentemente elevado, sin embargo, las hipótesis realizadas a lo largo del proyecto son conservadoras. Además, el resultado cuadra con el plan de negocio, ya que se obtienen beneficios netos positivos desde el segundo año. Esto quiere decir que la inversión sería muy rentable para el inversor, justificando la misma.

3. Payback

El payback estimado de la inversión es de 1,86 años. De nuevo, este resultado cuadra con el plan de negocio, ya que según las proyecciones obtendríamos beneficios a partir del segundo año.

Con toda esta información podemos visualizar la proyección de los beneficios netos y acumulados de *Pedidata*:

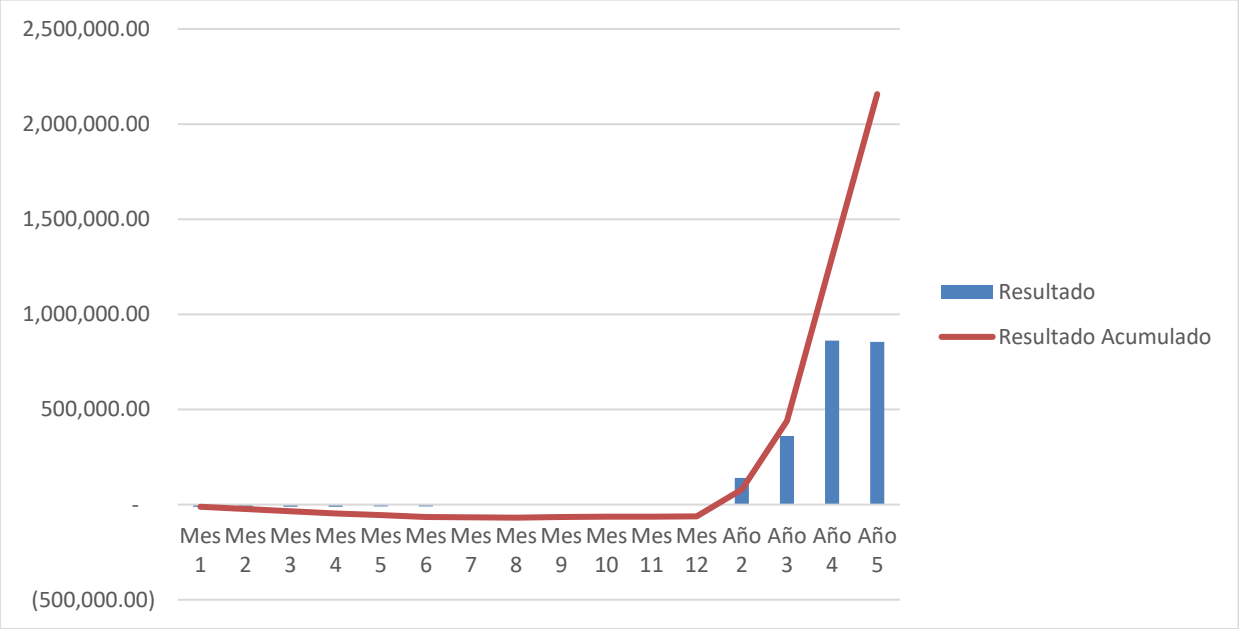


Ilustración 20: Proyección de beneficios (desglose mensual año 1). Fuente: Propia

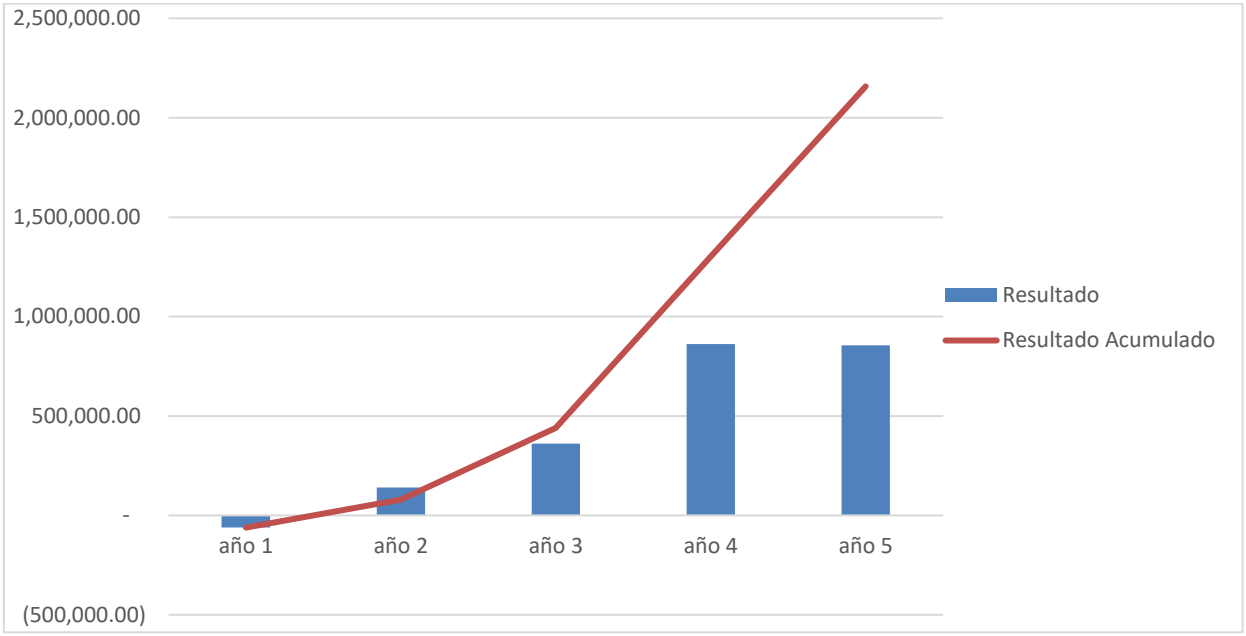


Ilustración 21: Proyección de beneficios. Fuente: Propia

CONCLUSIONES

CONCLUSIÓN SOBRE EL PROYECTO

Con el Business Plan desarrollado, en base a los resultados obtenidos, se puede concluir que de cara al aspecto financiero se trata de una buena oportunidad, tanto hoy en día como en el futuro. El cambio de las tendencias en el uso de la tecnología para el estudio y monitorización de la salud infantil indica que cada vez se monitoriza más el cuidado de la salud con la ayuda de aplicaciones y tecnologías. Además, con los constantes avances en medicina, es importante mantenerse informado, y el fácil acceso a información en medios alternativos a los profesionales, hace necesaria la existencia de fuentes de datos fiables para asegurar que las medidas de atención sean las correctas.

Es por esto por lo que el empleo de soluciones como el ML puede suponer la diferencia entre unos datos fiables y unos datos erróneos. Con la tendencia de las empresas a la adopción de esta tecnología para obtener 'insights' relevantes, así como para poder predecir e identificar tendencias, encontramos un mercado cambiante y creciente en el que penetrar con la solución propuesta.

En lo que se refiere a la factibilidad técnica, se puede observar que se trata de un servicio bien estructurado, cuyos resultados son observables en el producto mínimo viable. Teniendo en cuenta la estructura de costes, hay que prestar especial atención al contratar el personal, ya que, al ser un servicio, son el principal activo, por tanto, hay que atender con detalle a los perfiles descritos para poder desarrollar de manera eficiente el negocio. A pesar de la estructura empresarial que se ha decidido, es importante que exista una colaboración entre todos los miembros del equipo, ya que favorece el crecimiento y mejora continuos.

Con el objetivo de optimizar la estructura del negocio, se decide externalizar la labor de marketing, gestoría y abogados, esto permitirá centrarse en las áreas que son más esenciales dentro de *Pedidata* como los son el desarrollo de software, para poder crear, mantener y desplegar las soluciones *cloud* que necesitemos en nuestros proyectos, recruiting, para asegurarse que la capacidad de los consultores es la adecuada para desarrollar el servicio y poder prevenir posibles problemas relacionados con las altas rotaciones que se dan en el mercado y consultoría, la cual es la esencia del servicio y sin la que este carecería de sentido.

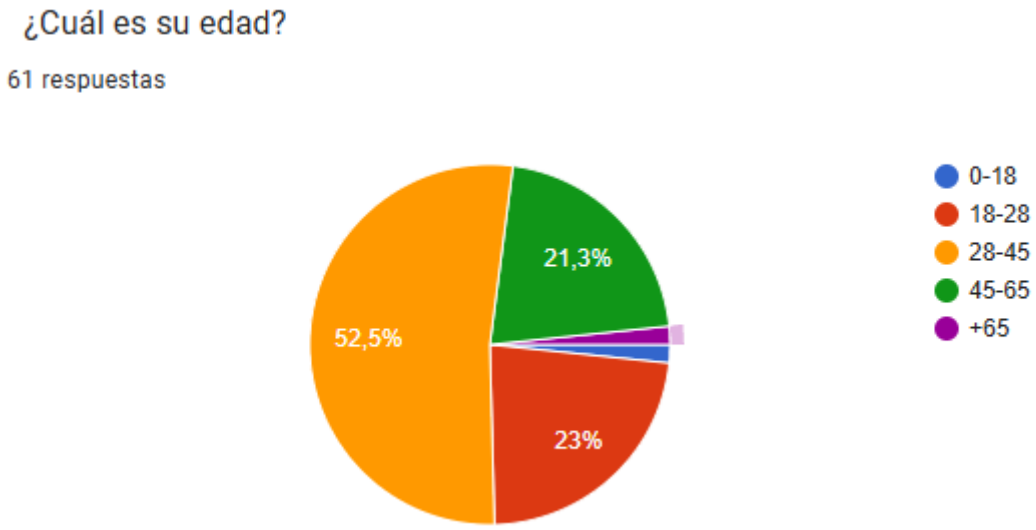
Teniendo en cuenta todo este contexto, se puede concluir que se trata de un proyecto factible desde el punto de vista técnico y financiero y atractivo desde el punto de vista comercial.

ANEXOS

ANEXO A – ENCUESTA SOBRE LAS TENDENCIAS EN LOS CUIDADOS INFANTILES Y USO DE TECNOLOGÍA

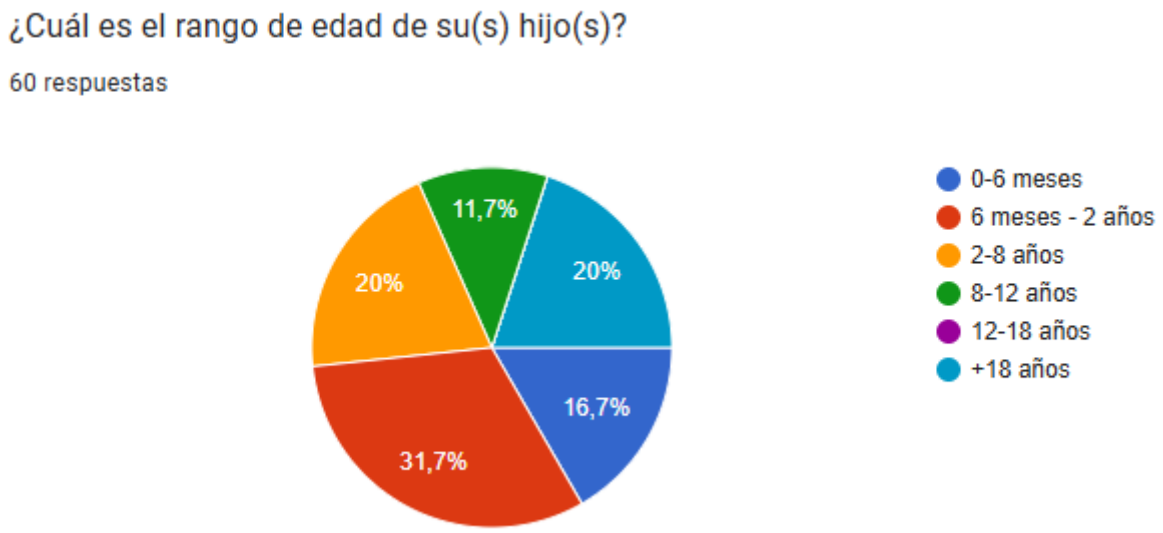
Se ha realizado una encuesta a una muestra de más de 50 personas en formato online y durante 10 días con la intención de validar la viabilidad comercial del servicio, estudiando tendencias y cambios y validando información adquirida por otras fuentes.

En primer lugar, se hace un estudio demográfico de la muestra, esto nos ayudará a ver las tendencias que existen, y cómo van cambiando con el tiempo en cuanto al empleo de la tecnología y el cuidado infantil.

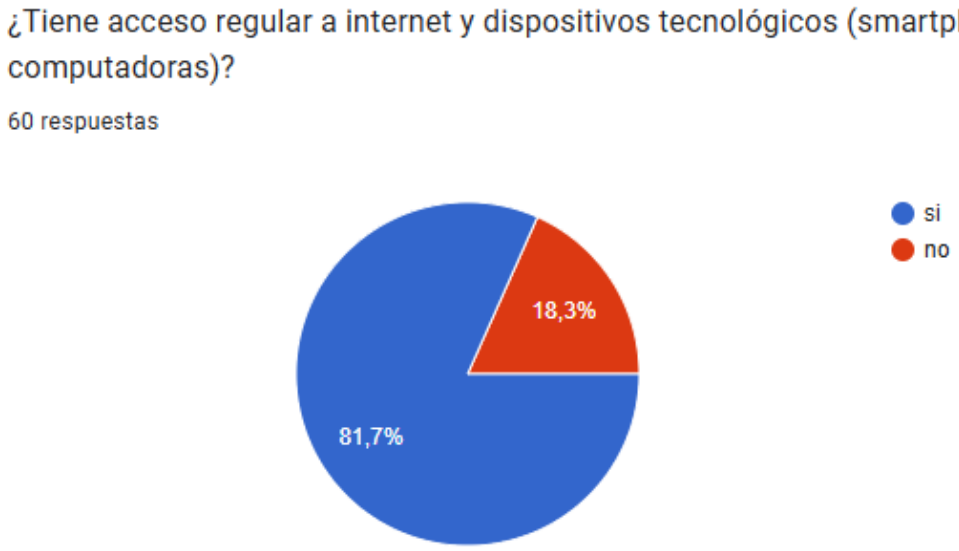


Como se puede observar el gran grueso de la muestra se encuentra entre las edades de 28 y 45 años, seguido de 18 y 28 años y 45-65 años. Esto nos indica que la gran mayoría en la muestra son jóvenes en los que se podrían ver reflejados los cambios relacionados con los avances tecnológicos y médicos de los últimos años.

Posteriormente se preguntó por la edad de los hijos de la muestra, de esta manera podemos ver si se trata de cuidados infantiles u otros.

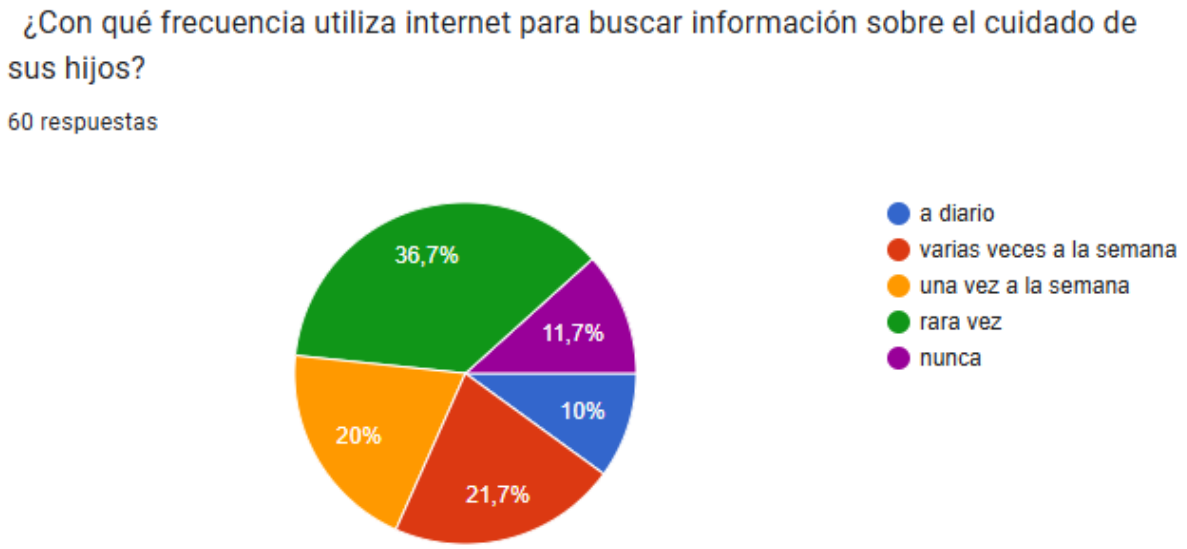


Como se puede observar en los datos obtenidos, más del 50% de los hijos estaban por debajo de la edad de 2 años y más del 70% por debajo de los 8 años, lo que acota las respuestas al cuidado infantil en su mayoría.



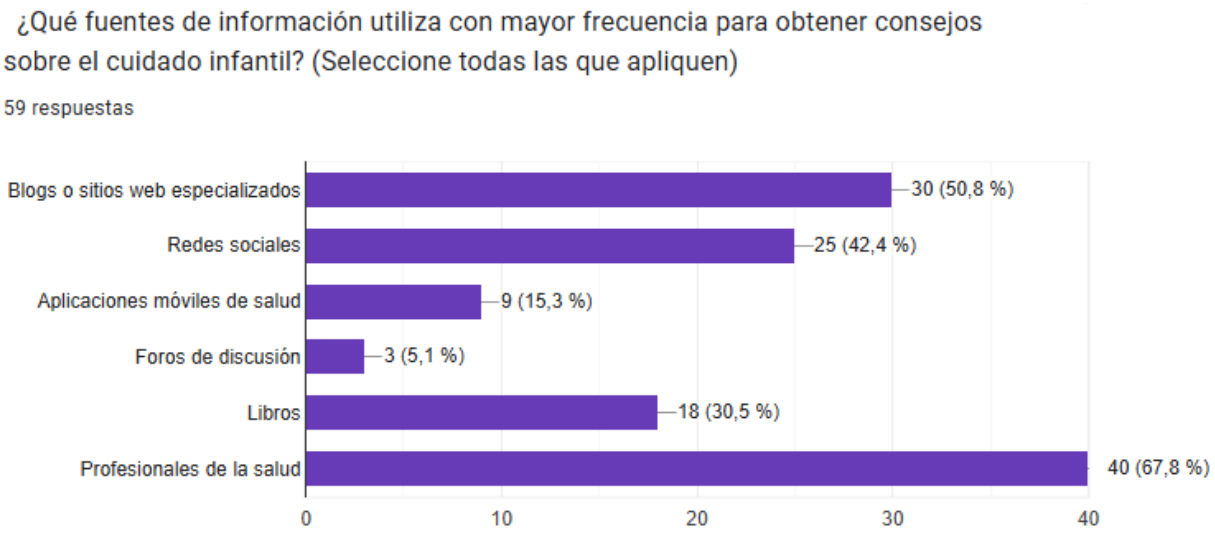
Observando los datos, llegamos a la conclusión que más del 80% de los encuestados tienen fácil acceso a internet, de donde pueden obtener información acerca del cuidado de sus hijos con facilidad.

Después se analizó si los encuestados empleaban internet para obtener información del cuidado de sus hijos.



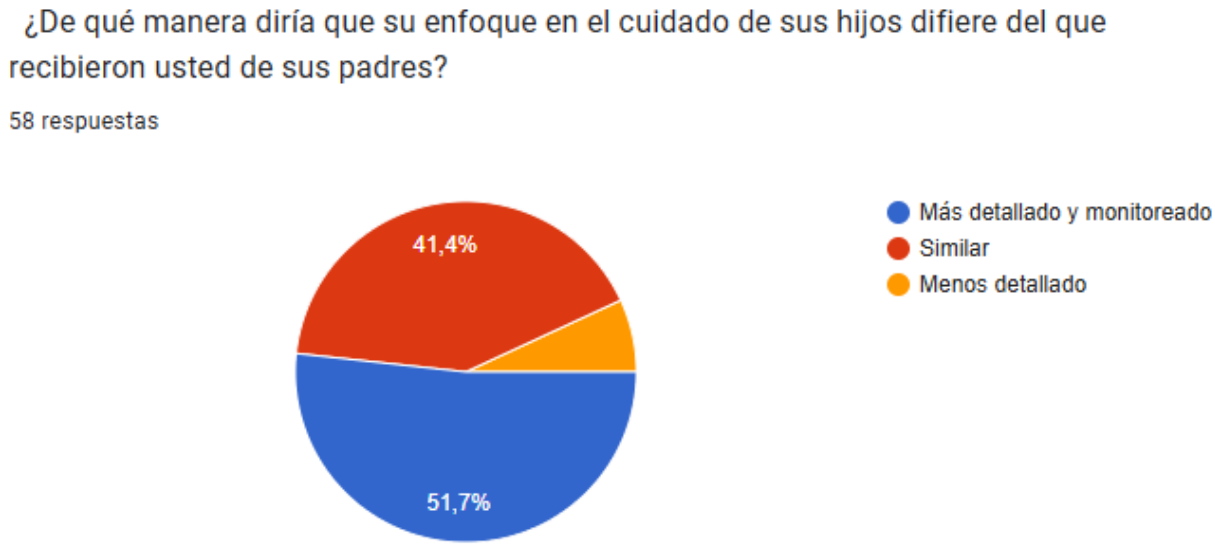
Como se puede observar por las respuestas, más del 50% emplean internet con frecuencia para obtener información adicional sobre el cuidado de sus hijos.

Y de entre las fuentes de internet, las más empleadas son:



Como se puede observar, de entre las principales fuentes de información, más del 50% de la gente encuestada emplea blogs o sitios web especializados y un 42% redes sociales. Esto demuestra que existe una gran confianza a la hora de buscar información en internet.

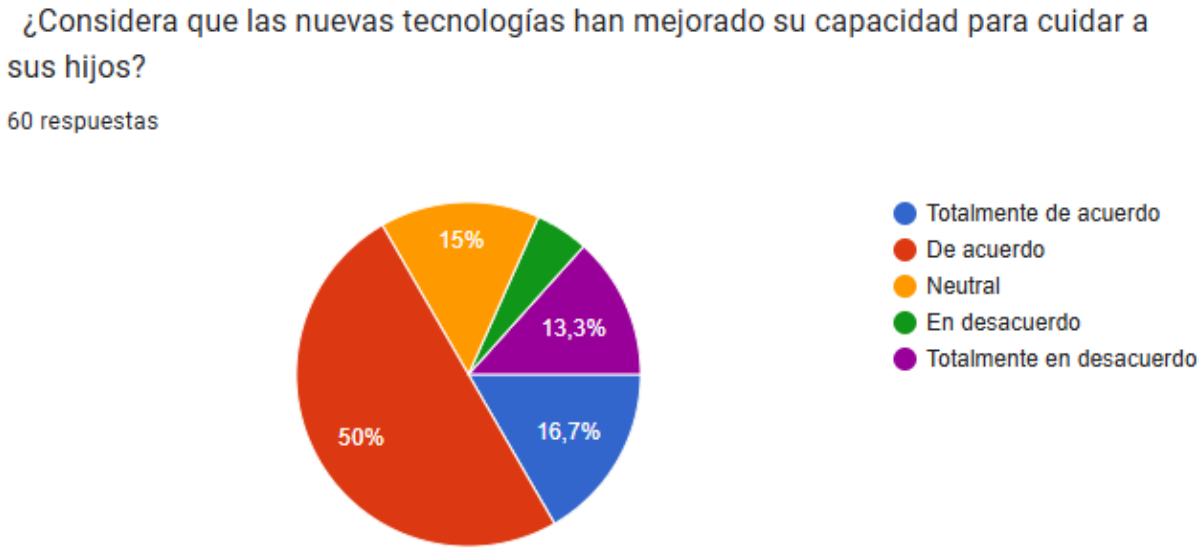
Posteriormente se trató de comparar el cuidado infantil realizado por la muestra con respecto a la percepción del cuidado infantil en el pasado.



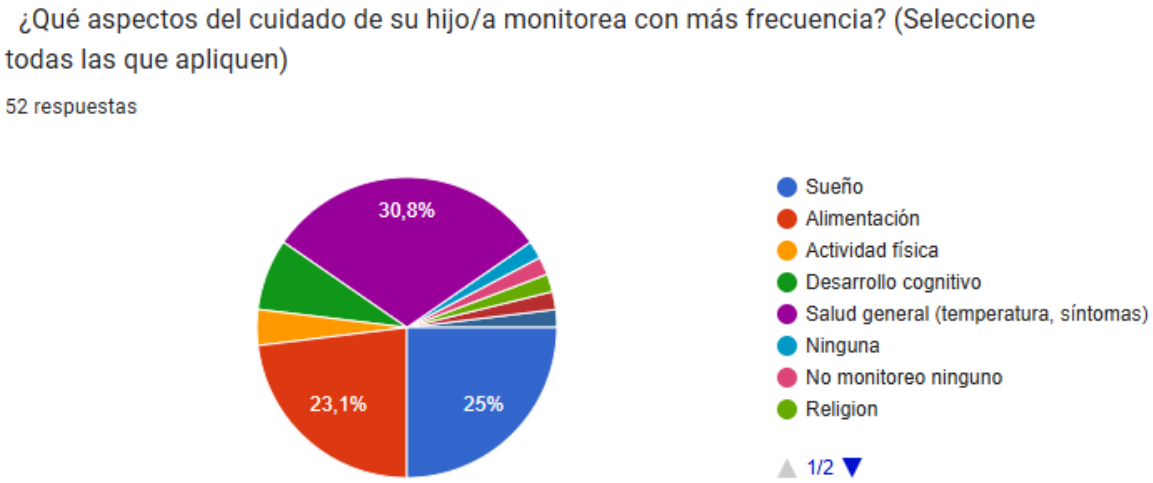
Como se puede observar, más del 50% tiene la percepción de que hoy en día el cuidado infantil es más detallado y monitoreado.



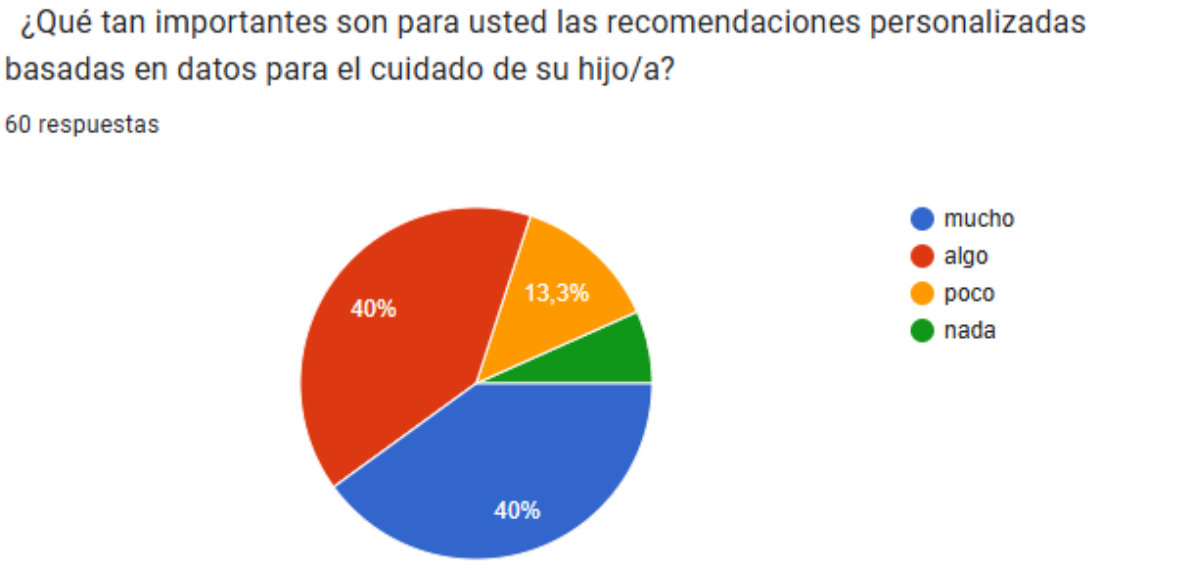
Como se puede observar, en relación con los resultados obtenidos anteriormente respecto al uso de internet para obtener información adicional, más del 88% coincide en que anteriormente no se usaban estas fuentes y tecnologías, validando la tendencia actual y el cambio de paradigma.



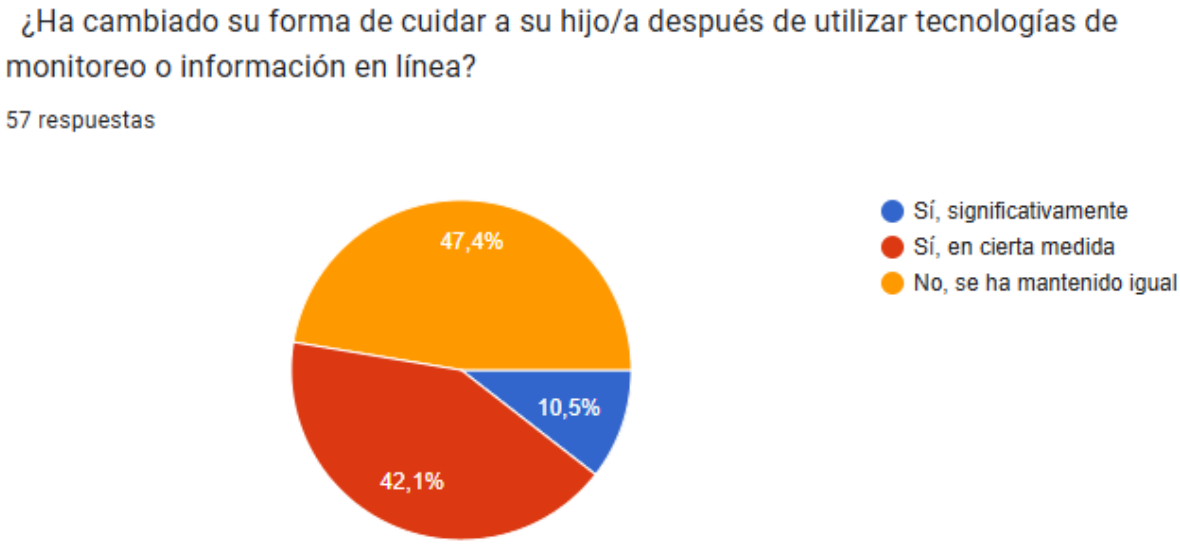
Además, más del 65% están de acuerdo en que el uso de la tecnología ha facilitado el cuidado de sus hijos.



Entre los aspectos más monitorizados por los encuestados empleando el uso de tecnología están la alimentación y la salud general. Esto justifica la existencia de empresas dedicadas al cuidado infantil mediante el uso de tecnología, lo que valida la idea de que el mercado de estas empresas de salud está en aumento.



Como se puede observar, los encuestados valoran el hecho de que las recomendaciones recibidas de forma personalizada puedan estar basadas en datos.

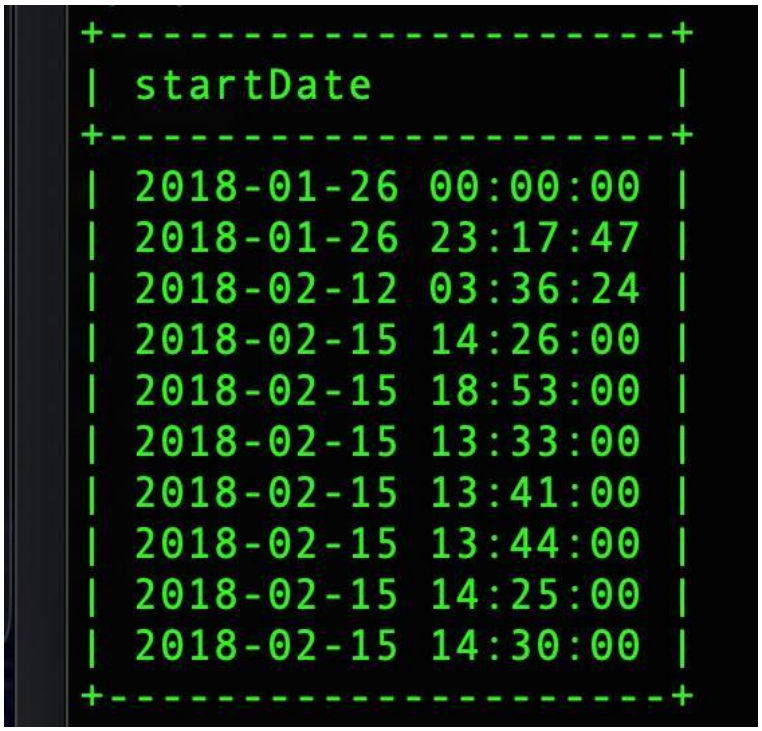


Además, más del 50% de los encuestados admite que su forma de cuidar a sus hijos ha cambiado con el uso de tecnologías e información de internet. De nuevo confirmando el cambio de paradigma.

ANEXO B – DOCUMENTACIÓN DEL ARCHIVO DE DATOS DE LACTAPP

Export feeding tracker data

Data exported:
TrackerID User data Including:
userID
username
bornDate
city (no country available?)
numSons
dueDate Baby data Including:
baby1ID
name
genre
bornDate
premature
gestationWeek
bornHeight
Baby2 data
For tandem breasteeding, two babies at the same supported. Not more than two. Same information as Baby Data. For quering non tandem breastfeeding (e.g., two or more babies not at the same time) group information by userID looking for different baby1ID.
startDate
Formatted date for starting the breastfeeding



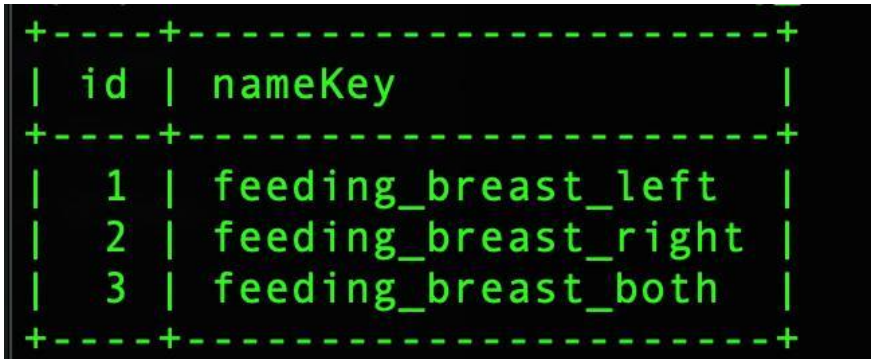
startDate
2018-01-26 00:00:00
2018-01-26 23:17:47
2018-02-12 03:36:24
2018-02-15 14:26:00
2018-02-15 18:53:00
2018-02-15 13:33:00
2018-02-15 13:41:00
2018-02-15 13:44:00
2018-02-15 14:25:00
2018-02-15 14:30:00

endDate

For ending the breastfeeding, same format as the start date.

breast

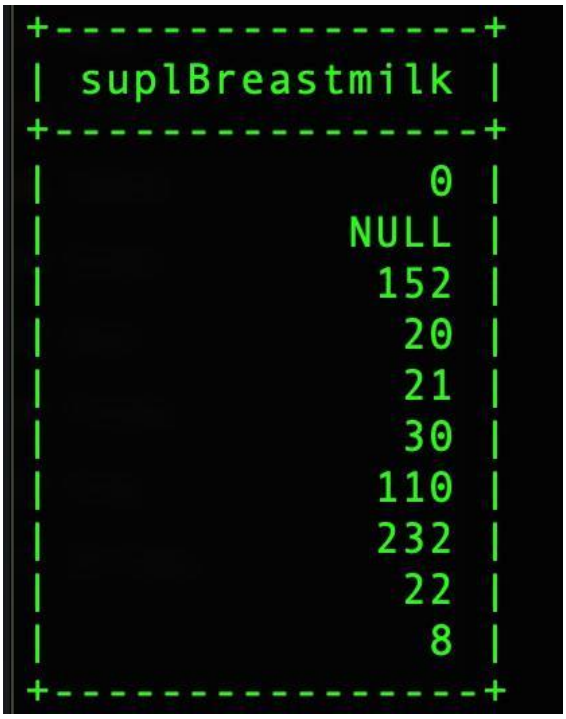
Breast used (may be unknown) from feeding_breast table.



id	nameKey
1	feeding_breast_left
2	feeding_breast_right
3	feeding_breast_both

suplBreastmilk

Quantity of supl milk used, if any (in ml??)



