ESCUELA TECNICA SUPERIOR DE INGENIERÍA (ICAI)

MÁSTER EN BIG DATA, TECNOLOGIA Y ANALITCA AVANZADA

TRABAJO FIN DE MÁSTER

Análisis de Patrones y Predicción de Comportamiento de Usuarios a través de Técnicas de Agrupamiento

Autor: Patricio Avila Pérez-Grovas

Director: Carlos Morrás Ruiz-Falcó

Madrid

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

Análisis de Patrones y Predicción de Comportamiento de Usuarios a través de Técnicas de Agrupamiento

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

curso académico 2022/23 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.: Patricio Avila Perez-Grovas Fecha: 27/ 06/2023

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

Fdo.: Carlos Morrás Ruiz-Falcó Fecha: ……/ ……/ ……

ESCUELA TECNICA SUPERIOR DE INGENIERÍA (ICAI)

MÁSTER EN BIG DATA, TECNOLOGIA Y ANALITCA AVANZADA

TRABAJO FIN DE MÁSTER

Análisis de Patrones y Predicción de Comportamiento de Usuarios a través de Técnicas de Agrupamiento

Autor: Patricio Avila Pérez-Grovas

Director: Carlos Morrás Ruiz-Falcó

Madrid

Agradecimientos

Quisiera expresar mi más sincero agradecimiento a todos los que han estado conmigo durante este desafiante pero gratificante viaje que ha sido mi máster. A mi director de máster, a mis estimados profesores y compañeros, su guía y consejos han sido invaluables y han hecho una gran diferencia en mi aprendizaje y crecimiento.

A mis queridos padres, a mis hermanos y a Sofia, su apoyo incondicional ha sido mi ancla en los momentos más difíciles y su alegría, mi motivación para superar cada reto. Todo su amor y apoyo han sido una parte integral de este logro. De corazón, gracias.

Resumen

Este proyecto se centra en la exploración, análisis y extracción de información relevante de un conjunto de datos compuesto por la interacción cientos de usuarios anónimos de una empresa. El conjunto de datos incluye una amplia gama de información, incluyendo detalles de la comunicación del usuario a lo largo de un año, como la identificación del usuario y del mensaje, el momento exacto en que se envió y se recibió cada mensaje, entre otros datos relevantes.

El objetivo principal del proyecto es analizar toda la información relevante de este conjunto de datos, utilizando técnicas de análisis de datos avanzadas. Principalmente, nos enfocaremos en la aplicación de técnicas de agrupamiento para categorizar a los usuarios según su forma de uso.

Además, se realizará un análisis detallado de los datos para estimar y predecir los horarios de los usuarios basándonos en su comportamiento pasado. Este enfoque nos permitirá identificar patrones y tendencias en los hábitos de comunicación de los usuarios, lo que a su vez podría proporcionar información valiosa sobre sus preferencias y necesidades.

Este estudio no solo proporcionará una visión más profunda del comportamiento de los usuarios, sino que también podría ser una herramienta valiosa para la empresa en la toma de decisiones estratégicas y en la mejora de sus servicios. Al entender mejor a sus usuarios, la empresa puede diseñar e implementar soluciones más eficientes y efectivas para satisfacer sus necesidades.

Palabras clave:

Aprendizaje automático, agrupamiento, predicciones, productividad, comportamiento.

Abstract

This project focuses on the exploration, analysis, and extraction of relevant information from a dataset composed of the interaction between hundreds of anonymous users from a company. The dataset includes a wide range of information, including user communication details over a year, such as user and message identification, the exact moment each message was sent and received, among other relevant data.

The main goal of the project is to acquire insights information from this dataset, using advanced and sophisticated data analysis techniques. Mainly, we will focus on the application of clustering techniques to categorize users based on their usage.

In addition, a detailed analysis of the data will be performed to estimate and predict users' schedules based on their past behavior. This approach will allow us to identify patterns and trends in user communication habits, which in turn could provide valuable information about their preferences and needs.

This study will not only provide a deeper insight into user behavior but could also be a valuable tool for the company in making strategic decisions and improving its services. By better understanding their users, the company can design and implement more efficient and effective solutions to meet their needs.

Keywords:

Machine learning, clustering, predicting, productivity, behavior.

Índice de la memoria

Índice de ilustraciones

Índice de tablas

1.- INTRODUCCIÓN

1.1 Contexto

En la era digital actual, donde los datos son una fuente invaluable de información y conocimiento, la capacidad para analizar y entender estos datos se ha convertido en una habilidad esencial. Las empresas generan y recolectan enormes cantidades cada día, desde transacciones comerciales, registros de clientes, hasta interacciones en redes sociales y patrones de comportamiento del usuario. Entre estos, los datos proporcionados por la interacción entre los usuarios son particularmente ricos en información y nos pueden ayudar a profundizar sobre sus preferencias, comportamientos y necesidades.

Un área intrigante de estudio es el análisis de los patrones de uso de los usuarios basado en los datos de comunicación. La comunicación es una parte integral de las actividades diarias de los usuarios y puede revelar mucho sobre sus hábitos, preferencias y comportamientos. Al analizar estos datos, es posible agrupar a los usuarios según su actividad, predecir sus horarios y entender mejor cómo interactúan con la empresa.

Sin embargo, el análisis de estos datos puede ser un desafío debido a su complejidad.
Requiere técnicas sofisticadas de análisis de datos, incluyendo técnicas de agrupación y análisis temporal, así como una comprensión profunda de los patrones de comportamiento del usuario.

1.2 Declaración del problema

A pesar de la abundancia de datos de la comunicación entre los usuarios, las empresas a menudo encuentran dificultades para extraer información útil y accionable de estos datos. Dado el volumen, la velocidad y la variedad, se necesitan enfoques y herramientas sofisticados para analizar e interpretar estos datos de manera eficaz. Además, la naturaleza dinámica y cambiante de los patrones de comportamiento de los usuarios añade otra capa de complejidad al análisis.

En particular, hay dos desafíos principales que este proyecto pretende abordar. Primero, cómo agrupar o segmentar a los usuarios según sus patrones de uso y comportamiento. La segmentación de los usuarios es una tarea fundamental en muchas aplicaciones, desde la personalización de los servicios hasta la toma de decisiones estratégicas. Sin embargo, los métodos actuales de segmentación de usuarios a menudo no tienen en cuenta la evolución de los comportamientos de los usuarios a lo largo del tiempo.

Por lo que el segundo desafío es la estimación y predicción de los horarios de los usuarios basándose en el uso pasado. A pesar de la importancia de esta tarea, hay poca investigación

sobre cómo predecir los horarios de los usuarios de manera eficaz y precisa para casos como este. Estos desafíos señalan una brecha en la práctica de análisis de datos, y subrayan la necesidad de métodos más efectivos para analizar y entender los datos de comunicación de los usuarios.

1.3 Objetivos del estudio

El objetivo general, como se mencionó anteriormente, es lograr identificar los diferentes tipos de usuarios dentro de la compañía para posteriormente comprender sus patrones de uso y lograr predecir su futura actividad.

Para una explicación más detallada, los objetivos específicos son los siguientes:

1. **Segmentar a los usuarios en base a su comportamiento de uso**: Utilizando técnicas de agrupación, el estudio buscará identificar grupos de usuarios que exhiban patrones de comportamiento similares.

2. **Identificar las variables más significativas para cada grupo**: El estudio se esforzará por identificar las características o variables que son más distintivas o significativas para cada grupo de usuarios.

3. **Estimar y predecir los horarios de los usuarios**: Usando técnicas de predicción y machine learning, el estudio buscará estimar y predecir los horarios de actividad de los usuarios en base a su comportamiento pasado.

1.4 Hipótesis

Es importante plantear una declaración clara y concisa de los resultados que se esperan obtener; ya que ayuda a guiar el diseño de la herramienta o modelo y a decidir qué datos se deben usar y analizar. Finalmente, se definieron dos hipótesis; una para el problema de agrupación y otra para la predicción de actividad, las cuales son las siguientes:

H1: A pesar de los cambios en los patrones de uso y comportamiento en la interacción entre los usuarios a lo largo del tiempo; se logra una efectiva segmentación de estos en grupos distintos.

H2: Las variables relacionadas con el comportamiento de uso y comunicación son predictores significativos de la actividad de los usuarios, permitiendo estimar y predecir con precisión sus horarios de actividad.

Estas hipótesis formuladas, H1 y H2, actuarán como nuestra brújula a lo largo de este estudio, orientando nuestras técnicas de análisis y ayudándonos a abordar de manera efectiva el

problema planteado. En la medida en que estas hipótesis se prueben ciertas, nos permitirán desentrañar los patrones subyacentes en los datos y, por lo tanto, aportar soluciones más informadas y sólidas para los desafíos que enfrentamos.

1.5 Metodología

Al igual que con la hipótesis, es esencial definir una metodología para cualquier tipo de proyecto; ya que otorga una estructura y dirección que proporciona un marco claro para el trabajo estableciendo qué y cómo se hará. La metodología por seguir es la siguiente:

1. **Descripción del Conjunto de Datos**: Se creará un directorio de los diferentes conjuntos de datos donde podamos apreciar cada variable, el tipo de dato y lo que representa. Con esto se entiende mejor el contexto general del data set y es de gran ayuda para la selección del modelo a utilizar.
2. **Procesamiento de Datos**: En esta etapa se realizarán las operaciones necesarias con los datos para que el formato sea apto para los modelos que se utilizaran.
3. **Métodos de Segmentación de Usuarios**: Se utilizarán técnicas de machine learning no supervisado como K-means y clustering jerárquica para segmentar a los usuarios en grupos basados en sus patrones de uso y comportamiento. Este paso requeriría una evaluación cuidadosa de la cantidad adecuada de clusters y la validación de los mismos para asegurar que son significativos y útiles.
4. **Identificación de Variables Significativas**: Identificar qué variables son significativamente diferentes entre los grupos de usuarios. Esto ayuda a caracterizar los grupos y entender qué factores contribuyen a la pertenencia a un grupo.
5. **Predicción de Horarios de Actividad:** Usando métodos de machine learning supervisado como regresión y árboles de decisión; predecir los horarios de actividad de los usuarios basándote en sus características y patrones de uso.
6. **Análisis de resultados:** Se interpretarán los resultados obtenidos por los diferentes modelos, se comentarán los cambios que se hayan hecho a lo largo de la ejecución de los mismo y finalmente como podrían ayudar en un futuro.

1.6 Importancia del estudio

Este estudio es fundamental en varias dimensiones. En primer lugar, al profundizar en nuestra comprensión de los patrones de comportamiento de los empleados, podemos facilitar un entorno de trabajo más eficiente y productivo. Un conocimiento detallado de los clusters de empleados y su evolución temporal puede permitir un mejor diseño y adaptación de políticas internas, mejorando así el rendimiento del equipo y potencialmente aumentando la retención de los empleados.

En segundo lugar, este estudio tiene el potencial de impactar en la eficiencia operativa de la empresa. Al predecir los horarios de actividad de los usuarios, podemos proporcionar una

gestión más efectiva de los recursos, optimizando así los procesos internos y posiblemente reduciendo costos.

Finalmente, este estudio tiene relevancia más allá de su aplicación inmediata. Al proponer y evaluar una metodología para analizar y predecir el comportamiento del usuario, estamos contribuyendo al crecimiento del campo de la ciencia de datos aplicada a los estudios de usuarios. Este trabajo podría, por lo tanto, proporcionar un valioso punto de referencia para futuras investigaciones en esta área y en campos relacionados. Con estas consideraciones, la importancia de este estudio es multifacética y ampliamente relevante, con implicaciones prácticas y teóricas significativas.

2.- DEFINICIÓN DEL PROYECTO

2.1 Estado del arte

En el contexto de análisis de comportamiento de empleados y gestión de recursos humanos, han surgido diversas metodologías y herramientas. El estado actual de la técnica abarca desde técnicas de análisis estadístico y de minería de datos hasta métodos de aprendizaje automático y de inteligencia artificial.

En el análisis estadístico, los estudios han empleado métodos de agrupamiento y de predicción para identificar grupos de empleados con comportamientos similares y para seguir su evolución a lo largo del tiempo. Por otro lado, en la minería de datos, los enfoques han incluido la extracción de patrones y la identificación de asociaciones para descubrir conexiones subyacentes entre diferentes variables.

El aprendizaje automático se ha convertido en un área de creciente interés, con investigaciones que han aplicado algoritmos de aprendizaje supervisado y no supervisado para predecir comportamientos y patrones de los empleados. Los modelos de aprendizaje profundo, como las redes neuronales, también han demostrado ser útiles en ciertos contextos, particularmente en el análisis de grandes volúmenes de datos.

Finalmente, las técnicas de inteligencia artificial y la adopción de sistemas de soporte a la decisión han mostrado un crecimiento significativo. Estas técnicas se han empleado para optimizar la asignación de recursos y para mejorar la toma de decisiones en la gestión de recursos humanos.

Aunque estos métodos han demostrado ser efectivos, la investigación en este campo continúa, con un constante esfuerzo por desarrollar métodos más sofisticados y precisos. Nuestro estudio se enmarca dentro de este estado del arte, con el objetivo de explorar nuevas vías y contribuir a la evolución continua de este campo.

2.2 Marco teórico

El estudio se basa en varios conceptos teóricos clave que provienen de la ciencia de datos y la gestión de recursos humanos.

Uno de los conceptos fundamentales en este proyecto es la **clusterización** o **agrupamiento**. En el aprendizaje automático, la clusterización se refiere al proceso de agrupar un conjunto de objetos de tal manera que los objetos en el mismo grupo (llamado clúster) son más similares entre sí que con aquellos en otros grupos. En el contexto de nuestro estudio, esto se refiere a la agrupación de empleados basada en patrones similares de comportamiento.

El segundo concepto crucial es la **predicción**, que es la técnica para pronosticar un resultado futuro basándose en patrones y tendencias actuales y pasadas. En nuestro estudio, la predicción se refiere a la utilización de patrones de comportamiento pasados y presentes de los empleados para prever su comportamiento futuro.