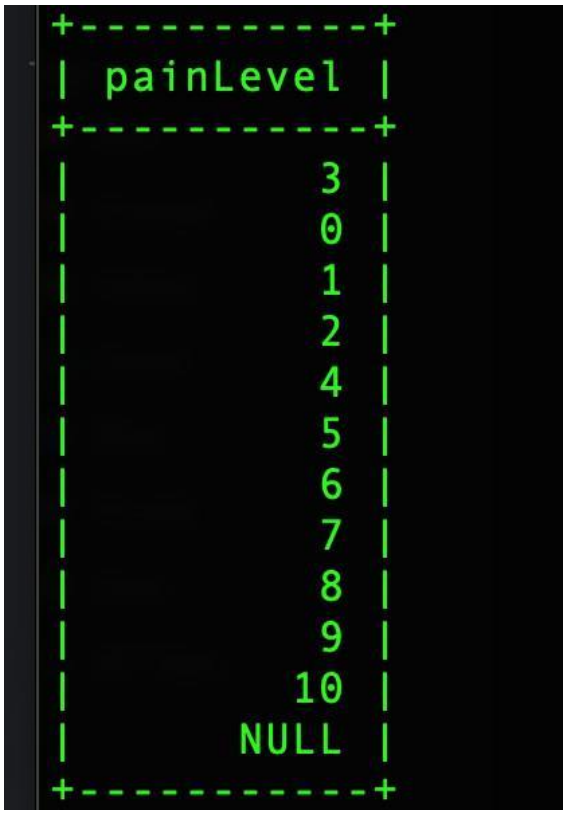
suplBreastmilk
0
NULL
152
20
21
30
110
232
22
8

suplFormula

Same as suplBreastmilk for formula milk.

painLevel

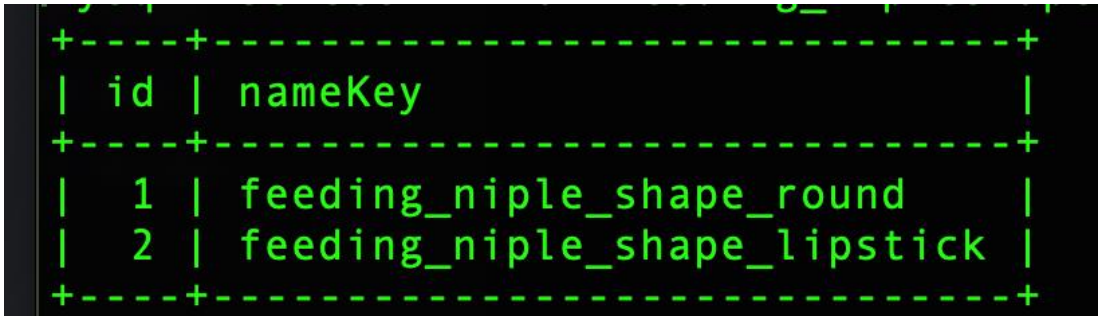
Ranges from 0-10 can be NULL (Unknown?)



painLevel
3
0
1
2
4
5
6
7
8
9
10
NULL

nipleShape

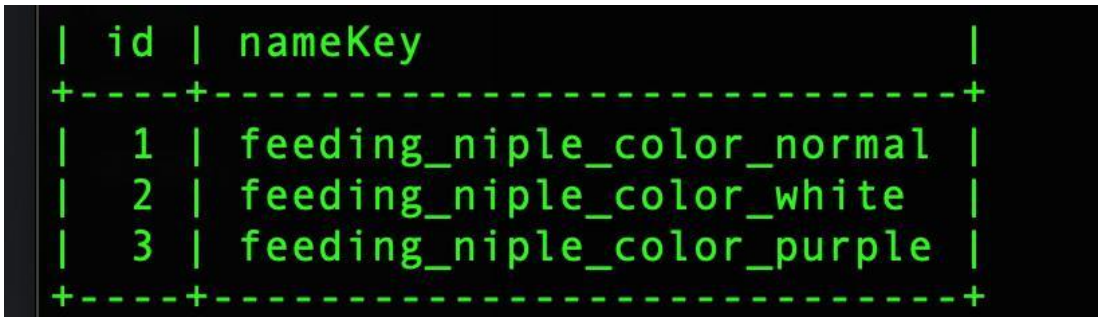
Nipple shape from feeding_nipleshape table, may be Unknown



id	nameKey
1	feeding_niple_shape_round
2	feeding_niple_shape_lipstick

feeding_niplecolor

Nipple color from feeding_niplecolor table, may be Unknown



id	nameKey
1	feeding_niple_color_normal
2	feeding_niple_color_white
3	feeding_niple_color_purple

milkPearl

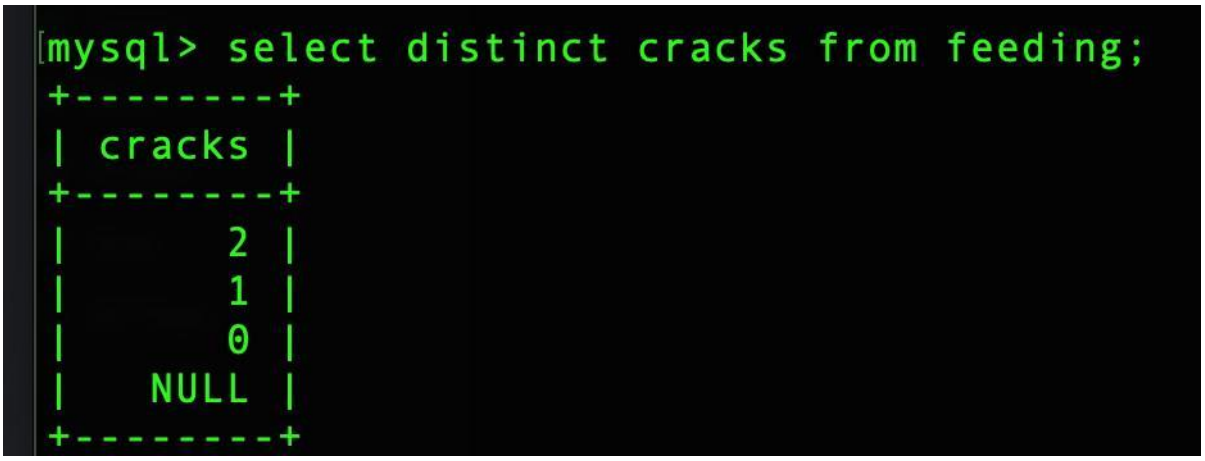
Present 1, absent 0 or unknown.



milkPearl
1
0
NULL

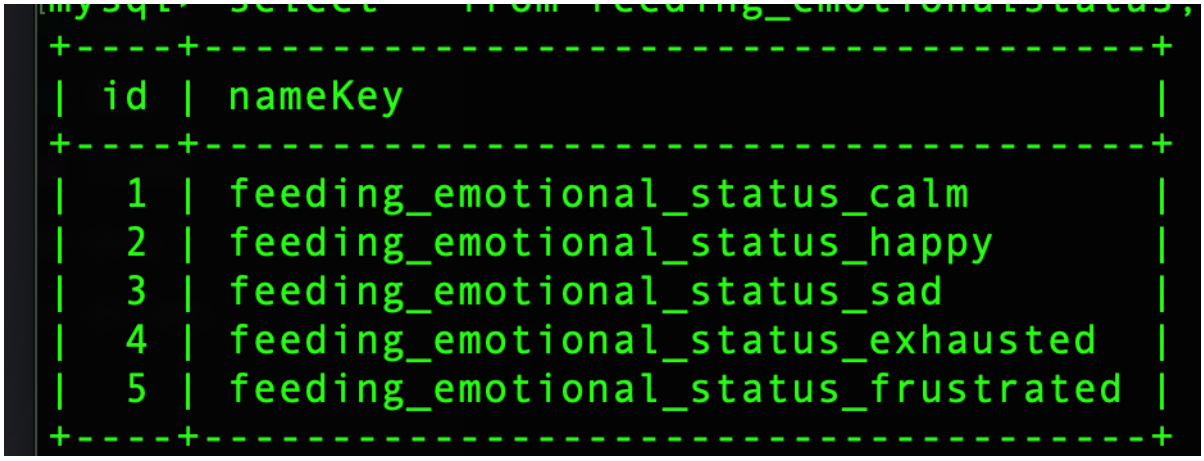
Cracks

Level of cracks, can be absent 0 or Unkown NULL



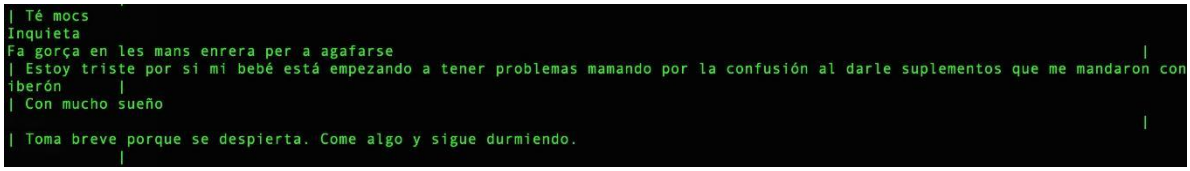
emotionalStatus

Relates to feeding_emotionalstatus table, can be unknown and have multiple values (e.g., user can be calm and happy).



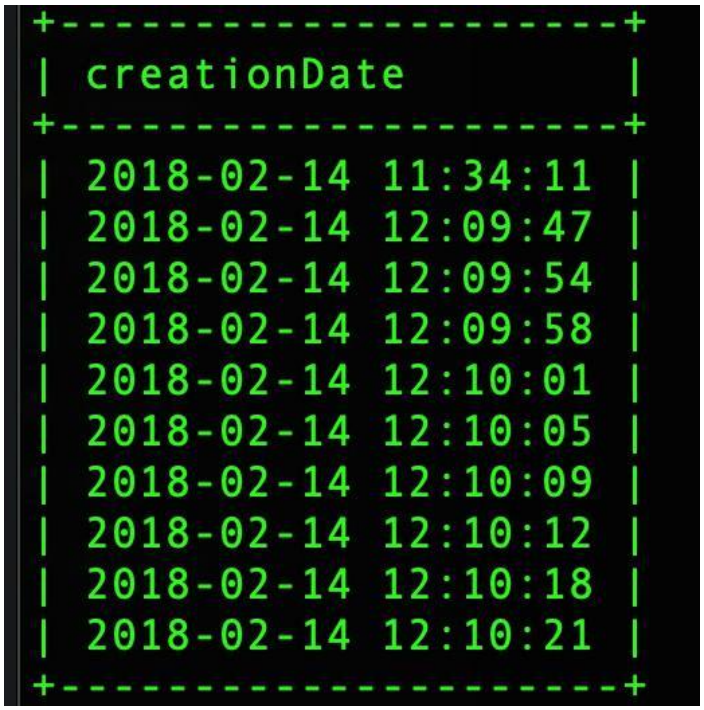
notes

Free text notes for the feeding.



creationDate

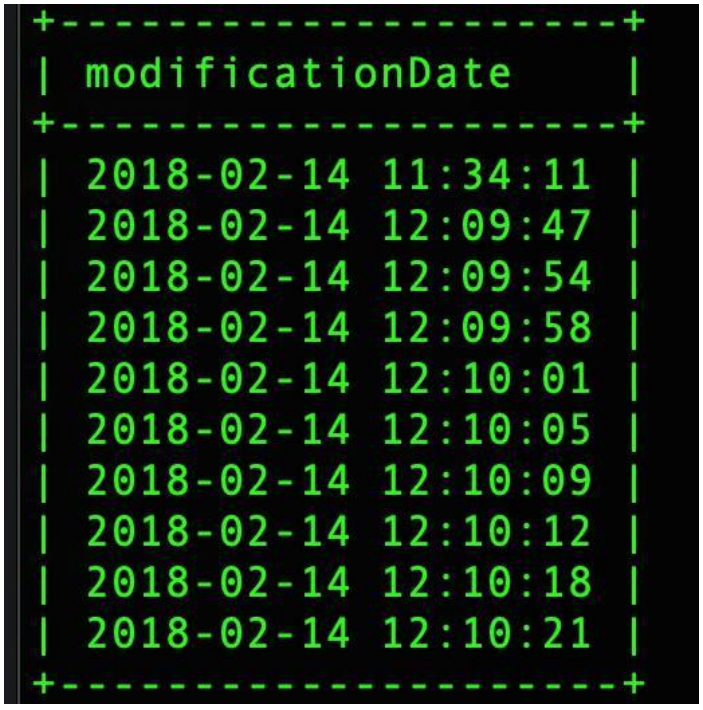
Tracker creation date



creationDate
2018-02-14 11:34:11
2018-02-14 12:09:47
2018-02-14 12:09:54
2018-02-14 12:09:58
2018-02-14 12:10:01
2018-02-14 12:10:05
2018-02-14 12:10:09
2018-02-14 12:10:12
2018-02-14 12:10:18
2018-02-14 12:10:21

modificationDate

Tracker update date



modificationDate
2018-02-14 11:34:11
2018-02-14 12:09:47
2018-02-14 12:09:54
2018-02-14 12:09:58
2018-02-14 12:10:01
2018-02-14 12:10:05
2018-02-14 12:10:09
2018-02-14 12:10:12
2018-02-14 12:10:18
2018-02-14 12:10:21

Program flow:

Build query joining information from user and first baby via JOIN.

Information on the second baby is included via a subquery in python, leaving information as NA (not available) if second baby does not exist or can not be found. Done using this procedure as MYSQL is not very efficient when combining inner and outer joins.

Information on rest of join tables (e.g., breast, nipple color and nipple shape) done via fixed dictionaries in python. Again MYSQL is not very efficient when combining inner and outer joins and in this case the join information is very small.

Information on feeding_emotionalstatus done splitting the string by ‘,’ character, removing trailing spaces and joining via a dictionary analogous to the previous ones. In this case the free format of the information does only enable this method.

User’s and babies names are obfuscated (e.g., replaced by ***) for privacy. We did not remove so field order is preserved and is easier to add them back if required.

Once all the information is present, it is exported to CSV.

ANEXO C - CÓDIGO DEL ANÁLISIS

```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
from sklearn.linear_model import LinearRegression
from scipy.stats import mannwhitneyu

```

```python
# carga de los datos original, se mantendrá como backup
file_path = 'data/breastfeeding_trackers.csv'
original_data = pd.read_csv(file_path, low_memory=False)
```

```python
data = original_data.copy()
```

```python
# Verify the data types
data.info()
```

```
    <class 'pandas.core.frame.DataFrame'>
    RangeIndex: 6872735 entries, 0 to 6872734
    Data columns (total 35 columns):
     #   Column                  Dtype
    ---  ------                  -----
```

0	trackID	int64
1	userID	int64
2	baby1ID	int64
3	baby2ID	float64
4	userName	object
5	bornDate	object
6	city	object
7	numSons	float64
8	dueDate	object
9	babyName	object
10	genre	object
11	bornDate.1	object
12	premature	int64
13	gestationWeek	float64
14	bornHeight	object
15	babyTandemName	object
16	babyTandemGenre	object
17	babyTandemBornDate	object
18	babyTandemPremature	float64
19	babyTandemGestationWeek	object
20	babyTandemBornHeight	object
21	startDate	object
22	endDate	object
23	breast	object
24	suplBreastmilk	object
25	suplFormula	object
26	painLevel	object
27	nipleShape	object
28	nipleColor	object
29	milkPearl	object
30	cracks	object
31	emotionalStatus	object
32	notes	object
33	creationDate	object
34	modificationDate	object

```
    dtypes: float64(4), int64(4), object(27)
    memory usage: 1.8+ GB
```

```python
# Eliminar las columnas censuradas
columns_to_drop = ['userName', 'babyName', 'babyTandemName',
'babyTandemGestationWeek', 'city', 'notes']
data = data.drop(columns=columns_to_drop)

```

```python
# Verify the data types
data.info()
```

```
    <class 'pandas.core.frame.DataFrame'>
    RangeIndex: 6872735 entries, 0 to 6872734
    Data columns (total 29 columns):
     #   Column               Dtype
    ---  ------               -----
     0   trackID              int64
     1   userID               int64
     2   baby1ID              int64
     3   baby2ID              float64
     4   bornDate             object
     5   numSons              float64
     6   dueDate              object
     7   genre                object
     8   bornDate.1           object
     9   premature            int64
     10  gestationWeek        float64
```

```
 11  bornHeight            object
 12  babyTandemGenre       object
 13  babyTandemBornDate    object
 14  babyTandemPremature   float64
 15  babyTandemBornHeight  object
 16  startDate             object
 17  endDate               object
 18  breast                object
 19  suplBreastmilk        object
 20  suplFormula           object
 21  painLevel             object
 22  nipleShape            object
 23  nipleColor            object
 24  milkPearl             object
 25  cracks                object
 26  emotionalStatus       object
 27  creationDate          object
 28  modificationDate      object
dtypes: float64(4), int64(4), object(21)
memory usage: 1.5+ GB
```

```python
# Definir las columnas críticas
critical_columns = ['trackID', 'userID', 'baby1ID', 'bornDate.1', 'startDate',
'endDate', 'creationDate', 'modificationDate']
```

```python
# Verificar valores nulos en las columnas críticas
print("Valores nulos antes de la limpieza:")
print(data[critical_columns].isnull().sum())
```

```
    Valores nulos antes de la limpieza:
    trackID                 0
    userID                  0
    baby1ID                 0
    bornDate.1              0
    startDate               0
    endDate             85874
    creationDate            0
    modificationDate        0
    dtype: int64
```

```python
# Eliminar filas con valores nulos en las columnas críticas
data = data.dropna(subset=critical_columns)
```

```python
# Verificar nuevamente que no haya valores nulos en las columnas críticas
print("\nValores nulos después de la limpieza:")
print(data[critical_columns].isnull().sum())

```

```
    Valores nulos después de la limpieza:
    trackID             0
    userID              0
    baby1ID             0
    bornDate.1          0
    startDate           0
    endDate             0
```

```
    creationDate        0
    modificationDate    0
    dtype: int64
```

```python
# Verificar el tipo de datos de la columna trackID
print("Tipo de dato de trackID:")
print(data['trackID'].dtype)

# Verificar si todos los valores en trackID son enteros
print("\n¿Todos los valores en trackID son enteros?")
are_all_integers = pd.api.types.is_integer_dtype(data['trackID'])
print(are_all_integers)

# Verificar si hay algún valor nulo en trackID
print("\n¿Hay algún valor nulo en trackID?")
has_nulls = data['trackID'].isnull().any()
print(has_nulls)

# Verificar si todos los valores en trackID son únicos
print("\n¿Todos los valores en trackID son únicos?")
are_all_unique = data['trackID'].is_unique
print(are_all_unique)
```

```
    Tipo de dato de trackID:
    int64

    ¿Todos los valores en trackID son enteros?
    True

    ¿Hay algún valor nulo en trackID?
    False
```

```
    ¿Todos los valores en trackID son únicos?
    True
```

```python
# Verificar el tipo de datos de la columna userID
print("Tipo de dato de userID:")
print(data['userID'].dtype)

# Verificar si todos los valores en userID son enteros
print("\n¿Todos los valores en userID son enteros?")
are_all_integers = pd.api.types.is_integer_dtype(data['userID'])
print(are_all_integers)

# Verificar si hay algún valor nulo en userID
print("\n¿Hay algún valor nulo en userID?")
has_nulls = data['userID'].isnull().any()
print(has_nulls)

# Mostrar una muestra de valores únicos para verificación
print("\nMuestra de valores únicos en userID:")
print(data['userID'].unique()[:10])  # Mostrando los primeros 10 valores únicos
como muestra
```

```
    Tipo de dato de userID:
    int64

    ¿Todos los valores en userID son enteros?
    True

    ¿Hay algún valor nulo en userID?
    False
```

```
    Muestra de valores únicos en userID:
    [16270  2738 27253 38462 18082 18464 26936 16932 34129 27887]
```

```python
# Verificar el tipo de datos de la columna baby1ID
print("Tipo de dato de baby1ID:")
print(data['baby1ID'].dtype)

# Verificar si todos los valores en baby1ID son enteros
print("\n¿Todos los valores en baby1ID son enteros?")
are_all_integers = pd.api.types.is_integer_dtype(data['baby1ID'])
print(are_all_integers)

# Verificar si hay algún valor nulo en baby1ID
print("\n¿Hay algún valor nulo en baby1ID?")
has_nulls = data['baby1ID'].isnull().any()
print(has_nulls)

# Verificar que cada userID pueda tener múltiples valores únicos de baby1ID
print("\nVerificación de múltiples valores únicos de baby1ID por userID:")
multiple_babies_per_user =
data.groupby('userID')['baby1ID'].nunique().value_counts().head(10)
print(multiple_babies_per_user)
```

```
    Tipo de dato de baby1ID:
    int64

    ¿Todos los valores en baby1ID son enteros?
    True

    ¿Hay algún valor nulo en baby1ID?
```

```
    False

    Verificación de múltiples valores únicos de baby1ID por userID:
    baby1ID
    1    74500
    2     2159
    3       58
    4        2
    5        1
    Name: count, dtype: int64
```

```python
# Calcular el número de usuarios únicos en el conjunto de datos actual para comporbar si la agrupacion anterior concuerda
unique_users_current = data['userID'].nunique()
print(f"Número de usuarios únicos en el conjunto de datos actual: {unique_users_current}")
```

```
    Número de usuarios únicos en el conjunto de datos actual: 76720
```

```python
# Verificar el tipo de datos de la columna baby2ID
print("Tipo de dato de baby2ID:")
print(data['baby2ID'].dtype)

# Verificar si todos los valores en baby2ID son enteros o ceros
print("\n¿Todos los valores en baby2ID son enteros o ceros?")
are_all_int_or_zero = (data['baby2ID'].apply(lambda x: isinstance(x, int) and x >= 0)).all()
print(are_all_int_or_zero)
```

```
# Verificar si hay algún valor nulo en baby2ID
print("\n¿Hay algún valor nulo en baby2ID?")
has_nulls_baby2ID = data['baby2ID'].isnull().any()
print(has_nulls_baby2ID)
```

```
    Tipo de dato de baby2ID:
    float64

    ¿Todos los valores en baby2ID son enteros o ceros?
    False

    ¿Hay algún valor nulo en baby2ID?
    True
```

```python
# Reemplazar valores nulos en baby2ID con 0
data['baby2ID'] = data['baby2ID'].fillna(0).astype(int)

# Verificar el tipo de datos de la columna baby2ID
print("Tipo de dato de baby2ID:")
print(data['baby2ID'].dtype)

# Verificar si todos los valores en baby2ID son enteros o ceros
print("\n¿Todos los valores en baby2ID son enteros o ceros?")
are_all_int_or_zero = (data['baby2ID'].apply(lambda x: isinstance(x, int) and x
>= 0)).all()
print(are_all_int_or_zero)

# Verificar si hay algún valor nulo en baby2ID
print("\n¿Hay algún valor nulo en baby2ID?")
has_nulls_baby2ID = data['baby2ID'].isnull().any()
```

```
print(has_nulls_baby2ID)

```

```
    Tipo de dato de baby2ID:
    int64

    ¿Todos los valores en baby2ID son enteros o ceros?
    True

    ¿Hay algún valor nulo en baby2ID?
    False
```

```python
# Verificar si hay algún valor nulo en bornDate
has_nulls_bornDate = data['bornDate'].isnull().any()
print(f"¿Hay algún valor nulo en bornDate? {has_nulls_bornDate}")

# Contar el número de valores nulos en bornDate
count_nulls_bornDate = data['bornDate'].isnull().sum()
print(f"Número de valores nulos en bornDate: {count_nulls_bornDate}")

# Reemplazar los valores nulos con 0
data['bornDate'] = data['bornDate'].fillna(0)

# Verificar si hay algún valor nulo en bornDate después del reemplazo
has_nulls_bornDate_after = data['bornDate'].isnull().any()
print(f"¿Hay algún valor nulo en bornDate después del reemplazo?
{has_nulls_bornDate_after}")

# Contar el número de valores nulos en bornDate después del reemplazo
count_nulls_bornDate_after = data['bornDate'].isnull().sum()
```

```
print(f"Número de valores nulos en bornDate después del reemplazo:
{count_nulls_bornDate_after}")

```

```
    ¿Hay algún valor nulo en bornDate? True
    Número de valores nulos en bornDate: 4168402
    ¿Hay algún valor nulo en bornDate después del reemplazo? False
    Número de valores nulos en bornDate después del reemplazo: 0
```

```python
# Verificar el tipo de datos de la columna numSons
print("Tipo de dato de numSons antes de la conversión:")
print(data['numSons'].dtype)

# Reemplazar valores nulos con 0
data['numSons'] = data['numSons'].fillna(0)

# Convertir la columna a tipo entero
data['numSons'] = data['numSons'].astype(int)

# Verificar si todos los valores en numSons son enteros o ceros
print("\n¿Todos los valores en numSons son enteros o ceros?")
are_all_int_or_zero = (data['numSons'].apply(lambda x: isinstance(x, int) and x
>= 0)).all()
print(are_all_int_or_zero)

# Verificar si hay algún valor nulo en numSons después del reemplazo
has_nulls_numSons_after = data['numSons'].isnull().any()
print(f"\n¿Hay algún valor nulo en numSons después del reemplazo?
{has_nulls_numSons_after}")

```

```
    Tipo de dato de numSons antes de la conversión:
    float64

    ¿Todos los valores en numSons son enteros o ceros?
    True

    ¿Hay algún valor nulo en numSons después del reemplazo? False



```python
# Verificar si hay algún valor nulo en dueDate
has_nulls_dueDate = data['dueDate'].isnull().any()
print(f"¿Hay algún valor nulo en dueDate? {has_nulls_dueDate}")

# Contar el número de valores nulos en dueDate
count_nulls_dueDate = data['dueDate'].isnull().sum()
print(f"Número de valores nulos en dueDate: {count_nulls_dueDate}")

# Reemplazar los valores nulos con 0
data['dueDate'] = data['dueDate'].fillna(0)

# Verificar si hay algún valor nulo en dueDate después del reemplazo
has_nulls_dueDate_after = data['dueDate'].isnull().any()
print(f"¿Hay algún valor nulo en dueDate después del reemplazo?
{has_nulls_dueDate_after}")

# Contar el número de valores nulos en dueDate después del reemplazo
count_nulls_dueDate_after = data['dueDate'].isnull().sum()
print(f"Número de valores nulos en dueDate después del reemplazo:
{count_nulls_dueDate_after}")
```

    ¿Hay algún valor nulo en dueDate? True
```

```
    Número de valores nulos en dueDate: 5716264
    ¿Hay algún valor nulo en dueDate después del reemplazo? False
    Número de valores nulos en dueDate después del reemplazo: 0
```

```python
# Verificar los valores únicos en la columna genre
unique_genres = data['genre'].unique()
print("Valores únicos en genre antes de la limpieza:")
print(unique_genres)
```

```
    Valores únicos en genre antes de la limpieza:
    ['F' 'M']
```

```python
# Verificar si hay algún valor nulo en bornDate.1
has_nulls_bornDate = data['bornDate.1'].isnull().any()
print(f"¿Hay algún valor nulo en bornDate.1? {has_nulls_bornDate}")
```

```
    ¿Hay algún valor nulo en bornDate.1? False
```

```python
# Verificar las filas donde premature es 0 pero gestationWeek tiene un valor distinto de 0
invalid_rows = data[(data['premature'] == 0) & (data['gestationWeek'] != 0)]
print(f"Número de filas con 'premature' 0 y 'gestationWeek' distinto de 0: {len(invalid_rows)}")
```

```
# Cambiar premature a 1 en estas filas
data.loc[(data['premature'] == 0) & (data['gestationWeek'] != 0), 'premature'] =
1

# Verificar las filas donde premature es 0 pero gestationWeek tiene un valor
distinto de 0 después del cambio
invalid_rows_after = data[(data['premature'] == 0) & (data['gestationWeek'] !=
0)]
print(f"Número de filas con 'premature' 0 y 'gestationWeek' distinto de 0
después del cambio: {len(invalid_rows_after)}")

# Reemplazar valores nulos en gestationWeek con 0 cuando premature es 1
data.loc[(data['premature'] == 1) & (data['gestationWeek'].isnull()),
'gestationWeek'] = 0

# Asegurarnos de que si premature es 0, gestationWeek sea 0
data.loc[data['premature'] == 0, 'gestationWeek'] = 0

# Mostrar los valores únicos de la columna gestationWeek
unique_gestation_weeks = data['gestationWeek'].unique()
print("Valores únicos en gestationWeek:")
print(unique_gestation_weeks)

# Verificar si hay algún valor nulo en gestationWeek
has_nulls_gestationWeek = data['gestationWeek'].isnull().any()
print(f"¿Hay algún valor nulo en gestationWeek? {has_nulls_gestationWeek}")

# Contar el número de valores nulos en gestationWeek
count_nulls_gestationWeek = data['gestationWeek'].isnull().sum()
print(f"Número de valores nulos en gestationWeek: {count_nulls_gestationWeek}")
```

```
    Número de filas con 'premature' 0 y 'gestationWeek' distinto de 0: 2747962
    Número de filas con 'premature' 0 y 'gestationWeek' distinto de 0 después
del cambio: 0
```

```
    Valores únicos en gestationWeek:
    [ 0. 20. 34. 31. 36. 35. 24. 32. 29. 28. 30. 23. 33. 26. 27. 21. 25. 22.]
    ¿Hay algún valor nulo en gestationWeek? False
    Número de valores nulos en gestationWeek: 0
```

```python
# Verificar el tipo de datos de la columna bornHeight
print("Tipo de dato de bornHeight antes de la conversión:")
print(data['bornHeight'].dtype)

# Reemplazar valores nulos con 0
data['bornHeight'] = data['bornHeight'].fillna(0)

# Eliminar espacios en blanco y convertir a NaN
data['bornHeight'] = pd.to_numeric(data['bornHeight'], errors='coerce')

# Reemplazar valores nulos con 0 nuevamente después de la conversión
data['bornHeight'] = data['bornHeight'].fillna(0)

# Asegurar que la columna es de tipo float
data['bornHeight'] = data['bornHeight'].astype(float)

# Verificar si todos los valores en bornHeight son decimales o ceros
print("\n¿Todos los valores en bornHeight son decimales o ceros?")
are_all_float_or_zero = (data['bornHeight'].apply(lambda x: isinstance(x, (int, float)) and x >= 0)).all()
print(are_all_float_or_zero)

# Verificar si hay algún valor nulo en bornHeight después del reemplazo
has_nulls_bornHeight_after = data['bornHeight'].isnull().any()
print(f"\n¿Hay algún valor nulo en bornHeight después del reemplazo? {has_nulls_bornHeight_after}")
```

```
```

    Tipo de dato de bornHeight antes de la conversión:
    object

    ¿Todos los valores en bornHeight son decimales o ceros?
    True

    ¿Hay algún valor nulo en bornHeight después del reemplazo? False


```python
# Verificar los tipos de datos antes de la conversión
print("Tipos de datos antes de la conversión:")
print(data[['startDate', 'endDate']].dtypes)
# Verificar si hay valores nulos en las columnas startDate y endDate
has_nulls_startDate = data['startDate'].isnull().any()
has_nulls_endDate = data['endDate'].isnull().any()
print(f"¿Hay algún valor nulo en startDate? {has_nulls_startDate}")
print(f"¿Hay algún valor nulo en endDate? {has_nulls_endDate}")

# Contar el número de valores nulos en las columnas startDate y endDate
count_nulls_startDate = data['startDate'].isnull().sum()
count_nulls_endDate = data['endDate'].isnull().sum()
print(f"Número de valores nulos en startDate: {count_nulls_startDate}")
print(f"Número de valores nulos en endDate: {count_nulls_endDate}")
```

    Tipos de datos antes de la conversión:
    startDate   object
    endDate   object
    dtype: object
    ¿Hay algún valor nulo en startDate? False
    ¿Hay algún valor nulo en endDate? False

```
    Número de valores nulos en startDate: 0
    Número de valores nulos en endDate: 0
```

```python
# Verificar si hay valores nulos en todas las columnas
null_values = data.isnull().sum()
print("Número de valores nulos en cada columna:")
print(null_values)

# Verificar si hay alguna columna con valores nulos
columns_with_nulls = null_values[null_values > 0]
print("\nColumnas con valores nulos:")
print(columns_with_nulls)
```

```
    Número de valores nulos en cada columna:
    trackID                      0
    userID                       0
    baby1ID                      0
    baby2ID                      0
    bornDate                     0
    numSons                      0
    dueDate                      0
    genre                        0
    bornDate.1                   0
    premature                    0
    gestationWeek                0
    bornHeight                   0
    babyTandemGenre        6776462
    babyTandemBornDate     6776462
    babyTandemPremature    6776462
    babyTandemBornHeight   6777342
    startDate                    0
```

```
    endDate                       0
    breast                        0
    suplBreastmilk                0
    suplFormula                   0
    painLevel                     0
    nipleShape                    0
    nipleColor                    0
    milkPearl                     0
    cracks                        0
    emotionalStatus               0
    creationDate                  0
    modificationDate              0
    dtype: int64

    Columnas con valores nulos:
    babyTandemGenre         6776462
    babyTandemBornDate      6776462
    babyTandemPremature     6776462
    babyTandemBornHeight    6777342
    dtype: int64
```

```python
# Verificar el tipo de datos de la columna babyTandemBornHeight
print("Tipo de dato de babyTandemBornHeight antes de la conversión:")
print(data['babyTandemBornHeight'].dtype)

# Reemplazar valores nulos con 0
data['babyTandemBornHeight'] = data['babyTandemBornHeight'].fillna(0)

# Eliminar espacios en blanco y convertir a NaN
data['babyTandemBornHeight'] = pd.to_numeric(data['babyTandemBornHeight'],
errors='coerce')
```

```
# Reemplazar valores nulos con 0 nuevamente después de la conversión
data['babyTandemBornHeight'] = data['babyTandemBornHeight'].fillna(0)

# Asegurar que la columna es de tipo float
data['babyTandemBornHeight'] = data['babyTandemBornHeight'].astype(float)

# Verificar si todos los valores en babyTandemBornHeight son decimales o ceros
print("\n¿Todos los valores en babyTandemBornHeight son decimales o ceros?")
are_all_float_or_zero_tandem_height = (data['babyTandemBornHeight'].apply(lambda
x: isinstance(x, (int, float)) and x >= 0)).all()
print(are_all_float_or_zero)

# Verificar si hay algún valor nulo en bornHeight después del reemplazo
has_nulls_babyTandemBornHeight_after =
data['babyTandemBornHeight'].isnull().any()
print(f"\n¿Hay algún valor nulo en babyTandemBornHeight después del reemplazo?
{has_nulls_babyTandemBornHeight_after}")
```

```
    Tipo de dato de babyTandemBornHeight antes de la conversión:
    object

    ¿Todos los valores en babyTandemBornHeight son decimales o ceros?
    True

    ¿Hay algún valor nulo en babyTandemBornHeight después del reemplazo? False
```

```python
# Definir las columnas de fecha con sus respectivos formatos
date_columns_format1 = ['bornDate', 'dueDate', 'bornDate.1',
'babyTandemBornDate']
date_columns_format2 = ['startDate', 'endDate', 'creationDate',
'modificationDate']
```

```
# Especificar los formatos esperados
date_format1 = '%Y-%m-%d'
date_format2 = '%Y-%m-%d %H:%M:%S'

# Convertir las columnas del primer grupo a datetime y mostrar el contenido
for col in date_columns_format1:
    if col in data.columns:
        data[col] = pd.to_datetime(data[col], format=date_format1,
errors='coerce')
        print(f"\nContenido de la columna {col} después de la conversión a
datetime:")
        print(data[col].head(10))  # Mostrar las primeras 10 filas para
inspección

# Convertir las columnas del segundo grupo a datetime y mostrar el contenido
for col in date_columns_format2:
    if col in data.columns:
        data[col] = pd.to_datetime(data[col], format=date_format2,
errors='coerce')
        print(f"\nContenido de la columna {col} después de la conversión a
datetime:")
        print(data[col].head(10))  # Mostrar las primeras 10 filas para
inspección

# Verificar los tipos de datos después de la conversión
print("\nTipos de datos después de la conversión a datetime:")
print(data.dtypes.loc[date_columns_format1 + date_columns_format2])
```

```
    Contenido de la columna bornDate después de la conversión a datetime:
    0    1987-03-21
    1    1987-03-21
    3    1987-03-21
```

```
4    1987-03-21
5    1987-03-21
6    1970-01-01
7    1985-05-06
10   1989-08-14
11   1978-09-09
12   1978-09-09
Name: bornDate, dtype: datetime64[ns]

Contenido de la columna dueDate después de la conversión a datetime:
0           NaT
1           NaT
3           NaT
4           NaT
5           NaT
6           NaT
7    1970-01-01
10   1970-01-01
11          NaT
12          NaT
Name: dueDate, dtype: datetime64[ns]

Contenido de la columna bornDate.1 después de la conversión a datetime:
0    2020-10-06
1    2020-10-06
3    2020-10-06
4    2020-10-06
5    2020-10-06
6    2018-08-04
7    2017-10-29
10   2018-03-18
11   2017-07-07
12   2017-07-07
Name: bornDate.1, dtype: datetime64[ns]
```

```
    Contenido de la columna babyTandemBornDate después de la conversión a
datetime:
    0    NaT
    1    NaT
    3    NaT
    4    NaT
    5    NaT
    6    NaT
    7    NaT
    10   NaT
    11   NaT
    12   NaT
    Name: babyTandemBornDate, dtype: datetime64[ns]

    Contenido de la columna startDate después de la conversión a datetime:
    0    2018-02-15 14:26:00
    1    2018-02-15 13:33:00
    3    2018-03-19 13:44:00
    4    2018-03-21 13:34:00
    5    2018-03-22 15:44:00
    6    2018-03-27 21:42:00
    7    2018-03-28 00:00:00
    10   2018-03-28 00:15:00
    11   2018-03-28 17:00:00
    12   2018-03-28 17:00:00
    Name: startDate, dtype: datetime64[ns]

    Contenido de la columna endDate después de la conversión a datetime:
    0    2018-02-15 15:26:00
    1    2018-02-15 13:38:00
    3    2018-03-19 13:44:00
    4    2018-03-21 14:34:00
    5    2018-03-22 16:44:00
    6    2018-03-27 21:42:00
    7    2018-03-28 00:27:00
```

```
    10   2018-03-28 00:15:00
    11   2018-03-28 08:30:00
    12   2018-03-28 08:30:00
    Name: endDate, dtype: datetime64[ns]

    Contenido de la columna creationDate después de la conversión a datetime:
    0    2018-02-15 12:27:33
    1    2018-02-15 12:34:11
    3    2018-03-19 11:45:09
    4    2018-03-20 16:40:25
    5    2018-03-22 13:46:05
    6    2018-03-27 18:42:45
    7    2018-03-27 21:18:30
    10   2018-03-27 22:22:08
    11   2018-03-28 09:24:39
    12   2018-03-28 09:26:03
    Name: creationDate, dtype: datetime64[ns]

    Contenido de la columna modificationDate después de la conversión a
datetime:
    0    2018-02-15 12:27:33
    1    2018-02-15 12:34:11
    3    2018-03-22 08:30:37
    4    2018-03-21 11:25:47
    5    2018-03-22 13:46:05
    6    2018-03-27 18:42:45
    7    2018-03-27 21:18:30
    10   2018-03-27 22:22:08
    11   2018-03-28 09:24:39
    12   2018-03-28 09:26:03
    Name: modificationDate, dtype: datetime64[ns]

    Tipos de datos después de la conversión a datetime:
    bornDate              datetime64[ns]
    dueDate               datetime64[ns]
```

```
    bornDate.1            datetime64[ns]
    babyTandemBornDate    datetime64[ns]
    startDate             datetime64[ns]
    endDate               datetime64[ns]
    creationDate          datetime64[ns]
    modificationDate      datetime64[ns]
    dtype: object
```

```python
data.info()
```

```
    <class 'pandas.core.frame.DataFrame'>
    Index: 6786861 entries, 0 to 6872734
    Data columns (total 29 columns):
     #   Column               Dtype
    ---  ------               -----
     0   trackID              int64
     1   userID               int64
     2   baby1ID              int64
     3   baby2ID              int64
     4   bornDate             datetime64[ns]
     5   numSons              int64
     6   dueDate              datetime64[ns]
     7   genre                object
     8   bornDate.1           datetime64[ns]
     9   premature            int64
     10  gestationWeek        float64
     11  bornHeight           float64
     12  babyTandemGenre      object
     13  babyTandemBornDate   datetime64[ns]
     14  babyTandemPremature  float64
     15  babyTandemBornHeight float64
```

```
 16  startDate             datetime64[ns]
 17  endDate               datetime64[ns]
 18  breast                object
 19  suplBreastmilk        object
 20  suplFormula           object
 21  painLevel             object
 22  nipleShape            object
 23  nipleColor            object
 24  milkPearl             object
 25  cracks                object
 26  emotionalStatus       object
 27  creationDate          datetime64[ns]
 28  modificationDate      datetime64[ns]
dtypes: datetime64[ns](8), float64(4), int64(6), object(11)
memory usage: 1.5+ GB
```

```python
##########################################################################################
########################################
```