Finalmente, la **gestión de recursos humanos** es un campo de estudio y práctica que se centra en la gestión eficiente y efectiva del personal de una organización. En nuestro estudio, este concepto se aplica para optimizar la gestión de los empleados basándose en los patrones de comportamiento identificados y las predicciones realizadas.

Estos conceptos forman el marco teórico en el que se basa nuestro estudio y proporcionan las bases para el diseño de nuestra metodología de investigación y el análisis de nuestros resultados.

2.3 Detalle de los datos

El desarrollo de un diccionario de datos es un aspecto crucial de cualquier proyecto relacionado al análisis de datos; ya que proporciona una breve descripción de todas las variables presentes en el conjunto de datos, incluyendo su descripción y el tipo de dato.

Para facilitar la comprensión de nuestro set de datos, definimos los siguientes diccionarios:

Addresses / Direcciones

Variable	Descripción	Tipo de dato
messageUuid	Identificador único del mensaje	Objeto
messageType	Tipo del mensaje	Categórica
address	Dirección asociada	Objeto
domain	Dominio relacionado	Objeto
name	Nombre asociado con la dirección	Objeto

Tabla 1 Data set de direcciones

Este data set es sobre todas las interacciones que se llevaron a cabo en el transcurso de tres meses. Por cada correo se crean al menos dos registros; uno de origen, otro de destino y en caso de ser necesario, se creará uno por cada dirección de correo que se tenga en copia. El contenido del data set será de gran ayuda para la agrupación ya que fácilmente podríamos segmentar los diferentes usuarios basándose en que tan activos sean en el ambiente laboral.

Emails / Correos

Variable	Descripción	Tipo de dato
uuid	Identificador único del email	Objeto

userId	Identificador del usuario	Objeto
is_recieved	Indica si el correo electrónico fue recibido (1: sí, 0: no)	Categórica
is_sent	Indica si el correo electrónico fue enviado (1: sí, 0: no)	Categórica
sent_datetime	Fecha y hora en que se envió el correo electrónico	Fecha/Hora
conversation_id	Identificador de la conversación	Objeto
has_attachments	Indica si el correo electrónico tiene archivos adjuntos (1: sí, 0: no)	Categórica
hour	Hora del día en que se envió o recibió el correo electrónico	Entero
week_day_txt	Día de la semana en que se envió o recibió el correo electrónico	Objeto
week_day	Día de la semana representado como un número	Entero
week_hour	Hora de la semana en que se envió o recibió el correo electrónico	Objeto
is_answered	Indica si el correo electrónico fue respondido (1: sí, 0: no)	Categórica
is_answered_not	Indica si el correo electrónico no fue respondido (1: sí, 0: no)	Categórica
is_answer	Indica si el correo electrónico es una respuesta (1: sí, 0: no)	Categórica
is_answer_not	Indica si el correo electrónico no es una respuesta (1: sí, 0: no)	Categórica
folder_name	Nombre de la carpeta donde se almacena el correo electrónico	Objeto
sent_date	Fecha en que se envió el correo electrónico	Fecha/Hora
attachments_num	Número de archivos adjuntos en el correo electrónico	Entero
no_working_hours	Indica si el correo electrónico fue enviado o recibido fuera del horario laboral (1: sí, 0: no)	Categórica
importance	Indica la importancia del correo electrónico (1: sí, 0: no)	Categórica
addressNum	Número asociado a la dirección del correo electrónico	Entero
iterPosition	Posición de la iteración	Entero
iterSize	Tamaño de la iteración	Entero
convIsSent	Indica si la conversación fue enviada (1: sí, 0: no)	Categórica
timeAnswer	Tiempo de respuesta al correo electrónico	Entero

Tabla 2 Data set de correos

Este data set nos proporciona detalles sobre cada correo enviado. A diferencia del data set de **addresses**, solo se crea un registro por cada correo. Sin embargo, se facilita información relevante sobre el mismo, como la fecha en que se envió, si hubo alguna respuesta y la importancia. Este data set también será esencial en la agrupación; sin embargo, también será indispensable en la predicción de patrones de actividad.

Users / Empleados

Variable	Descripción	Tipo de dato
userId	Identificador único del usuario	Objeto
displayName	Nombre mostrado del usuario	Objeto
userPrincipalName	Nombre principal del usuario	Objeto
usageLocation	Ubicación de uso del usuario	Objeto
companyName	Nombre de la empresa donde trabaja el usuario	Objeto
departmentName	Nombre del departamento donde trabaja el usuario	Objeto

Tabla 3 Data set de usuarios

Finalmente, esta ultimo data set facilita toda la información sobre los usuarios involucrados, independientemente del número de correos que se hayan mandado; incluyendo información demográfica y organizacionales.

Cabe mencionar que se realizaron algunas transformaciones a los tipos de datos para que se ajustaran correctamente a los modelos.

2.4 Diseño de investigación

El diseño de investigación para este estudio se estructurará en torno a la creación de un entorno de desarrollo en Visual Studio Code (VSC), una plataforma ampliamente reconocida por su capacidad para manejar una amplia variedad de lenguajes de programación y su enfoque en la eficiencia del desarrollador. Este entorno permitirá la implementación de modelos de machine learning necesarios para el análisis y clasificación de los datos de los usuarios.

Estos modelos de machine learning serán seleccionados y adaptados para obtener la máxima precisión y eficiencia posible en la identificación de patrones y tendencias en los datos. Se hará especial hincapié en los modelos capaces de agrupar a los usuarios según su comportamiento y prever sus horarios y patrones de uso. Esto incluirá técnicas de aprendizaje supervisado y no supervisado, según sea apropiado.

De igual manera, usaremos las librerías matplotlib y seaborn de Python para la visualización de los datos. Ambas son potentes librerias de visualización de datos que permite convertir grandes conjuntos de datos en representaciones gráficas intuitivas y fácilmente interpretables. Estas visualizaciones no sólo serán cruciales para el análisis de los resultados,

sino que también proporcionarán una manera efectiva de presentar los hallazgos a los interesados no técnicos.

En conjunto, este diseño de investigación combina técnicas de análisis de datos avanzadas con herramientas de desarrollo de alta calidad y visualizaciones de datos potentes, proporcionando un marco sólido y flexible para investigar y entender los patrones de uso de los usuarios.

3.- ANÁLISIS EXPLORATORIO

Este análisis exploratorio se enfoca en la extracción de insights de nuestros sets de datos. El objetivo es entender su estructura, extraer patrones importantes y relaciones entre las variables; lo que nos proporcionara un mejor contexto para el análisis posterior, ya que este proceso no solo permite obtener una visión preliminar y descriptiva de los datos, sino que también prepara el terreno para la modelización y el análisis predictivo. Este proceso es crucial para identificar los pasos necesarios en la etapa de preprocesamiento de datos, asegurando que los datos estén en un formato adecuado y limpio antes de ser introducidos en los algoritmos de machine learning.

Los objetivos son:

- **Facilitar la comprensión de los datos**: Las visualizaciones son una forma eficaz de entender grandes cantidades de datos complejos. Las librerías de visualización de datos nos permiten crear gráficos que pueden revelar patrones, tendencias y relaciones que no se verían fácilmente en los datos brutos.

- **Generar insights**: Las visualizaciones pueden ayudarnos a generar insights útiles y formular hipótesis para un análisis más profundo. Por ejemplo, un histograma puede revelar la distribución de los datos, mientras que un diagrama de dispersión puede revelar correlaciones entre variables.

- **Comunicar los resultados**: Las visualizaciones son una excelente manera de comunicar los resultados del análisis a los demás. Un gráfico bien diseñado puede transmitir la esencia de los datos de manera más eficaz que las descripciones textuales.

No obstante, es importante reconocer que nuestra exploración no ha agotado todas las posibles dimensiones y aspectos de los datos. Hemos optado por centrarnos en las características que consideramos más relevantes y esenciales para el propósito de nuestro estudio. A pesar de esto, quedan aún más aspectos de los datos que podrían ser objeto de investigación adicional y que podrían arrojar aún más luz sobre los patrones y tendencias subyacentes. Por lo tanto, recomendamos considerar este análisis como un punto de partida sólido y no como una exploración exhaustiva de todos los posibles aspectos de los datos.

3.1 Set de direcciones

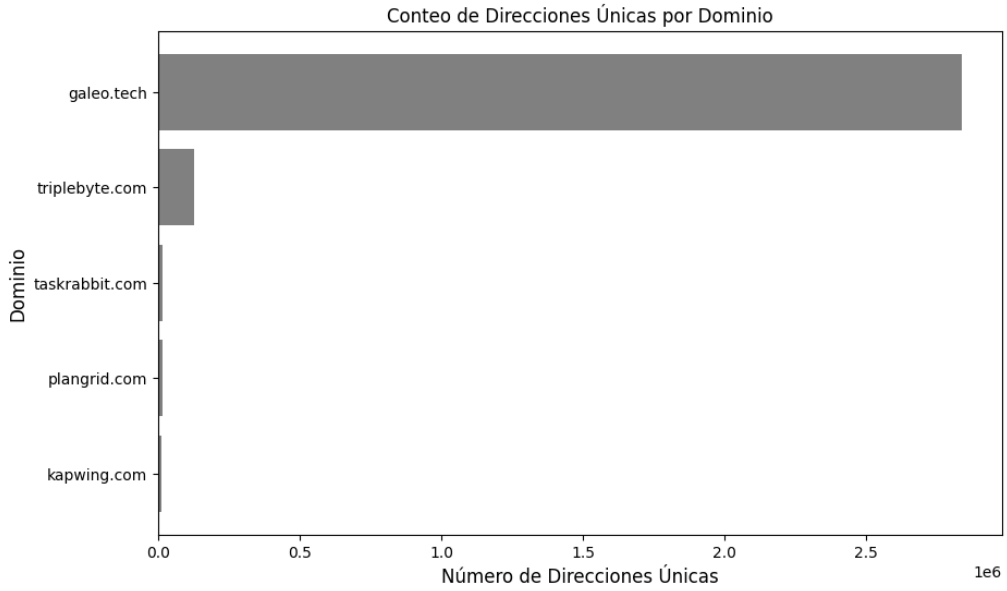

Ilustración 1.1 Conteo de direcciones únicas

Como se esperaba con este gráfico, el dominio dominante dentro de todas las interacciones es el de Galeo; sin embargo, consideramos que era importante resaltar que otros dominios son populares o frecuentes dentro de la empresa.

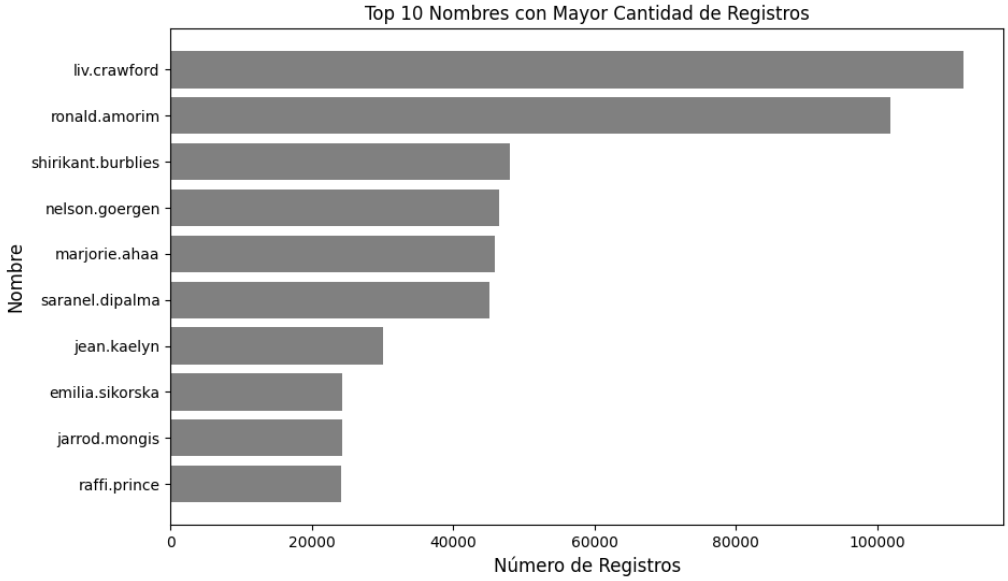

Ilustración 2 Usuarios con mayor cantidad de registros

En este otro gráfico, consideramos que era importante identificar cuáles son los usuarios que más interactúan; ya sea por correos enviados, recibidos o en copia, hay una gran diferencia

entre los primeros dos usuarios que con el resto. Datos como estos serán esenciales para la agrupación ya que buscamos segmentar entre los usuarios con mucha actividad y con los que tienen poca.

3.2 Set de correos

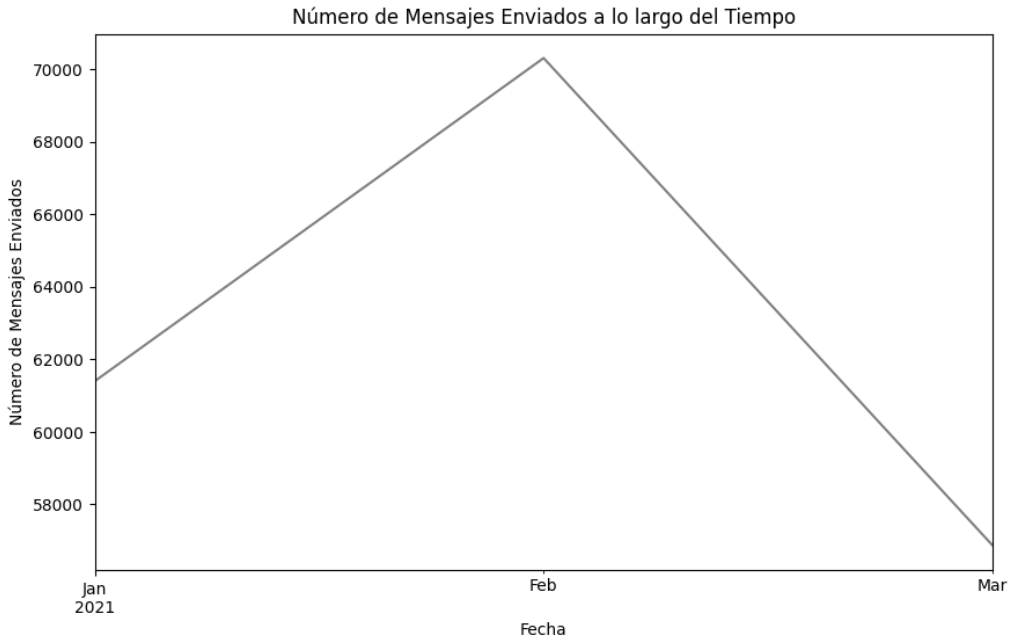

Ilustración 3 Mensajes enviados a lo largo del tiempo

Esta grafica nos muestra cómo ha cambiado la cantidad de correos a lo largo del tiempo. Por el considerable incremento en febrero, podemos asumir que hubo un periodo de alta actividad o una serie de eventos que requirieron una comunicación más intensiva. La disminución en marzo podría sugerir una vuelta a un ritmo de trabajo más estándar o la conclusión de dichos eventos. Este tipo de patrones temporales puede proporcionar información valiosa sobre los ciclos operativos y las demandas de comunicación dentro de la organización.