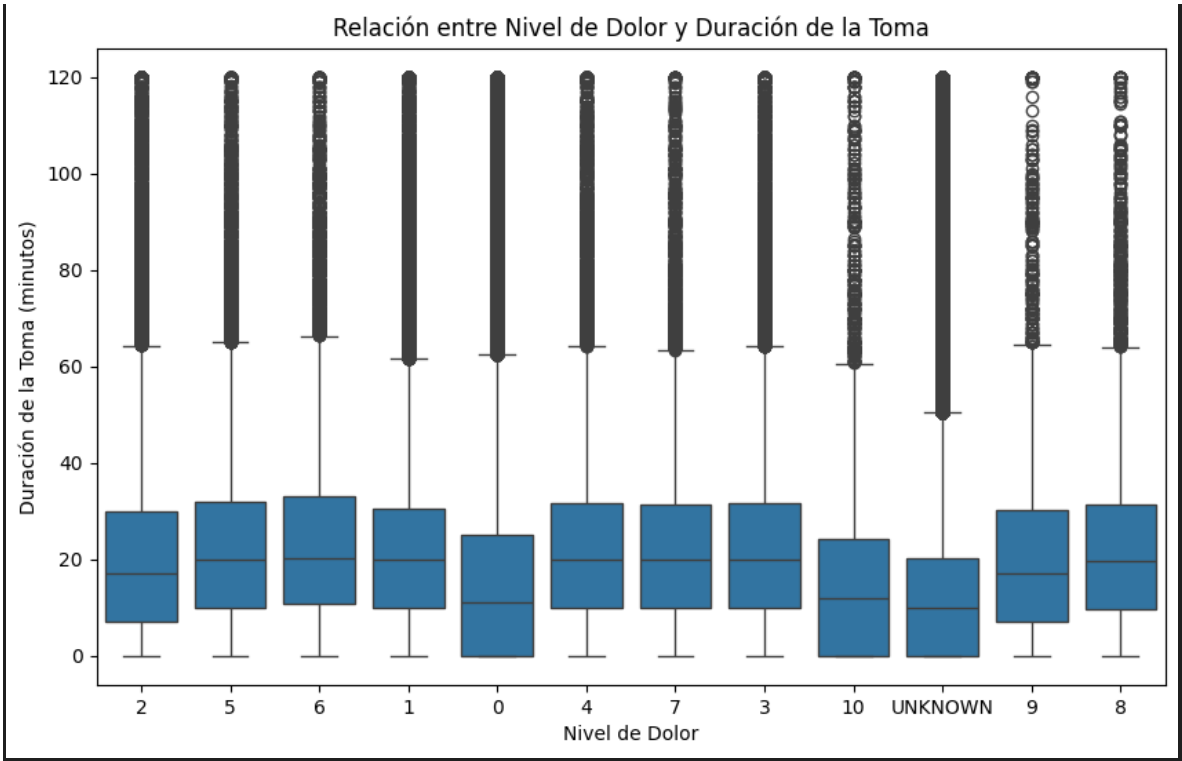
```python
# Revisar el cálculo de la duración de las tomas
data['duration'] = (data['endDate'] - data['startDate']).dt.total_seconds() /
60  # Duración en minutos
# Crear una variable específica para la muestra y filtrar valores de duración
anómalos (e.g., duración mayor a 2 horas)
data_sample_painLevel_duration = data[(data['duration'] >= 0) &
(data['duration'] <= 120)]  # Filtrar duración mayor a 2 horas
print(data[['startDate', 'endDate', 'duration']].head(20))
```

```
             startDate             endDate  duration
0  2018-02-15 14:26:00 2018-02-15 15:26:00      60.0
```

```
    1  2018-02-15 13:33:00 2018-02-15 13:38:00       5.0
    3  2018-03-19 13:44:00 2018-03-19 13:44:00       0.0
    4  2018-03-21 13:34:00 2018-03-21 14:34:00      60.0
    5  2018-03-22 15:44:00 2018-03-22 16:44:00      60.0
    6  2018-03-27 21:42:00 2018-03-27 21:42:00       0.0
    7  2018-03-28 00:00:00 2018-03-28 00:27:00      27.0
    10 2018-03-28 00:15:00 2018-03-28 00:15:00       0.0
    11 2018-03-28 17:00:00 2018-03-28 08:30:00    -510.0
    12 2018-03-28 17:00:00 2018-03-28 08:30:00    -510.0
    13 2018-03-28 12:30:00 2018-03-28 12:54:00      24.0
    14 2018-03-28 13:52:00 2018-03-28 13:52:00       0.0
    15 2018-03-28 15:54:00 2018-03-28 15:54:00       0.0
    16 2018-03-28 15:54:00 2018-03-28 15:54:00       0.0
    17 2018-03-28 19:55:00 2018-03-28 19:55:00       0.0
    18 2018-03-28 16:56:00 2018-03-28 16:56:00       0.0
    19 2018-03-28 15:00:00 2018-03-28 15:16:00      16.0
    20 2018-03-28 16:00:00 2018-03-28 16:27:00      27.0
    21 2018-03-28 18:15:00 2018-03-28 18:15:00       0.0
    22 2018-03-28 20:10:00 2018-03-28 20:20:00      10.0
```

```python
# Visualizar el gráfico con la muestra filtrada
plt.figure(figsize=(10, 6))
sns.boxplot(x='painLevel', y='duration', data=data_sample_painLevel_duration)
plt.title('Relación entre Nivel de Dolor y Duración de la Toma')
plt.xlabel('Nivel de Dolor')
plt.ylabel('Duración de la Toma (minutos)')
plt.show()

```




```python
# Filtrar los valores 'UNKNOWN' en painLevel y convertir a numérico
data_sample_painLevel_duration_filtered =
data_sample_painLevel_duration[data_sample_painLevel_duration['painLevel'] !=
'UNKNOWN'].copy()
data_sample_painLevel_duration_filtered['painLevel'] =
pd.to_numeric(data_sample_painLevel_duration_filtered['painLevel'],
errors='coerce')

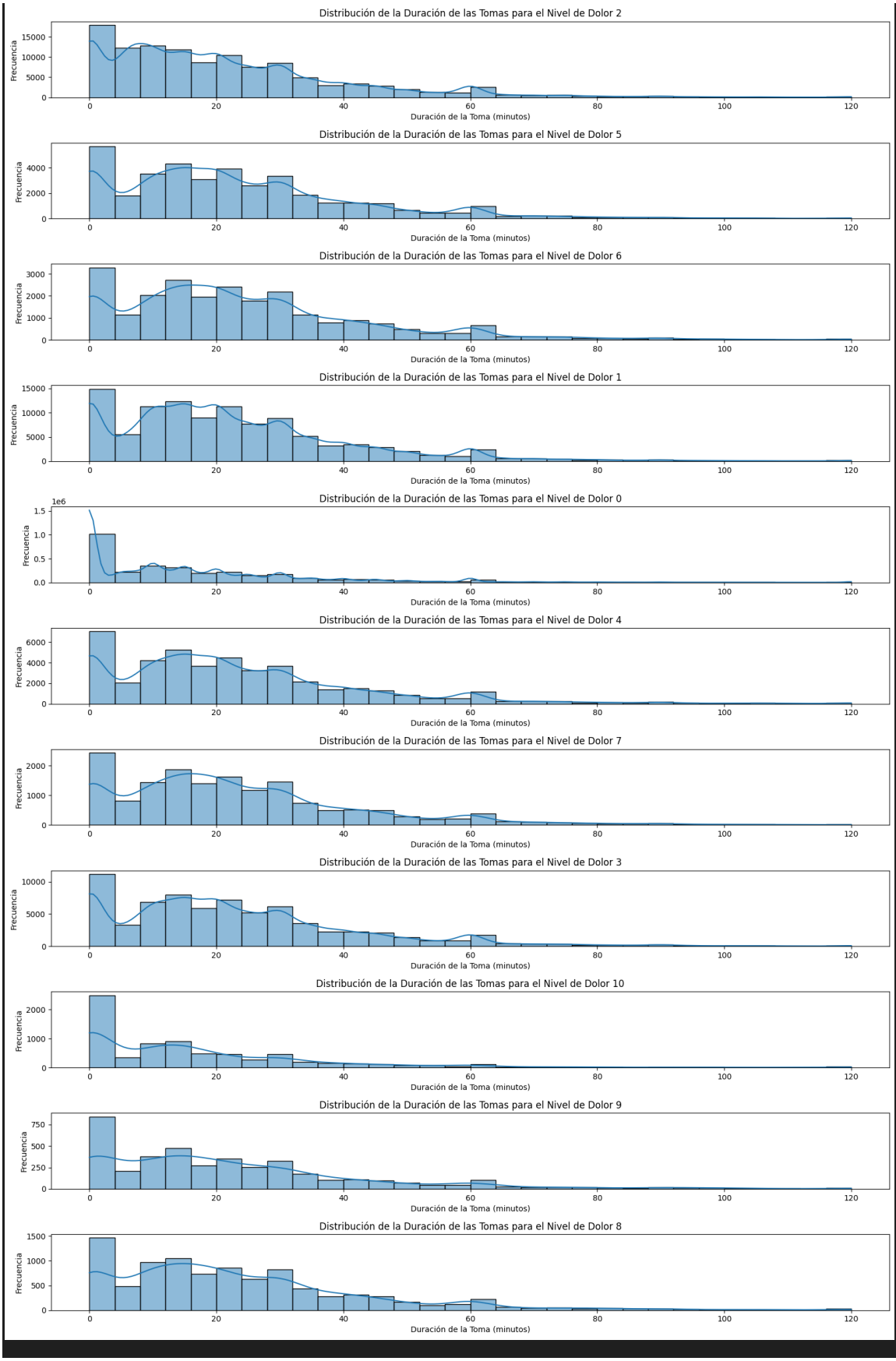
# Eliminar filas con valores nulos en painLevel
data_sample_painLevel_duration_filtered =
data_sample_painLevel_duration_filtered.dropna(subset=['painLevel'])

# Obtener los niveles únicos de dolor
pain_levels = data_sample_painLevel_duration_filtered['painLevel'].unique()

# Crear un subplot para cada nivel de dolor
```

```
plt.figure(figsize=(16, 24))
for i, level in enumerate(pain_levels):
    plt.subplot(len(pain_levels), 1, i + 1)
    sns.histplot(data=data_sample_painLevel_duration_filtered[data_sample_painLe
vel_duration_filtered['painLevel'] == level]['duration'], kde=True, bins=30)
    plt.title(f'Distribución de la Duración de las Tomas para el Nivel de Dolor
{int(level)}')
    plt.xlabel('Duración de la Toma (minutos)')
    plt.ylabel('Frecuencia')

plt.tight_layout()
plt.show()
```


Distribución de la Duración de las Tomas para el Nivel de Dolor 2
Distribución de la Duración de las Tomas para el Nivel de Dolor 5
Distribución de la Duración de las Tomas para el Nivel de Dolor 6
Distribución de la Duración de las Tomas para el Nivel de Dolor 1
Distribución de la Duración de las Tomas para el Nivel de Dolor 0
Distribución de la Duración de las Tomas para el Nivel de Dolor 4
Distribución de la Duración de las Tomas para el Nivel de Dolor 7
Distribución de la Duración de las Tomas para el Nivel de Dolor 3
Distribución de la Duración de las Tomas para el Nivel de Dolor 10
Distribución de la Duración de las Tomas para el Nivel de Dolor 9
Distribución de la Duración de las Tomas para el Nivel de Dolor 8
Frecuencia
Duración de la Toma (minutos)
1e6

Para estudiar la relación con el resto de los trackers, en primer lugar se filtran los valores, eliminando los desconocidos, posteriormente, se filtra la duración para valores positivos y menores de 2h. finalmente se ordena el nivel de dolor de manera categórica.

```python
# Crear una copia de los datos originales para no modificarlos directamente
data_copy = data.copy()

# Verificar y limpiar los valores 'UNKNOWN' en las variables de interés en la copia
data_copy.loc[:, 'nipleShape'] = data_copy['nipleShape'].replace('UNKNOWN', pd.NA)
data_copy.loc[:, 'nipleColor'] = data_copy['nipleColor'].replace('UNKNOWN', pd.NA)
data_copy.loc[:, 'milkPearl'] = data_copy['milkPearl'].replace('UNKNOWN', pd.NA)
data_copy.loc[:, 'cracks'] = data_copy['cracks'].replace('UNKNOWN', pd.NA)
data_copy.loc[:, 'painLevel'] = data_copy['painLevel'].replace('UNKNOWN', pd.NA)

# Filtrar duración entre 0 y 120 minutos en la copia
data_copy = data_copy[(data_copy['duration'] >= 0) & (data_copy['duration'] <= 120)]

# Convertir painLevel a categórico y ordenar los valores de 0 a 10
pain_level_order = [str(i) for i in range(11)]
data_copy.loc[:, 'painLevel'] = pd.Categorical(data_copy['painLevel'], categories=pain_level_order, ordered=True)

# Mostrar el tipo de datos y valores únicos
print("Valores únicos en nipleShape:")
print(data_copy['nipleShape'].unique())

print("\nValores únicos en nippleColor:")
```

```
print(data_copy['nipleColor'].unique())

print("\nValores únicos en milkPearl:")
print(data_copy['milkPearl'].unique())

print("\nValores únicos en cracks:")
print(data_copy['cracks'].unique())

print("\nValores únicos en painLevel:")
print(data_copy['painLevel'].unique())
```

```
    Valores únicos en nipleShape:
    [<NA> 'feeding_niple_shape_round' 'feeding_niple_shape_lipstick']

    Valores únicos en nippleColor:
    [<NA> 'feeding_niple_color_white' 'feeding_niple_color_purple'
     'feeding_niple_color_normal']

    Valores únicos en milkPearl:
    [<NA> '1' '0']

    Valores únicos en cracks:
    ['0' '1' <NA>]

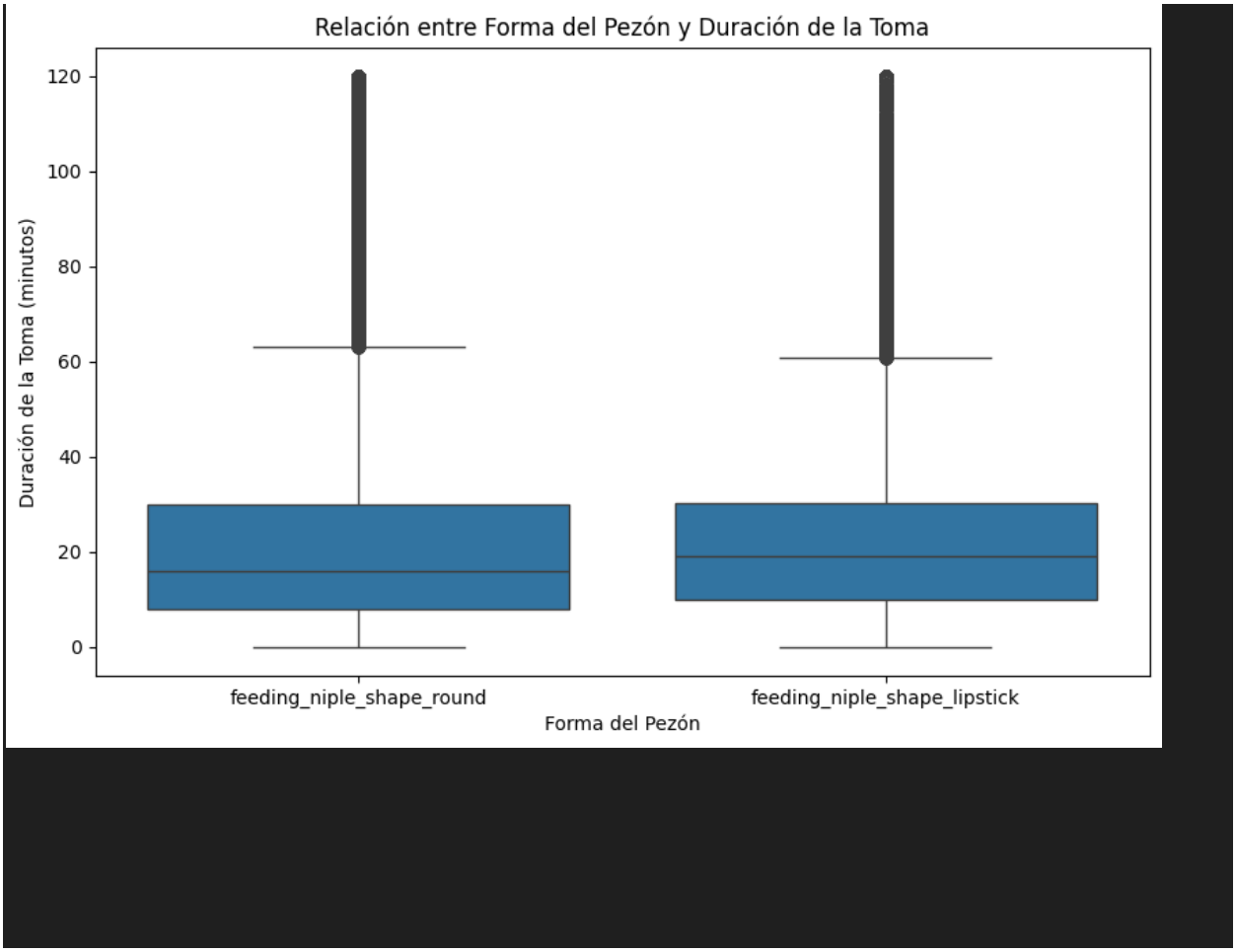
    Valores únicos en painLevel:
    ['2', '5', '6', '1', '0', ..., '3', '10', NaN, '9', '8']
    Length: 12
    Categories (11, object): ['0' < '1' < '2' < '3' ... '7' < '8' < '9' < '10']
```

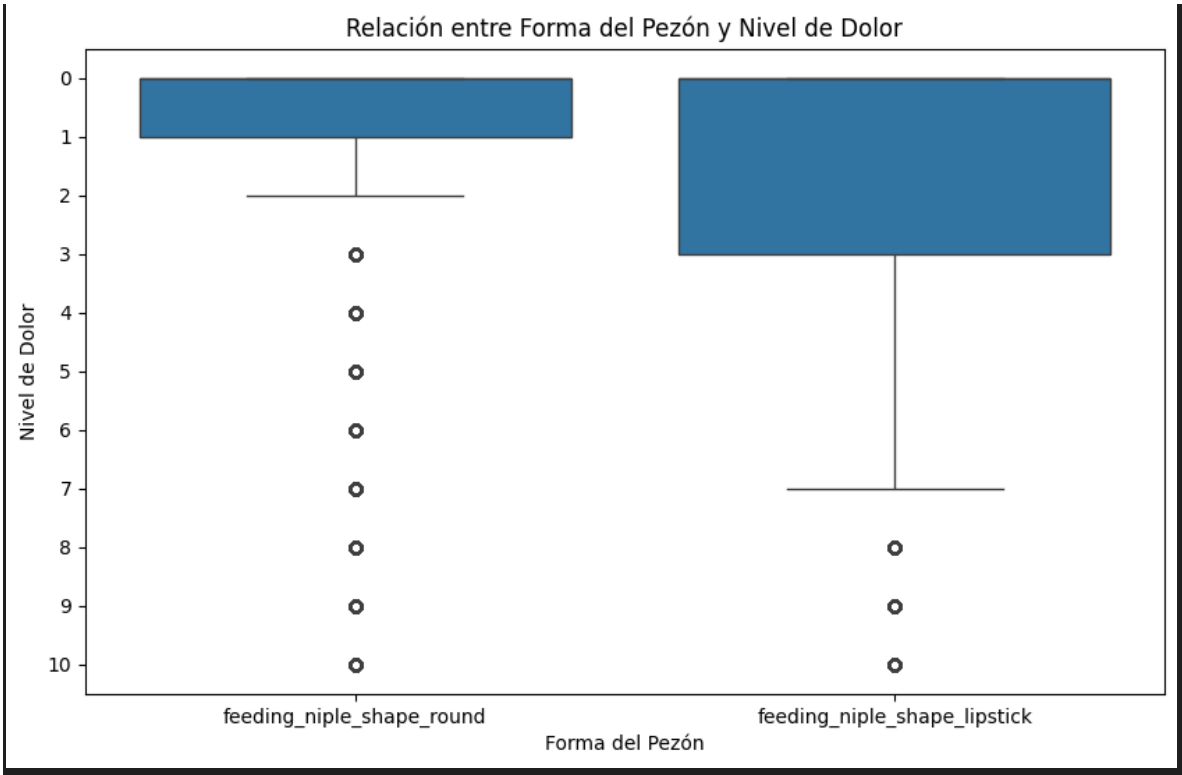
una vez filtrados, se dibujan los diagramas de cajas para estudiar si existen relaciones

```python
# Filtrar datos para nipleShape y painLevel y asignar a una variable específica
data_nipleShape_filtered = data_copy.dropna(subset=['nipleShape', 'painLevel'])

# Boxplot para nipleShape y duración
plt.figure(figsize=(10, 6))
sns.boxplot(x='nipleShape', y='duration', data=data_nipleShape_filtered)
plt.title('Relación entre Forma del Pezón y Duración de la Toma')
plt.xlabel('Forma del Pezón')
plt.ylabel('Duración de la Toma (minutos)')
plt.show()

# Boxplot para nipleShape y nivel de dolor
plt.figure(figsize=(10, 6))
sns.boxplot(x='nipleShape', y='painLevel', data=data_nipleShape_filtered)
plt.title('Relación entre Forma del Pezón y Nivel de Dolor')
plt.xlabel('Forma del Pezón')
plt.ylabel('Nivel de Dolor')
plt.show()
```


Relación entre Forma del Pezón y Duración de la Toma
Duración de la Toma (minutos)
120
100
80
60
40
20
0
feeding_niple_shape_round
feeding_niple_shape_lipstick
Forma del Pezón

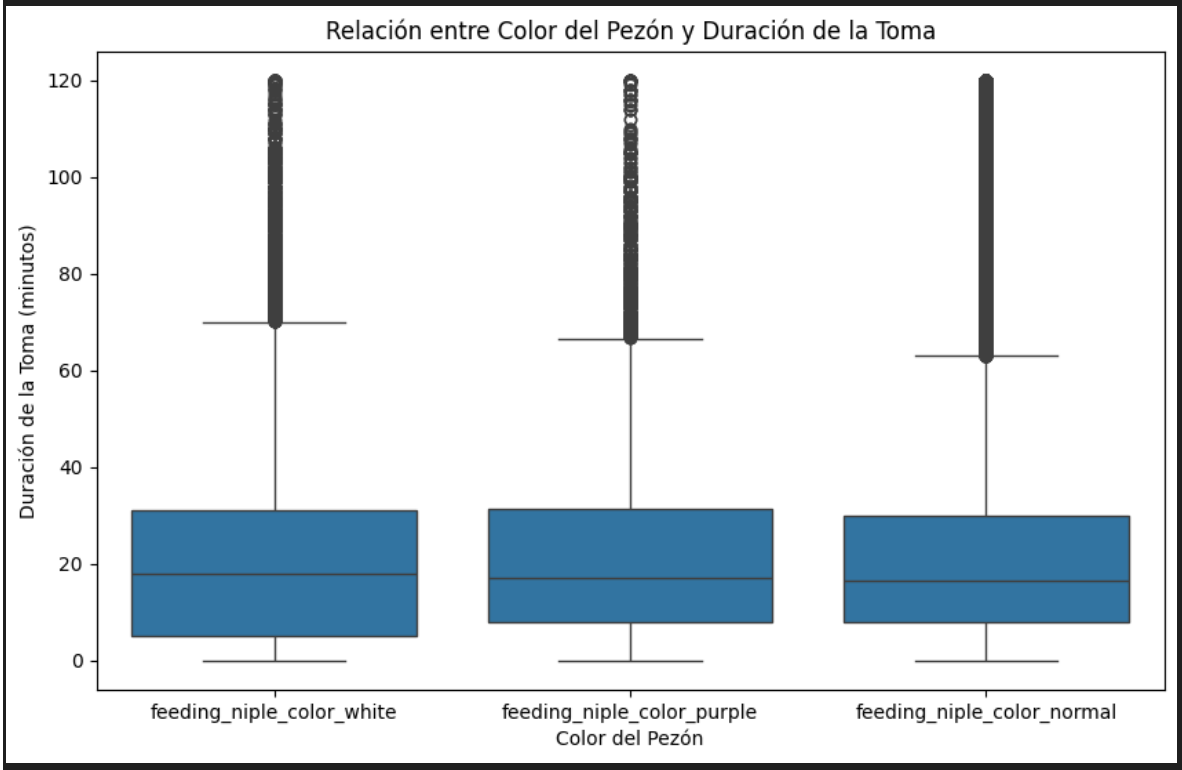


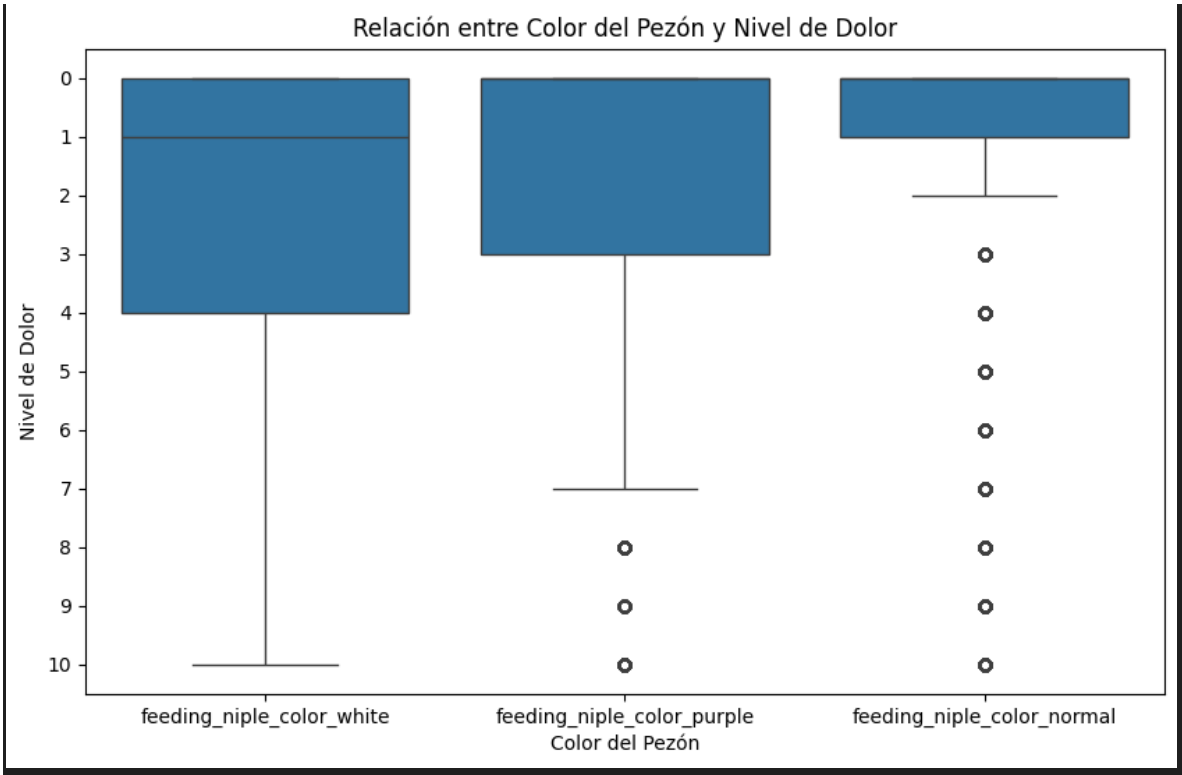

```python
# Filtrar datos para nippleColor y painLevel y asignar a una variable específica
data_nippleColor_filtered = data_copy.dropna(subset=['nipleColor', 'painLevel'])

# Boxplot para nippleColor y duración
plt.figure(figsize=(10, 6))
sns.boxplot(x='nipleColor', y='duration', data=data_nippleColor_filtered)
plt.title('Relación entre Color del Pezón y Duración de la Toma')
plt.xlabel('Color del Pezón')
plt.ylabel('Duración de la Toma (minutos)')
plt.show()

# Boxplot para nippleColor y nivel de dolor
plt.figure(figsize=(10, 6))
sns.boxplot(x='nipleColor', y='painLevel', data=data_nippleColor_filtered)
plt.title('Relación entre Color del Pezón y Nivel de Dolor')
plt.xlabel('Color del Pezón')
```

```
plt.ylabel('Nivel de Dolor')
plt.show()
```
```


```

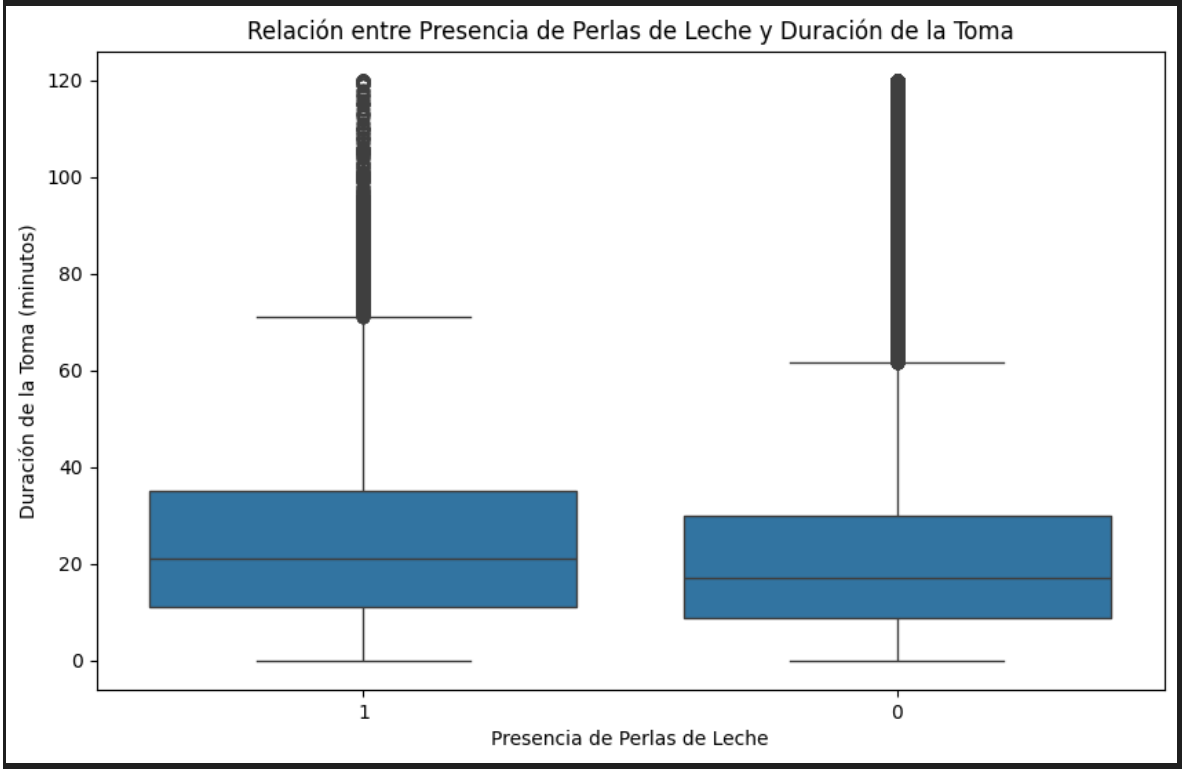


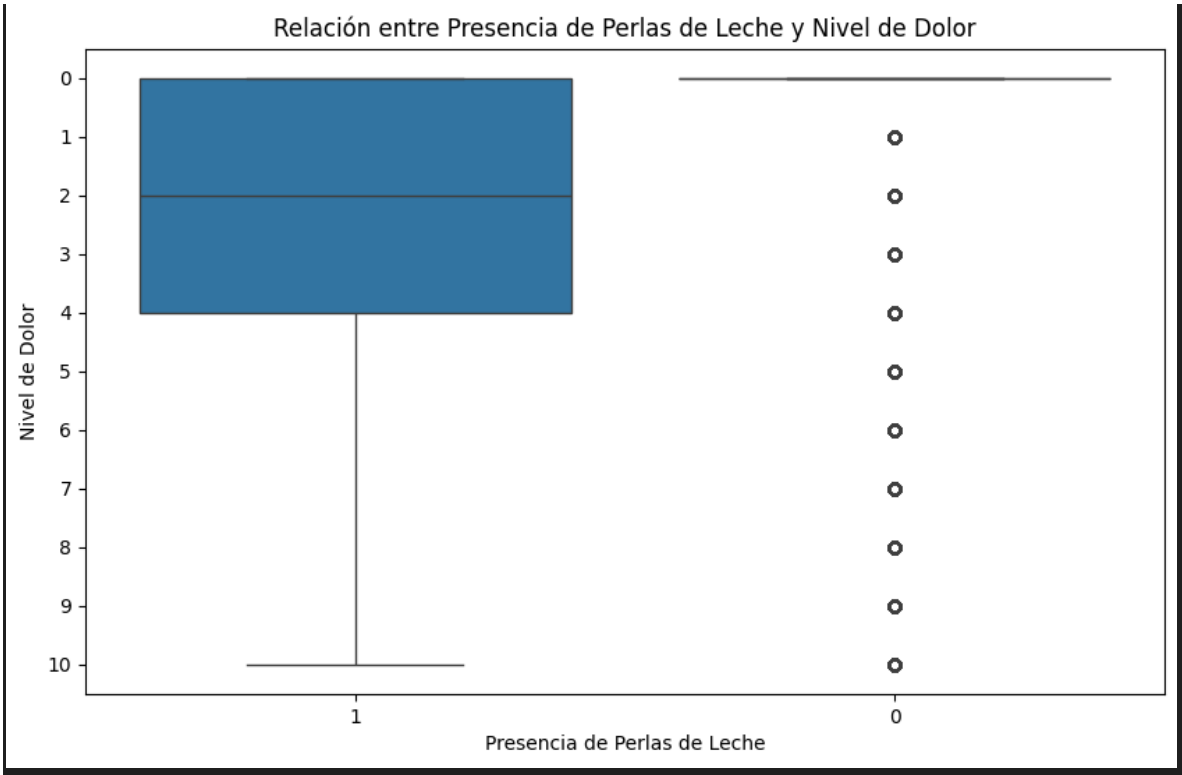

```python
# Filtrar datos para milkPearl y painLevel y asignar a una variable específica
data_milkPearl_filtered = data_copy.dropna(subset=['milkPearl', 'painLevel'])

# Boxplot para milkPearl y duración
plt.figure(figsize=(10, 6))
sns.boxplot(x='milkPearl', y='duration', data=data_milkPearl_filtered)
plt.title('Relación entre Presencia de Perlas de Leche y Duración de la Toma')
plt.xlabel('Presencia de Perlas de Leche')
plt.ylabel('Duración de la Toma (minutos)')
plt.show()

# Boxplot para milkPearl y nivel de dolor
plt.figure(figsize=(10, 6))
sns.boxplot(x='milkPearl', y='painLevel', data=data_milkPearl_filtered)
plt.title('Relación entre Presencia de Perlas de Leche y Nivel de Dolor')
plt.xlabel('Presencia de Perlas de Leche')
```

```
plt.ylabel('Nivel de Dolor')
plt.show()
```
```


```

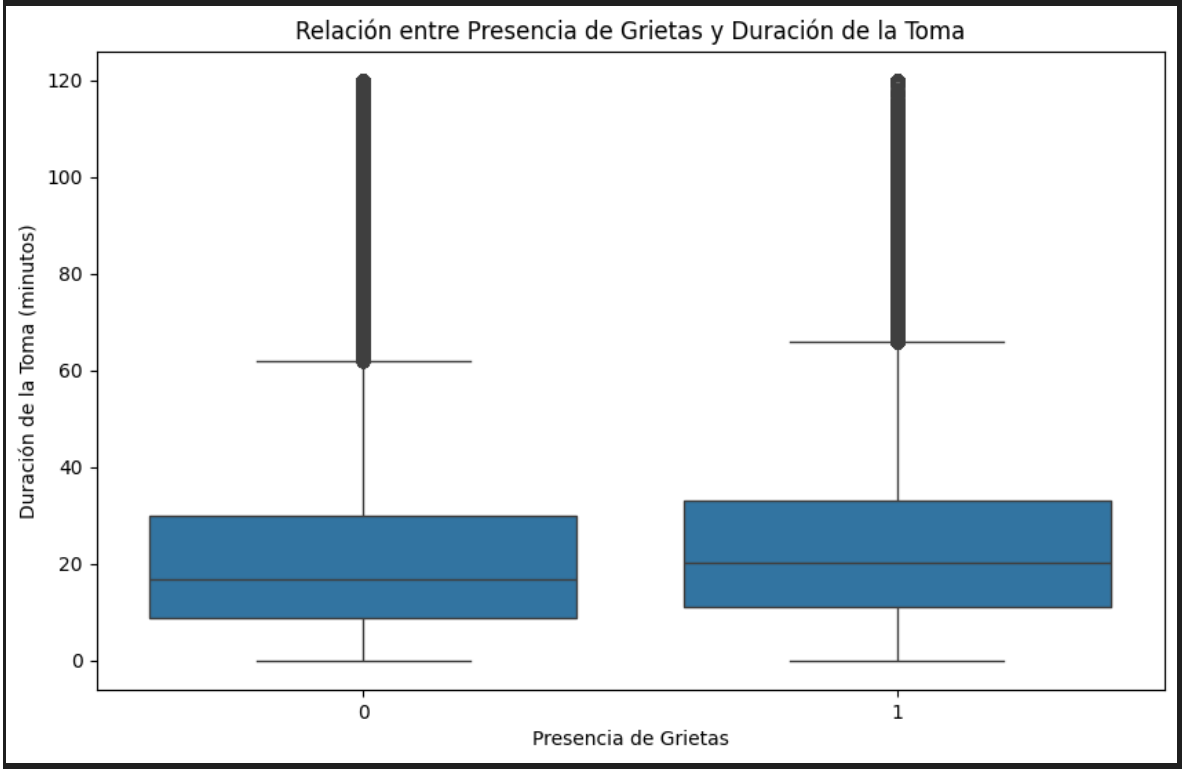


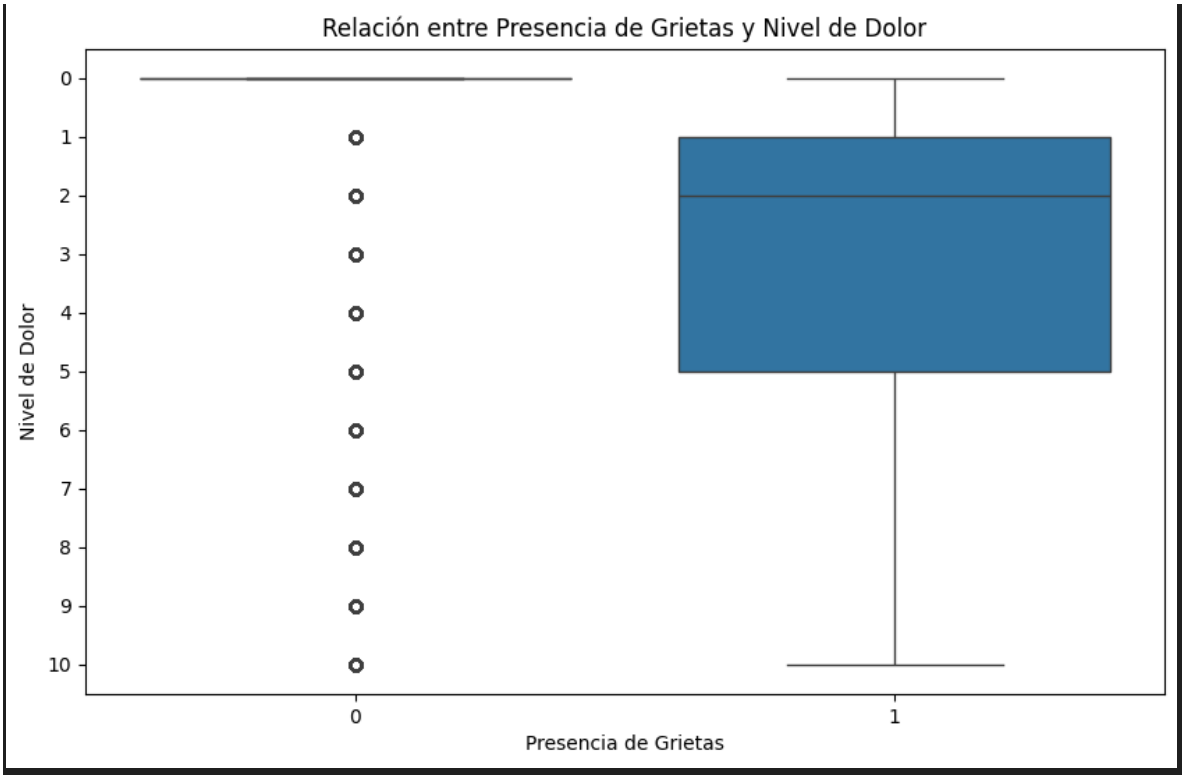

```python
# Filtrar datos para cracks y painLevel y asignar a una variable específica
data_cracks_filtered = data_copy.dropna(subset=['cracks', 'painLevel'])

# Boxplot para cracks y duración
plt.figure(figsize=(10, 6))
sns.boxplot(x='cracks', y='duration', data=data_cracks_filtered)
plt.title('Relación entre Presencia de Grietas y Duración de la Toma')
plt.xlabel('Presencia de Grietas')
plt.ylabel('Duración de la Toma (minutos)')
plt.show()

# Boxplot para cracks y nivel de dolor
plt.figure(figsize=(10, 6))
sns.boxplot(x='cracks', y='painLevel', data=data_cracks_filtered)
plt.title('Relación entre Presencia de Grietas y Nivel de Dolor')
plt.xlabel('Presencia de Grietas')
```

```
plt.ylabel('Nivel de Dolor')
plt.show()
```
```


```




```

```python
# Crear una copia de los datos originales para no modificarlos directamente
data_genre = data.copy()
```

```python
# Verificar y limpiar los valores 'UNKNOWN' en las variables de interés en la copia
data_genre.loc[:, 'nipleShape'] = data_genre['nipleShape'].replace('UNKNOWN', pd.NA)
data_genre.loc[:, 'nipleColor'] = data_genre['nipleColor'].replace('UNKNOWN', pd.NA)
data_genre.loc[:, 'milkPearl'] = data_genre['milkPearl'].replace('UNKNOWN', pd.NA)
data_genre.loc[:, 'cracks'] = data_genre['cracks'].replace('UNKNOWN', pd.NA)
data_genre.loc[:, 'painLevel'] = data_genre['painLevel'].replace('UNKNOWN', pd.NA)

# Convertir painLevel a numérico (float) para asegurar ordenación correcta
data_genre.loc[:, 'painLevel'] = pd.to_numeric(data_genre['painLevel'], errors='coerce')
```
```

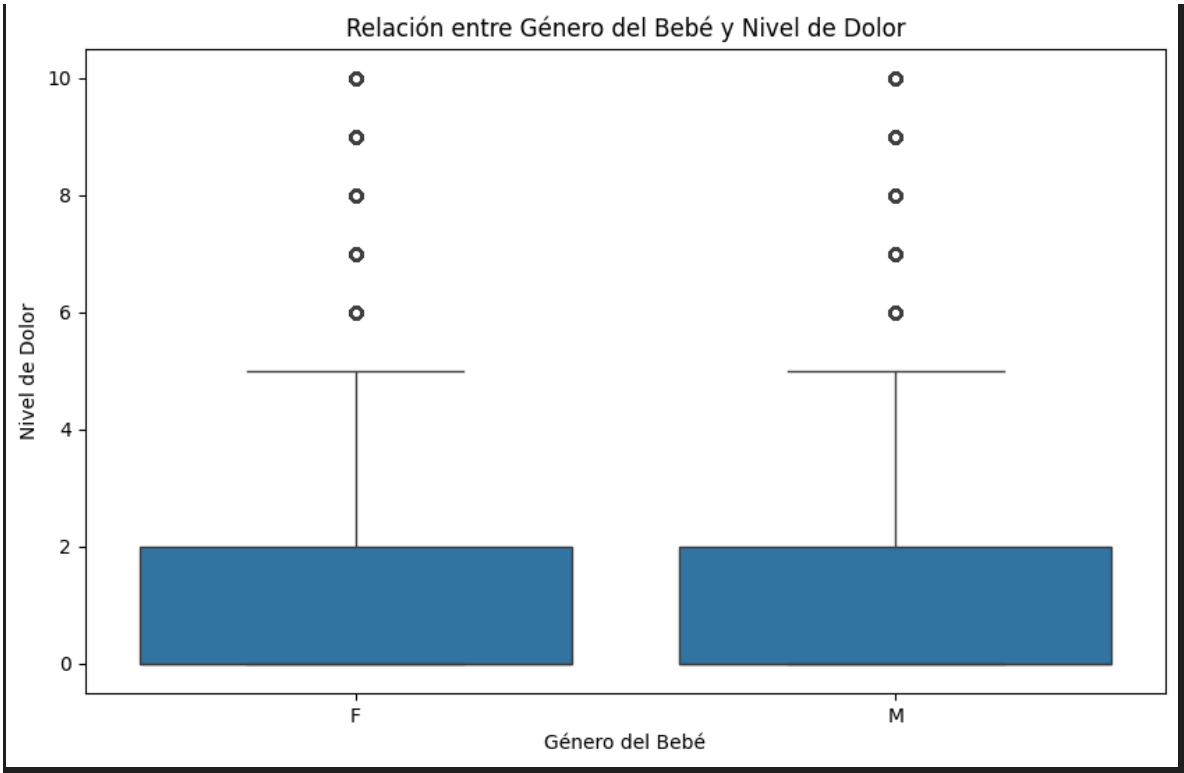
```
# Filtrar datos relevantes
data_genre_filtered = data_genre.dropna(subset=['genre', 'nipleShape',
'nipleColor', 'milkPearl', 'cracks', 'painLevel'])

# Mostrar la cantidad de datos por cada combinación de género y nivel de dolor
print("\nCantidad de datos por cada combinación de género y nivel de dolor:")
print(data_genre_filtered.groupby(['genre', 'painLevel']).size())
```

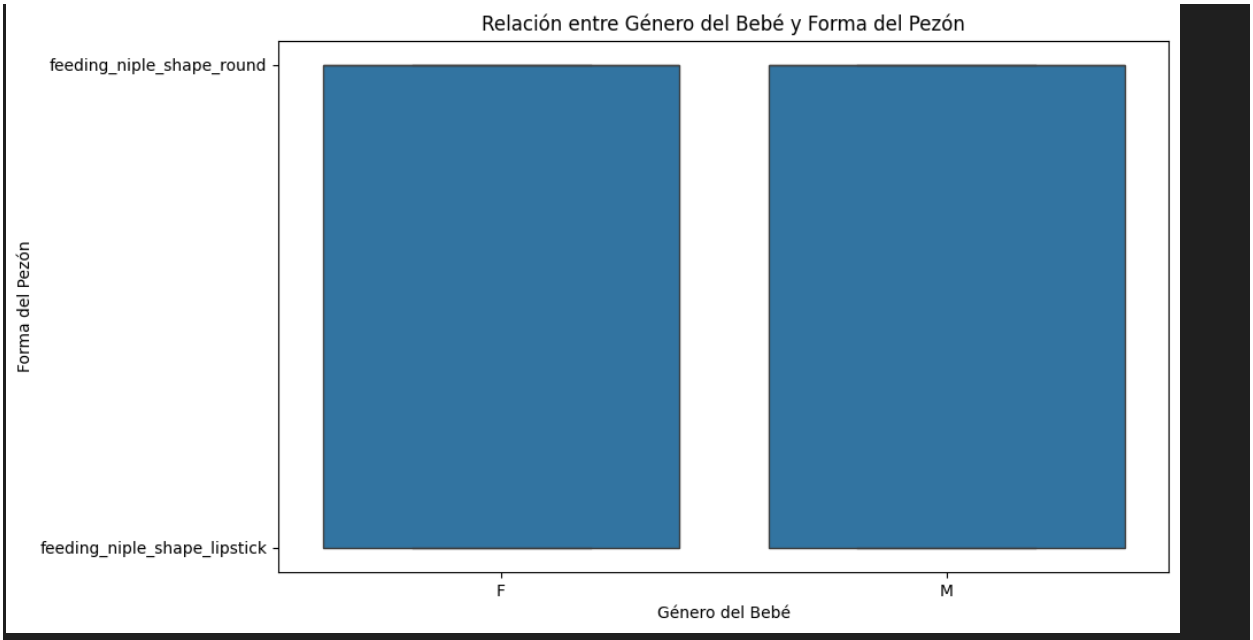
```
    Cantidad de datos por cada combinación de género y nivel de dolor:
    genre  painLevel
    F      0.0          167145
           1.0           25063
           2.0           28435
           3.0           14245
           4.0            8121
           5.0            6325
           6.0            4422
           7.0            2925
           8.0            1652
           9.0             545
           10.0           1128
    M      0.0          166170
           1.0           27303
           2.0           27730
           3.0           16370
           4.0            9728
           5.0            9063
           6.0            4952
           7.0            3183
           8.0            1747
           9.0             672
           10.0            605
    dtype: int64
```

Se puede concluir que las muestras son significativas y relevantes

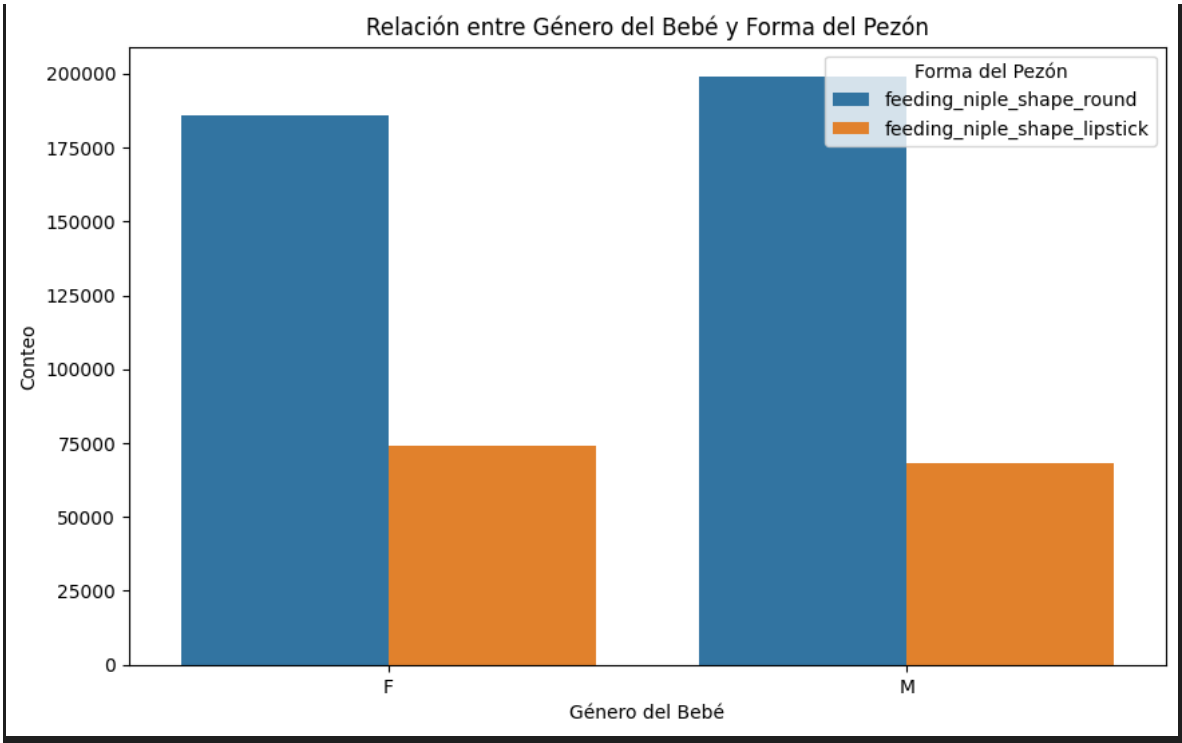
```python
# Boxplot para género y nivel de dolor
plt.figure(figsize=(10, 6))
sns.boxplot(x='genre', y='painLevel', data=data_genre_filtered)
plt.title('Relación entre Género del Bebé y Nivel de Dolor')
plt.xlabel('Género del Bebé')
plt.ylabel('Nivel de Dolor')
plt.show()
```



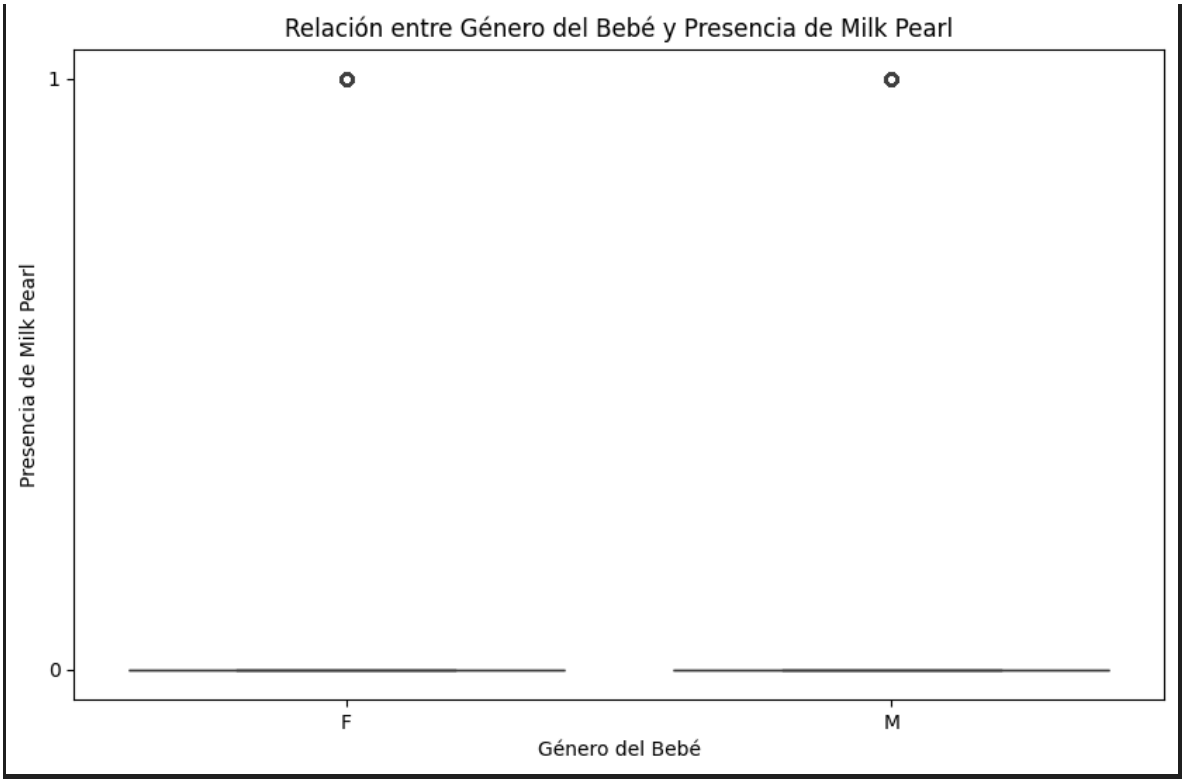

```python
# Boxplot para género y forma del pezón
plt.figure(figsize=(10, 6))
sns.boxplot(x='genre', y='nipleShape', data=data_genre_filtered)
plt.title('Relación entre Género del Bebé y Forma del Pezón')
plt.xlabel('Género del Bebé')
plt.ylabel('Forma del Pezón')
plt.show()
```



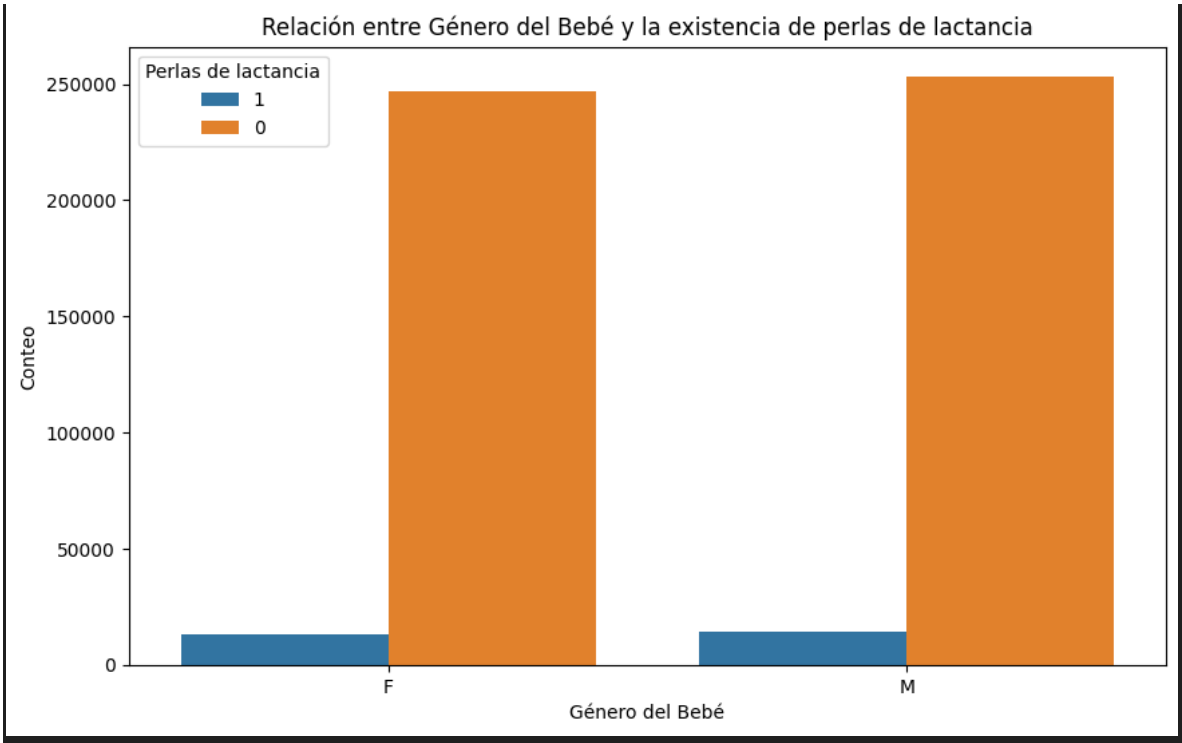

```python
plt.figure(figsize=(10, 6))
sns.countplot(x='genre', hue='nipleShape', data=data_genre_filtered)
plt.title('Relación entre Género del Bebé y Forma del Pezón')
plt.xlabel('Género del Bebé')
plt.ylabel('Conteo')
plt.legend(title='Forma del Pezón')
plt.show()
```



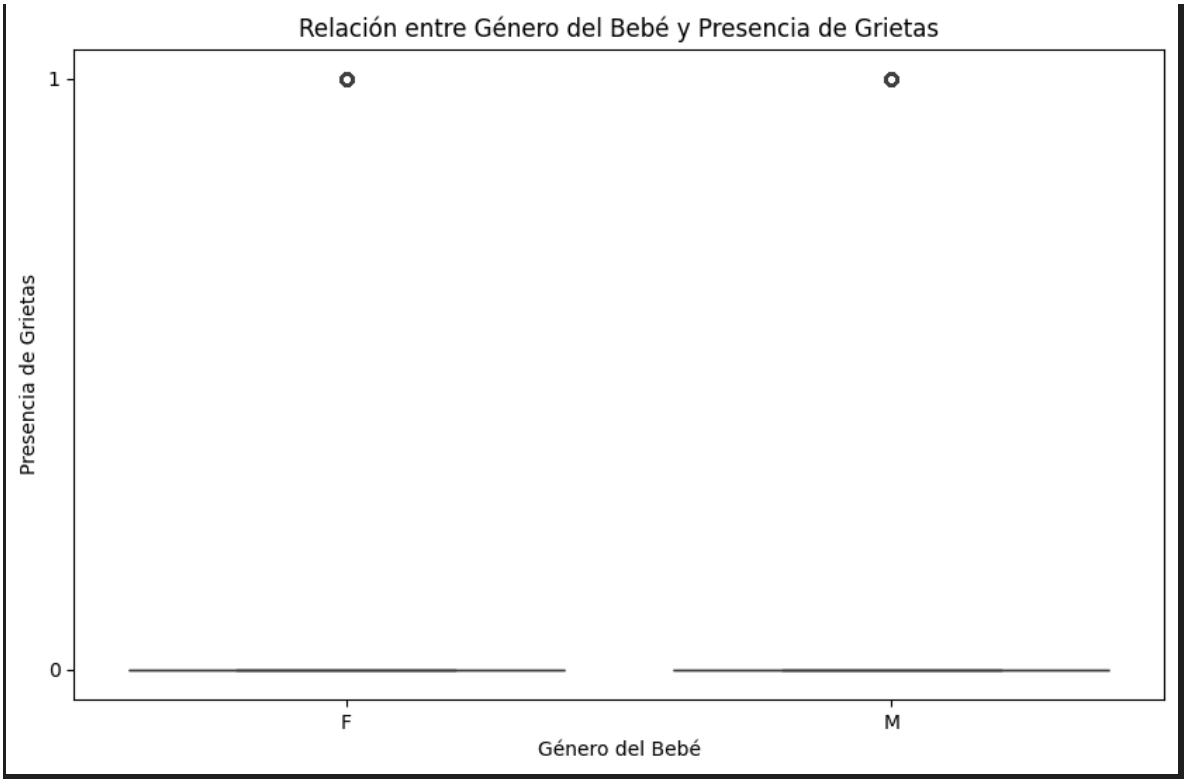

```python
# Boxplot para género y milk pearl
plt.figure(figsize=(10, 6))
sns.boxplot(x='genre', y='milkPearl', data=data_genre_filtered)
plt.title('Relación entre Género del Bebé y Presencia de Milk Pearl')
plt.xlabel('Género del Bebé')
plt.ylabel('Presencia de Milk Pearl')
plt.show()
```




```python
plt.figure(figsize=(10, 6))
sns.countplot(x='genre', hue='milkPearl', data=data_genre_filtered)
plt.title('Relación entre Género del Bebé y la existencia de perlas de
lactancia')
plt.xlabel('Género del Bebé')
plt.ylabel('Conteo')
plt.legend(title='Perlas de lactancia')
plt.show()
```

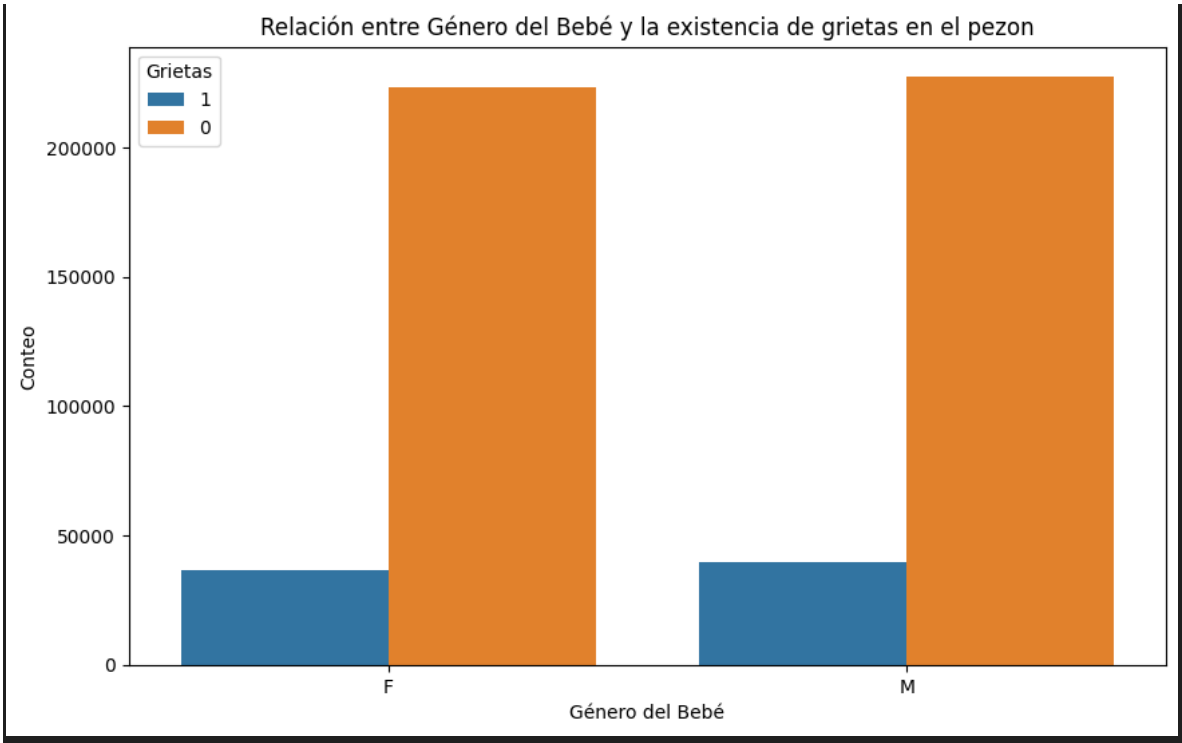



```python
# Boxplot para género y grietas
plt.figure(figsize=(10, 6))
sns.boxplot(x='genre', y='cracks', data=data_genre_filtered)
plt.title('Relación entre Género del Bebé y Presencia de Grietas')
plt.xlabel('Género del Bebé')
plt.ylabel('Presencia de Grietas')
plt.show()
```



De nuevo, no parece ser buena opción el gráfico de cajas.

```python
plt.figure(figsize=(10, 6))
sns.countplot(x='genre', hue='cracks', data=data_genre_filtered)
plt.title('Relación entre Género del Bebé y la existencia de grietas en el pezon')
plt.xlabel('Género del Bebé')
plt.ylabel('Conteo')
plt.legend(title='Grietas')
plt.show()
```




```python
# Convertir las columnas de leche suplementaria a numéricas
data['suplBreastmilk'] = pd.to_numeric(data['suplBreastmilk'], errors='coerce')
data['suplFormula'] = pd.to_numeric(data['suplFormula'], errors='coerce')

# Crear la variable suplementaria de leche materna
data['suplBreastmilkBool'] = data['suplBreastmilk'].apply(lambda x: True if
pd.notnull(x) and x != 0 else False)

# Crear la variable suplementaria de fórmula
data['suplFormulaBool'] = data['suplFormula'].apply(lambda x: True if
pd.notnull(x) and x != 0 else False)

# Crear la variable suplementaria general
data['suplAnyBool'] = data['suplBreastmilkBool'] | data['suplFormulaBool']

# Verificar las nuevas columnas
print(data[['suplBreastmilk', 'suplFormula', 'suplBreastmilkBool',
'suplFormulaBool', 'suplAnyBool']].head())
```

```
   suplBreastmilk  suplFormula  suplBreastmilkBool  suplFormulaBool  \
0             0.0          0.0               False            False
1             0.0          0.0               False            False
3             0.0          0.0               False            False
4             0.0          0.0               False            False
5             0.0          0.0               False            False
```

```
      suplAnyBool
   0        False
   1        False
   3        False
   4        False
   5        False
```

```python
```

```python
# Reemplazar valores 'UNKNOWN' y 'unknown' por NaN en las variables de interés
data['nipleShape'] = data['nipleShape'].replace(['UNKNOWN', 'unknown'], pd.NA)
data['nipleColor'] = data['nipleColor'].replace(['UNKNOWN', 'unknown'], pd.NA)
data['milkPearl'] = data['milkPearl'].replace(['UNKNOWN', 'unknown'], pd.NA)
data['cracks'] = data['cracks'].replace(['UNKNOWN', 'unknown'], pd.NA)
data['painLevel'] = data['painLevel'].replace(['UNKNOWN', 'unknown'], pd.NA)

# Convertir painLevel a numérico (float) para asegurar ordenación correcta
data['painLevel'] = pd.to_numeric(data['painLevel'], errors='coerce')

# Filtrar datos relevantes eliminando filas con NaN en las columnas de interés
data_filtered = data.dropna(subset=['nipleShape', 'nipleColor', 'milkPearl',
'cracks', 'painLevel', 'suplAnyBool'])

# Convertir painLevel a categórico y ordenar los valores de 0 a 10
pain_level_order = [str(i) for i in range(11)]
data_filtered['painLevel'] = pd.Categorical(data_filtered['painLevel'],
categories=pain_level_order, ordered=True)
```

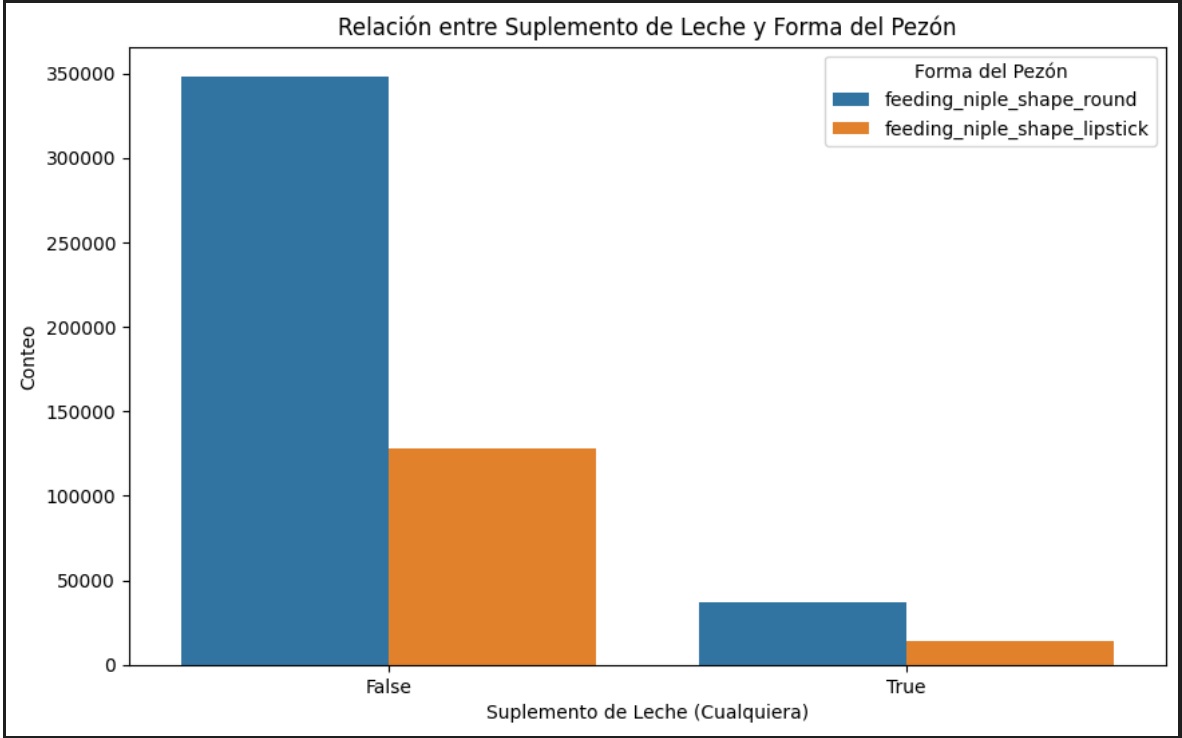
```python
# Crear una copia de los datos originales para no modificarlos directamente
data_genre = data.copy()
```

```python
# Conteo de muestras
conteo_suplAnyBool_nipleShape = data_filtered.groupby(['suplAnyBool',
'nipleShape']).size().reset_index(name='count')
print("\nConteo de muestras por `suplAnyBool` y `nipleShape`:")
print(conteo_suplAnyBool_nipleShape)
```