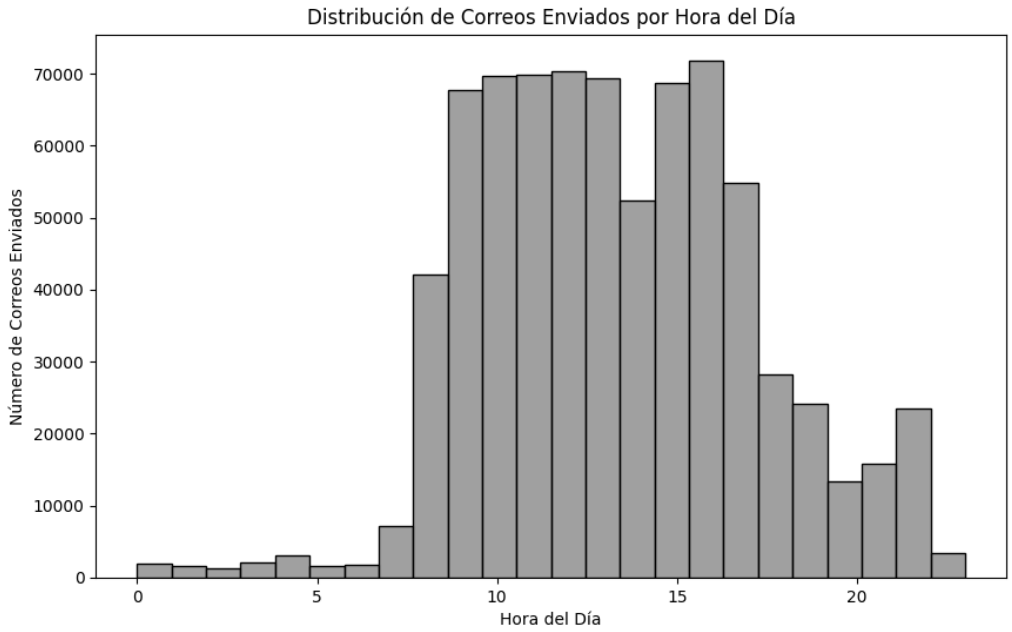

Ilustración 4 Distribución de correos

Esta grafica nos muestra un patrón interesante. Se observa un volumen considerable de correos enviados fuera del horario laboral común, lo que podría indicar una cultura laboral que se extiende más allá de las horas de trabajo estándar. Esta observación es importante, ya que puede tener implicaciones en términos de equilibrio entre el trabajo y la vida personal, eficiencia laboral y la satisfacción de los empleados. Es posible que se requiera una exploración más profunda para comprender mejor las causas y las implicaciones de esta tendencia.

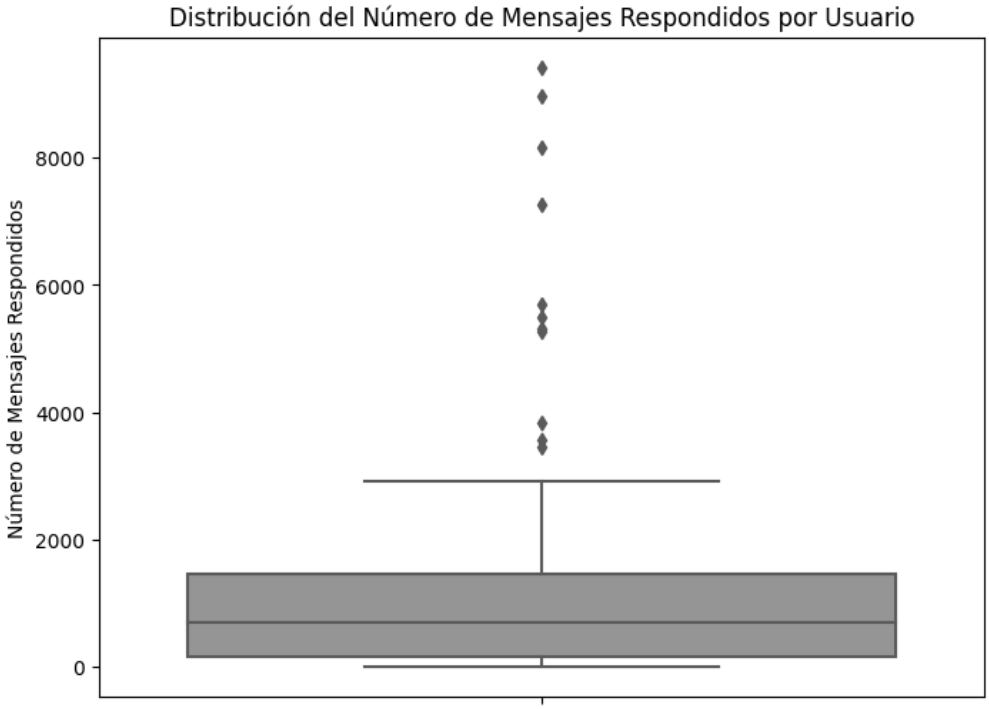

Ilustración 5 Mensajes respondidos por usuario

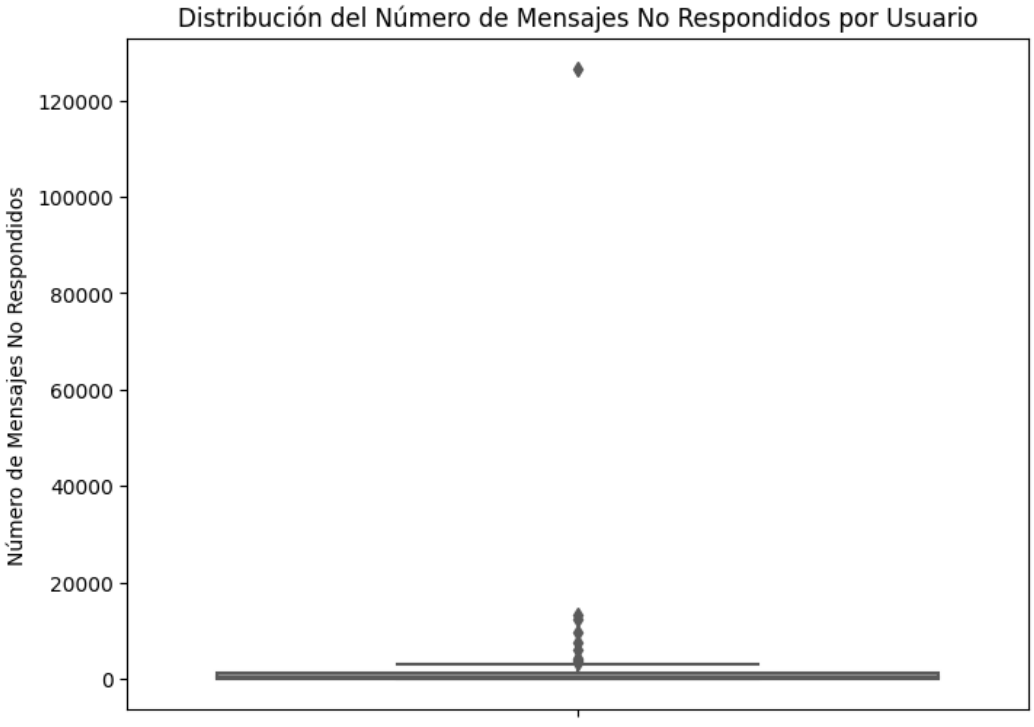

Ilustración 6 Mensajes no respondidos por usuario

Estas últimas dos graficas destacan una tendencia clara en el comportamiento de respuesta a los mensajes. Se observa que la cantidad de mensajes respondidos supera significativamente

a los no respondidos, lo que indica un alto nivel de compromiso entre los usuarios en la plataforma.

Respecto a los datos atípicos, es interesante notar qué en los mensajes respondidos, los datos atípicos están más distribuidos, lo que puede indicar usuarios individuales o grupos que son excepcionalmente activos en la respuesta a los correos electrónicos. Por otro lado, el dato atípico notable en los mensajes no respondidos puede ser una señal de un usuario o grupo de usuarios que recibe una cantidad inusualmente alta de correos electrónicos que no responden, lo cual puede ser un punto de investigación adicional para entender las posibles razones detrás de este comportamiento.

3.3 Set de usuarios

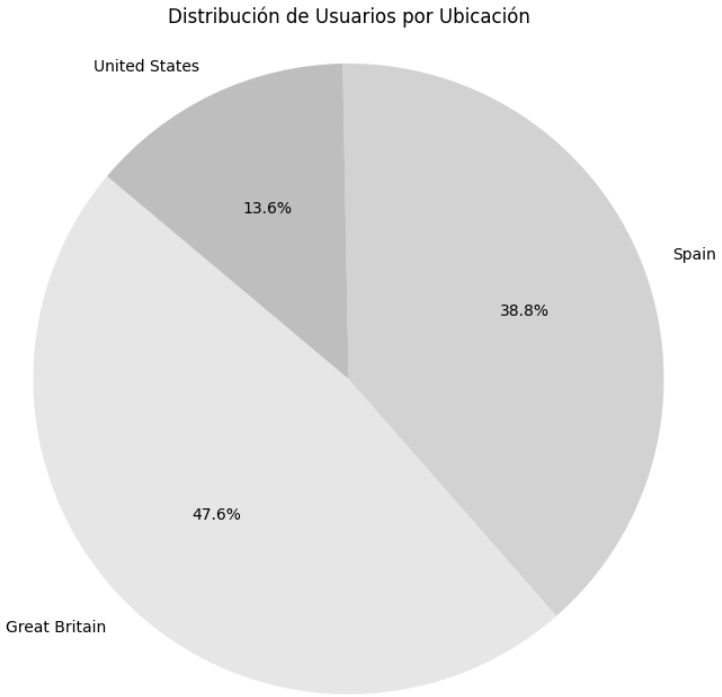

Ilustración 7 Distribución por ubicación

Ahora, esta grafica nos brinda un panorama claro de la presencia geográfica de los usuarios. Un gran porcentaje de los usuarios se ubican en Gran Bretaña, lo cual refleja que es el mercado predominante en la base de usuarios de la empresa. Le sigue España con un número significativo de usuarios, indicando que es otro mercado importante para la empresa. Por otro lado, se observa una menor presencia en Estados Unidos, lo que podría sugerir una oportunidad para expandirse y aumentar la base de usuarios en esa región. También es importante destacar que este desglose por ubicación puede tener implicaciones en el diseño de estrategias de negocio y operaciones, como la atención al cliente y la gestión de la diversidad cultural.

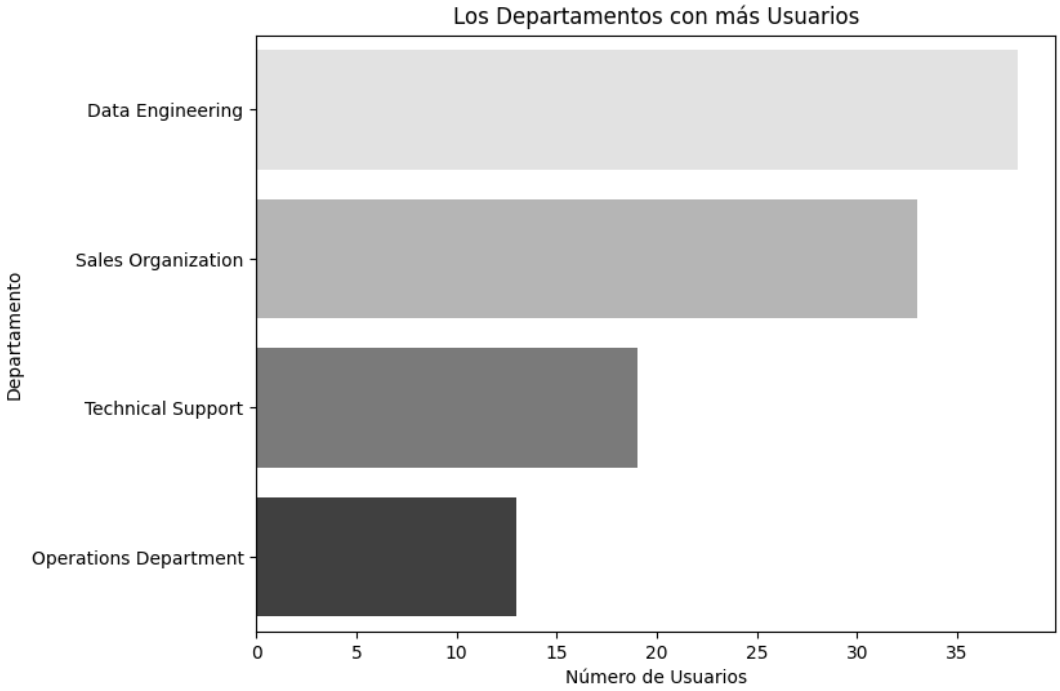

Ilustración 8 Departamentos con más usuarios

En este caso observamos la estructura interna de la empresa a través de la lente de sus usuarios. Observamos que el departamento de Ingeniería de Datos tiene la mayor cantidad de usuarios, lo cual es coherente con el enfoque tecnológico y orientado a los datos de la empresa. Le sigue la organización de ventas, que también tiene un alto número de usuarios, indicando su papel crucial en la generación de ingresos. El soporte técnico y el departamento de operaciones siguen en cantidad de usuarios, reforzando el hecho de que estos son componentes clave en cualquier organización de TI. En resumen, esta distribución nos da una idea de cómo se distribuyen los recursos humanos en la organización y en qué áreas la empresa está enfocando sus esfuerzos.

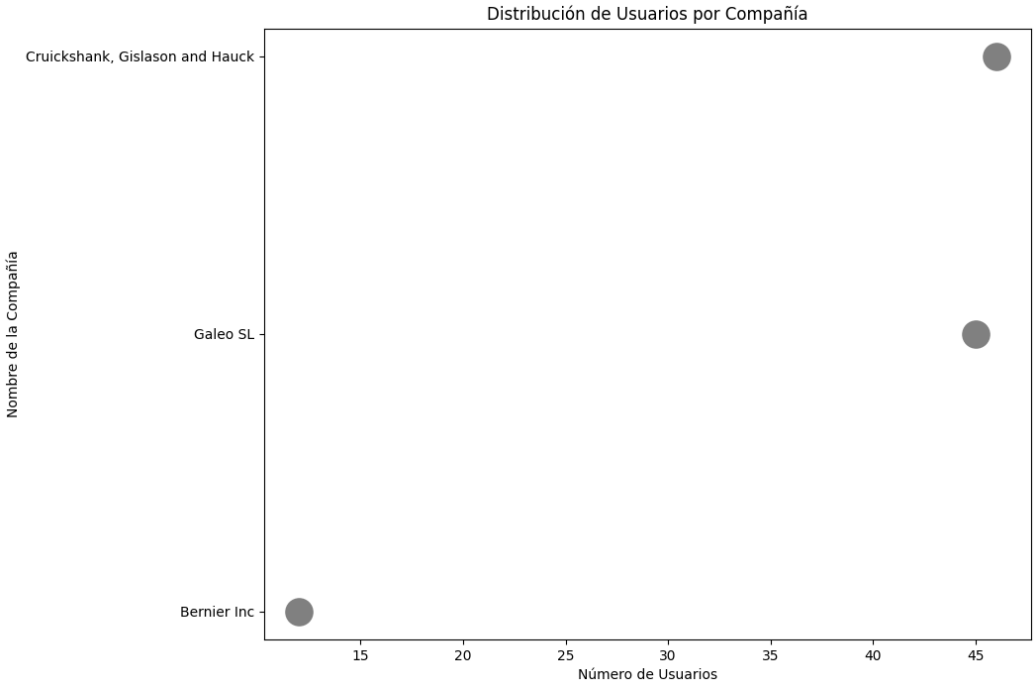

Ilustración 9 Usuarios por compañia

Finalmente, esta última grafica nos ofrece una perspectiva de la distribución de usuarios a nivel de compañías asociadas o subsidiarias. Observamos que la empresa Cruickshank, Gislason and Hauck encabeza la lista, seguida por Galeo y Bernier. Esto podría indicar una amplia colaboración o afiliación con Cruickshank, Gislason and Hauck, lo que resulta en un mayor número de usuarios asociados. Las otras dos compañías, Galeo y Bernier, aunque con menos usuarios, también representan una proporción significativa. Estos hallazgos pueden ser indicativos de las relaciones de negocio y la estructura de la empresa, y proporcionan un marco de referencia para entender el flujo de comunicación y la red de colaboración entre estas entidades.

A pesar de haber obtenido insights valiosos, aún quedan muchas posibilidades y preguntas abiertas que pueden ser respondidas a medida que continuamos con el modelado y la extracción de características.

En la próxima sección, comenzaremos a construir y afinar un modelo que busque alcanzar nuestros objetivos establecidos en la hipótesis. Utilizaremos la visión obtenida a través de este análisis para informar nuestras decisiones en este proceso. El fin último de nuestro esfuerzo es crear un modelo preciso que permita anticipar el comportamiento de respuesta y ayudar a optimizar las estrategias de comunicación de la empresa.

4.- DESCRIPCIÓN DE LAS HERRAMIENTAS

Finalmente abordaremos la esencia del estudio; la creación de nuestros modelos. Como se mencionó anteriormente, usaremos estos modelos para segmentar a nuestros usuarios basándonos en su actividad, identificar las variables más importantes que determinan la pertenencia a estos grupos, examinar la evolución de estos grupos a lo largo del tiempo y, finalmente, predecir los horarios de los usuarios. Al aplicar las potentes técnicas de aprendizaje automático a nuestros conjuntos de datos, buscamos obtener una visión más detallada y predictiva de la conducta de nuestros usuarios. Cada paso de este proceso será descrito y explicado en detalle, mostrando no sólo los resultados, sino también la metodología que seguimos para obtenerlos. Esperamos que este viaje a través de los intrincados caminos del aprendizaje automático ilustre la capacidad de estas herramientas para revelar patrones y predecir comportamientos futuros.

4.1 Entorno de ejecución

Para la creación de nuestros modelos, utilizaremos un entorno local de Anaconda con Python 3.9 como lenguaje de programación. Haremos uso de varias bibliotecas de Python de gran utilidad en este ámbito, como Pandas para la manipulación de datos, Scikit-Learn para el desarrollo de nuestros modelos de aprendizaje automático, Numpy para las operaciones matemáticas y las librerías mencionadas anteriormente Matplotlib junto con Seaborn para la visualización de los datos La combinación de estas librerías nos permite tener un entorno muy completo y potente para llevar a cabo nuestra investigación y conseguir los resultados que buscamos.

4.2 Agrupación de usuarios

En nuestro esfuerzo por encontrar los clusters óptimos en nuestros datos, hemos explorado dos enfoques populares y eficaces: K-Means y Agrupación Jerárquica. Aunque estos métodos comparten ciertos principios, también presentan diferencias significativas que los hacen únicos en su rendimiento y resultados. K-Means es un algoritmo de agrupamiento particional que se enfoca en minimizar la varianza dentro del cluster, sin embargo, requiere que se especifique el número de clusters de antemano, lo que puede ser una limitación si no se tiene una idea previa de la estructura de los datos. Por otro lado, la agrupación jerárquica, un enfoque de agrupamiento aglomerativo, ofrece una visualización intuitiva de la estructura de los datos a través del dendrograma y permite la identificación de la cantidad óptima de clusters de forma más directa.

4.2.1 Preparación de los datos

Para obtener una comprensión más detallada de los comportamientos de los usuarios en relación con los correos electrónicos, es esencial desglosar y examinar varios aspectos. Por lo tanto, procedemos a calcular métricas adicionales a nivel de usuario. No sólo consideramos el número total de correos electrónicos enviados y recibidos, sino que también diferenciamos entre la importancia de los correos electrónicos enviados y recibidos. Esta distinción nos proporciona una visión más clara de la importancia relativa que los usuarios asignan a los correos electrónicos que envían y reciben.

Además, separamos el número de conversaciones iniciadas por los usuarios, proporcionando una medida de la participación de cada usuario. Para entender más sobre la naturaleza de la correspondencia, también calculamos el número total de respuestas enviadas y recibidas, así como los correos electrónicos que no son respuestas, tanto enviados como recibidos. Esto nos permite entender mejor la proporción de la comunicación que está en el contexto de una conversación existente y la que está iniciando una nueva conversación.
Finalmente, consideramos la cantidad de correos electrónicos enviados fuera del horario laboral. Este último punto es especialmente relevante, ya que puede indicar situaciones en las que los usuarios pueden sentirse abrumados o estar trabajando horas extras, lo que podría afectar a su bienestar y productividad.

Combinamos todas estas métricas en un nuevo DataFrame de Pandas para obtener un retrato más matizado de los patrones de comunicación por correo electrónico de cada usuario, lo que nos permite realizar una segmentación más informada y significativa de los usuarios.

4.2.2 K-Means

Antes de proceder, los datos son preprocesados utilizando la técnica de Standard Scaler. Esta técnica transforma las características numéricas para que tengan una media de 0 y una desviación estándar de 1, asegurando que todas las características tengan el mismo peso automático.

Para determinar el número óptimo de clusters, se utiliza una combinación del método Elbow y el Silhouette Score. Se calcula y grafica el Sum of Squares (SSQ) para una gama de valores de K, y se elige el valor de K que causa un "codo" en la gráfica. Paralelamente, se calcula el Silhouette Score para la misma gama de valores de K, y se selecciona el valor que maximiza el Silhouette Score.

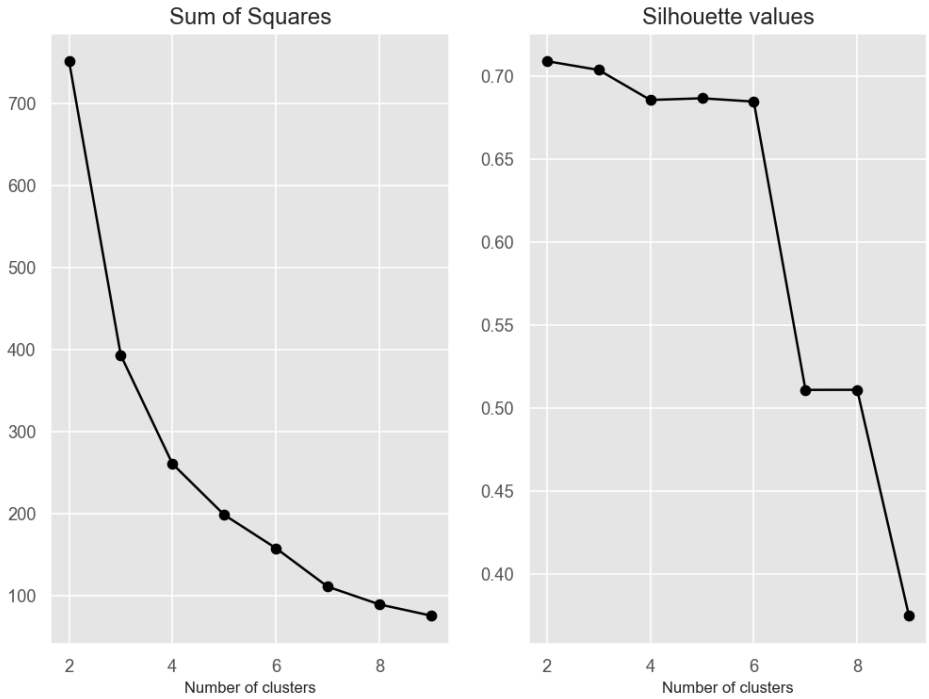

Ilustración 10 # de cluster por K-means

A continuación, se ajusta un modelo de clustering K-Means tomando en cuenta el número de clusters sugerido por nuestro modelo, en este caso, 3. También, y con el fin de visualizar los clusters en un espacio bidimensional, se realiza un Análisis de Componentes Principales (PCA); esto reduce el número de características a dos componentes principales que capturan la mayor cantidad de varianza en los datos. De igual manera, calculamos el silhouette coefficient lo cual nos ayuda a evaluar la calidad de agrupamiento; al obtener una media de 0.704, podemos considerar que es un resultado relativamente bueno.

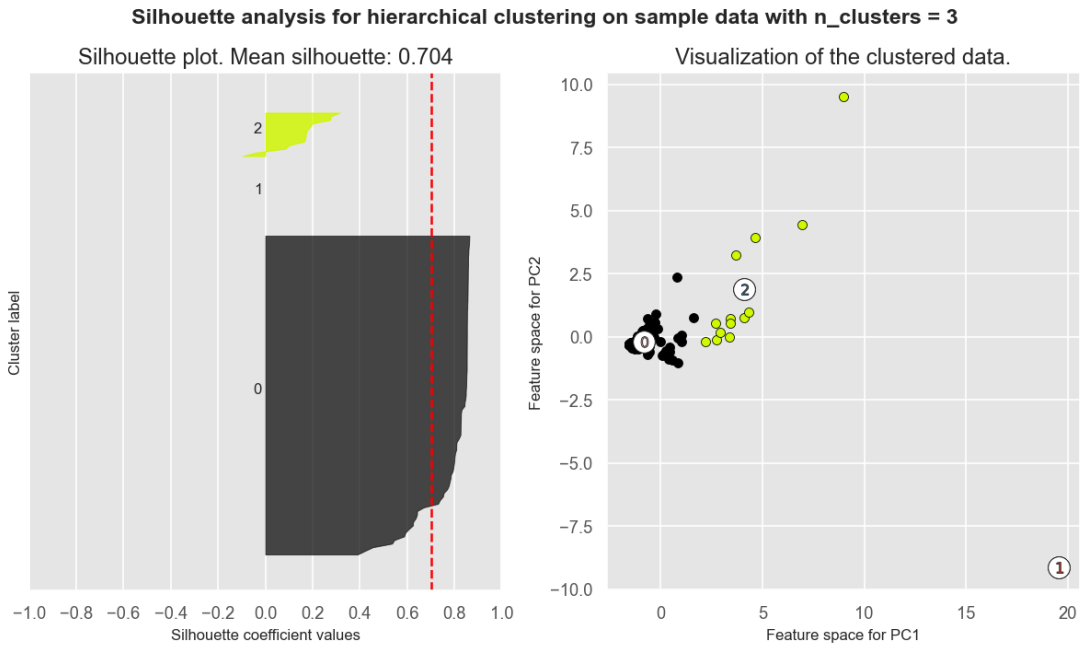

Ilustración 11 Clusters por k-means

Ya que tenemos nuestros clusters definidos, realizamos un conteo del número de usuarios en cada cluster, para obtener una visión general de cómo se distribuyen los usuarios entre los diferentes clusters; el resultado fue el siguiente:

Cluster:
1 89
2 13
3 1

Por último, se calculó la media de cada característica para cada cluster y se representó en un gráfico de barras. Con esto, tenemos una visión general de las características que diferencian a los clusters y podemos interpretar los resultados del análisis de agrupamiento.