En primer lugar, se representan las frecuencias de la nueva variable que se dan según la forma del pezón

```
# Gráfico de barras para `suplAnyBool` y `nipleShape`
plt.figure(figsize=(10, 6))
sns.countplot(x='suplAnyBool', hue='nipleShape', data=data_filtered)
plt.title('Relación entre Suplemento de Leche y Forma del Pezón')
plt.xlabel('Suplemento de Leche (Cualquiera)')
plt.ylabel('Conteo')
plt.legend(title='Forma del Pezón')
plt.show()
```

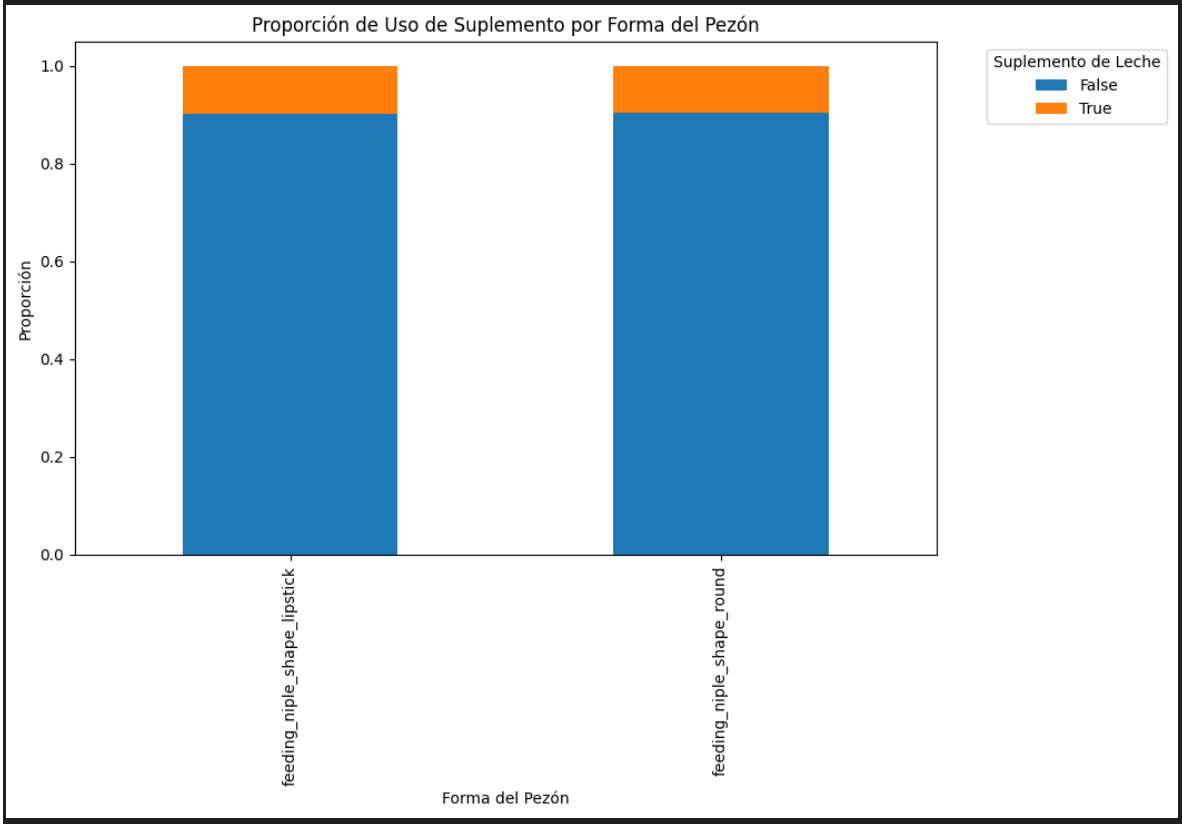
```
Conteo de muestras por `suplAnyBool` y `nipleShape`:
   suplAnyBool                    nipleShape   count
0        False  feeding_niple_shape_lipstick  128346
1        False     feeding_niple_shape_round  348414
2         True  feeding_niple_shape_lipstick   13904
3         True     feeding_niple_shape_round   36865
```



Dado que el gráfico es difícil de interpretar, se decide por normalizar las variables y representarlas en un gráfico de barras apiladas

```python
# Crear tabla de contingencia y calcular proporciones
nipleshape_contingency = pd.crosstab(data_filtered['nipleShape'],
data_filtered['suplAnyBool'], normalize='index')

# Gráfico de barras apiladas
nipleshape_contingency.plot(kind='bar', stacked=True, figsize=(10, 6))
plt.title('Proporción de Uso de Suplemento por Forma del Pezón')
plt.xlabel('Forma del Pezón')
plt.ylabel('Proporción')
plt.legend(title='Suplemento de Leche', bbox_to_anchor=(1.05, 1), loc='upper
left')
plt.show()
```

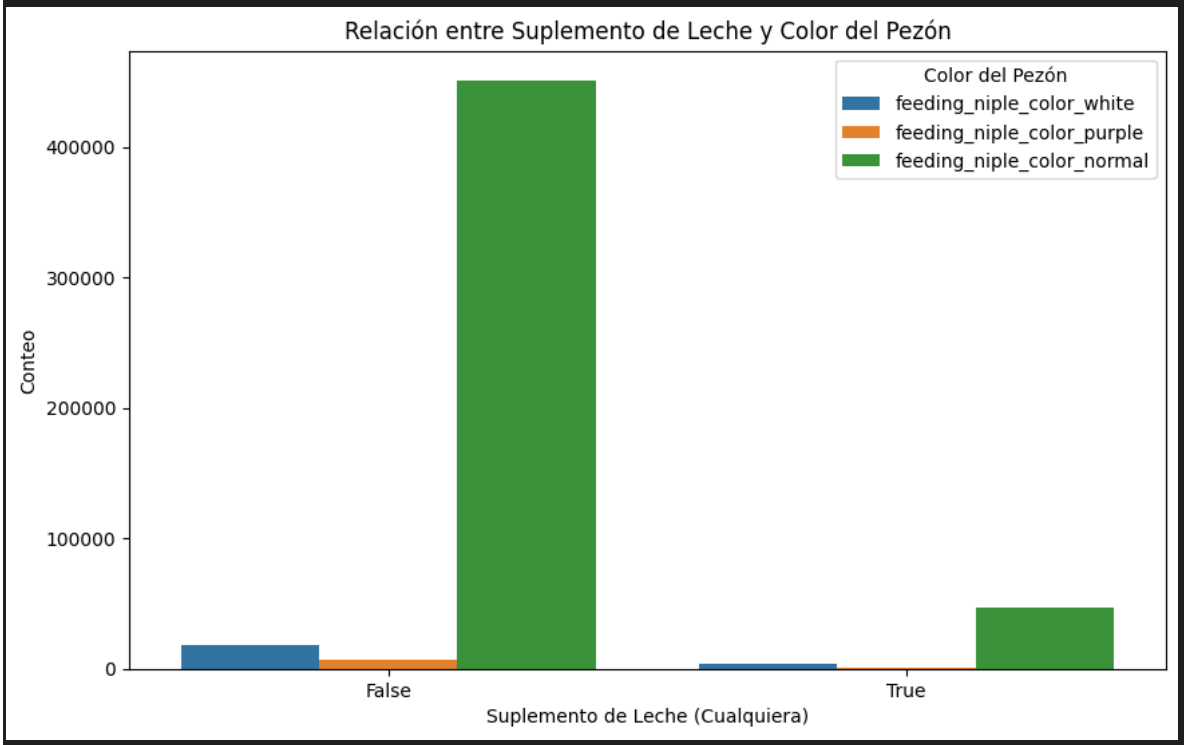



```python
# Conteo de muestras
conteo_suplAnyBool_nippleColor = data_filtered.groupby(['suplAnyBool',
'nipleColor']).size().reset_index(name='count')
print("\nConteo de muestras por `suplAnyBool` y `nipleColor`:")
print(conteo_suplAnyBool_nippleColor)

# Gráfico de barras para `suplAnyBool` y `nippleColor`
```

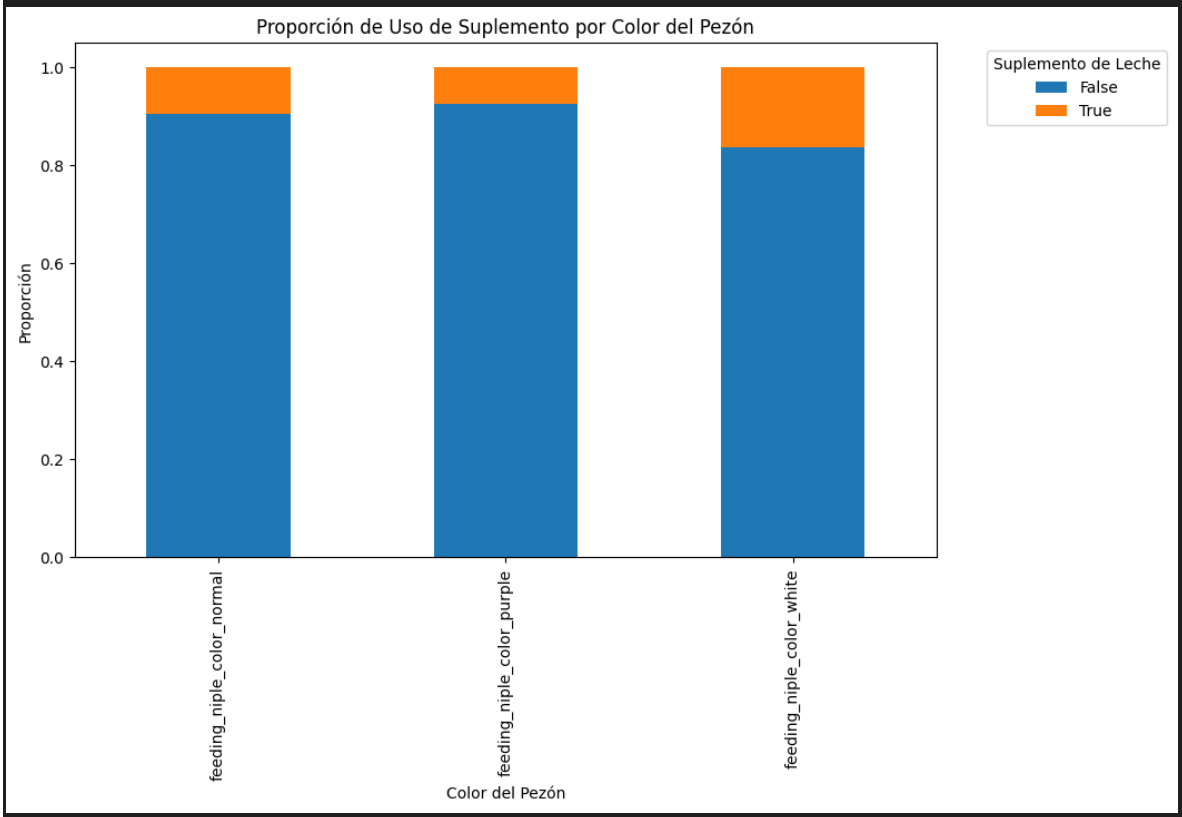
```
plt.figure(figsize=(10, 6))
sns.countplot(x='suplAnyBool', hue='nipleColor', data=data_filtered)
plt.title('Relación entre Suplemento de Leche y Color del Pezón')
plt.xlabel('Suplemento de Leche (Cualquiera)')
plt.ylabel('Conteo')
plt.legend(title='Color del Pezón')
plt.show()
```

```
Conteo de muestras por `suplAnyBool` y `nipleColor`:
   suplAnyBool                nipleColor   count
0        False  feeding_niple_color_normal  451045
1        False  feeding_niple_color_purple    7130
2        False   feeding_niple_color_white   18585
3         True  feeding_niple_color_normal   46569
4         True  feeding_niple_color_purple     575
5         True   feeding_niple_color_white    3625
```




```python
# Crear tabla de contingencia y calcular proporciones
nipplecolor_contingency = pd.crosstab(data_filtered['nipleColor'],
data_filtered['suplAnyBool'], normalize='index')
```

```
# Gráfico de barras apiladas
nipplecolor_contingency.plot(kind='bar', stacked=True, figsize=(10, 6))
plt.title('Proporción de Uso de Suplemento por Color del Pezón')
plt.xlabel('Color del Pezón')
plt.ylabel('Proporción')
plt.legend(title='Suplemento de Leche', bbox_to_anchor=(1.05, 1), loc='upper
left')
plt.show()
```

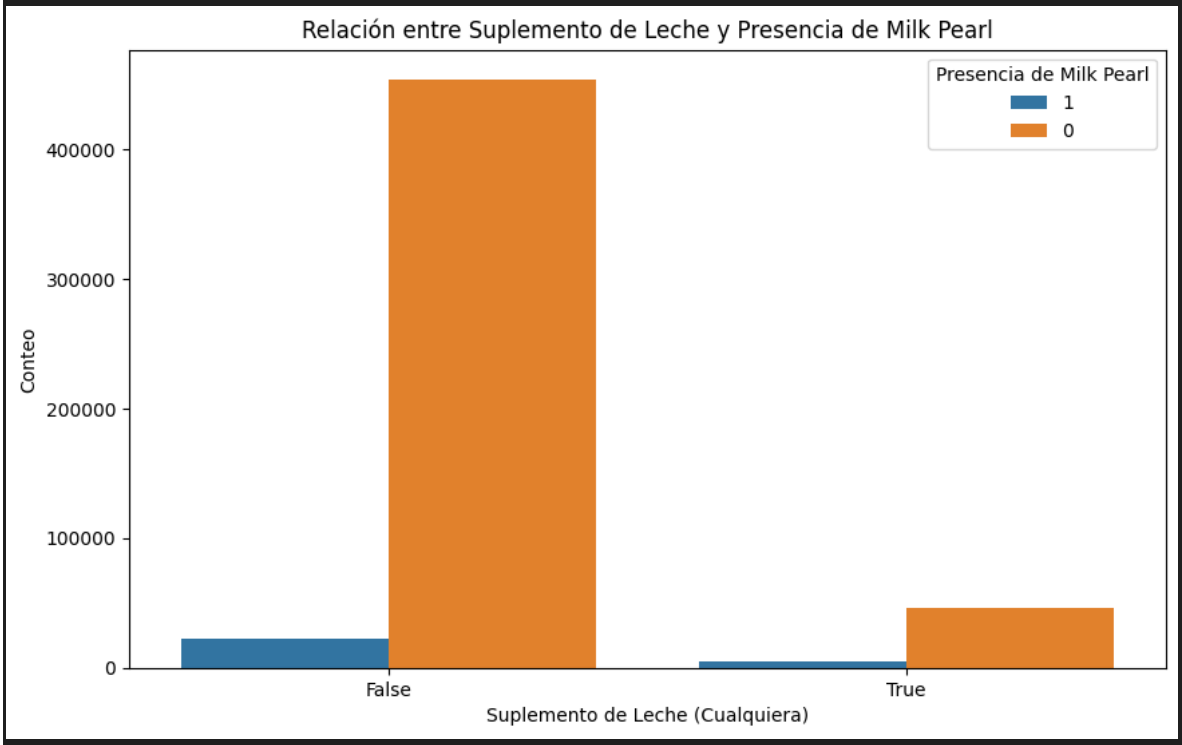



```python
# Conteo de muestras
conteo_suplAnyBool_milkPearl = data_filtered.groupby(['suplAnyBool',
'milkPearl']).size().reset_index(name='count')
print("\nConteo de muestras por `suplAnyBool` y `milkPearl`:")
print(conteo_suplAnyBool_milkPearl)

# Gráfico de barras para `suplAnyBool` y `milkPearl`
plt.figure(figsize=(10, 6))
sns.countplot(x='suplAnyBool', hue='milkPearl', data=data_filtered)
plt.title('Relación entre Suplemento de Leche y Presencia de Milk Pearl')
plt.xlabel('Suplemento de Leche (Cualquiera)')
plt.ylabel('Conteo')
plt.legend(title='Presencia de Milk Pearl')
plt.show()
```

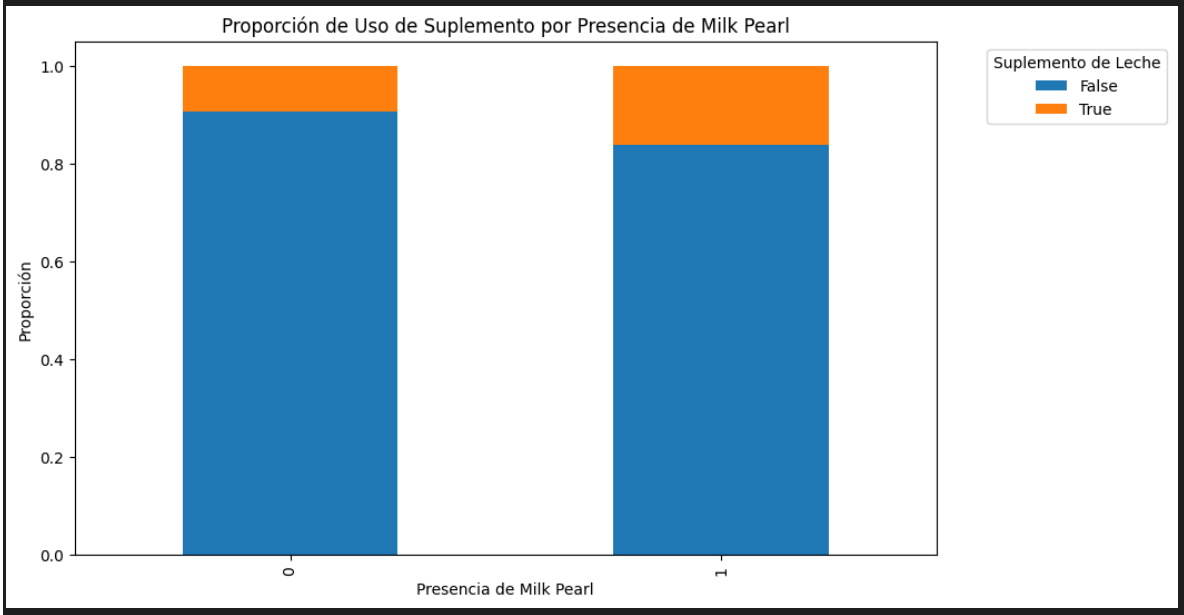
```
```

```
Conteo de muestras por `suplAnyBool` y `milkPearl`:
   suplAnyBool milkPearl   count
0        False         0  454041
1        False         1   22719
2         True         0   46375
3         True         1    4394
```




```python
# Crear tabla de contingencia y calcular proporciones
milkpearl_contingency = pd.crosstab(data_filtered['milkPearl'],
data_filtered['suplAnyBool'], normalize='index')

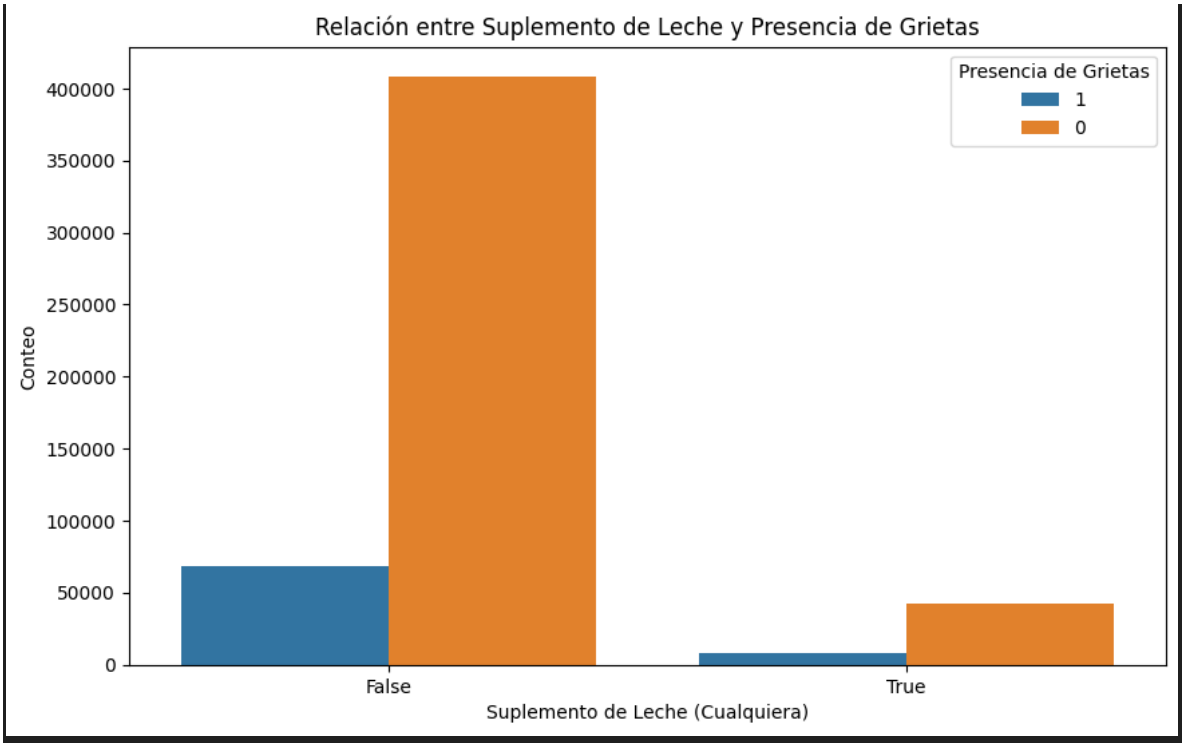
# Gráfico de barras apiladas
milkpearl_contingency.plot(kind='bar', stacked=True, figsize=(10, 6))
plt.title('Proporción de Uso de Suplemento por Presencia de Milk Pearl')
plt.xlabel('Presencia de Milk Pearl')
plt.ylabel('Proporción')
plt.legend(title='Suplemento de Leche', bbox_to_anchor=(1.05, 1), loc='upper
left')
plt.show()
```




```python
# Conteo de muestras
conteo_suplAnyBool_cracks = data_filtered.groupby(['suplAnyBool',
'cracks']).size().reset_index(name='count')
print("\nConteo de muestras por `suplAnyBool` y `cracks`:")
print(conteo_suplAnyBool_cracks)

# Gráfico de barras para `suplAnyBool` y `cracks`
plt.figure(figsize=(10, 6))
sns.countplot(x='suplAnyBool', hue='cracks', data=data_filtered)
plt.title('Relación entre Suplemento de Leche y Presencia de Grietas')
plt.xlabel('Suplemento de Leche (Cualquiera)')
plt.ylabel('Conteo')
plt.legend(title='Presencia de Grietas')
plt.show()
```

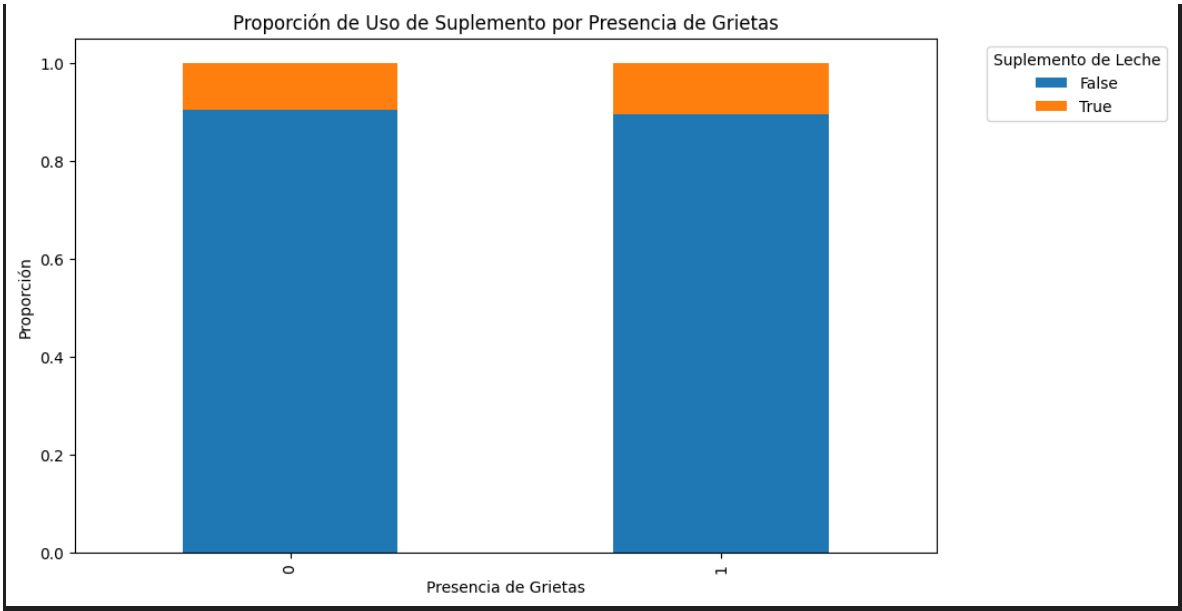
```
Conteo de muestras por `suplAnyBool` y `cracks`:
   suplAnyBool cracks   count
0        False      0  408547
1        False      1   68213
2         True      0   42726
3         True      1    8043
```




```python

# Crear tabla de contingencia y calcular proporciones
cracks_contingency = pd.crosstab(data_filtered['cracks'],
data_filtered['suplAnyBool'], normalize='index')

# Gráfico de barras apiladas
cracks_contingency.plot(kind='bar', stacked=True, figsize=(10, 6))
plt.title('Proporción de Uso de Suplemento por Presencia de Grietas')
plt.xlabel('Presencia de Grietas')
plt.ylabel('Proporción')
plt.legend(title='Suplemento de Leche', bbox_to_anchor=(1.05, 1), loc='upper
left')
plt.show()
```




```python
# Crear la variable 'hora_toma' con horas y minutos
data['hora_toma'] = data['startDate'].dt.hour + data['startDate'].dt.minute / 60

# Verificar los primeros registros
print(data[['startDate', 'hora_toma']].head())
```

```
              startDate  hora_toma
    0 2018-02-15 14:26:00  14.433333
    1 2018-02-15 13:33:00  13.550000
    3 2018-03-19 13:44:00  13.733333
    4 2018-03-21 13:34:00  13.566667
    5 2018-03-22 15:44:00  15.733333
```

```python
# Reemplazar valores 'unknown' por NaN en painLevel
data['painLevel'] = data['painLevel'].replace(['unknown', 'UNKNOWN'], pd.NA)

# Convertir painLevel a numérico (float) para asegurar ordenación correcta
data['painLevel'] = pd.to_numeric(data['painLevel'], errors='coerce')

# Filtrar datos relevantes eliminando filas con NaN en painLevel
data_filtered = data.dropna(subset=['painLevel', 'hora_toma'])
```

```python
# Crear el gráfico de dispersión
plt.figure(figsize=(12, 6))
```

```
sns.scatterplot(x='hora_toma', y='painLevel', data=data_filtered)

# Configurar las marcas en el eje X de 0 a 24
plt.xticks(ticks=range(25))

# Configurar el título y las etiquetas de los ejes
plt.title('Relación entre Nivel de Dolor y Hora del Día')
plt.xlabel('Hora del Día')
plt.ylabel('Nivel de Dolor')

# Mostrar el gráfico
plt.show()
```



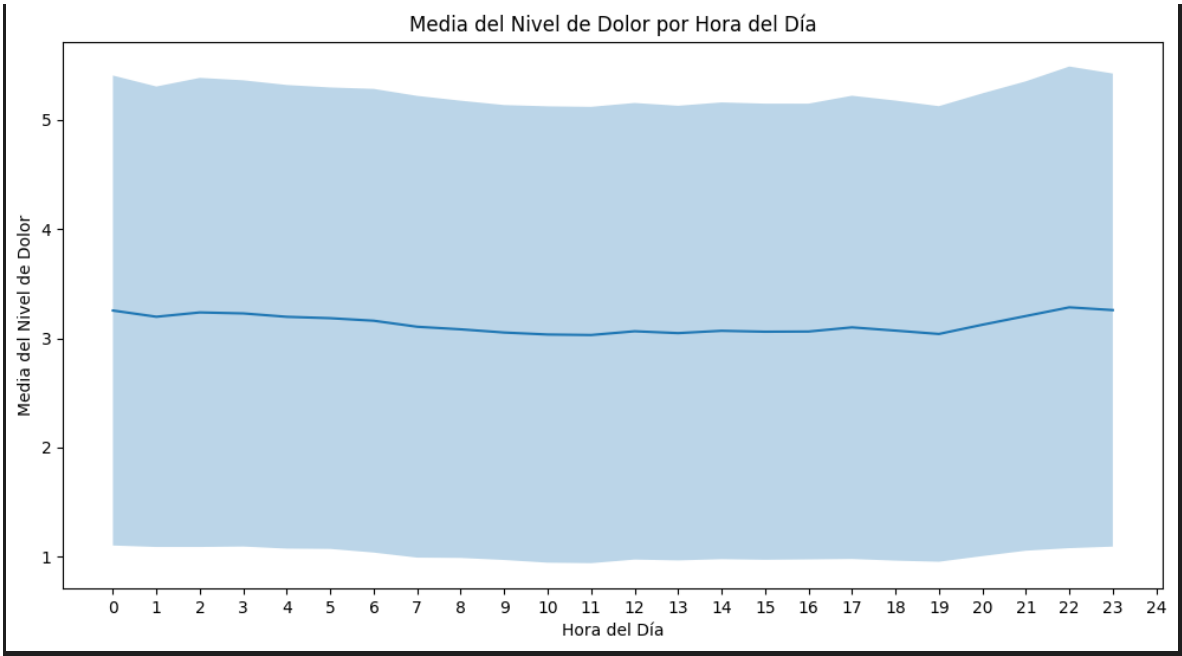

```python
# Convertir painLevel a numérico y crear una copia del DataFrame filtrado
data_filtered = data.copy()
data_filtered['painLevel'] = pd.to_numeric(data_filtered['painLevel'],
errors='coerce')

# Verificar si hay valores NaN después de la conversión y eliminarlos
data_filtered = data_filtered.dropna(subset=['painLevel'])
# Filtrar los valores de dolor que sean 0
data_filtered = data_filtered[data_filtered['painLevel'] != 0]
```

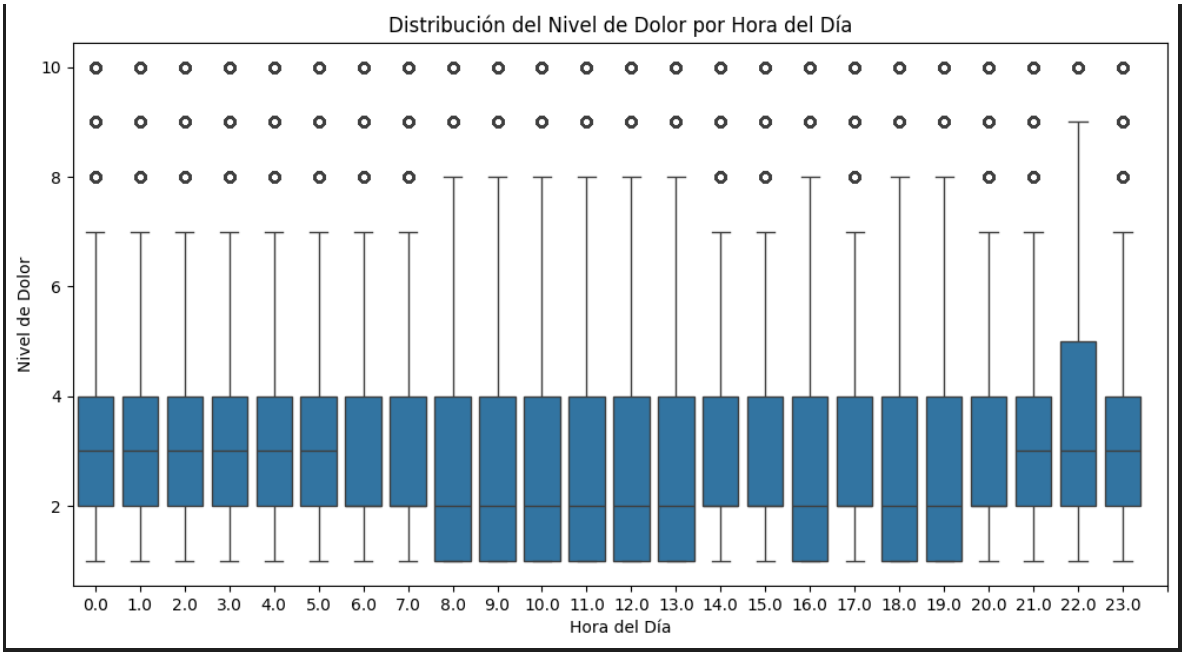
```python
# Calcular la media y la desviación estándar del nivel de dolor por hora del día
pain_stats_by_hour =
data_filtered.groupby(data_filtered['startDate'].dt.hour)['painLevel'].agg(['mea
n', 'std']).reset_index()
print(pain_stats_by_hour)
```

	startDate	mean	std
0	0.0	3.253854	2.152543
1	1.0	3.196847	2.109203
2	2.0	3.236299	2.148697
3	3.0	3.227280	2.135470
4	4.0	3.196274	2.123576
5	5.0	3.183576	2.113603
6	6.0	3.160281	2.123794
7	7.0	3.105080	2.114929
8	8.0	3.081656	2.094276
9	9.0	3.052329	2.083744
10	10.0	3.034087	2.090255
11	11.0	3.029570	2.090185
12	12.0	3.063803	2.091345
13	13.0	3.046971	2.082487
14	14.0	3.068155	2.092049
15	15.0	3.059665	2.089369
16	16.0	3.061412	2.087490
17	17.0	3.099601	2.121806
18	18.0	3.069836	2.107239
19	19.0	3.039119	2.086739
20	20.0	3.123665	2.119553
21	21.0	3.203510	2.149934
22	22.0	3.283091	2.206071
23	23.0	3.257860	2.166979

```python
# Gráfico de línea para la media del nivel de dolor por hora del día
plt.figure(figsize=(12, 6))
sns.lineplot(x=pain_stats_by_hour['startDate'], y=pain_stats_by_hour['mean'])
plt.fill_between(pain_stats_by_hour['startDate'], pain_stats_by_hour['mean'] -
pain_stats_by_hour['std'], pain_stats_by_hour['mean'] +
pain_stats_by_hour['std'], alpha=0.3)
plt.title('Media del Nivel de Dolor por Hora del Día')
plt.xlabel('Hora del Día')
plt.ylabel('Media del Nivel de Dolor')
plt.xticks(ticks=range(25))
plt.show()
```



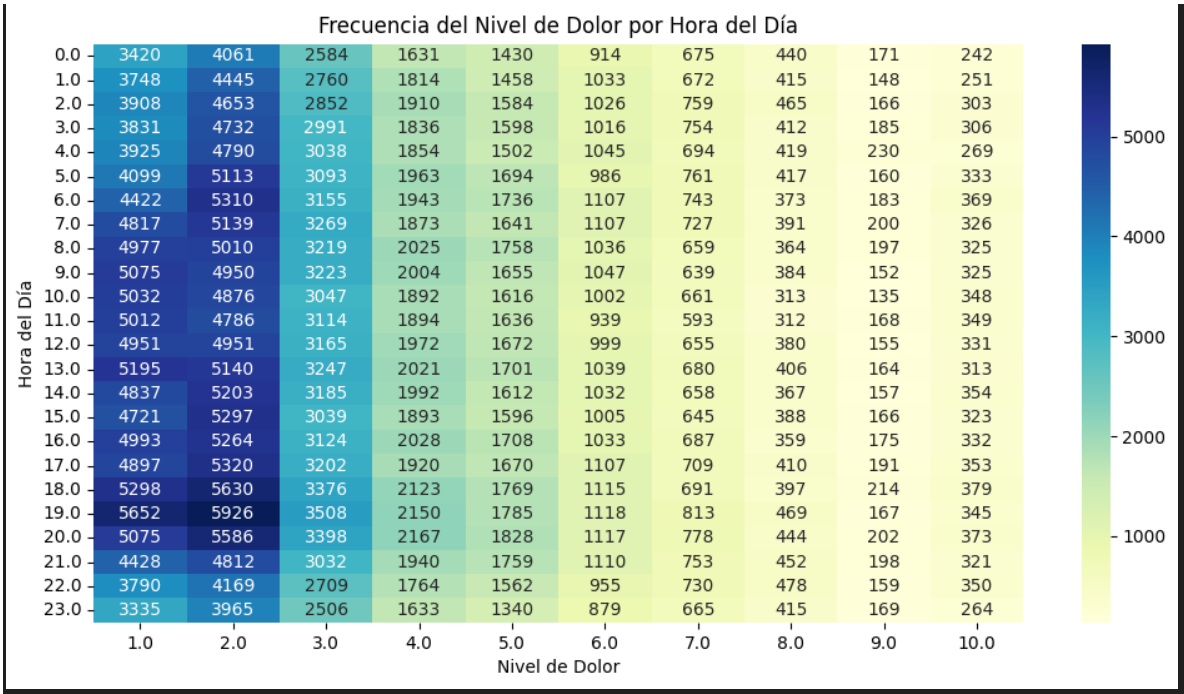

```python
# Crear boxplots para la distribución del nivel de dolor por hora del día
plt.figure(figsize=(12, 6))
sns.boxplot(x=data_filtered['startDate'].dt.hour, y='painLevel',
data=data_filtered)
plt.title('Distribución del Nivel de Dolor por Hora del Día')
plt.xlabel('Hora del Día')
plt.ylabel('Nivel de Dolor')
plt.xticks(ticks=range(25))
plt.show()
```




```python
# Crear una tabla de frecuencia
heatmap_data = pd.crosstab(data_filtered['startDate'].dt.hour,
data_filtered['painLevel'])

# Crear el heatmap
plt.figure(figsize=(12, 6))
sns.heatmap(heatmap_data, cmap='YlGnBu', annot=True, fmt='d')
plt.title('Frecuencia del Nivel de Dolor por Hora del Día')
plt.xlabel('Nivel de Dolor')
plt.ylabel('Hora del Día')
plt.show()

```




```python
# Crear la variable 'hora_toma' con horas y minutos
data['hora_toma'] = data['startDate'].dt.hour + data['startDate'].dt.minute / 60

# Verificar los primeros registros
print(data[['startDate', 'hora_toma']].head())
```

```
             startDate  hora_toma
0  2018-02-15 14:26:00  14.433333
1  2018-02-15 13:33:00  13.550000
3  2018-03-19 13:44:00  13.733333
4  2018-03-21 13:34:00  13.566667
5  2018-03-22 15:44:00  15.733333
```

```python
# Crear la variable 'edad_bebé_dias' restando la fecha de nacimiento de la fecha de inicio de la toma
data['edad_bebé_dias'] = (data['startDate'] - data['bornDate.1']).dt.days

# Verificar los primeros registros
print(data[['startDate', 'bornDate.1', 'edad_bebé_dias']].head())
```

```
             startDate bornDate.1  edad_bebé_dias
0  2018-02-15 14:26:00 2020-10-06          -964.0
1  2018-02-15 13:33:00 2020-10-06          -964.0
3  2018-03-19 13:44:00 2020-10-06          -932.0
4  2018-03-21 13:34:00 2020-10-06          -930.0
```

```
    5 2018-03-22 15:44:00 2020-10-06          -929.0
```

```python
# Filtrar los valores anómalos, si existen
# Por ejemplo, eliminando edades negativas o extremadamente altas
data_filtered_age = data[(data['edad_bebé_dias'] >= 0) & (data['edad_bebé_dias']
<= 500)]  # Filtrar edades en un rango razonable (por ejemplo, 0 a 10 años)

# Calcular los valores máximos y mínimos de la edad del bebé
max_age = data_filtered_age['edad_bebé_dias'].max()
min_age = data_filtered_age['edad_bebé_dias'].min()

# Mostrar los valores máximo y mínimo
print(f"Edad máxima del bebé (en días): {max_age}")
print(f"Edad mínima del bebé (en días): {min_age}")

# Verificar la distribución de la nueva variable de edad
print(data_filtered_age['edad_bebé_dias'].describe())
```

```
    Edad máxima del bebé (en días): 500.0
    Edad mínima del bebé (en días): 0.0
    count    6.725817e+06
    mean     7.790072e+01
    std      7.531921e+01
    min      0.000000e+00
    25%      2.300000e+01
    50%      5.500000e+01
    75%      1.090000e+02
    max      5.000000e+02
    Name: edad_bebé_dias, dtype: float64
```

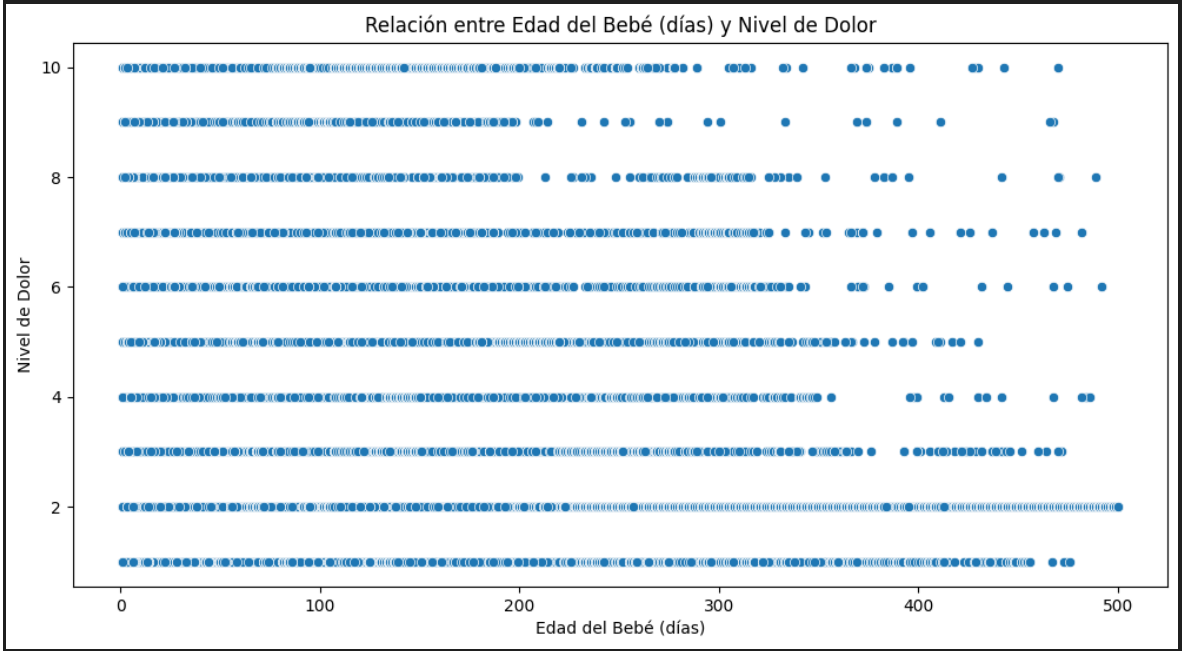
```python
# Convertir painLevel a numérico y eliminar valores irrelevantes
data_filtered_age['painLevel'] = pd.to_numeric(data_filtered_age['painLevel'],
errors='coerce')

# Eliminar registros con painLevel NaN
data_filtered_age = data_filtered_age.dropna(subset=['painLevel'])

# Filtrar los valores de dolor y edad que sean 0
data_filtered_age = data_filtered_age[(data_filtered_age['painLevel'] != 0) &
(data_filtered_age['edad_bebé_dias'] != 0)]
```

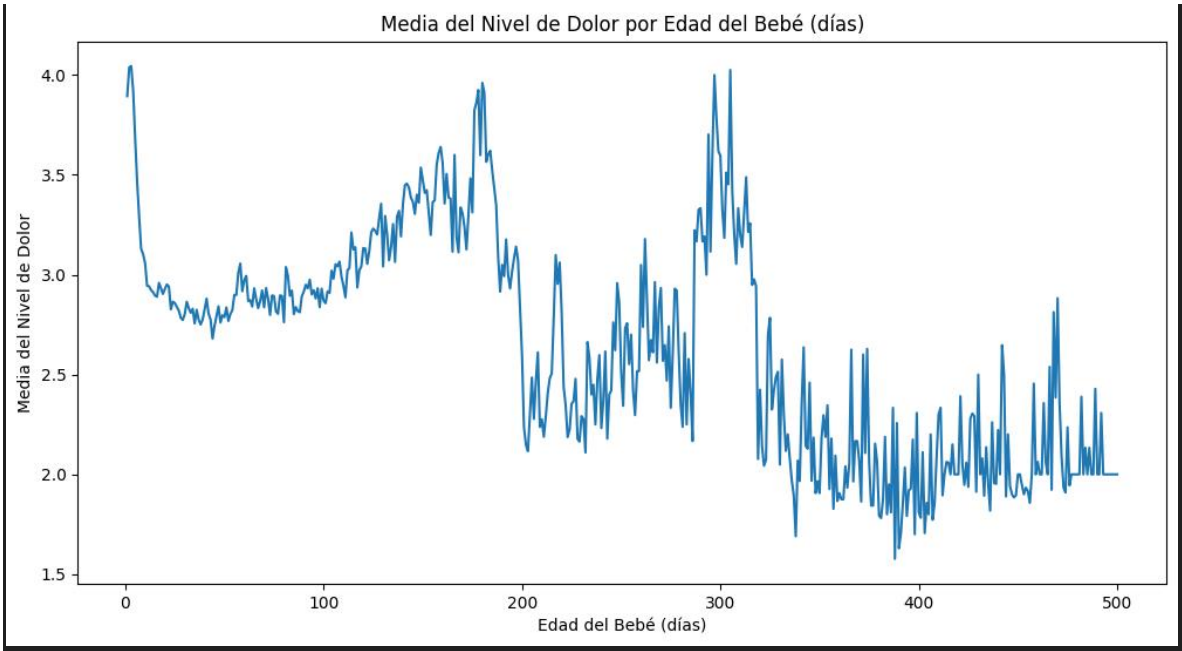
```python
# Crear un gráfico de dispersión de edad del bebé y nivel de dolor
```

```
plt.figure(figsize=(12, 6))
sns.scatterplot(x='edad_bebé_dias', y='painLevel', data=data_filtered_age)
plt.title('Relación entre Edad del Bebé (días) y Nivel de Dolor')
plt.xlabel('Edad del Bebé (días)')
plt.ylabel('Nivel de Dolor')
plt.show()
```




```python
# Calcular la media del nivel de dolor por edad del bebé en días
pain_stats_by_age =
data_filtered_age.groupby('edad_bebé_dias')['painLevel'].mean().reset_index()

# Crear un gráfico de línea para la media del nivel de dolor por edad del bebé
plt.figure(figsize=(12, 6))
sns.lineplot(x='edad_bebé_dias', y='painLevel', data=pain_stats_by_age)
plt.title('Media del Nivel de Dolor por Edad del Bebé (días)')
plt.xlabel('Edad del Bebé (días)')
plt.ylabel('Media del Nivel de Dolor')
plt.show()
```




```python
# Preparar los datos para la regresión
X = pain_stats_by_age[['edad_bebé_dias']]  # Variable independiente
y = pain_stats_by_age['painLevel']         # Variable dependiente

# Crear y ajustar el modelo de regresión lineal
model = LinearRegression()
model.fit(X, y)

# Predecir los valores de y usando el modelo ajustado
y_pred = model.predict(X)
```

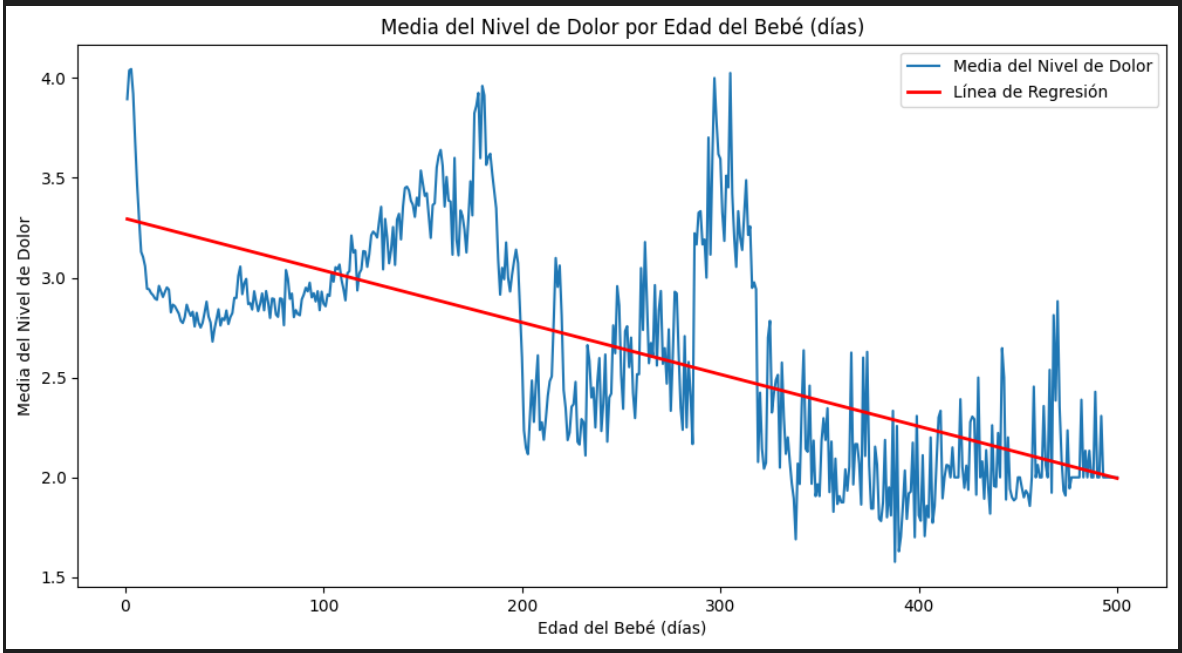
```python
# Crear el gráfico de línea para la media del nivel de dolor por edad del bebé
plt.figure(figsize=(12, 6))
sns.lineplot(x='edad_bebé_dias', y='painLevel', data=pain_stats_by_age,
label='Media del Nivel de Dolor')

# Dibujar la línea de regresión
plt.plot(pain_stats_by_age['edad_bebé_dias'], y_pred, color='red', linewidth=2,
label='Línea de Regresión')

# Configurar el gráfico
plt.title('Media del Nivel de Dolor por Edad del Bebé (días)')
plt.xlabel('Edad del Bebé (días)')
plt.ylabel('Media del Nivel de Dolor')
plt.legend()
```

```
plt.show()

# Mostrar los coeficientes de la regresión
print(f"Coeficiente de la regresión lineal: {model.coef_[0]}")
print(f"Intercepción de la regresión lineal: {model.intercept_}")
print(f"R^2: {model.score(X, y)}")
```




```
Coeficiente de la regresión lineal: -0.002602527114002052
Intercepción de la regresión lineal: 3.296576848345022
R^2: 0.46579576931807976
```

```python
```

ANEXO D - ANÁLISIS DE DATOS EXTERNO (NATALIA GORRÍN, CRISTINA SALAZAR, BEATRIZ MARTÍN)



ANÁLISIS DE DATOS

Natalia Gorrín Moreno
Cristina Salazar Diego
Beatriz Martín Cinos

Tabla de contenidos

1. ***PREPROCESADO DE LOS DATOS*** ***4***

trackID *4*
UserID *4*
Baby1ID *4*
Baby2ID *5*
bornDateMom *6*
numSons *7*
Premature *10*
gestationWeek *10*
bornHeight *11*
babyTandemGenre *13*
babyTandemBornDate *13*
babyTandemPremature *13*
babyTandemGestationWeek *14*
babyTandemBornHeight *14*
StartDate *15*
endDate *15*
breast *16*
suplBreastMilk *17*
suplFormula *18*
painLevel *19*
nipleShape *19*
nipleColor *20*
milkPearl *20*
Cracks *21*
emotionalStatus *21*
Notes *22*
creationDate *22*
modificationDate *23*
city *23*

2. ***ANÁLISIS DE RESULTADOS*** ***24***

2.1. Gráficas de frecuencias 24

Promedio de la frecuencia de tomas por edad del bebé *24*

Influencia de la diferencia de edad en la frecuencia de las tomas en tándem *25*

Promedio de la frecuencia de tomas por edad del primer bebé *25*

Promedio de tomas por edad del segundo bebe (tándem) *26*

Frecuencia de Tomas a lo largo del tiempo *27*

2.2. Gráficas de Países 27

Número de Países *27*

Cantidad de tomas por usuario en cada país *28*

Número de mujeres registradas por país *28*

2.3. Relaciones con la Vuelta al Trabajo 29

Relación entre la Edad promedio del Bebé y Número Promedio de Tomas por Usuario añadiendo la vuelta al trabajo *29*

Relación entre la edad Promedio del Bebé y Número promedio de tomas por País añadiendo la vuelta al trabajo *30*

Relación entre la edad Promedio del Bebé y Número promedio de tomas por estado emocional añadiendo la vuelta al trabajo *30*

2.4. Gráficas de emociones 31

Estado emocional *31*

Análisis de sentimientos negativos/tristes: exhausted, frustrated y sad *32*

2.5. Análisis de las Notas 34

Word Cloud *34*

Análisis de sentimientos *34*

Histograma de palabras más frecuentes *34*

Correlación de variables numéricas *35*

3. *CONCLUSIONES Y PRÓXIMOS PASOS* *36*

1. PREPROCESADO DE LOS DATOS

Se han realizado dos procesados de los datos en crudo. El primer procesado se ha enfocado en borrar datos erróneos, sobre todo aquellos que vinieran mal registrados, y en un segundo procesado de los datos se han eliminado todos aquellos valores y columnas que no fuesen relevantes por ser nulos, entre otros, haciendo una limpieza más exhaustiva y manteniendo solo los que estuvieran totalmente correctos para su posterior análisis. Se ha partido de un dataset de 6 millones de registros y los datos finales con los que se han realizado los análisis han sido aproximadamente 300.000 (separando posteriormente tándem de no tándem).

Las variables estudiadas y su correspondiente tratamiento han sido:

trackID

Es el identificador único para cada registro de seguimiento, son tomas únicas.

UserID

Identificación de las usuarias. En total, se han registrado 76,783 mujeres.

Baby1ID

Identifica al primer bebé de la madre. El número total de Baby1ID supera el de las madres, indicando que algunas madres han registrado a más de un bebé bajo este identificador. En total, 2231 usuarias tienen más de un bebé registrado bajo Baby1ID.

Además de registrar múltiples bebés bajo el identificador Baby1ID, se observa en la imagen que algunas usuarias han registrado a diferentes bebés en distintos periodos de tiempo.

```
userID  baby1ID
16270   16878        [16880, 16879, 50683]
23116   16334              [48168, 16334]
42556   35361        [37915, 35361, 34822]
        37915              [34822, 35361]
102185  286818           [286825, 286818]
        286825     [64605, 286825, 286818]
102612  196864           [196865, 196864]
198765  127114           [127114, 127115]
207995  140072           [140072, 140073]
211164  139559           [139559, 139586]
212714  141075           [141075, 141076]
231585  159721           [159721, 159722]
257217  184897           [184897, 184896]
263474  190933           [190933, 190932]
420349  363420           [363419, 363420]
521749  450632           [450632, 450630]
525751  447672           [447672, 447671]
Name: baby2ID, dtype: object
```

Es necesario recolocar estos registros de tal manera que no haya un mismo Baby1ID repetido por madre y garantizar la coherencia, aunque antes de hacerlo, es crucial revisar las demás columnas del dataset.

Baby2ID

En este caso, el total de usuarias que están registrando más de un bebe bajo la columna baby2ID son 17.

Se observa, como ocurría con Baby1ID, que hay ciertas madres que tienen repetido el baby1ID en el baby2ID. Para solventar este problema, se ha eliminado el baby2ID que estuviese repetido, manteniendo solo el resto.

Se eliminan los registros que tienen 4 bebes simultáneamente porque se considera imposible.

```
userID  baby1ID
16270   16878        [16880, 16879, 50683]
23116   16334              [48168, 16334]
42556   35361       [37915, 35361, 34822]
        37915              [34822, 35361]
102185  286818           [286825, 286818]
        286825    [64605, 286825, 286818]
102612  196864           [196865, 196864]
198765  127114           [127114, 127115]
207995  140072           [140072, 140073]
211164  139559           [139559, 139586]
212714  141075           [141075, 141076]
231585  159721           [159721, 159722]
257217  184897           [184897, 184896]
263474  190933           [190933, 190932]
420349  363420           [363419, 363420]
521749  450632           [450632, 450630]
525751  447672           [447672, 447671]
Name: baby2ID, dtype: object
```

En una segunda fase de procesamiento, más exhaustiva, se procedió a eliminar todos los valores nulos, mejorando así la calidad y la integridad de los datos. Después de procesar esta variable (borrar los registros erróneos y recolocar los IDs de manera única para cada madre, queda así el registro:

```
userID  baby1ID
102612  196864     [196865.0]
197437  370451     [125817.0]
221057  413883     [149038.0]
270243  197439     [199006.0]
310379  248523     [248524.0]
330915  303184     [303185.0]
339088  282022     [282023.0]
503662  428925     [429490.0]
535439  457019     [457020.0]
Name: baby2ID, dtype: object
```

bornDateMom

Se observa que hay 4M de registros con la fecha de nacimiento de la madre nula. A partir de esta columna, se crea una nueva columna que se llama ***momAge***, que se calcula restando la fecha actual menos la fecha de nacimiento de la madre.

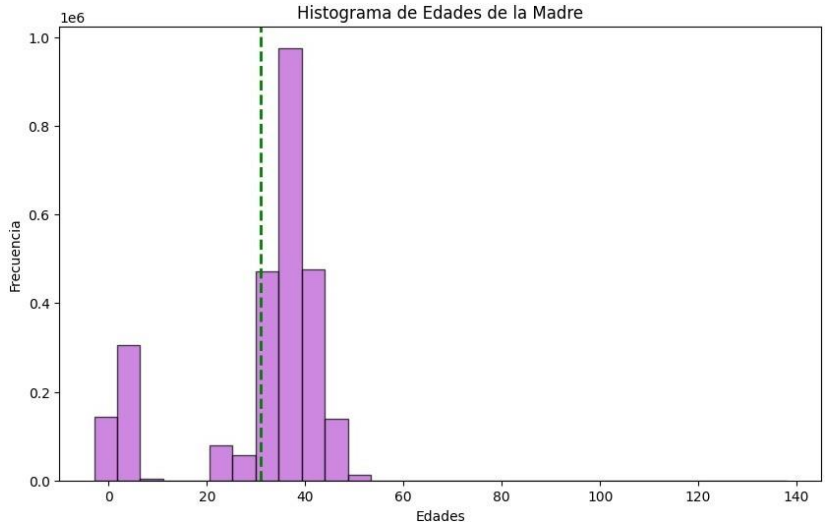
Se observa, a partir del histograma de la imagen, que los registros más frecuentes son las edades de 35-40 años. Por otro lado, hay edades en torno a 0 y también hay registros negativos, lo cual no es posible y se tienen que tratar. Se procede a eliminar estos datos incorrectos.

26.0	5199
23.0	3547
6.0	3244
8.0	2532
22.0	1659
50.0	1549
21.0	393
138.0	362
17.0	311
54.0	279
51.0	271
7.0	254
9.0	68
20.0	47
15.0	37
53.0	33
18.0	16
19.0	12
10.0	7
60.0	5
52.0	4
13.0	4
16.0	4
55.0	3
11.0	3
63.0	2
12.0	1
57.0	1
-3.0	1

momAge

Se han borrado las edades de las madres fuera del rango de 15 a 55 ya que se han considerado edades muy poco probables, y también las que tienen

valores nulos.



numSons

Esta variable representa el número de hijos, aunque no se tiene claro si hijos totales de una madre o el número de hijos registrados en la app. Hay valores que llegan hasta el millón, lo cual es un error, así como valores de 105 hijos por madre. Se limitaron los registros válidos a aquellos con cinco hijos o menos. Este es el conteo inicial de la variable antes de procesarla:

```
numSons
0.0            3680891
1.0            2589828
2.0             280885
3.0              31282
4.0               4384
10.0               479
5.0                280
11.0               114
619925729.0         27
7.0                 13
8.0                  8
118.0                7
6.0                  5
105.0                4
5100.0               2
689203058.0          2
13.0                 1
Name: count, dtype: int64
```

Se borran los registros mencionados (que están claramente mal) y se dejan únicamente aquellos que tienen un número menor o igual a 5 hijos.

dueDate

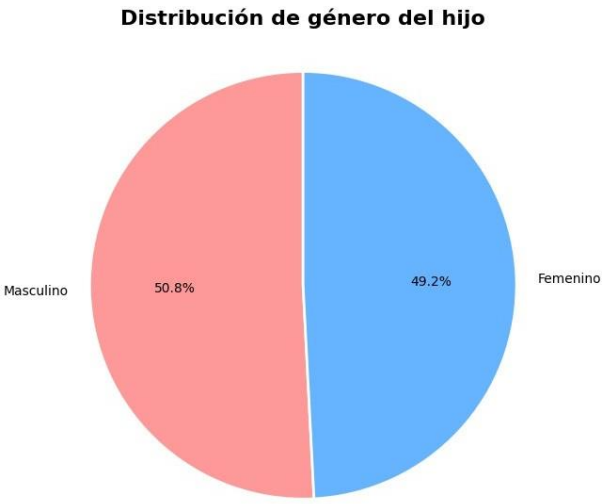
Esta variable representa la fecha estimada de parto, la fecha en la que "se sale de cuentas". Hay muchos registros, como se observa en la imagen, que tienen fechas de 1670. Estas fechas no tienen sentido ya que son fechas incluso previas al nacimiento de la madre.

```
dueDate
2019-03-08    18096
2020-10-03     7023
1970-01-01     6652
2021-04-02     6123
2023-03-30     5770
               ...
2018-12-23        1
2018-09-06        1
2018-12-08        1
2018-11-21        1
2024-01-21        1
Name: count, Length: 1872, dtype: int64
```

Se calcula el porcentaje de registros erróneos, considerando erróneos aquellos en los que hay una diferencia desde que se sale de cuentas hasta el parto superior a 5 meses arriba o abajo, ya que serían partos prematuros o postérminos. El porcentaje de errores se considera del 85.03%.

Genre

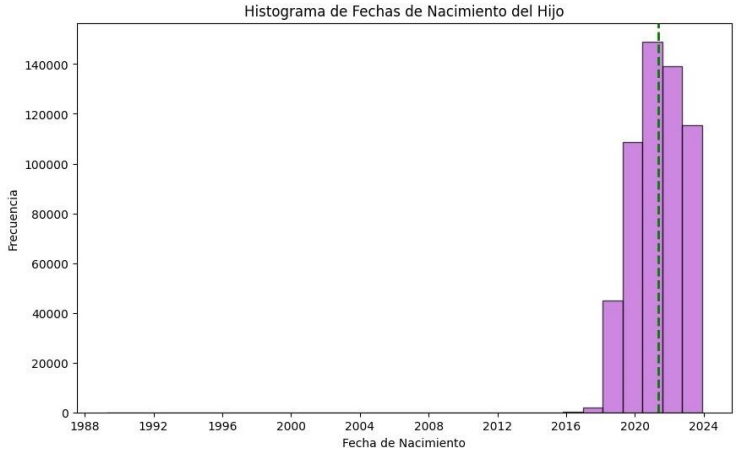
La distribución del género del hijo es bastante balanceada, con un 50.8% para masculino y otro 46.2% para femenino, como se muestra en el pie chart:



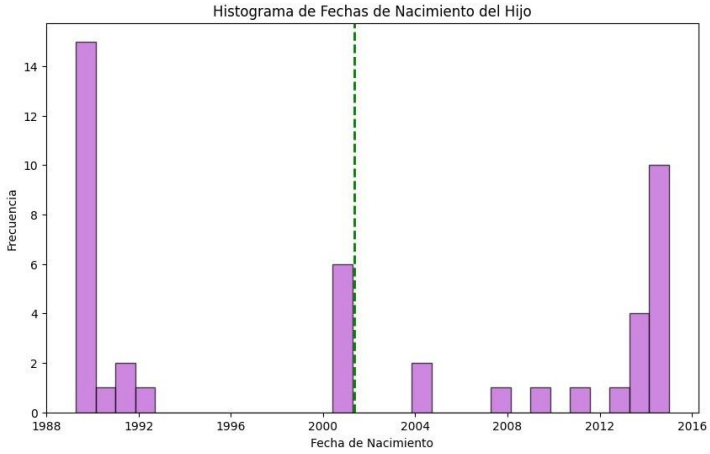
bornDateBaby1

Es la fecha de nacimiento del primer bebé. Como se aprecia el histograma obtenido de las fechas de nacimiento del *bornDateBaby1*, la frecuencia por

fecha de nacimiento es muy superior en los años que van desde el 2017 a 2024:



Al estar la escala en millones no se aprecia bien la frecuencia del histograma anterior, por lo que se procede a hacer un zoom para ver todos los valores entre los años 1688 y 2016:



En un segundo preprocesado, se borran todas aquellas fechas que están fuera del rango seleccionado, y nos quedamos solo con un registro, que es el siguiente:

	trackID	userID	baby1ID	baby2ID	city	genre	bornDateBaby1	premature	bornHeight	babyTandemGenre	...	creationDate	momAge	registered_num_sons	suplBreastmilk_c
448156	479175	136007	86724	NO_TANDEM	Barcelona	F	1996-06-02	0	47.0	NaN	...	2019-10-01 18:53:49	57.0	1	

1 rows × 34 columns

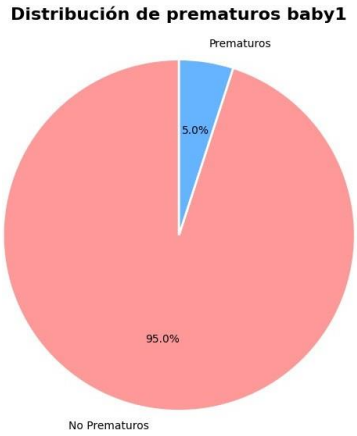
Durante esta segunda etapa de procesamiento, se plantea una metodología para corregir las fechas de nacimiento incorrectas de los bebés, comparándolas con la fecha de StartDate, que marca el inicio de la toma o registro. Para implementar esto, se organiza la información por el par madre-

hijo y se identifica la fecha mínima de StartDate para cada bebé, que corresponde a la primera toma registrada. Esta fecha se compara con la fecha de nacimiento del bebé, considerando un intervalo máximo aceptable de dos años.

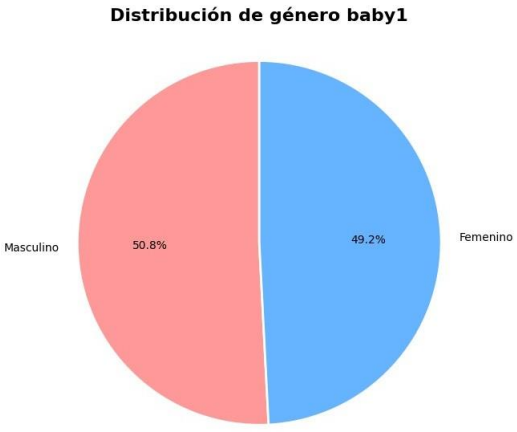
Además, se ha decidido separar los datos entre los casos de Tandem y no Tandem, con el objetivo de mantener más precisión en la información y sacar conclusiones en la siguiente sección.

Premature

Los bebes prematuros representan un 5% del total de bebés:



De los cuales la mayoría son masculinos, un total de 50.8%, frente a 46.2%:



gestationWeek

Al analizar la distribución de los valores para la semana de gestación mediante un recuento, se observa una notable cantidad de registros con un valor de 0. Considerando que las otras semanas registradas varían de la 22 a la 36, y que el umbral para un bebé considerado no prematuro es de 37

semanas, se hipotetiza que el valor 0 podría representar a bebés de 37 semanas o más. Sin embargo, esta suposición no está completamente verificada.

Contrariamente a esta hipótesis, si realmente el valor 0 en la semana de gestación indicara que el bebé no es prematuro, todos los registros donde premature = 0 deberían mostrar una gestationWeek de 0. Sin embargo, este no es el caso, ya que existen 8,366

registros donde la gestationWeek es 0 pero premature es 1, indicando inconsistencias que necesitan revisión adicional.

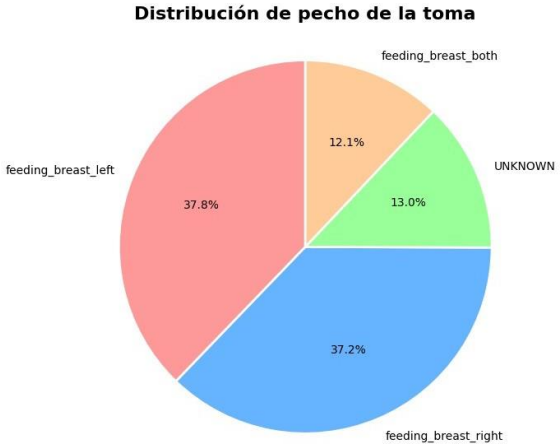
```
gestationWeek
0.0     3788477
36.0     134416
35.0      87208
34.0      30185
33.0      28542
30.0      14697
20.0      13588
28.0       4963
32.0       4947
31.0       3432
27.0       2896
25.0       2577
29.0       1776
26.0        597
23.0        548
21.0         29
24.0         19
22.0          3
Name: count, dtype: int64
```

Adicionalmente, se realiza una verificación de las semanas de gestación mínimas y máximas para los registros categorizados tanto como premature = 1 como premature = 0. Sorprendentemente, para ambas categorías, las semanas de gestación registradas oscilan entre un mínimo de 21 y un máximo de 26 semanas. Este hallazgo sugiere un error en los datos o una posible malinterpretación de los valores ingresados, ya que estos rangos no son coherentes con las definiciones estándar de prematuridad ni de gestación completa.

bornHeight

En esta variable hay muchos valores incorrectos:

- valores que incluyen el string "cm" al lado del número
- valores que tienen una coma en vez de un punto para separar los cm de los mm



Se obtiene para la variable *bornHeight* que:

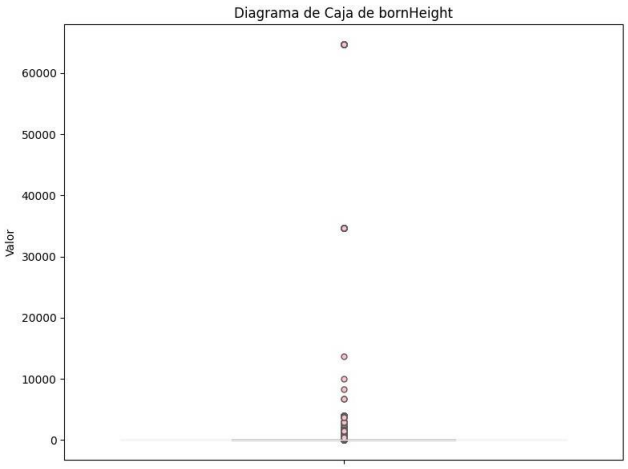
- el valor mínimo es 0.0 cm
- el valor máximo es 8303.21 cm
- la media 45.64087113577482 cm
- la mediana 47.0 cm
- la desviación estándar es 35.37238362761846 cm

Como se puede observar, el mínimo y máximo son valores que no tienen sentido, por lo que se saca un rango de bebés cuya altura se considere errónea por estar debajo de los límites considerados "normales" para recién nacidos.

En el caso de bebés prematuros, hay 16628 que tienen una altura por debajo de 25cm y 22601 que tienen una altura superior a 45cm.

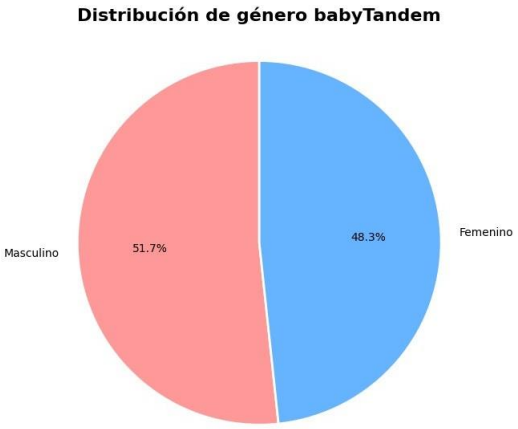
En el caso de bebés no prematuros, hay 262112 bebés con una altura por debajo de 40cm y 2587 bebés con una altura superior a 60cm.

En total, el número de registros que se considera están en un rango fuera de los intervalos normales es de 8.51%. Tras dibujar el boxplot de esta variable, se observa que hay muchos valores atípicos en el extremo superior del intervalo:


Diagrama de Caja de bornHeight
Valor
60000
50000
40000
30000
20000
10000
0

babyTandemGenre

Para la variable género del segundo bebe, aquel con el que se hace tándem (**babyTandemGenre)**, se observa que, de nuevo, hay una distribución balanceada de clases femenina y masculina como cabría esperar.