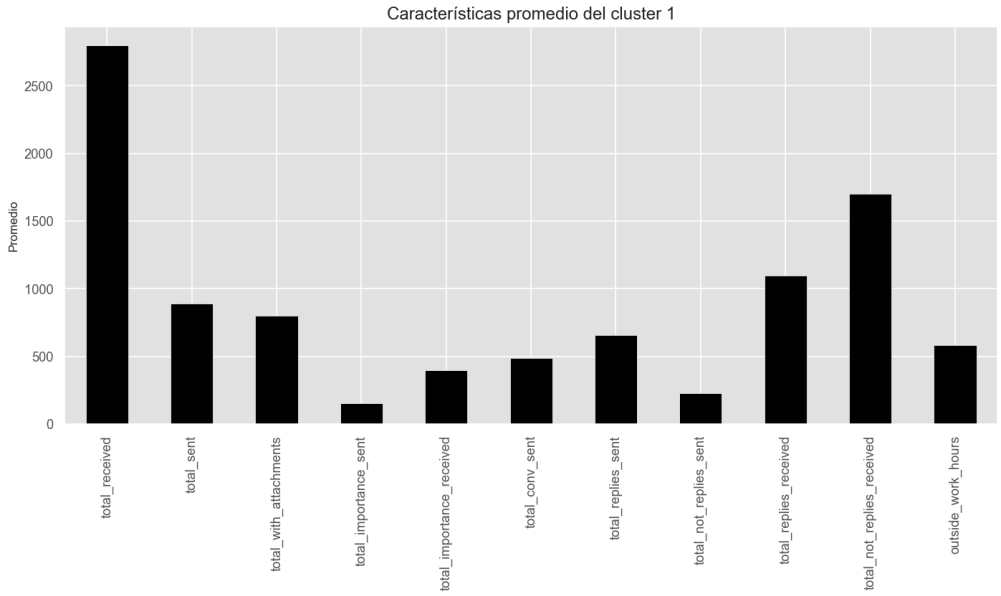

Ilustración 12 Cluster 1 K-means

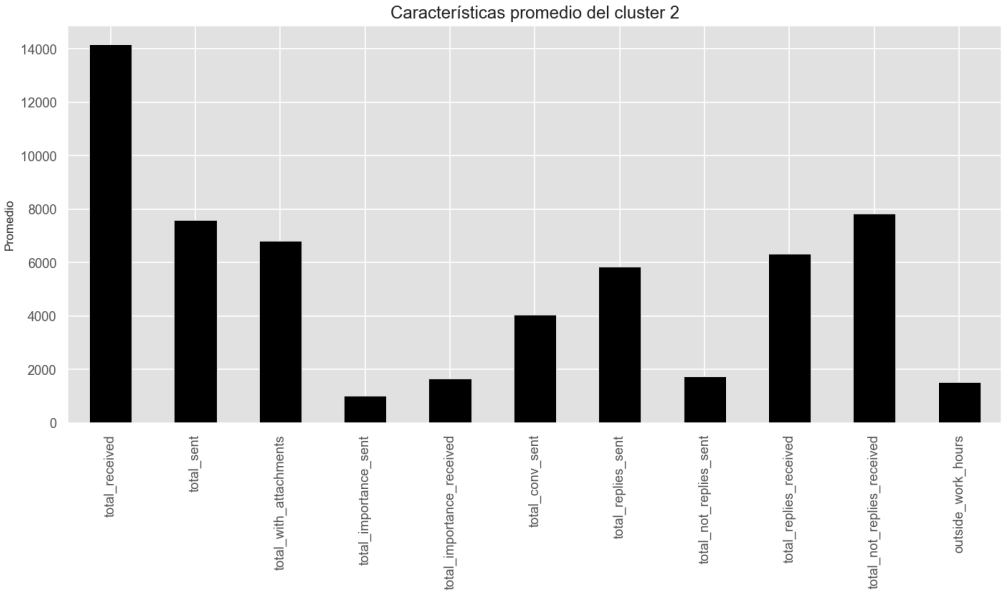

Ilustración 13 Cluster 2 K-means

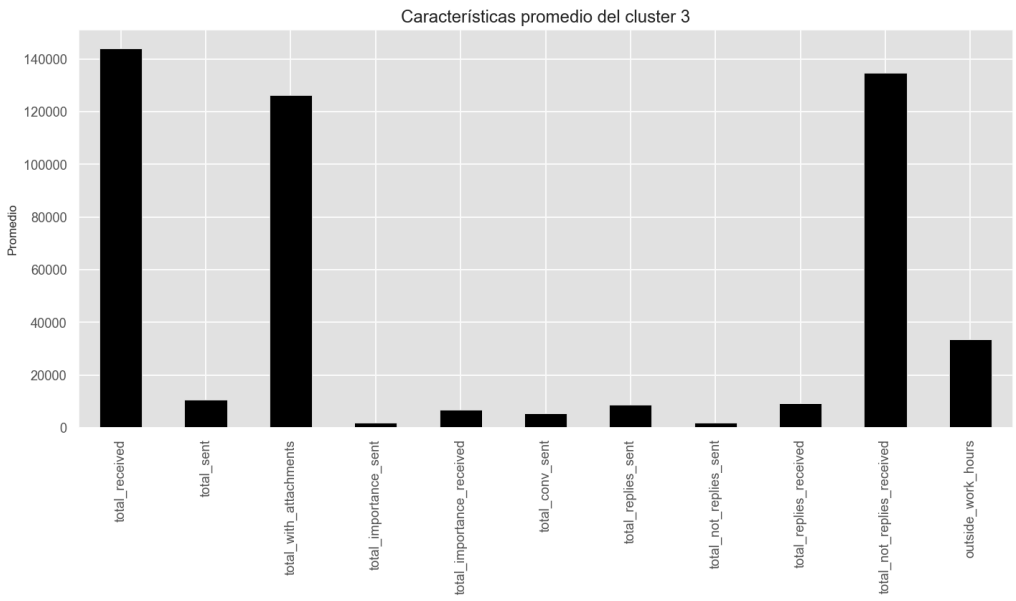

Ilustración 14 Cluster 3 K-means

El hecho de que exista un cluster con solo un miembro podría suponer un error en la agrupación o el uso de un numero de clusters no optimo. No obstante, aunque los valores parezcan outliers, consideramos que es probable que el miembro del cluster 3 pueda estar en una posición de alto rango o de liderazgo en una organización grande o que este al mando de varios proyectos o ser el punto de contacto principal para ciertos proyectos o tareas, lo que justificaría la enorme cantidad de correos recibidos.

Antes de profundizar más en los resultados obtenidos con el método de agrupación de K-means, creemos que es pertinente explorar otros métodos de agrupación para asegurarnos de que estamos obteniendo una representación precisa y útil de nuestros datos. Como se mencionó anteriormente, estamos interesados en aplicar la agrupación jerárquica, ya que esta técnica ofrece un enfoque diferente y puede revelar relaciones y estructuras en los datos que no son inmediatamente evidentes con el método de K-means. La agrupación jerárquica es particularmente útil porque presenta una visión visual clara del agrupamiento en forma de dendrograma, permitiendo una mejor interpretación de la similitud entre grupos. Además, a diferencia del método de K-means, no se necesita especificar el número de grupos de antemano, lo que puede proporcionar una perspectiva diferente de la agrupación óptima.

Por lo tanto, probaremos dicho metodo en nuestros datos y compararemos sus resultados con los obtenidos en este último modelo para asegurarnos de que estamos haciendo uso completo de nuestras herramientas de análisis y obteniendo la visión más completa posible de nuestros datos.

4.2.3 Agrupación Jerárquica

Al igual que con K-means, primero es necesario normalizar los datos utilizando Standard Scaler. Este paso es crucial para garantizar que todas las características tengan el mismo peso en los cálculos de distancia durante la agrupación. Por otro lado, usamos el método Linkage; el cual se utiliza para realizar la agrupación jerárquica. Este método toma como entrada la matriz de características normalizada y el método de enlace a utilizar. En este caso, también se utiliza el método de 'ward', que minimiza la suma de las diferencias cuadradas dentro de todos los clusters. El resultado es una matriz que representa la jerarquía de agrupación en forma de enlaces entre clusters, la cual es visualizada con la ayuda de un dendograma o diagrama de arbol.

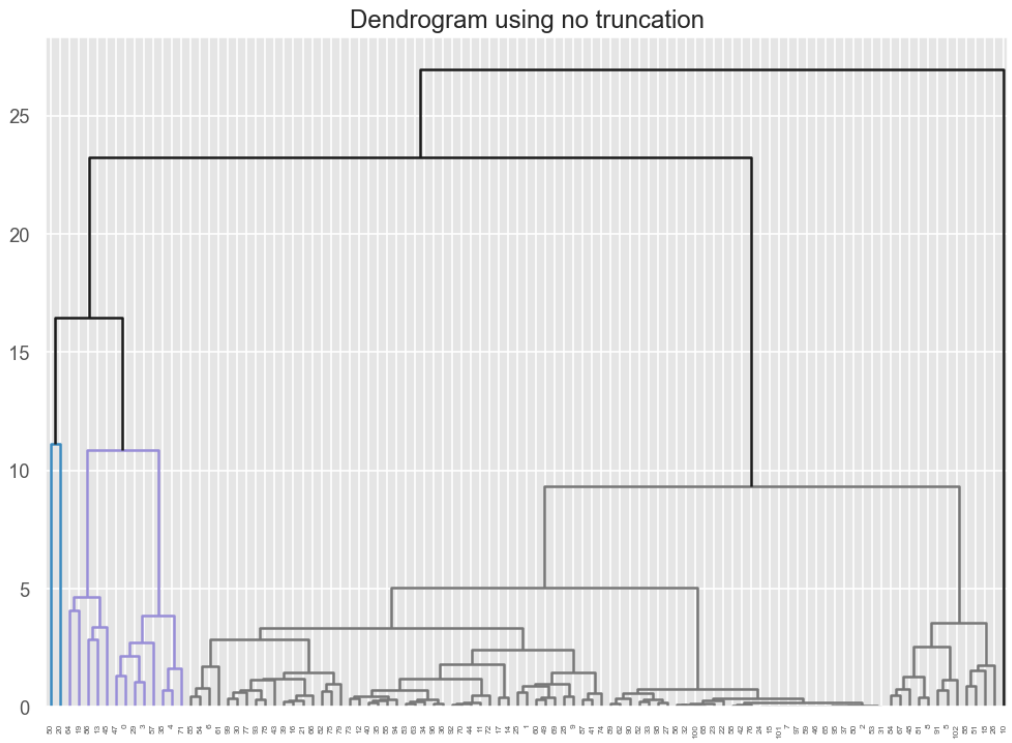

Ilustración 15 Dendograma sin truncar

Las visualizaciones pueden parecer un poco complejas o difíciles de interpretar; por lo que pueden variar en términos de truncación para ofrecer una vista simplificada.

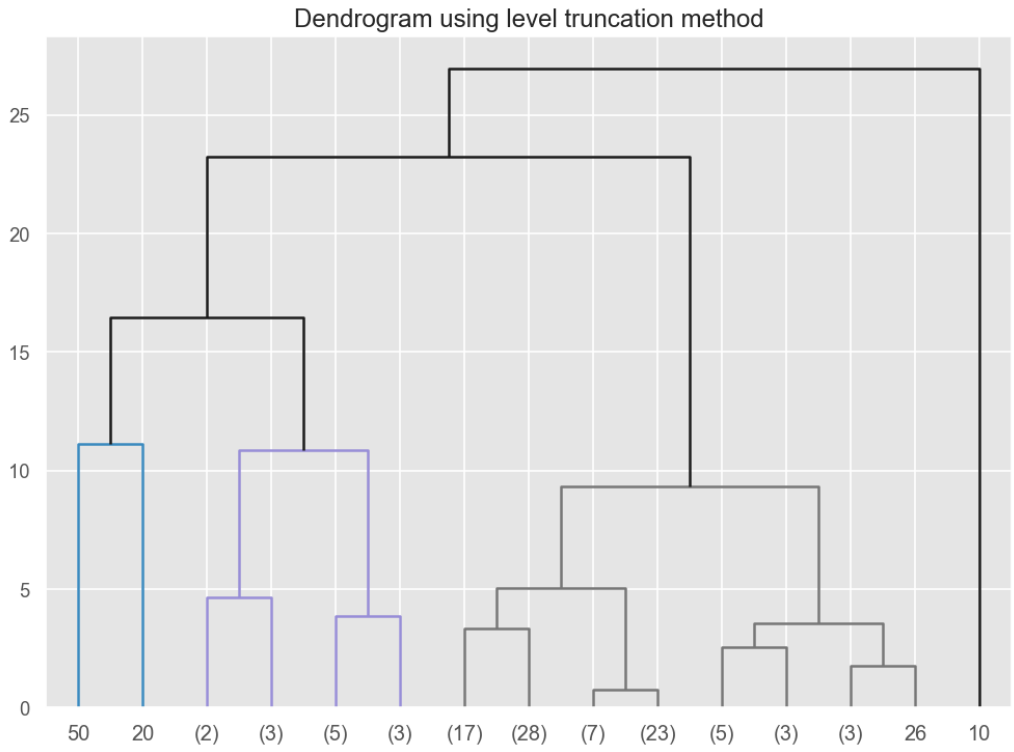

Ilustración 16 Dendograma truncado

A simple vista, se puede interpretar que hay más de 3 clusters, en contraste con el método de K-means. Por otro lado, para determinar el numero optimo haremos uso de la técnica Cut-Tree ya que proporciona una forma de convertir esta representación jerárquica en una asignación de puntos de datos a clusters individuales. Esto se hace especificando una altura en el dendrograma, y "cortando" el dendrograma en ese punto. En este caso, lo haremos en el punto 15; el cual consideramos que era un punto razonable ya que es un punto medio.
Finalmente obtenemos un numero óptimo de 4 clusters; con los cuales trabajaremos para intentar obtener un mejor resultado que con el modelo pasado.

Estos son los valores que obtuvimos por cluster:

Cluster:
1 13
2 87
3 1
4 2

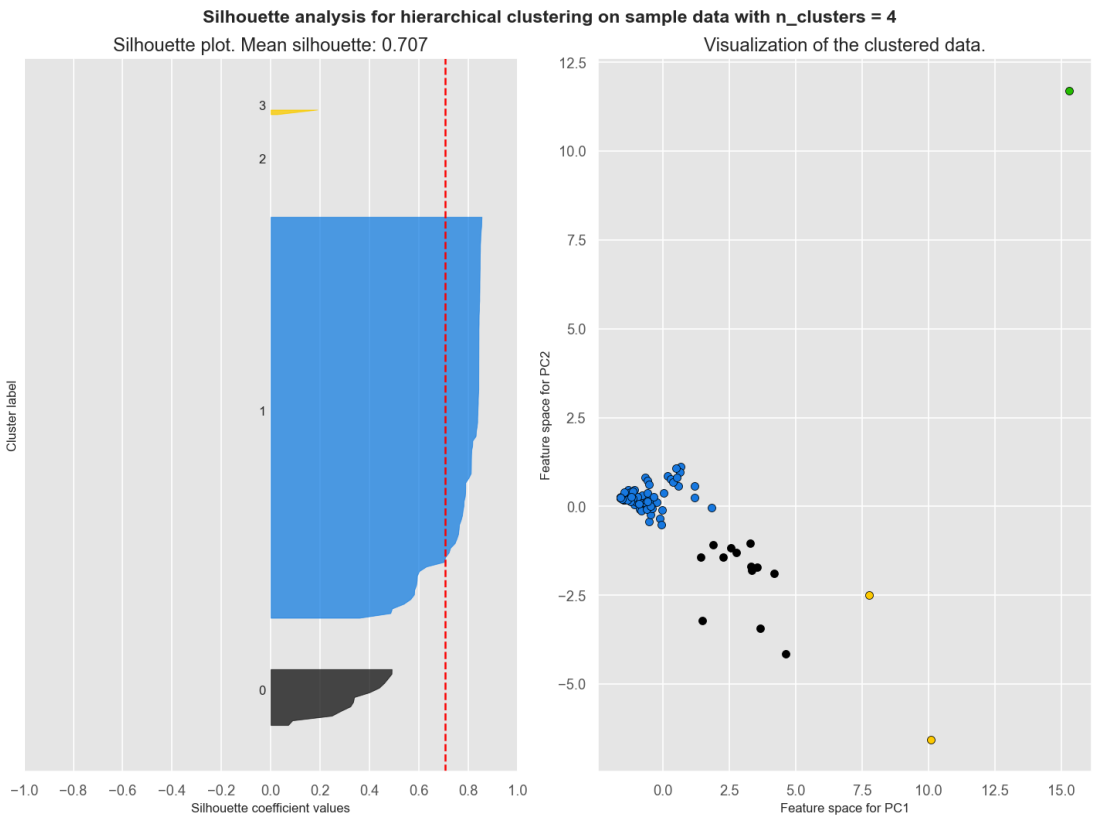

Ilustración 17 Clusters jerarquicos

Nuevamente, utilizamos el método de la silueta para validar la calidad de la agrupación. Con un silhouette coefficient de 0.707 sugiere que los clusters están relativamente bien separados. A diferencia del K-means, hay un cuarto cluster de dos miembros. Este cambio en el número de clusters puede ser útil para capturar patrones más sutiles o diferencias más matizadas entre los grupos, permitiendo una comprensión más detallada y precisa de los datos; justificando por eso que se decidió explorar un enfoque diferente con la agrupación jerárquica.

Finalmente, se examinan las características promedio de cada cluster:

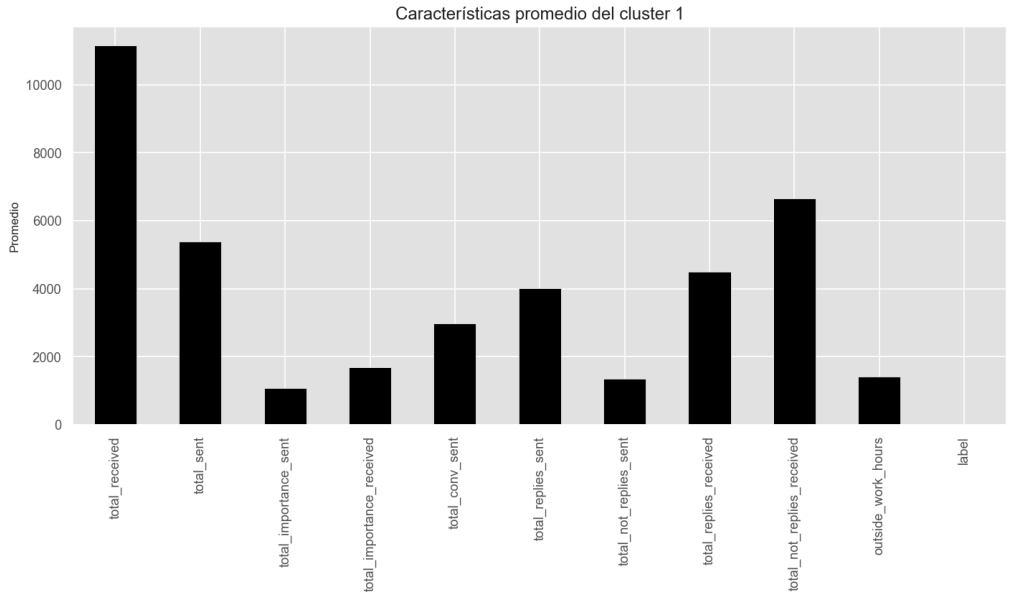

Ilustración 18 Cluster 1 jerarquico

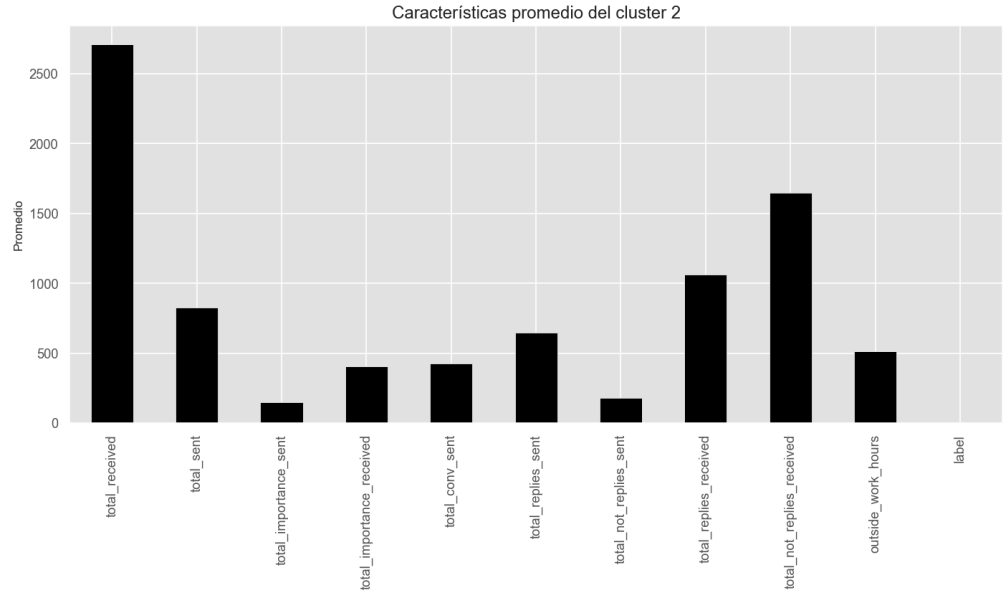

Ilustración 19 Cluster 2 jerarquico

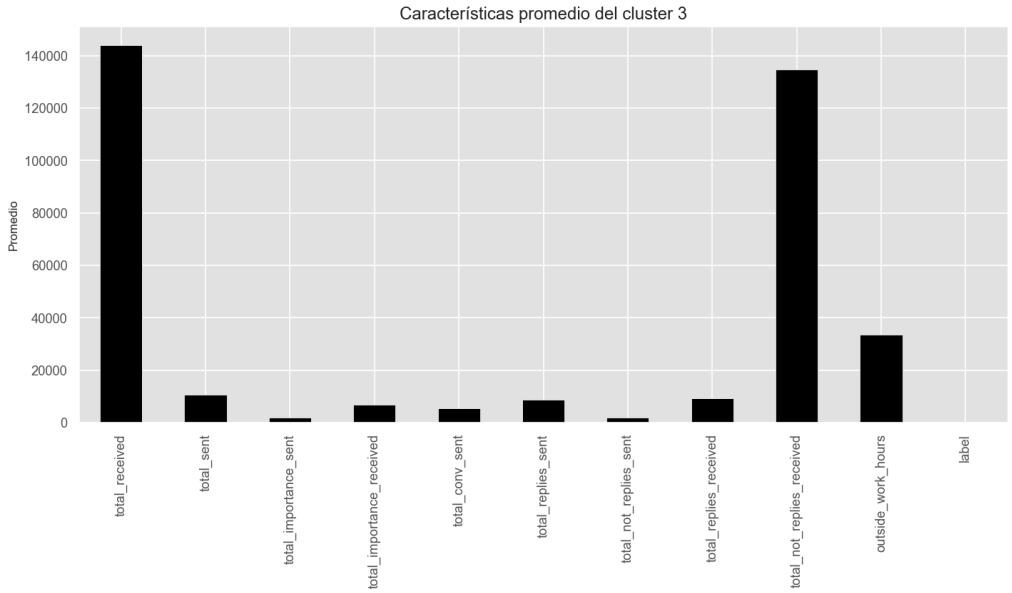

Ilustración 20 Cluster 3 jerarquico

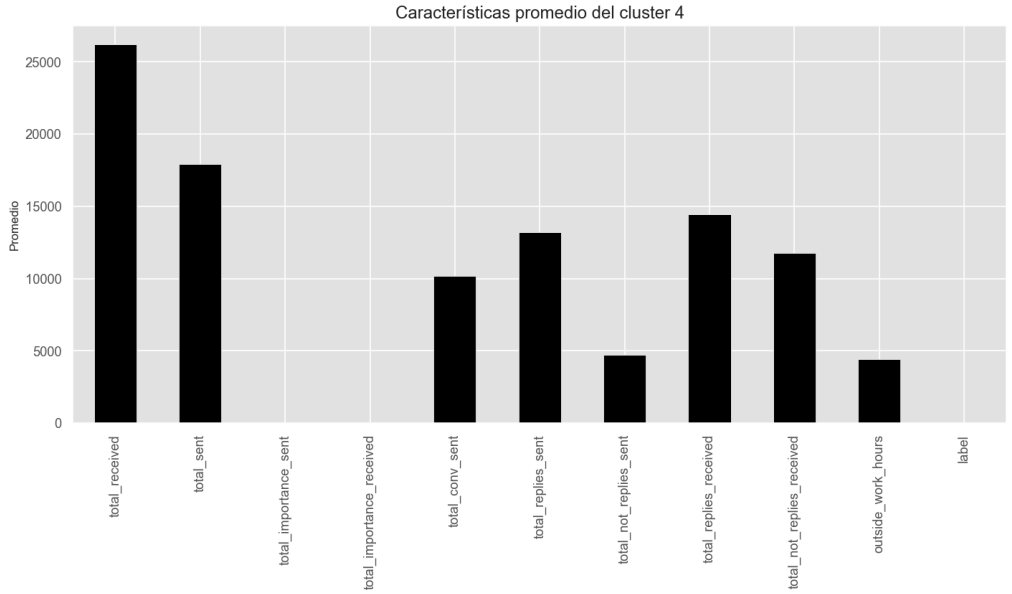

Ilustración 21 Cluster 4 jerarquico

Comparando esto con los resultados del modelo pasado, parece que el modelo de agrupación jerárquica ha podido identificar un grupo adicional que era menos visible en el análisis K-means; sin embargo, cabe resaltar que se resalta por un aspecto realmente distintivo es que el "total_importance_sent" y el "total_importance_received" son cero. Esto sugiere que estos individuos, aunque muy activos en términos de comunicación por correo electrónico, no marcan ningún correo electrónico como importante o no reciben correos electrónicos marcados como importantes. Esto sugiere que podrían ejercer roles como el de servicio al cliente o de servicio técnico; lo cual implica mucha comunicación, inclusive fuera de horario de trabajo, pero sin implicar una significativa importancia.

4.2.4 Características principales de los clusters

Finalmente, después de haber considerado tanto el método de agrupación K-means como el de agrupación jerárquica, hemos decidido optar por los resultados proporcionados por este último. La razón principal de esta elección es la capacidad de la agrupación jerárquica para identificar un cuarto cluster distintivo, lo que permite una mayor granularidad y entendimiento de los patrones de comportamiento de los usuarios.

Podríamos caracterizar los cuatro clusters de la siguiente manera:

# Cluster	Descripción
1	Representa a los trabajadores de nivel intermedio, que están bastante activos en las conversaciones por correo electrónico.
2	Representa a los trabajadores de nivel básico, con una actividad de correo electrónico menor en comparación con otros grupos.
3	Corresponde a un único individuo que recibe una cantidad extraordinariamente alta de correos electrónicos, posiblemente pertenecientes a roles de alta dirección.
4	Corresponde a roles que requieren una gran cantidad de comunicación y coordinación, posiblemente gerentes de proyecto, roles de servicio al cliente, roles de soporte técnico o roles de coordinación de eventos

Tabla 4 Descripción de los clusters

Estos resultados nos ofrecen una visión muy útil de la comunicación entre los usuarios. Sin embargo, este análisis es solo una parte del rompecabezas. En la siguiente etapa de nuestro trabajo, realizaremos una predicción de los horarios de cada cluster.

4.3 Predicción de horarios

Para optimizar la eficiencia y la productividad en cualquier organización, es crucial entender los patrones de comportamiento de los distintos grupos de personas que la conforman. En nuestro caso, habiendo identificado distintos clusters a través de nuestros modelos de agrupación, queremos ahora entender sus patrones de actividad en términos de horarios. Predecir los horarios de cada cluster nos permitirá entender cuándo cada grupo está más activo y receptivo, lo que puede llevar a mejoras en la comunicación, la asignación de tareas y la colaboración general. Asimismo, estas predicciones pueden ser útiles para adaptar los

horarios de reuniones y otros eventos a los patrones de actividad de cada grupo, maximizando así la participación y el compromiso. También nos servirá para identificar potenciales problemas, por ejemplo, si un grupo está trabajando constantemente fuera del horario laboral estándar, lo que podría ser un signo de sobrecarga de trabajo o mala gestión del tiempo. En conclusión, la predicción de los horarios de cada cluster nos proporcionará una herramienta valiosa para mejorar la dinámica de trabajo y el bienestar de los miembros de la organización.