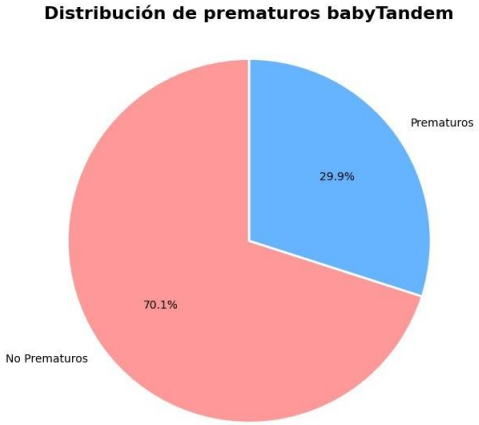


babyTandemBornDate

En la evaluación de los datos, se observa que la fecha de nacimiento de los bebés registrados en tándem (*babyTandemBornDate*) presenta una mejora significativa en términos de precisión en comparación con la fecha de nacimiento del primer bebé. El rango de estas fechas de nacimiento abarca desde el año 2013 hasta el año 2023, lo que indica una correcta consistencia y precisión en el registro de las fechas para los bebés en tándem.

babyTandemPremature

Esta variable es booleana para dividir entre los bebés del tándem (es decir el segundo) que son o no prematuros. El porcentaje de bebés prematuros en la variable babyTandem es del 26.6%:



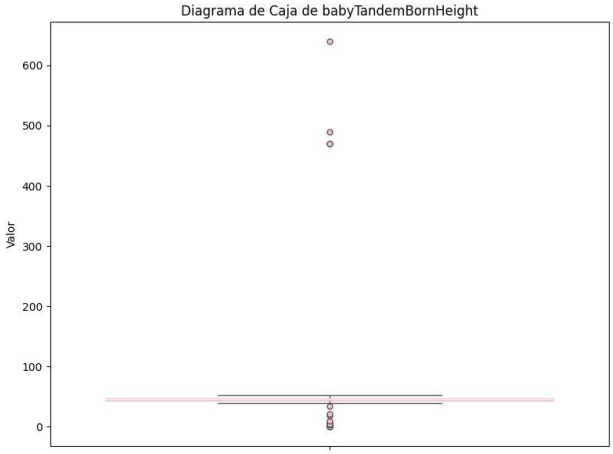
babyTandemGestationWeek

Esta variable se borra ya que los datos que contiene están anonimizados.

babyTandemBornHeight

En esta variable, que representa la longitud en cm del bebé con el que se hace tándem, se realiza el mismo procesado que con el *BornBabyHeight*, se eliminan aquellos valores que por ser demasiado pequeños (inferiores a 25cm) o demasiado grandes (superiores a 60cm) no tienen sentido.

En este caso, al haber muchos valores a 0, el número de bebés fuera de los intervalos es 6833, lo cual representa un porcentaje del total de registros que están fuera de los intervalos de 63.35%.



Durante la segunda fase de procesamiento de datos, al aplicar un filtro sobre la altura de los bebés en tándem, se constató que sólo quedan 13 registros válidos. Dado que este número es considerablemente bajo, se tomó la decisión de separar estos registros en un

dataframe independiente para diferenciar claramente entre los bebés en tándem y los que no están en tándem.

StartDate

La variable StartDate refleja el inicio de la toma o registro. Se ha identificado que únicamente 11 registros presentan un valor nulo para esta variable. Al comparar el *StartDate* con el *EndDate* y el *CreationDate*, se ha observado una correlación entre estas fechas, lo que necesitará una revisión más detallada de las temporalidades en las siguientes variables.

	startDate	endDate	creationDate
0	2018-03-27 21:42:00	2018-03-27 21:42:00	2018-03-27 18:42:45
1	2018-03-28 00:00:00	2018-03-28 00:27:00	2018-03-27 21:18:30
2	2018-03-28 00:25:00	NaN	2018-03-27 21:27:02
3	2018-03-28 00:15:00	NaN	2018-03-27 21:27:53
4	2018-03-28 00:15:00	2018-03-28 00:15:00	2018-03-27 22:22:08

endDate

La variable que registra el fin de la toma presenta un número significativo de valores nulos (85.872), lo cual es común dado que muchas usuarias solo anotan el inicio de la toma y omiten el final. Para abordar esta omisión, se ha identificado que muchas usuarias especifican el fin de la toma en la sección de comentarios. Para incorporar estos detalles en la columna *endDate*, se han utilizado expresiones regulares (RegEx) para detectar una lista de palabras clave pertinentes.

Se ha establecido que la duración máxima de una toma, asumiendo que no haya errores de registro, es de dos horas. Sin embargo, solo el 5.46% de los registros superan este intervalo, siendo la duración media de una toma de 45 minutos.

Durante el segundo procesamiento de los datos, se encontraron 3.780 registros donde la fecha de *endDate* es anterior a la de *startDate*, indicando un error.

En estos casos, los registros han sido eliminados. Una alternativa observada fue que en muchas instancias el *endDate* está simplemente intercambiado con el *startDate* debido a un error de registro, lo cual sugeriría la posibilidad de invertir estos valores. No obstante, tal ajuste podría llevar a interpretaciones erróneas y aleatorias de los datos, lo que no se considera adecuado sin una verificación más rigurosa, por lo que esta corrección no se ha llevado a cabo.

	startDate	endDate
155	2018-04-14 04:47:44	2018-04-14 04:35:58
163	2018-04-14 20:26:16	2018-04-14 20:22:23
170	2018-04-15 15:58:55	2018-04-15 15:39:19
177	2018-04-16 04:41:59	2018-04-16 02:50:17
185	2018-04-16 21:26:55	2018-04-16 21:19:05
...	...	...
230091	2023-11-26 03:04:20	2023-11-25 10:49:20
230534	2023-11-27 07:45:12	2023-11-27 07:14:12
230575	2023-11-27 11:23:33	2023-11-27 10:15:33
230667	2023-11-27 19:23:46	2023-11-27 18:55:46
230732	2023-11-28 02:30:53	2023-11-28 02:29:53

3768 rows × 2 columns

Durante la segunda fase de procesamiento de datos, se eliminaron todos los valores nulos. Posteriormente, se intentó imputar los valores nulos de la variable *endDate* utilizando el texto registrado en las notas, conforme se mencionó anteriormente. Sin embargo, este enfoque no resultó viable dado que no se encontraron registros en las notas que especificaran el *endDate*.

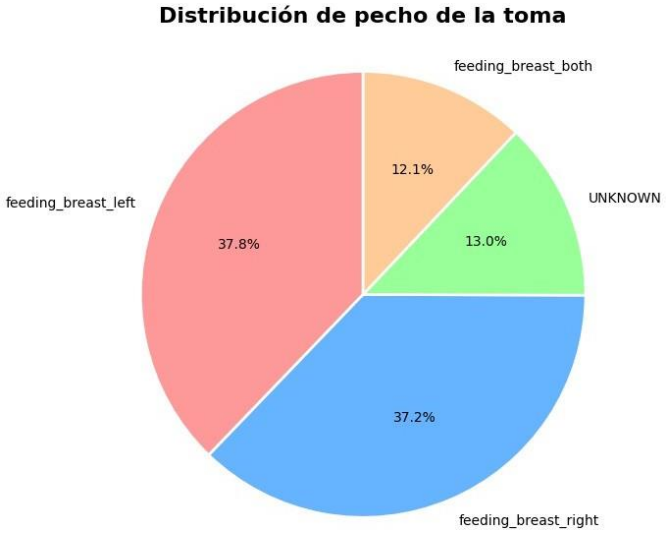
	endDate	notes
5836	NaN	Muy lleno el pecho
64696	NaN	??
64758	NaN	?? 40ml
65326	NaN	Chupon 3 oz
65466	NaN	Chupon
65638	NaN	Chupon
65815	NaN	Chupon
65928	NaN	Chupon
75885	NaN	Ambos pechos
151526	NaN	Muy poco de ambas
151547	NaN	Muy poco
151617	NaN	Poco
151720	NaN	Muy poco
151733	NaN	Poco (90)
151814	NaN	Muy poco
151993	NaN	80 de cada
152183	NaN	80

Finalmente se comprueba si hay registros con *startDate* posteriores al *endDate* y se confirma que hay 3768.

breast

La variable *breast* es una variable categórica que hace referencia a si el pecho es el derecho, el izquierdo o ambos.

- Pecho derecho representa un 37.2%
- Pecho izquierdo representa un 37.8%
- Ambos pechos representan un 12.1%
- Hay un 13% de datos que son “unkown”

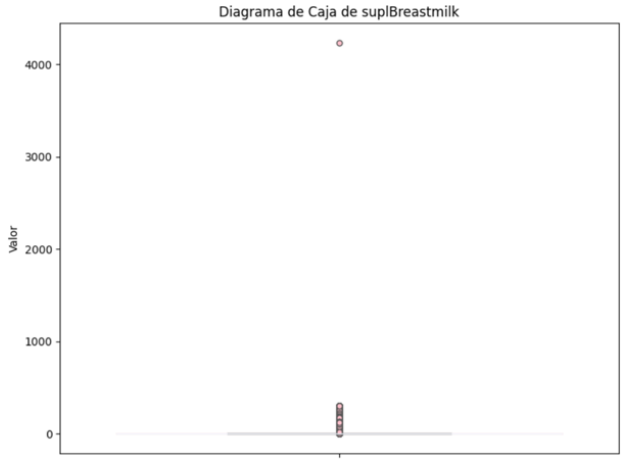


suplBreastMilk

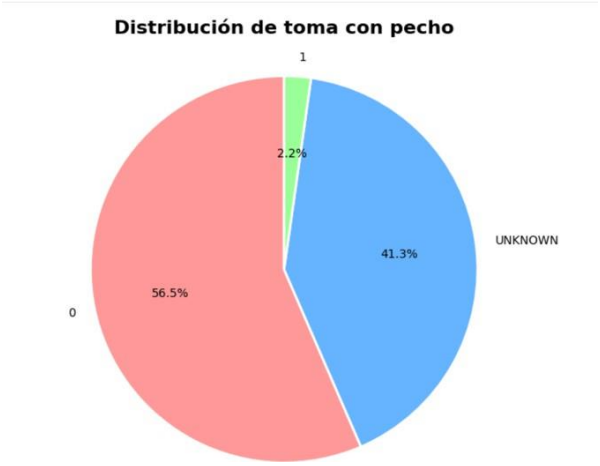
Variable numérica que indica la cantidad de leche de pecho que se da. Coincide que aquellos registros (madres) que no dan fórmula (suplBreastMilk = 0), son aquellos que dan el pecho, como se esperaba.

Consideramos los registros con valores de 'breastmilk' entre 5 y 150 ml, registros válidos. Tenemos un total de 4025261. El porcentaje total dentro de la variable de registros en el rango considerado no erróneo es 58.56%.

Del diagrama de caja observamos una dispersión baja de valores dentro de lo que es el boxplot y bastantes atípicos del lado superior (derecho).



Esta variable se convierte en categórica, dando como salida si da leche de pecho o no, para facilitar su posterior estudio.



En un segundo preprocesado se procede a borrar los valores desconocidos de la variable
suplBreastMilk.

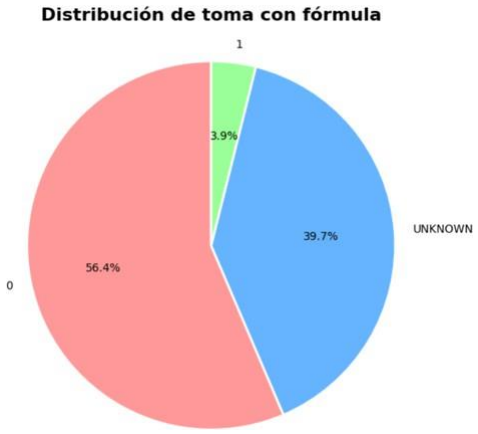
suplFormula

Variable numérica que representa la cantidad de leche suplementada con fórmula. Primero se analiza los valores que son 0 y los unknown:

- Cantidad de veces que *suplbreastmil*k es igual a 0.0: 3882568
- Cantidad de veces que *suplFormula* es diferente de 0.0 o de 'UNKNOWN': 2724660.

Se observa que los registros con valores dentro del rango considerado normal (entre 5 y 150 ml) son 4078635 lo cual supone un porcentaje de registros dentro del rango considerado no erróneo del 56.37%.

Esta variable también se convierte a categórica y se observa que no hay registros que tengan un 0 tanto para la toma dada con leche de pecho como con leche de fórmula. Lo que sí ocurre es que nos encontramos registros que sean *unknown* en alguna de ambas variables y 0 en la otra, lo cual se podría interpretar como que en aquella en la que hay un *unknown* realmente es la que se está dando pecho, es decir ese *unkown* sería un 1.


Distribución de toma con fórmula
1
3.9%
39.7%
UNKNOWN
56.4%
0

En un segundo preprocesado se procede a borrar los valores desconocidos de la variable
suplFormula.

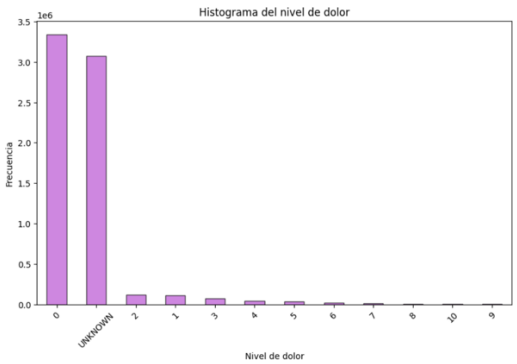
painLevel

Para la variable ***painLevel***, que varía entre 0 y 10 en función del nivel de dolor, se observa a partir del histograma una frecuencia muy superior en niveles de dolor de 0 y unknown.

Para un nivel de dolor 0 hay, 3340807 mujeres en total, siendo la clase mayoritaria.

```
painLevel
0          3340807
UNKNOWN    3077360
2           119480
1           109807
3            74185
4            46319
5            39378
6            24930
7            16818
8             9679
10            7781
9             4213
Name: count, dtype: int64
```

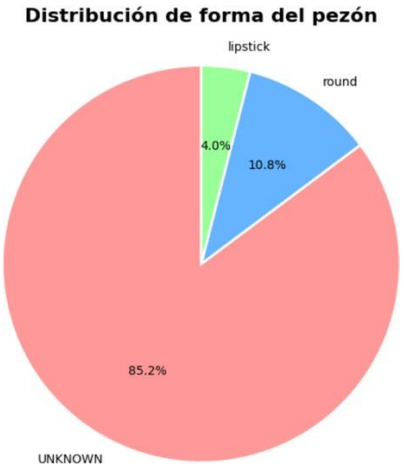
A continuación se representa el histograma de dolor:



nipleShape

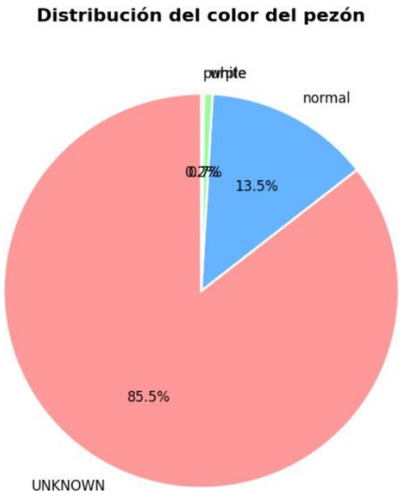
Esta variable es categórica, con tres posibles valores de salida: pezón

redondo(round), lipstick o desconocido. Como se observa a partir del gráfico de tarta, el valor *unkown* es el más frecuente, con un 85.2% de la distribución total.



nipleColor

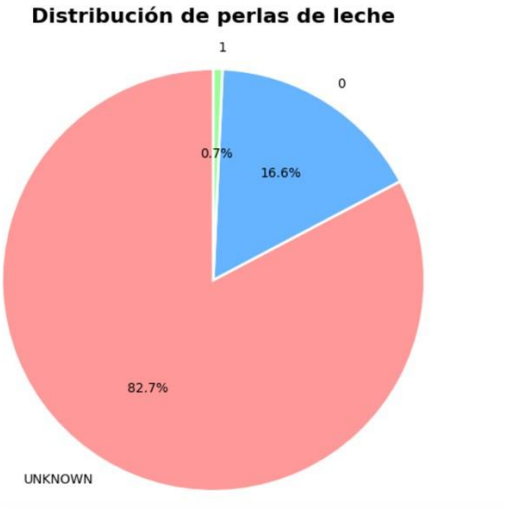
La variable *nipplecolor* es también una variable categórica, que toma 4 potenciales valores: normal, purple, white y unkown. Como se aprecia en el pie chart, los valores unkown son los mayoritarios, con un 85.5% del total:



milkPearl

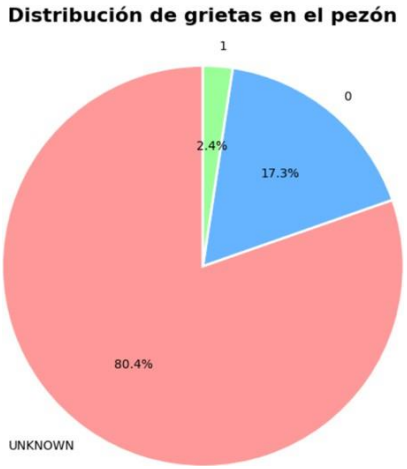
La variable milkPearl es una categórica que varía entre 0 y 1, teniendo también muchos valores unknown (82.7%). Hace referencia a si las madres tienen

coágulos, también conocidos como “perlas de leche”.



Cracks

La variable cracks (grietas en el pezón) es una categórica que varía entre 0 y 1, habiendo también muchos valores unknown, pero está ligeramente más registrada que la variable de coágulos de leche por las usuarias.



emotionalStatus

En la variable *emotionalStatus* tenemos una lista de categorías de estado emocional para cada madre donde cada una de ellas tiene una o más

```
feeding_emotional_status_calm - feeding_emotional_status_happy - feeding_emotional_status_exhausted - feeding_emotional_status_frustrated
feeding_emotional_status_calm - feeding_emotional_status_exhausted - feeding_emotional_status_happy - feeding_emotional_status_frustrated - feeding_emotional_status_sad
feeding_emotional_status_happy - feeding_emotional_status_exhausted - feeding_emotional_status_frustrated
feeding_emotional_status_calm - feeding_emotional_status_happy - feeding_emotional_status_sad - feeding_emotional_status_exhausted - feeding_emotional_status_frustrated
```

emociones, como se aprecia en la foto:

Durante el procesado, se divide la columna en cada una de las emociones y se pasa con un *one hot encoding* a categórica 1-0:

feeding_emotional_status_calm_emotionalStatus	feeding_emotional_status_exhausted_emotionalStatus	feeding_emotional_status_frustrated_emotionalStatus	feeding_emotional_status_happy_emotionalStatus
0	0	0	0
0	1	0	0
0	1	0	0
0	1	0	0
0	0	0	0
0	0	0	1
0	0	0	1
0	0	0	1
0	0	0	1
0	0	0	1
0	0	0	1
0	0	0	1
0	0	0	1
0	0	0	0
0	1	0	0

Se borra la columna original que tiene el array con todas las emociones y además los valores unkown.

Notes

En la variable notas, la madre incluye su valoración de aquello que considera: desde sentimiento de cómo se encuentra, tiempo que ha durado la toma, si el bebé está tomando algún tipo de medicina, etc.

A partir de esta columna, que contiene información muy valiosa de diverso tipo, se intenta sacar más material para otras variables. Por ejemplo, muchas madres incluyen el horario de la toma en notas pero cada una lo anota de una manera diferente, por lo que se intentan buscar patrones e incluir dichos datos en las columnas correspondientes.

	notes
25	Té mocs\n\nquieta\n\nFa gorça en les mans enrera per a agafarse
26	Estoy triste por si mi bebé está empezando a tener problemas mamando por la confusión al darle suplementos que me mandaron con biberón
36	Con mucho sueño \n
41	Toma breve porque se despierta. Come algo y sigue durmiendo.
47	Mala noche \n
54	Aunque extraño cuando no pide. Me agobia un poco cuando hay mas demanda\n
66	La toma terminó a las 22:22\n\n\n
68	Termina toma 23:30 h. \n
79	el pecho izquierdo gotea bastante
80	dolor sólo al cogerse
86	Rot sin leche
93	retuerce el pezón dentro de la boca
101	Vomito\n
102	Erupto sin vomito
104	A veces mama tranquilo, otras con los ojos como platos, otras parece que se esta peleando con el pecho.\nA veces 20min otras 3min\nMama generalmente en cada toma de una sola teta, cada vez una, a veces de las dos. Le gusta mas la izquierda y es que yo noto que la derecha tiene menos leche, no se si eso se podra.\nA demanda, pero no dejo mas de 4h sin comer.\nRegurgita mucho.\nA veces si le insisto mucho o hago dos tomas mas seguidas me vomita leche.
130	mucho cólico
147	Comió inquieta, no quiso biberón
148	Inquieta, parece tener hambre pero al ofrecerle biberón no lo quiere
161	Despues de dormir 5h30 seguidas! Excepcional
168	se queda dormido en pecho.
173	Sin erupto y con vomito aigualit
210	me ardeun poco el pecho
218	Lloraba al acabar y lo he sacado del pecho y se ha calmado
221	Le he quitado pecho
222	Le he quitado pecho
232	3 onzas

En un segundo preprocesado se borran todos los valores nulos de la variable notes.

creationDate

Esta variable representa la fecha de registro en la app. Es una variable que se intuye que es de tipo *log* y todos sus valores son aparentemente correctos, siendo:

- La fecha mínima: '2018-03-27 18:42:45'

- La fecha máxima: '2023-11-28 06:18:02'

modificationDate

Es la fecha de modificación del track. Al tratarse de una variable que no aporta información y además sus valores coinciden con *creationDate*, se ha procedido a borrarla.

city

La variable city incluye la ubicación geográfica, pero su registro presenta una gran variabilidad debido a las diversas formas en que las madres la han reportado, lo que ha representado un desafío significativo durante el preprocesamiento. Para abordar esto, se intentó utilizar una API de geolocalización, pero los resultados no fueron satisfactorios debido a la diversidad en los datos, que incluyen referencias a pueblos, ciudades y datos incorrectos. Como medida inicial, se eliminaron los datos erróneos, dejando el análisis completo de la variable city como última parte.

Tras la depuración de datos erróneos en otras variables, el número de registros válidos en la variable "city" se redujo significativamente, resultando en 44.611 registros. Con la ayuda de ChatGPT, se extrajo el país correspondiente a cada ciudad registrada y se añadió una nueva columna, “country”. Finalmente, se siguiendo la lógica de dividir la información en dos conjuntos de datos independientes, uno para casos de tándem y otro para no tándem.

En la variable creada "country", el total de registros disminuyó a 43.132, debido a la exclusión de valores considerados erróneos, concretamente aquellos que tenían en el registro los siguientes strings: ["zw sx f", "sjl", "z", "nan", "24 de mayo"].

2. ANÁLISIS DE RESULTADOS

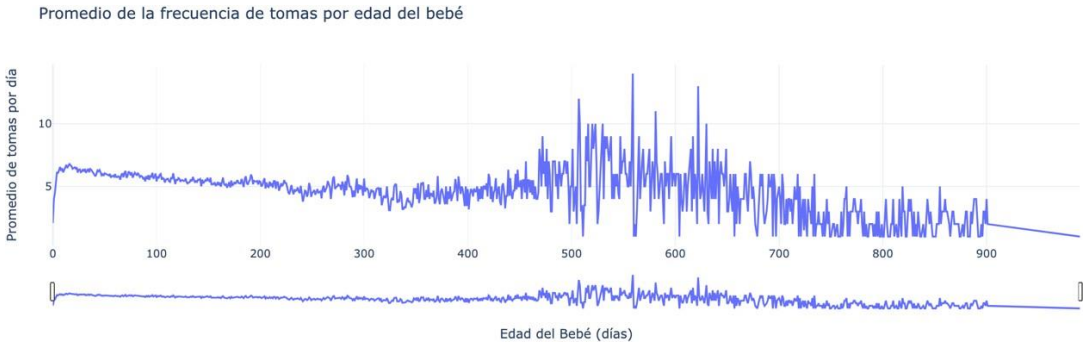
2.1. Gráficas de frecuencias

Promedio de la frecuencia de tomas por edad del bebé

Se comienza graficando los datos obtenidos, en concreto, el primer gráfico que se saca es el promedio de la frecuencia de tomas por edad del bebe.

En dicho gráfico, se observa una tendencia clara y marcada hacia una disminución en la frecuencia de tomas a medida que aumenta la edad del bebé. Observándose que:

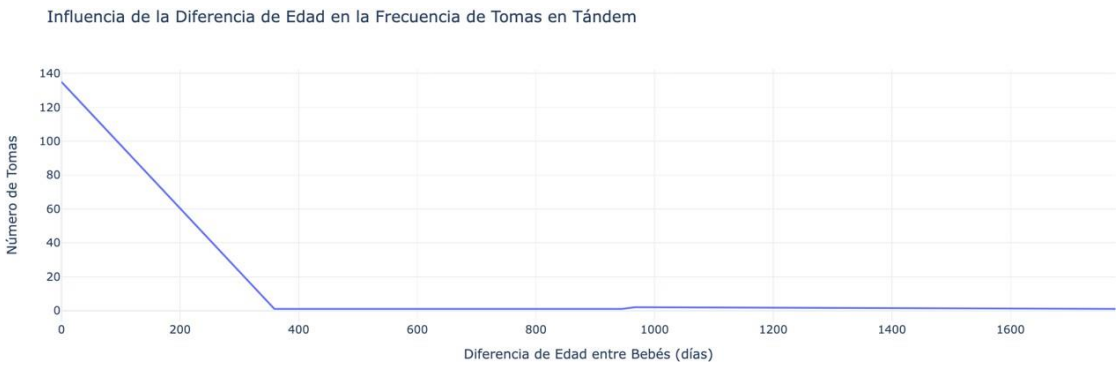
- Durante los primeros días de vida, la frecuencia de las tomas es notablemente alta, lo que es consistente con las necesidades nutricionales intensas de los recién nacidos y la frecuencia alimentaria recomendada.
- A medida que el gráfico avanza, particularmente hasta aproximadamente los 300 días de edad, se aprecia que la variabilidad en la cantidad de tomas por día es relativamente baja, lo que sugiere un patrón de alimentación más uniforme. Este período podría correlacionarse con una fase de lactancia exclusiva seguida por la introducción gradual de alimentos complementarios, manteniendo aún una dependencia significativa de la lactancia.
- Posteriormente, entre los 300 y los 700 días, la variabilidad de la frecuencia de tomas por día comienza a aumentar. Este aumento puede atribuirse a diferencias individuales en la duración de la lactancia extendida y en la introducción de dietas más variadas y sólidas en los niños a medida que crecen. Factores como las decisiones personales de las madres, consejos pediátricos, y posiblemente la influencia de regresar al trabajo , podrían explicar la mayor dispersión observada en este segmento del gráfico.
- Finalmente, a partir de los 700 días de edad del bebé, se observa que la variabilidad en la frecuencia de tomas disminuye nuevamente. Esto podría indicar una normalización en los patrones de alimentación a medida que los niños se acercan y superan los dos años, con una dieta cada vez más dominada por alimentos sólidos y una disminución gradual en la lactancia materna como principal fuente de nutrición.


Promedio de la frecuencia de tomas por edad del bebé
Promedio de tomas por día
10
5
0
100
200
300
400
500
600
700
800
900
Edad del Bebé (días)

Influencia de la diferencia de edad en la frecuencia de las tomas en tándem

Para visualizar la relación entre las edades de dos bebés que están siendo alimentados en tándem y cómo sus edades relativas impactan en la frecuencia de las tomas, se usa un enfoque que refleja las diferencia en sus edades en el momento de cada toma. El siguiente gráfico muestra cómo dicha diferencia afecta a la cantidad de tomas que se realizan en tándem. Se observa:

- **Alta frecuencia en diferencias de edad pequeñas:** cuando los bebés tienen edades muy similares, es posible que ambos tengan necesidades nutricionales comparables debido a que se encuentran en etapas de desarrollo similares. Esto podría facilitar una mayor frecuencia de tomas en tándem.
- **Disminución con mayores diferencias de edad:** a medida que la diferencia de edad se amplía, es probable que los bebés estén en etapas de desarrollo muy distintas. Por ejemplo, un bebé podría estar comenzando con alimentos sólidos mientras que el otro todavía depende completamente de la leche materna. Esto puede hacer que la coordinación de las tomas en tándem sea menos frecuente.
- **Estabilización de tomas:** la estabilización del número de tomas en un nivel bajo con diferencias de edad mayores puede deberse a un límite en la conveniencia o necesidad de realizar tomas en tándem cuando uno de los bebés ha superado significativamente la etapa de lactancia intensiva, mientras que el otro aún podría estar en ella.



Promedio de la frecuencia de tomas por edad del primer bebé

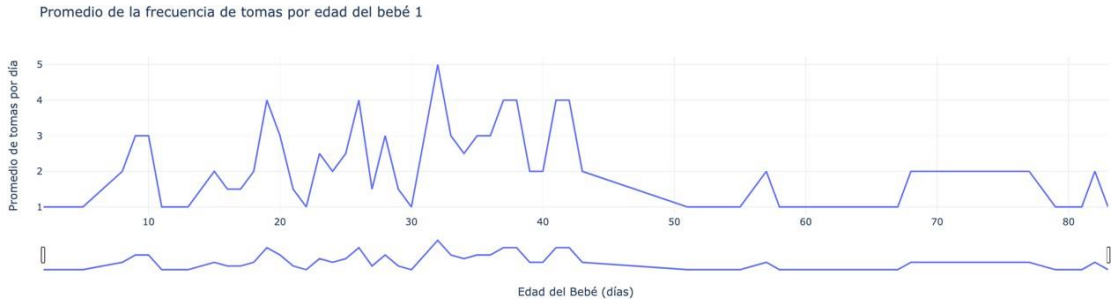
Se analiza también el promedio de la frecuencia de tomas por edad del primer bebé, de donde se observa:

Fluctuaciones iniciales: aparecen una serie de fluctuaciones en la frecuencia de tomas durante los primeros 50 días de vida del bebé. Estas variaciones podrían atribuirse a la adaptación inicial tanto de la madre como del bebé a la lactancia, así como a la respuesta del bebé a su crecimiento inicial y aumento de las necesidades nutricionales.

- **Estabilización y variabilidad:** entre los días 10 y 50, se observa una tendencia general hacia una estabilización con picos ocasionales, lo que indica que mientras algunos bebés pueden establecer un patrón regular de alimentación, otros

pueden experimentar variaciones debido a factores como el crecimiento, la salud general y el comportamiento de alimentación.

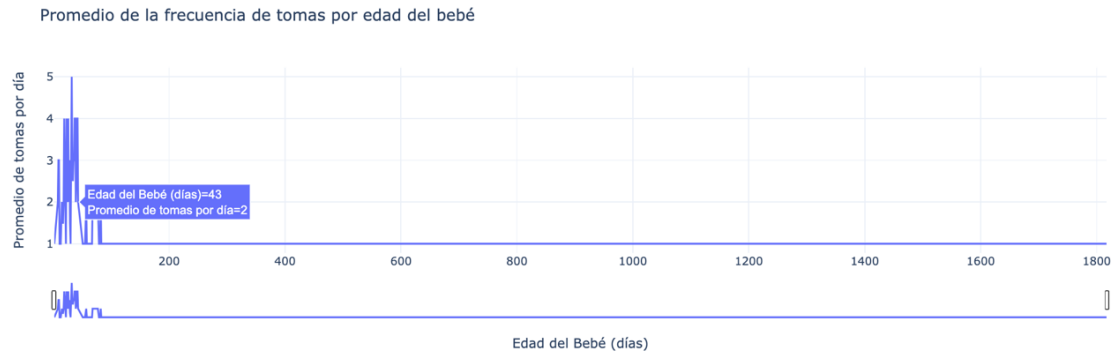
- **Descenso notable:** aproximadamente a partir del día 50, la gráfica muestra un descenso en la frecuencia de las tomas, alcanzando un mínimo alrededor del día 60. Esta disminución podría estar relacionada con la introducción de alimentos sólidos o una mayor capacidad del bebé, lo que permite intervalos más largos entre las tomas.
- **Estabilidad posterior con fluctuaciones menores:** después del descenso, la frecuencia de tomas se estabiliza en un nivel más bajo con algunas fluctuaciones menores hacia el final del período observado. Este patrón puede indicar la adaptación a un régimen de alimentación más regular y menos frecuente a medida que el bebé crece.



Promedio de tomas por edad del segundo bebe (tándem)

En esta gráfica se observa un comportamiento diferente debido a un menor número de datos con:

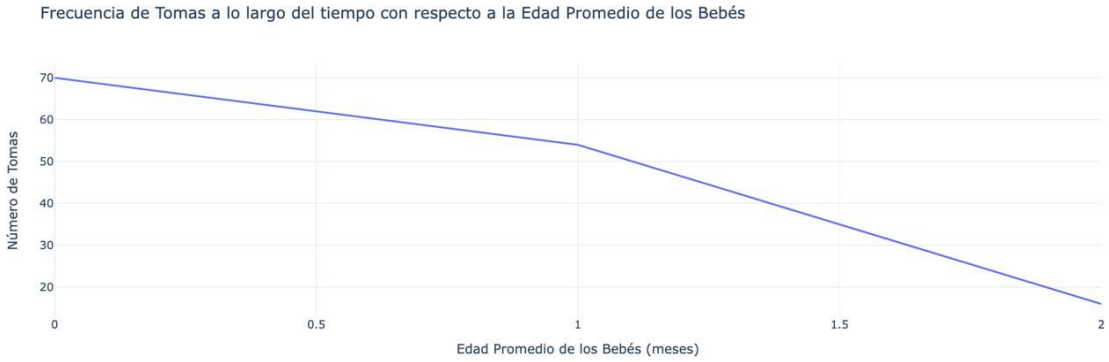
- **Altos picos iniciales:** observamos que la gráfica comienza con una serie de picos relativamente altos en las primeras etapas, dentro de los primeros 200 días aproximadamente. Estos picos indican una mayor frecuencia de alimentación, lo cual es común en recién nacidos y bebés jóvenes que requieren alimentación más frecuente debido a estómagos pequeños y altas demandas de crecimiento.
- **Disminución marcada:** después de los primeros 200 días, la gráfica muestra una disminución considerable en la frecuencia promedio de tomas por día, estabilizándose en un nivel más bajo. Este cambio puede estar relacionado con el desarrollo del bebé, que incluye la introducción de alimentos sólidos y un menor requerimiento de tomas de leche exclusivamente.
- **Estabilidad con mínimas fluctuaciones:** más allá de los 400 días, el gráfico se estabiliza notablemente con algunas fluctuaciones menores. Esto puede indicar que los patrones de alimentación se han estabilizado a medida que el bebé crece y se adapta a un régimen de alimentación que incluye tanto sólidos como, posiblemente, menos tomas de leche.



Frecuencia de Tomas a lo largo del tiempo

Como se observa en la gráfica, hay una tendencia decreciente continua en el número de tomas a medida que los bebés crecen. Inicialmente, la pendiente es relativamente suave, lo que muestra una disminución gradual en la frecuencia de tomas desde el nacimiento hasta el primer año de vida.

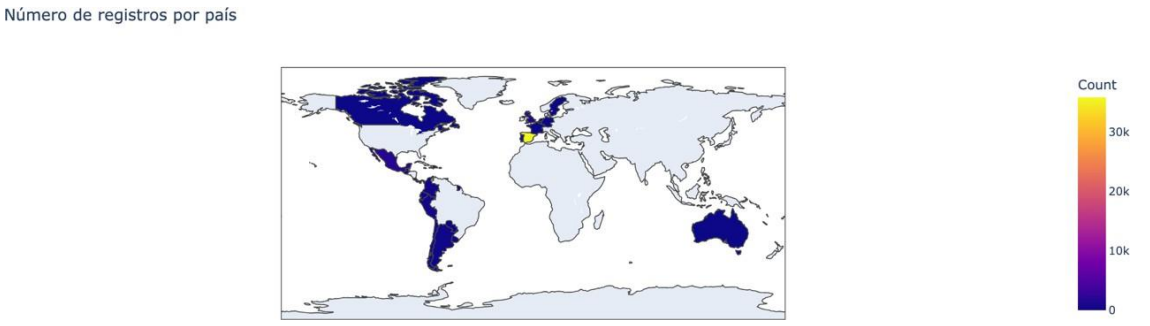
Alrededor de los 12 meses (1 año), se observa un cambio notable en la pendiente, que se vuelve más pronunciada indicando que la disminución en el número de tomas se acelera a partir del primer año. Esta fase podría coincidir con cambios significativos en la dieta del bebé, como la introducción más completa de alimentos sólidos y una reducción en la dependencia de la leche materna o fórmula.



2.2. Gráficas de Países

Número de Países

Se analiza por medio de mapas coropléticos de Python el número de registros que se obtiene por país y, como es esperado, en España el volumen de registros es el más alto. Se obtienen registros para México, USA, Colombia, Costa Rica, Panamá, Perú, Ecuador, Francia, Alemania, UK, Noruega, Países Bajos y Australia.