Debido a consideraciones prácticas tanto organizativas como computacionales, hemos decidido abordar la predicción de los horarios para cada cluster de manera individualizada. Al realizar el análisis de un cluster a la vez, podemos centrar nuestra atención y recursos en entender y modelar los patrones de actividad de ese grupo en particular, lo que a su vez puede resultar en modelos de predicción más precisos y útiles. Esta estrategia también nos permitirá manejar de manera más eficiente los recursos de cómputo y memoria disponibles, ya que al trabajar con un subconjunto de datos a la vez, reducimos la carga computacional. Además, el enfoque por etapas nos proporcionará la flexibilidad para adaptar y refinar nuestro enfoque a medida que avanzamos en el análisis de cada cluster.

4.3.1 Preparación de los datos

Para cada cluster, realizamos un conjunto de transformaciones a los datos para prepararlos para el modelado. Comenzamos por aislar cada cluster, tomando solo aquellos registros en nuestro DataFrame que corresponden al cluster en cuestión.

A continuación, eliminamos una serie de columnas que no se necesitan para el modelo. Estas incluyen 'uuid', 'userId', 'conversation_id', 'sent_datetime', 'sent_date', y 'week_hour'. Estas columnas se eliminaron debido a su naturaleza única para cada registro ('uuid', 'userId', 'conversation_id'), porque se duplican la información ya presente en otras columnas ('sent_datetime', 'sent_date') o porque representan información que ya hemos tomado en cuenta de otra manera ('week_hour').

Posteriormente, realizamos una codificación One-Hot en las columnas categóricas. La codificación One-Hot es un proceso mediante el cual las categorías de una variable categórica se convierten en nuevas columnas en el DataFrame, y se les asigna un valor de 1 o 0 dependiendo de si esa categoría es la correcta para el registro en cuestión. Esto se hace para permitir que nuestro modelo maneje estas variables categóricas.

Decidimos usar el modelo de Random Forest Regressor, que es un método de ensemble learning. En el aprendizaje de ensemble, combinamos varios algoritmos de aprendizaje automático para obtener un modelo más fuerte y más preciso. El Random Forest es un modelo que incorpora muchos árboles de decisión diferentes para hacer sus predicciones, y se considera que tiene un rendimiento excelente y es robusto ante el overfitting. Esto se debe a que el modelo se beneficia de la "sabiduría de la multitud"; en lugar de depender de las predicciones de un solo árbol de decisión, el Random Forest toma en cuenta las predicciones de todos los árboles dentro del bosque para hacer su predicción final. Esta técnica de

ensemble learning nos permite obtener un modelo que es capaz de capturar patrones complejos en los datos y proporciona una mayor precisión en las predicciones.

4.3.2 Cluster 1

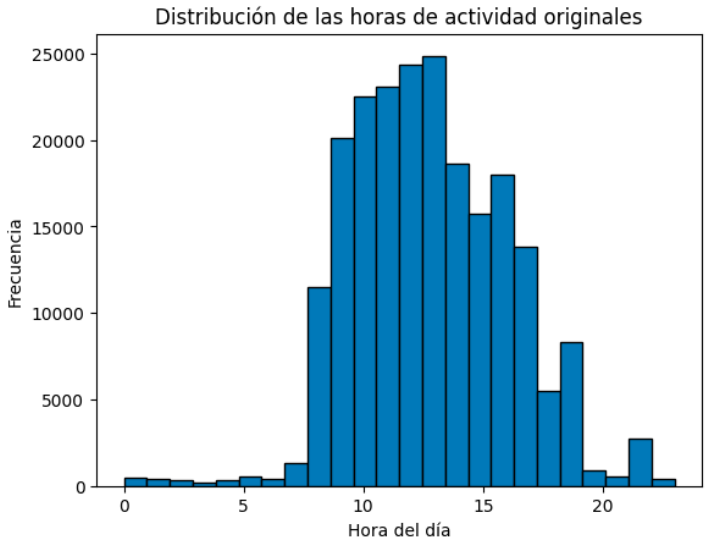

Ilustración 22 Distribución original cluster 1

La distribución original de este clúster revela un patrón de actividad notable que puede proporcionar una visión del comportamiento de trabajo de los usuarios intermedios. La actividad comienza a aumentar a las 7 de la mañana, posiblemente indicando el comienzo del día laboral para la mayoría. Esta actividad se incrementa hasta alcanzar un pico a la 1 de la tarde, lo cual puede representar un período de productividad máxima en medio de su jornada laboral.

No obstante, tras este pico de productividad, es apreciable una disminución significativa a partir de las 2 de la tarde, usualmente la hora del descanso para comer, hasta las 8 de la noche. Este período de descenso de actividad refleja posiblemente el agotamiento natural y la finalización de las tareas diarias que suelen acompañar al final de la jornada laboral.
Lo que resulta especialmente intrigante es el ligero aumento de actividad observado a las 10 de la noche. Este patrón indica que algunos usuarios podrían estar extendiendo su jornada laboral más allá del horario convencional. Este comportamiento podría estar motivado por plazos de trabajo ajustados, la necesidad de completar tareas pendientes o, quizás, una preferencia personal por trabajar durante horas nocturnas. En cualquier caso, este hallazgo sugiere la existencia de un compromiso y una flexibilidad notables entre los usuarios intermedios, que están dispuestos a trabajar más allá de las horas tradicionales para cumplir con sus responsabilidades.

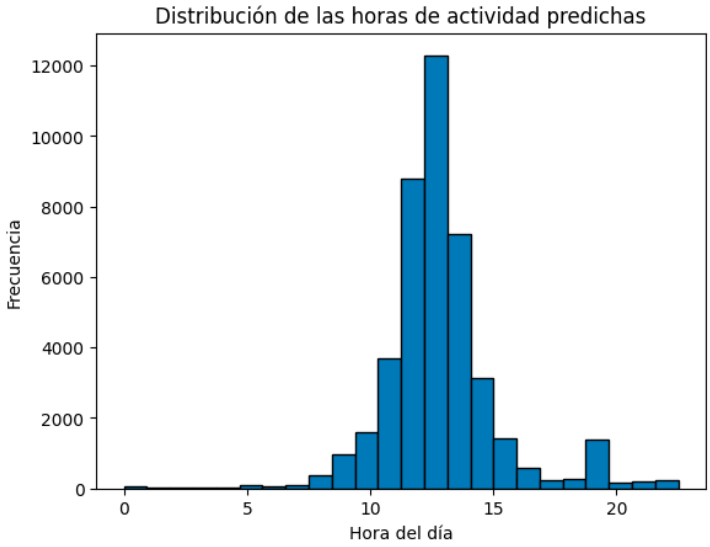

Ilustración 23 Distribución predicha cluster 1

En las predicciones generadas por nuestro modelo, se observa que los usuarios intermedios inician su actividad temprano, comenzando a las 7 de la mañana, con un incremento más gradual y simétrico de la actividad en comparación con los datos originales. Curiosamente, se identifica un cambio en el punto máximo de productividad, situándose entre las 12 y las 3 de la tarde, con un pico evidente a la 1 de la tarde. Similar a la distribución original, las predicciones reflejan una disminución en la actividad al acercarse las 8 de la noche, aunque se observa un pequeño pero notable aumento en este punto. Este comportamiento podría interpretarse como una señal de que algunos usuarios intermedios se esfuerzan en terminar tareas pendientes o preparar informes antes de finalizar su jornada.

En contraste con los datos originales, la actividad durante las horas fuera del horario de trabajo es prácticamente inexistente en nuestras predicciones. Este hecho sugiere que nuestro modelo anticipa una tendencia hacia un mayor equilibrio entre el trabajo y la vida personal entre los usuarios intermedios. En conclusión, el modelo predice un escenario donde los usuarios se esfuerzan durante la jornada laboral, pero también valoran y respetan su tiempo personal fuera del horario de trabajo.

4.3.3 Cluster 2

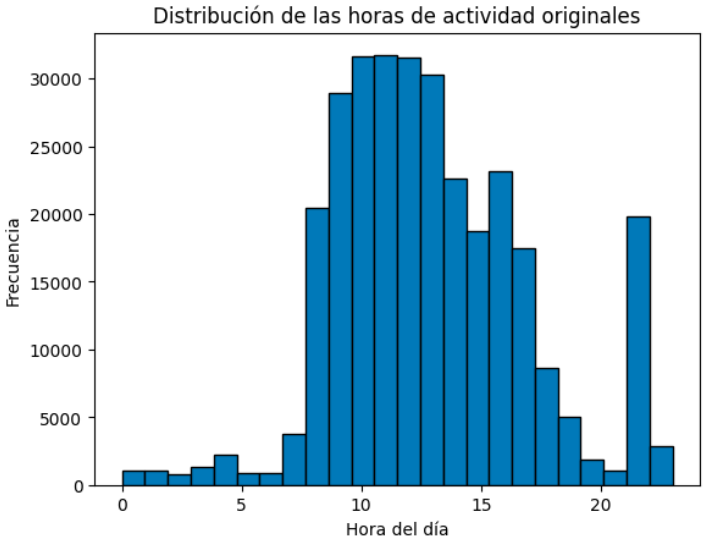

Ilustración 24 Distribución original cluster 2

Al examinar la distribución de la actividad en el clúster 2, tomando en cuenta que corresponde a usuarios de nivel básico. La actividad notable se inicia una hora más tarde en comparación con otros clústeres, lo que podría indicar un inicio de jornada menos exigente o la fase de aprendizaje en la que estos usuarios se encuentran para organizar y planificar eficientemente su día de trabajo. Observamos un declive uniforme de la actividad durante el día, lo que sugiere una distribución constante del trabajo.

Sin embargo, también podría ser indicativo de una menor urgencia o presión en sus roles, típica de posiciones más avanzadas. Notablemente, este clúster registra una alta cantidad de actividad fuera del horario de trabajo, con más actividad a las 21 horas que a las 17. Esto puede interpretarse de varias maneras. Por un lado, podría ser un testimonio de compromiso y dedicación, con usuarios dispuestos a trabajar más allá del horario convencional para cumplir con sus tareas. Sin embargo, también podría ser indicativo de desafíos en la gestión del tiempo o sobrecarga de trabajo. Este patrón invita a una consideración más profunda, sugiriendo que los gerentes o supervisores podrían necesitar explorar estrategias de apoyo para garantizar que estos empleados de nivel básico estén manejando sus responsabilidades de manera efectiva.

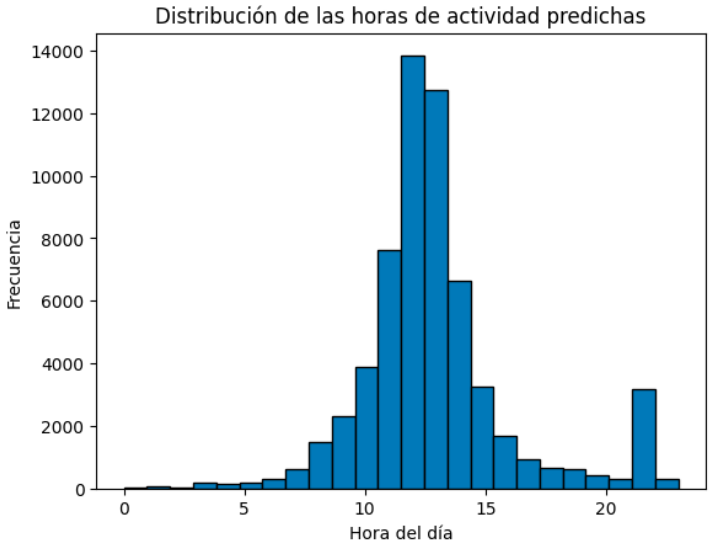

Ilustración 25 Distribución predicha cluster 2

En cuanto a las predicciones generadas por nuestro modelo para el clúster 2, se observa una distribución de la actividad más simétrica a lo largo del día. La actividad comienza en torno a las 7 y se prolonga hasta las 8, alcanzando su punto máximo a las 2 de la tarde. Este patrón puede indicar una asignación más eficiente de las tareas a lo largo del día, o quizás una mejor adaptación de estos usuarios de nivel básico a sus roles laborales y a la dinámica de su jornada laboral.

No obstante, lo que resulta especialmente interesante es la persistente actividad fuera de horas de trabajo. Se nota un volumen de actividad considerable a las 10 de la noche, superando incluso la actividad durante las horas de la tarde. Esto podría sugerir que los usuarios de nivel básico de este clúster todavía pueden estar enfrentando desafíos que les lleva a trabajar más allá del horario convencional. Esta es una observación importante que debería ser considerada al diseñar estrategias de apoyo y formación para estos usuarios.

4.3.4 Cluster 3

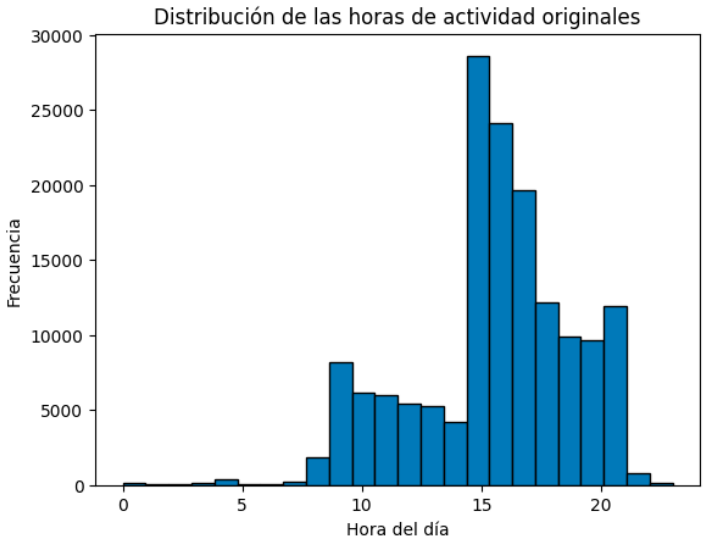

Ilustración 26 Distribución original cluster 3

El tercer clúster se distingue claramente de los otros en varios aspectos, y esto es posiblemente atribuible a su composición de roles directivos. La actividad comienza notablemente más tarde, siendo las 9 de la mañana el inicio de la jornada laboral en lugar del amanecer temprano que se observa en otros grupos. Una vez iniciada, la actividad se mantiene constante hasta las 3 de la tarde, momento en el cual se registra un aumento significativo, casi triplicando la actividad en comparación con las horas previas.

Dado que este clúster se compone en su mayoría de usuarios en roles de liderazgo, es posible que este aumento en la actividad esté relacionado con las tareas de revisión diaria, la toma de decisiones importantes o la gestión de pendientes de alto nivel. Esta es una característica singular de este grupo, resaltando la naturaleza intensiva de las responsabilidades de liderazgo durante el horario laboral.

Más allá de las 3 de la tarde, la actividad comienza a disminuir, aunque no desciende al nivel que se observa antes de las 3. Esto sugiere que, incluso después del pico de actividad, los roles directivos todavía tienen tareas importantes a realizar. Sin embargo, después de las 8 de la noche, se observa una disminución significativa en la actividad, señalando un respeto marcado por el horario de trabajo y posiblemente un énfasis en mantener un equilibrio saludable entre el trabajo y la vida personal. Esta es una observación crucial que demuestra cómo, a pesar de las responsabilidades y la presión inherentes a los roles de liderazgo, la importancia de respetar el tiempo personal sigue siendo primordial.

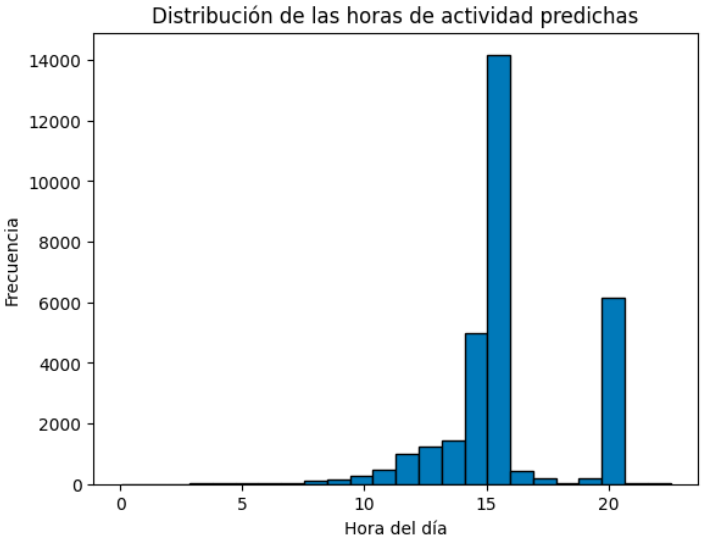

Ilustración 27 Distribución predicha cluster 3

En las predicciones generadas por nuestro modelo, se observa una dinámica de trabajo que comienza más tarde, a las 10 de la mañana. La actividad incrementa progresivamente hasta las 3 de la tarde, momento en el que se presenta un aumento significativo, multiplicando por más de dos veces la cantidad de trabajo realizado.

Sin embargo, lo que resalta en este patrón es el descenso abrupto después de las 3 de la tarde, con la actividad cayendo a niveles casi nulos. Este patrón sugiere que el modelo predice un enfoque de trabajo intensivo en un periodo de tiempo limitado para los roles directivos, seguido de un marcado periodo de calma en el que quizás se prioricen las tareas de bajo rendimiento o las reuniones de equipo.

No obstante, otro detalle interesante es el repunte que se observa a las 8 de la noche. Este incremento considerable puede interpretarse como un esfuerzo para finalizar tareas, hacer revisiones de última hora o preparar la agenda para el día siguiente. Finalmente, después de las 8 de la noche, la actividad se reduce a cero, subrayando una vez más el compromiso con el respeto del tiempo personal fuera del horario laboral. Estos patrones reflejan el equilibrio que los roles directivos tratan de mantener entre las demandas del trabajo y la necesidad de tiempo personal.

4.3.5 Cluster 4

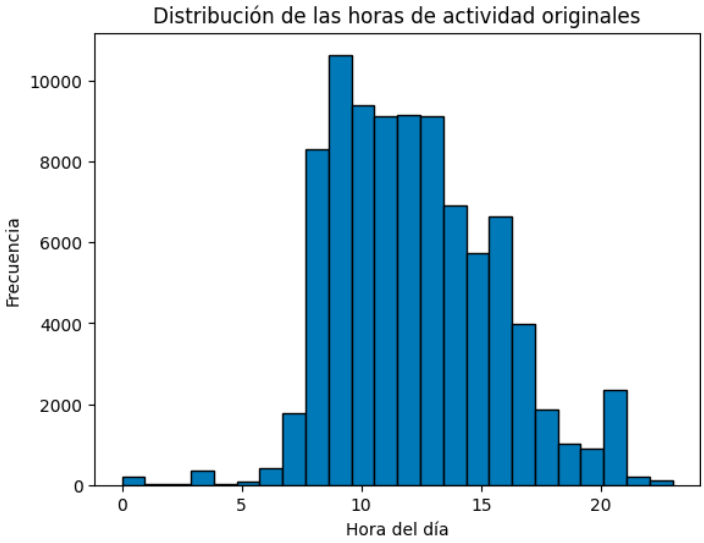

Ilustración 28 Distribución original cluster 4

En el cuarto y último clúster, la actividad laboral comienza temprano, a las 7 de la mañana, con un incremento significativo a las 8. La cúspide de actividad se da a las 9 de la mañana, momento en el cual quizás se realizan las tareas más cruciales del día o se llevan a cabo reuniones de seguimiento.

A partir de las 10, se aprecia un ligero descenso, pero la actividad se estabiliza y se mantiene constante hasta la 1 de la tarde, indicando un periodo sostenido de trabajo continuo. Esta actividad constante durante el transcurso de la mañana puede estar asociada con la naturaleza de los roles presentes en este clúster, que podrían incluir gerentes de proyecto, roles de servicio al cliente, roles de soporte técnico o coordinadores de eventos, todos los cuales requieren una alta capacidad de comunicación y coordinación.

Posteriormente, hay un descenso en la actividad a partir de las 2 y otro nuevamente a las 4, lo que podría corresponder al tiempo reservado para tareas de menor intensidad o para prepararse para el cierre del día. Finalmente, después de una reducción más significativa, a las 8 de la noche, se percibe un repunte, que posiblemente corresponde a tareas de finalización del día o preparación para el día siguiente.

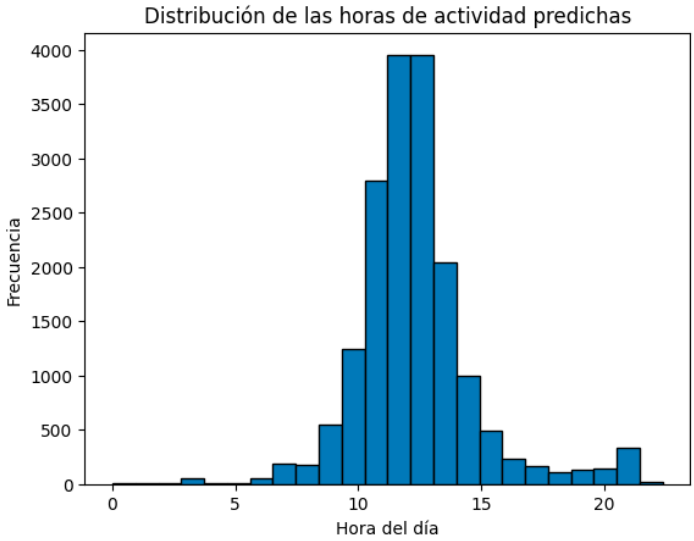

Ilustración 29 Distribución predicha cluster 4

En esta última predicción de nuestro modelo para el cuarto clúster, el cual se caracteriza por roles que requieren una comunicación y coordinación, se observa una distribución de actividad comenzando a las 7 de la mañana, la actividad aumenta gradualmente, lo cual puede representar la preparación y planificación de la jornada, la respuesta a comunicaciones pendientes del día anterior y la coordinación temprana con otros miembros del equipo o con clientes.

El pico de actividad se alcanza al mediodía, momento en que es posible que estas funciones de roles de coordinación y comunicación sean más demandantes, dada la necesidad de proporcionar actualizaciones, resolver consultas y mantener la cohesión del equipo o del proyecto.

La disminución constante de la actividad a partir de las 2 de la tarde puede ser indicativa de una transición hacia tareas de menor intensidad o la conclusión de las actividades del día. Aunque la actividad disminuye hasta las 8 de la noche, hay un pequeño repunte a las 9, similar al de la distribución original. Este repunte puede ser interpretado como el tiempo dedicado a finalizar las últimas tareas del día, tal vez revisar y responder correos electrónicos o preparar tareas para el día siguiente. Sin embargo, el bajo nivel de actividad después de las 9 indica un fuerte respeto por el equilibrio entre el trabajo y la vida personal, lo que es esencial en roles con una alta demanda de comunicación y coordinación para evitar el agotamiento y mantener la eficiencia.

5.- ANÁLISIS DE RESULTADOS

A lo largo de nuestro análisis, hemos llevado a cabo un profundo proceso de agrupación y predicción para comprender mejor los patrones de comportamiento de los usuarios de la empresa Galeo. Mediante el uso de técnicas de aprendizaje automático, como el K-Means para la agrupación y el Random Forest Regressor para la predicción, hemos conseguido describir y prever las tendencias de actividad de cuatro grupos distintos de usuarios.

El primer grupo, constituido por usuarios intermedios, mostró una actividad matutina temprana con un máximo alrededor de las 2 de la tarde y una disminución constante hacia el final de la jornada laboral, pero con un pequeño repunte de actividad nocturna. Nuestro modelo predijo una tendencia similar, con una mayor concentración de la actividad hacia el final de la jornada y una casi nula actividad fuera de horario laboral.

Para el segundo grupo, formado por usuarios de nivel básico, la actividad significativa se iniciaba una hora más tarde que el primer grupo, y el decrecimiento de actividad era más constante. Nuestro modelo predijo una tendencia de actividad más equilibrada a lo largo del día, aunque todavía con una significativa actividad fuera del horario de trabajo.

En el tercer grupo, donde se encontraban roles directivos, la actividad se mantenía bastante constante a lo largo del día, con un aumento significativo alrededor de las 3 de la tarde. El modelo predijo un patrón similar, aunque con una actividad casi nula después de las 3 de la tarde y un nuevo repunte a las 8 de la noche, reflejando quizás la necesidad de estos roles de liderazgo de terminar tareas al final del día.

Finalmente, el cuarto grupo, compuesto por roles que requieren una gran cantidad de comunicación y coordinación, mostró un pico temprano de actividad a las 9 de la mañana, seguido de un nivel de actividad bastante constante durante el día. El modelo predijo un patrón más simétrico, con una subida gradual hacia el mediodía y un descenso gradual hacia el final del día.

En resumen, nuestros hallazgos aportan un valioso entendimiento sobre cómo diferentes roles interactúan con la plataforma a lo largo del día y cómo estas interacciones pueden cambiar con el tiempo. Es importante destacar que estos patrones de actividad no son estáticos, y por lo tanto, las predicciones deberían ser revisadas y actualizadas regularmente para seguir siendo relevantes.

Sin embargo, estos hallazgos pueden servir como base para acciones futuras: desde la personalización de la interacción con la plataforma hasta la optimización de procesos laborales, lo que podría resultar en una mayor productividad y un mejor equilibrio entre el trabajo y la vida personal para los usuarios de la plataforma. A medida que continuamos recopilando y analizando datos, podremos afinar aún más estos modelos y proporcionar recomendaciones más personalizadas para cada tipo de usuario.

6.- CONCLUSIONES

En este proyecto, nos hemos embarcado en una exploración profunda y rigurosa del comportamiento de los usuarios de una plataforma de productividad, empleando técnicas avanzadas de aprendizaje automático y análisis de datos. Comenzamos nuestro viaje con un detallado análisis exploratorio de los datos, lo que nos permitió adquirir una visión clara de las características y comportamientos subyacentes de los usuarios. Este entendimiento fue crucial para la etapa de preprocesamiento y limpieza de los datos, donde transformamos y estructuramos nuestros datos para que fueran aptos para el modelado.

En la fase de modelado, implementamos un algoritmo de agrupación K-Means, que nos permitió segmentar a los usuarios en cuatro grupos distintos, cada uno con sus propias características y patrones de comportamiento. A través de esta segmentación, pudimos identificar roles distintivos entre los usuarios, desde los más básicos hasta los roles de alta gerencia, cada uno con sus propias dinámicas de trabajo y necesidades de comunicación.
La implementación del algoritmo Random Forest Regressor nos permitió crear un modelo predictivo robusto y efectivo, capaz de anticipar la actividad futura de los usuarios en la plataforma. Esta predicción proporciona información valiosa que puede utilizarse para optimizar la interacción y la experiencia del usuario, personalizando aún más la plataforma para satisfacer las necesidades específicas de cada grupo de usuarios.

Los aprendizajes adquiridos a lo largo de este proyecto son numerosos, pero uno de los más destacados es el valor de un análisis detallado y una segmentación precisa de los usuarios. Entender las diferencias entre los usuarios y su comportamiento en la plataforma ha demostrado ser fundamental para la creación de un modelo predictivo eficaz.

En conclusión, este proyecto ha sido un valioso viaje a través del análisis de datos, el aprendizaje automático y la comprensión del comportamiento del usuario. Los resultados obtenidos y las habilidades adquiridas en este proceso no sólo contribuyen a nuestro conocimiento en el campo de la ciencia de datos, sino que también abren nuevos caminos para la mejora continua y la personalización de las plataformas digitales.