Cantidad de tomas por usuario en cada país

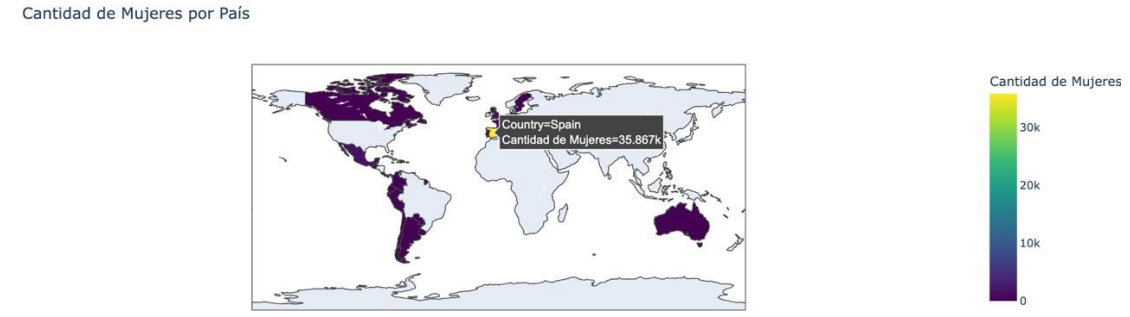
Posteriormente, se ha realizado un análisis sobre la cantidad media de tomas por usuario en cada país. A partir del mapa generado, se observa que el Reino Unido presenta la mayor frecuencia media de tomas, con un promedio de 1246 tomas por usuario. Este dato podría indicar una utilización más intensiva de la aplicación por parte de las usuarias en dicho país. Por otro lado, en España, donde se esperaría una frecuencia más alta debido al alto volumen de registros, la media de tomas por usuario es aproximadamente de 800. Este análisis nos ayuda a entender las diferencias en el uso de la aplicación entre países y puede ser indicativo de variaciones en las prácticas de lactancia o en la adopción de la tecnología en diferentes contextos culturales y socioeconómicos.



Número de mujeres registradas por país

En cuanto al número de mujeres registradas por país, se observa que España

lidera la lista con un total de 35.867 mujeres inscritas, lo cual es esperado dado que es el país en el que se fundó la aplicación y dada la popularidad de la aplicación en este mercado.



2.3. Relaciones con la Vuelta al Trabajo

Relación entre la Edad promedio del Bebé y Número Promedio de Tomas por Usuario añadiendo la vuelta al trabajo

El scatterplot generado muestra la relación entre la edad promedio del bebé (en meses) y el número promedio de tomas por usuario.

La mayoría de los puntos se concentran en la parte inferior izquierda del gráfico, indicando que la mayoría de las tomas ocurren cuando los bebés son más pequeños, especialmente en los primeros 5 meses de vida. Existe una disminución clara en el número promedio de tomas a medida que la edad del bebé aumenta, lo que es esperado ya que los bebés comienzan a introducir alimentos sólidos y dependen menos de la leche materna o fórmula.

La línea roja vertical en los 4 meses marca el punto de "Vuelta al Trabajo" en países como España. Se puede observar que hay una disminución en el número de tomas registradas justo después de este punto, aunque varias usuarias siguen registrando un número considerable de tomas. Algunos puntos notables se encuentran a la derecha de la línea, lo que indica que algunas usuarias continúan con un número relativamente alto de tomas incluso después de la vuelta al trabajo, lo que puede asociarse con más facilidades de teletrabajo, lactancia y conciliación. Sin embargo, la noción de que las tomas son más frecuentes en los primeros meses de vida del bebé y tienden a

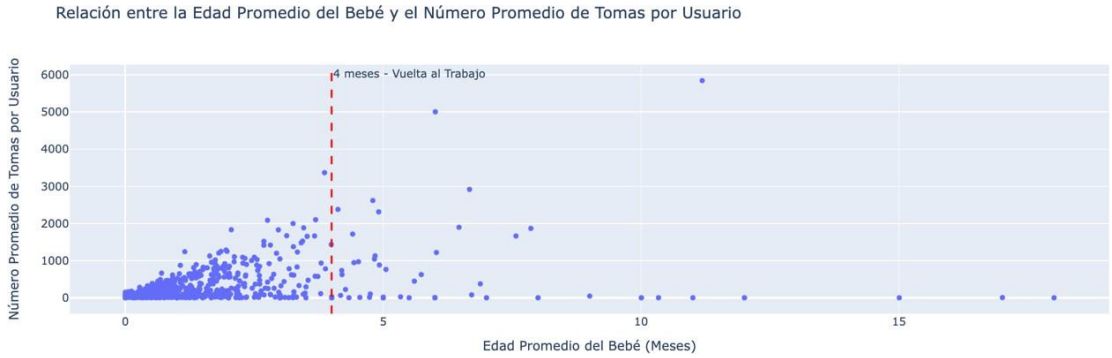
disminuir conforme el bebé crece, es clara.

Aunque hay una tendencia general a la disminución, la "Vuelta al Trabajo" no parece ser un punto de corte drástico para todas las mamás, lo que sugiere variabilidad en cómo los diferentes usuarios manejan la lactancia o alimentación después de volver al trabajo.

Es posible que algunos usuarios tengan la capacidad de continuar con tomas frecuentes a través de la extracción de leche materna o el ajuste de horarios, mientras que otros pueden necesitar reducir la cantidad de tomas.

Variabilidad entre Usuarios:

La dispersión de puntos a lo largo de la gráfica, especialmente con usuarias que siguen teniendo un alto número de tomas después de los 4 meses, indica diferencias en las prácticas de lactancia, apoyo disponible, y posiblemente en las políticas laborales que afectan a los usuarios de manera diversa.



Relación entre la edad Promedio del Bebé y Número promedio de tomas por País añadiendo la vuelta al trabajo

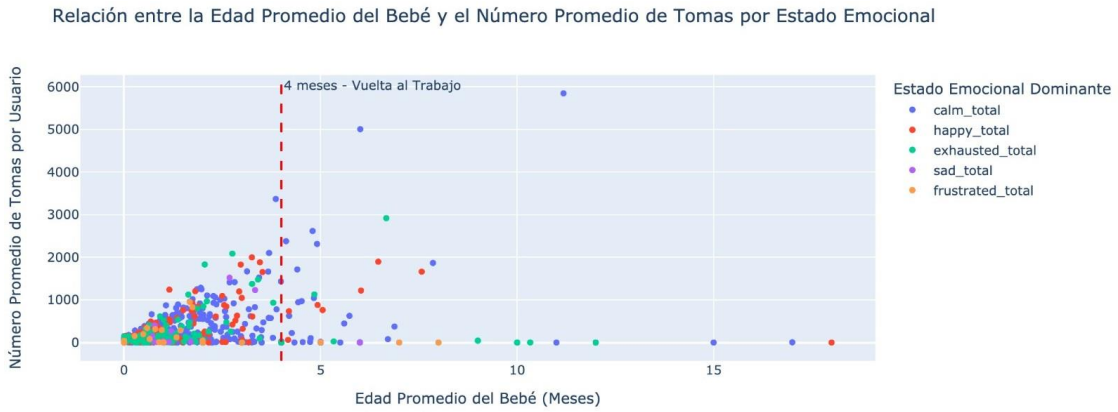
Al representar el mismo gráfico, pero con un matiz distinto para cada país, se observa claramente que España (representada en azul) tiene el promedio más alto de tomas, (como cabe esperar ya que es el país donde se lanza inicialmente la app).



Relación entre la edad Promedio del Bebé y Número promedio de tomas por estado emocional añadiendo la vuelta al trabajo

Al representar los datos con diferentes colores según el estado emocional

dominante, se observa que al inicio predomina claramente el estado "*exhausted*" (en verde). En etapas posteriores, se aprecia una tendencia hacia el estado "*happy*" (en rojo). Para bebés con una edad promedio superior a los 5 meses, solo se observan unos pocos puntos con colores "atípicos".



2.4. Gráficas de emociones

Estado emocional

Para analizar cómo es el estado emocional de la madre se analiza con un gráfico de plotly una *time series* que representa el número de tomas según la edad del bebé (días) y el estado emocional de la madre. De las gráficas se pueden sacar varias conclusiones:

- La mayoría de las tomas se concentran en los primeros 100 días de vida del bebé. Después de este periodo, la cantidad de tomas disminuye significativamente, independientemente del estado emocional reportado por la madre.

- El estado emocional de "calm" (calma) es predominante en cuanto al número de tomas, especialmente en los primeros 50 días. Esto podría sugerir que cuando las madres están calmadas, es más probable que realicen más tomas.

- Los estados "exhausted" (agotada) y "happy" (contenta) muestran un pico inicial similar pero disminuyen más rápidamente que el estado "calm". Los estados "frustated" (frustada) y "sad" (triste) tienen un perfil más bajo en comparación con los otros estados, lo que sugiere que estas emociones son menos reportadas o tienen menos impacto en el número de tomas.

- Claramente, el estado "exhausted" muestra una disminución bastante rápida después de los primeros días, lo que podría reflejar una adaptación de las madres al cuidado del bebé o una reducción de la percepción de agotamiento conforme se normaliza la rutina de cuidados.

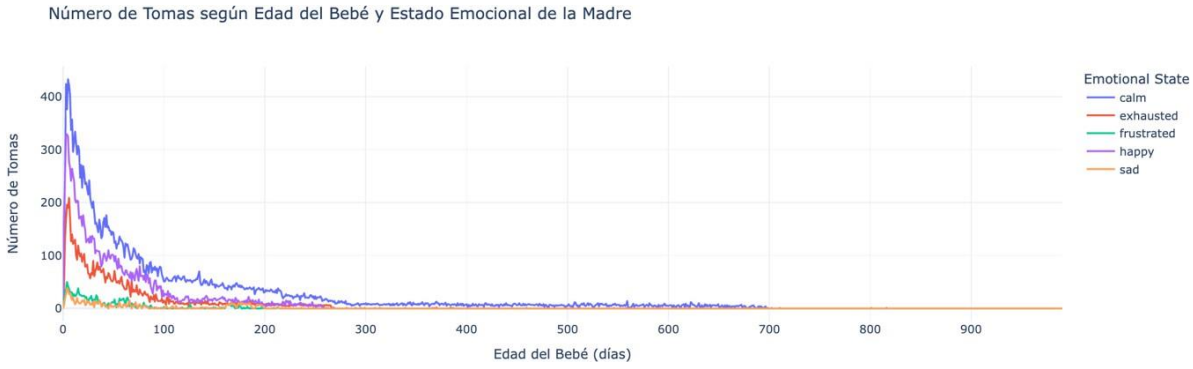
Observando que más allá de los 700 días desde el nacimiento del bebé solo se

reportan los estados emocionales de "calm" (calma), "sad" (tristeza) y "exhausted" (agotamiento), podemos derivar conclusiones adicionales y considerar su relevancia en el contexto de la crianza prolongada y la lactancia extendida:

- Persistencia de la Lactancia: la continuidad de la lactancia más allá de los 700 días es menos común, pero la presencia de registros sugiere que algunas madres continúan amamantando a sus hijos durante este período extendido. Esto es notable dado que supera las recomendaciones típicas de duración de la lactancia que suelen centrarse en los primeros dos años.

Estados emocionales predominantes en etapas avanzadas:

- La presencia exclusiva de los estados "calm", "sad", y "exhausted" podría indicar patrones emocionales específicos asociados con la lactancia de mayor duración. Que "calm" sea uno de los estados reportados sugiere que las madres que continúan con la lactancia por un período prolongado podrían experimentar esta actividad como un momento de calma y conexión con sus hijos, a pesar de la posible fatiga o desafíos emocionales.
- La presencia de "sad" y "exhausted" podría reflejar los desafíos emocionales y físicos que implica la lactancia extendida. Estas emociones pueden estar relacionadas con la demanda constante de cuidado y atención que requiere un niño de más de dos años, además de gestionar otras responsabilidades y aspectos de la vida personal y profesional.



Análisis de sentimientos negativos/tristes: exhausted, frustrated y sad

En esta gráfica de barras apiladas se muestran los diferentes estados de las madres para los estados emocionales *exhaustas*, *frustradas* y *tristes* teniendo en cuenta diferentes variables:

Para aquellas madres que no tienen grietas (grietas = 0), tenemos aproximadamente que:

- 76% de las madres están exhaustas
- 15% de las madres están frustradas

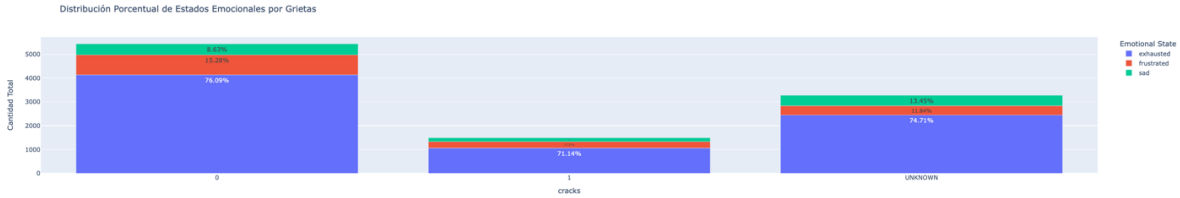
- 8% de las madres están tristes

Para aquellas madres que si tienen grietas (grietas = 1), tenemos aproximadamente que:

- 71% de las madres están exhaustas
- 17% de las madres están frustradas
- 11% de las madres están tristes

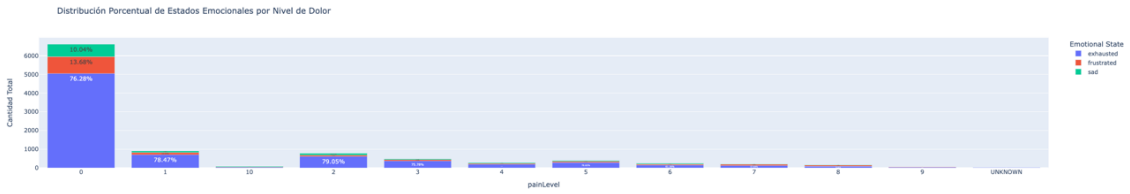
Para aquellas madres donde no se sabe si hay grietas, tenemos aproximadamente que:

- 74% de las madres están exhaustas
- 11% de las madres están frustradas
- 13% de las madres están tristes



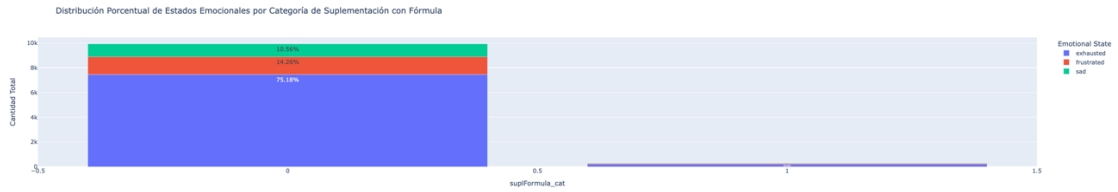
De estos datos podemos concluir que las madres que tienen grietas parecen estar más frustradas y tristes, frente a aquellas donde aparecen las grietas, que tienen un nivel de frustración y tristeza menor. En todos los grupos el sentimiento predominante es exhaustas.

Al hacer una gráfica de barras apiladas para ver la distribución de estados emocionales en función de cada nivel de dolor, se observa que el estado más predominante en todos los niveles es el de ***exhaustas***, que aumenta ligeramente (en puntos porcentuales) a medida que el nivel de dolor avanza:



También se analiza la distribución porcentual de estados emocionales por categoría de suplementación con fórmula y, aunque hay bastantes más registros sin fórmula (*suplFormula 0*), se observa que en aquellas madres que si dan fórmula, el estado emocional es 10 puntos porcentuales más bajo en ***exhaustas***, pero parecen estar más *frustradas*. Esto se puede relacionar con que, de manera natural no puedan dar el pecho o no tengan

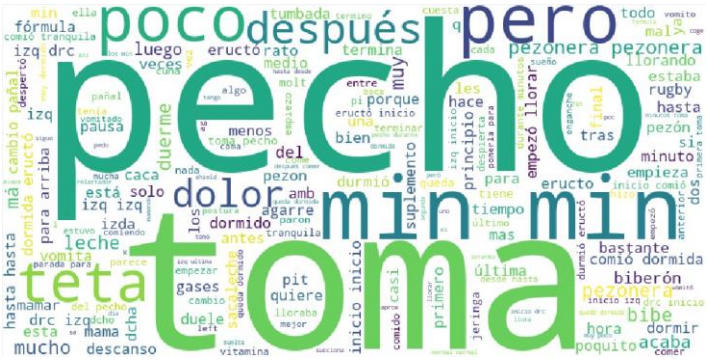
suficiente leche, tengan que suplementar y esto les haga estar frustradas.



2.5. Análisis de las Notas

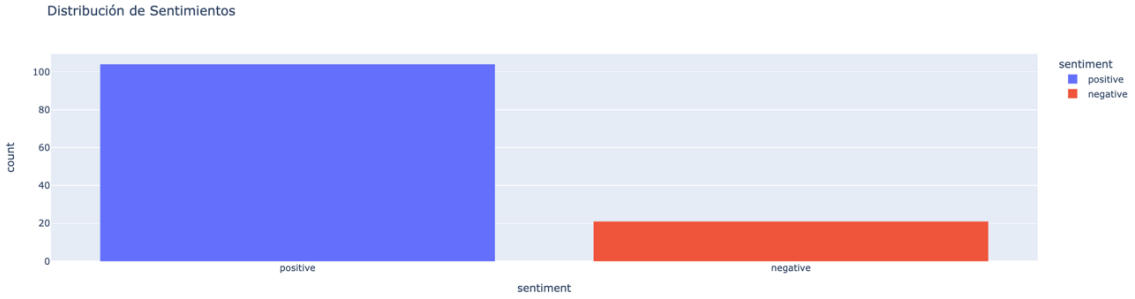
Word Cloud

A partir de las notas, se obtiene la nube de palabras con más frecuencias. Se visualiza que palabras como *pecho*, *toma* o *min* son las más abundantes.



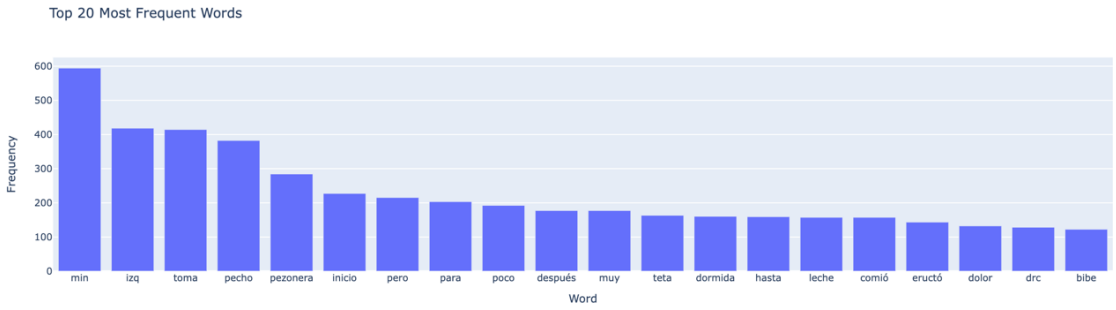
Análisis de sentimientos

A partir de las notas, se observa una predominancia de sentimientos positivos, representados en azul vs. los rojos.



Histograma de palabras más frecuentes

En el histograma presentado, se enumeran las 20 palabras más frecuentes en orden de importancia. Entre ellas, la palabra "min" destaca notablemente, lo cual se atribuye a su uso frecuente en las notas donde las madres anotan la duración de las tomas de lactancia, expresada en minutos, así como detalles adicionales sobre si la alimentación se realizó desde el pecho izquierdo o derecho.

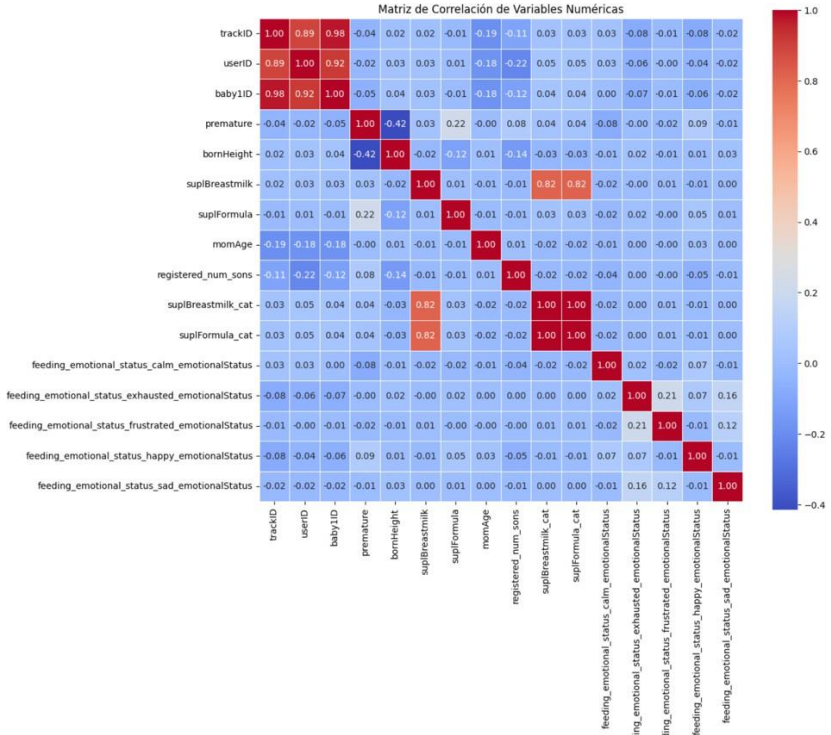


Correlación de variables numéricas

Se ha generado un histograma de correlación entre las variables numéricas para estudiar las posibles colinealidades entre ellas. La hipótesis inicial planteaba que variables como la leche suplementada y la leche materna (de pecho) estarían negativamente correlacionadas, es decir, que un aumento en una implicaría una disminución en la otra. Sin embargo, los resultados de la matriz de correlación no respaldan esta hipótesis.

Las variables que sí parece que presentan una correlación significativa son:

- Las variables categóricas creadas con su respectiva variable numérica (como cabría esperar)
- La altura del bebé con su condición de prematuridad, mostrando una correlación ligera de 0.42.


Matriz de Correlación de Variables Numéricas
trackID
userID
baby1ID
premature
bornHeight
suplBreastmilk
suplFormula
momAge
registered_num_sons
suplBreastmilk_cat
suplFormula_cat
feeding_emotional_status_calm_emotionalStatus
feeding_emotional_status_exhausted_emotionalStatus
feeding_emotional_status_frustrated_emotionalStatus
feeding_emotional_status_happy_emotionalStatus
feeding_emotional_status_sad_emotionalStatus
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

3. CONCLUSIONES Y PRÓXIMOS PASOS

Como conclusiones, cabe destacar el carácter fundamentalmente exploratorio del trabajo, con un enfoque significativo en la limpieza y el procesamiento de datos. Este proceso ha sido esencial para asegurar la calidad y la fiabilidad de los datos, dado que se trata de información registrada directamente por las madres en la app. Enfrentarnos a datos erróneos o mal registrados ha sido un desafío recurrente, lo que ha requerido un esfuerzo considerable en la limpieza y validación de dichos datos para poder llevar a cabo el análisis.

La investigación incluye diversos aspectos del comportamiento de la lactancia y el estado emocional de las madres, desde la evolución de la frecuencia de las tomas a medida que el bebé crece hasta el impacto de la vuelta al trabajo y el análisis emocional a través de las notas registradas. La exploración geográfica mediante mapas ha permitido visualizar diferencias por país en los patrones de lactancia, mientras que el análisis de la correlación de variables numéricas ha asegurado que todas las variables incluidas en el estudio aportan información relevante y no hay colinealidades.

Uno de los hallazgos más interesantes ha sido la identificación de patrones emocionales predominantes y su variación a lo largo del tiempo, así como la influencia de factores como el dolor asociado a grietas en el pezón y la suplementación con fórmula en el estado emocional de las madres. Se ha visto cómo, para aquellas madres con grietas, el sentimiento predominante es la frustración y en aquellas en las que la tristeza aparece nada más dar a luz (depresión postparto), el sentimiento es mantenido a medida que el bebé crece. Estos resultados proporcionan una comprensión más profunda de los desafíos y experiencias emocionales a

los que se enfrentan las madres durante la lactancia.

Este análisis también ha proporcionado una base sólida para implementar mejoras en la recogida de datos en futuros pasos. Se han identificado áreas clave donde la precisión y la consistencia de los datos pueden mejorarse, y se plantea como potencial mejora el rellenado de datos en la app mediante desplegables para evitar valores nulos o erróneos. Esta medida, así como la clarificación de cada variable a insertar, facilitará análisis más precisos y útiles en el futuro.

Un siguiente paso crucial a partir de estos hallazgos será desarrollar un modelo predictivo que permita anticipar el abandono de la lactancia por parte de las madres. Este modelo podría utilizarse para diseñar intervenciones más efectivas y personalizadas, apoyando a las madres en su decisión de continuar con la lactancia y mejorando así los resultados tanto para ellas como para sus bebés.

REFERENCIAS

TABLAS

Tabla 1: Evolución Market Share. Fuente: Propia 50
Tabla 2: Proyectos por mes en los primeros 5 años. Fuente: Propia.....51
Tabla 3: Coste por servicio y proveedor. Fuente: Propia 56
Tabla 4: Coste por servicio normalizado. Fuente: Propia 57
Tabla 5: Beneficios anuales proyectados. Fuente: propia 60
Tabla 6: Porcentaje de MS y Número de partners por año de proyecto. Fuente: Propia 61
Tabla 7: Personal contratado sin tiempo parcial. Fuente: Propia.....96
Tabla 8: Personal contratado con tiempo parcial. Fuente: Propia.....96
Tabla 9: Ponderaciones de factores de localización. Fuente: Propia 97
Tabla 10: Locales elegidos. Fuente: propia.....99
Tabla 11: Propiedades de los locales normalizadas. Fuente: Propia.....99
Tabla 12: Matriz ponderación para selección de local. Fuente: Propia 99
Tabla 13: Tabla de riesgos. Fuente: Propia 102
Tabla 14: Probabilidad e Importancia de los Riesgos 103
Tabla 15: Matriz de Riesgos (impacto y probabilidad).....104
Tabla 16: Distribución del salario de desarrollo de negocio. Fuente: Propia.....124
Tabla 20: Plan de negocio - hipótesis 1. Fuente: Propia 134
Tabla 18: Distribución alternativa de ventas año 1. Fuente: Propia.....134
Tabla 19: Ajuste de personal para estudio del plan de negocio. Fuente: Propia.....134
Tabla 20: Mobiliario necesario para cada empleado por área. Fuente: Propia 135
Tabla 21: Cálculo del CAPEX. Fuente: Propia.....136
Tabla 22: Cuenta de pérdidas y ganancias tras revisión de ventas año 1. Fuente: propia.....5
Tabla 23: Pagos de interés por el préstamo bancario. Fuente: Propia.....4
Tabla 24: Resultado del ejercicio. Fuente: Propia.....3

ILUSTRACIONES

Ilustración 1:mapa de stakeholders. Fuente: propia 18
Ilustración 2: Business Model Canvas. Fuente: propia 24
Ilustración 3: Análisis PESTEL. Fuente: propia 33
Ilustración 4: Análisis DAFO. Fuente: Propia 47
Ilustración 5: Número de empresas en el sector sanitario 2014-2023. Fuente: propia. Datos: INE 48
Ilustración 6: TAM, SAM, SOM. Fuente: propia 51
Ilustración 7: Crecimiento de Empresas sector sanitario. Fuente: Propia 52
Ilustración 8: Número de empresas partner/proyectos por año. Fuente: Propia 52
Ilustración 9: Objetivos del proyecto. Fuente: Propia 71
Ilustración 10: Diagrama de Gantt del proyecto. Fuente: Propia 72
Ilustración 11: Relación entre forma de pezón y nivel de dolor. Fuente: Propia 79
Ilustración 12: Relación entre color de pezón y nivel de dolor. Fuente: Propia 80
Ilustración 13: Relación entre existencia de perlas de lactancia y nivel de dolor. Fuente: Propia 80
Ilustración 14: Relación entre existencia de grietas en pezón y nivel de dolor. Fuente: Propia ..81
Ilustración 15:Proporción de uso de suplemento de alimentación y color del pezón. Fuente: propia 83
Ilustración 16: Proporción de uso de suplemento de alimentación y existencia de perlas de lactancia. Fuente: propia 83
Ilustración 17. Regresión lineal aplicada a la media del nivel de dolor por la edad del bebé. Fuente: Propia 86
Ilustración 18: Logo Empresa. Fuente: Propia 113
Ilustración 19: Organigrama de la empresa. Fuente: Propia 120
Ilustración 20: Proyección de beneficios (desglose mensual año 1). Fuente: Propia.....3
Ilustración 21: Proyección de beneficios. Fuente: Propia 4

BIBLIOGRAFÍA

(ASPA), A. S. (15 de diciembre de 2023). *hhs*. Obtenido de https://www.hhs.gov/about/news/2023/12/15/hhs-to-improve-maternal-health-outcomes-with-new-cms-care-model-that-expands-access-to-services-other-proven-maternal-health-approaches.html

aliance, C. s. (s.f.). *Cloud security aliance*. Obtenido de https://cloudsecurityalliance.org/research/cloud-controls-matrix

Álvarez, A. (2021). *Nuevos desafíos, nuevos paradigmas en la salud infantil. Revista Indexia. Recuperado de* . Obtenido de https://revistaindexia.com/2021/04/08/nuevos-desafios-nuevos-paradigmas-en-la-salud-infantil/

Arranz, E. (6 de julio de 2022). *Fundacion Adecco*. Obtenido de https://fundacionadecco.org/notas-de-prensa/el-envejecimiento-se-dispara-en-espana-en-2022-hasta-el-1335-se-contabilizan-133-personas-mayores-de-64-anos-por-cada-100-menores-de-16/

avcontentteam. (6 de febrero de 2024). *analytics vidhya*. Obtenido de https://www.analyticsvidhya.com/blog/2023/04/machine-learning-for-businesses/

Bankinter, D. d. (21 de 06 de 2024). *Bankinter*. Obtenido de https://www.bankinter.com/blog/economia/previsiones-ipc-espana

BOE. (5 de diciembre de 2018). *Ley Orgánica 3/2018, de 5 de diciembre, de Protección de Datos Personales y garantía de los derechos digitales.* Obtenido de https://www.boe.es/buscar/doc.php?id=BOE-A-2018-16673

BROOKE AUXIER, M. A. (28 de JULY de 2020). *PEW RESEARCH CENTER* . Obtenido de https://www.pewresearch.org/internet/2020/07/28/parents-attitudes-and-experiences-related-to-digital-technology/

clinic, c. (25 de 01 de 2024). *cleveland clinic*.

cloud, G. (s.f.). *Google cloud and GDPR*. Obtenido de https://cloud.google.com/privacy/gdpr?hl=es-419

cloud, g. (s.f.). *google cloud compliance iso 27001*. Obtenido de https://cloud.google.com/security/compliance/iso-27001?hl=es-419

Comission, E. (s.f.). *commission.europa.eu*. Obtenido de https://commission.europa.eu/law/law-topic/data-protection/data-protection-eu_en

Company, M. &. (01 de agosto de 2023). *QuantumBlack*. Obtenido de https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-AIs-breakout-year

compliance, g. (s.f.). *google protection compliance*. Obtenido de https://business.safety.google/intl/es/compliance/

Díaz, R. (s.f.). *The Machine Learners*. Obtenido de https://www.themachinelearners.com/metricas-de-clasificacion/

Digital, M. (12 de agosto de 2020). *McKinsey Digital*. Obtenido de https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/how-six-companies-are-using-technology-and-data-to-transform-themselves

dobuss, A. (10 de 05 de 2024). *tecnooficinas*. Obtenido de https://tecno-oficinas.com/blog/post/metros-por-persona-oficina

Ergosistema. (s.f.). *Ergosistema*. Obtenido de https://ergosistema.com/mejores-sillas-ergonomicas/

España, A. M. (30 de 09 de 2022). *Garrigues*. Obtenido de https://www.garrigues.com/es_ES/noticia/publicada-boe-reforma-reduce-3000-euros-1-euro-capital-minimo-constituir-sociedad

Glassdoor. (1 de julio de 2024). *Glassdoor*. Obtenido de https://www.glassdoor.es/Sueldos/madrid-recruiter-sueldo-SRCH_IL.0,6_IM1030_KO7,16.htm

Glassdoor. (4 de julio de 2025). *Glassdoor*. Obtenido de https://www.glassdoor.es/Sueldos/madrid-consultant-sueldo-SRCH_IL.0,6_IM1030_KO7,17.htm

Glassdoor. (s.f.). *Glassdoor*. Obtenido de https://www.glassdoor.es/Sueldos/index.htm

Google. (s.f.). *Google cloud*. Obtenido de https://cloud.google.com/sustainability/region-carbon?_gl=1*vft943*_up*MQ..&gclid=1c8e64053c511c35b7670283f170e268&gclsrc=3p.ds#region-picker

Health, T. N. (2021-2022). *data resource center for child and adolecent health.* Obtenido de https://www.childhealthdata.org/learn-about-the-nsch/NSCH

hinomi. (s.f.). *hinomi*. Obtenido de https://eu.hinomi.co/es/products/hinomi-h1-classic-ergonomic-office-chair

hosteltur. (3 de 1 de 2019). *hosteltur*. Obtenido de https://www.hosteltur.com/125780_nueve-de-cada-diez-grandes-empresas-estan-invirtiendo-en-machine-learning.html

IBM. (s.f.). *IBM*. Obtenido de https://www.ibm.com/es-es/watson

IEC, I. (2022). *ISO*. Obtenido de https://www.iso.org/standard/27001

INE. (s.f.). *INE*. Obtenido de https://ine.es/index.htm

JACKSON, J. (12 de november de 2023). *leantime*. Obtenido de https://leantime.io/mastering-business-model-a-guide-to-completing-the-lean-canvas-model/

Korobeyko, T. (25 de enero de 2023). *itransition*. Obtenido de https://www.itransition.com/machine-learning/statistics

Lactapp. (s.f.). *Lactapp medical*. Obtenido de https://lactapp.es/.

Learning, I. M. (s.f.). *IBM ML*. Obtenido de https://cloud.ibm.com/catalog/services/watson-machine-learning

media, w. (6 de octubre de 2023). *webcastle*. Obtenido de https://webcastletech.com/blog/the-ai-advantage-leveraging-machine-learning-for-competitive-edge/

Microsoft. (s.f.). *Microsoft Azure*. Obtenido de https://azure.microsoft.com/es-es/

Mónica Machicado Sotoca, D. D.-J. (abril de 2018). *zonahospitalaria.com*. Obtenido de https://zonahospitalaria.com/pezones-blancos-durante-la-lactancia-causas-sintomas-y-como-aliviarlo/

myAnalog. (26 de marzo de 2019). *myAnalog*. Obtenido de https://www.analog.com/en/resources/technical-articles/guidelines-for-the-optomechanical-integration-of-heartrate-monitors-in-wearable-wrist-devices.html

Na8. (4 de noviembre de 2017). *Aprende ML*. Obtenido de https://www.aprendemachinelearning.com/principales-algoritmos-usados-en-machine-learning/

Oracle. (s.f.). *Oracle Health*. Obtenido de https://www.oracle.com/health/

Philips. (s.f.). *Philips HealthSuite*. Obtenido de https://www.philips.es/healthcare/innovation/about-health-suite

RUIZ, E. (6 de diciembre de 2023). *mi bebé y yo*. Obtenido de https://mibebéyyo.elmundo.es/bebés/alimentacion/lactancia/problemas-pezon

Salary.com. (s.f.). *Salary.Com*. Obtenido de https://www.salary.com/

Santos, D. (03 de julio de 2024). *Hubspot*. Obtenido de https://blog.hubspot.es/marketing/crear-plan-proyecto#:~:text=Un%20plan%20de%20proyecto%20es%20una%20herramienta%20para,llevar%C3%A1n%20a%20concluir%20el%20proyecto%20de%20manera%20eficaz.

SAS. (s.f.). *SAS*. Obtenido de https://www.sas.com/es_es/home.html

Services, C. f. (13 de diciembre de 2022). *cms.gov*. Obtenido de https://www.cms.gov/newsroom/press-releases/readout-cms-hosts-maternal-health-convening-leaders-across-government-industry

STAFF, C. (27 de MARZO de 2024). *COURSERA*. Obtenido de https://www.coursera.org/articles/what-is-machine-learning

STATES, I. -U. (s.f.). *IBM*. Obtenido de https://www.ibm.com/topics/machine-learning

Studio, I. W. (s.f.). *IBM Watson Studio*. Obtenido de https://www.ibm.com/products/watson-studio/pricing

Tamayo, M. c. (21 de febrero de 2022). *El Economista*. Obtenido de https://revistas.eleconomista.es/capital-privado/2022/febrero/el-sector-salud-motor-del-crecimiento-en-2022-PA10459565

Tang, B. L. (16 de junio de 2024). *ML: A Potential Therapeutic Tool to Facilitate Neonatal Therapeutic Decision Making. Pediatr Drugs 26, 355–363 (2024).* Obtenido de https://doi.org/10.1007/s40272-024-00638-6

technavio. (octubre de 2023). *technavio*. Obtenido de https://www.technavio.com/report/healthcare-analytics-market-industry-analysis

Torres, I. (5 de may de 2022). *ive consultores*. Obtenido de https://iveconsultores.com/precio-hora-consultoria-estrategica/

UPND. (s.f.). *upnd.org*. Obtenido de https://www.undp.org/es/sustainable-development-goals

webfx. (s.f.). *webfx*. Obtenido de https://www.webfx.com/martech/pricing/big-data-consulting/

Parliment, European (06 de agosto de 2023). *European Parliment.* Obtenido de https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence