COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

GRADO EN INGENIERÍA EN TECNOLOGÍAS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

PHENOMENOLOGICAL ANALYSIS OF NEURODIVERGENT PROFILES USING MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE ON VOICE DATA

Autor: Enrique Sanz Tur

Director: David Martín-Corral Calvo

Madrid

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

Phenomenological Analysis of Neurodivergent Profiles Using Machine Learning and Artificial Intelligence on Voice Data

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

curso académico 2024/25 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.: Enrique Sanz Tur Fecha: 13/06/2025

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

MARTIN CORRAL CALVO DAVID - 03909768K

Digitally signed by MARTIN CORRAL CALVO DAVID - 03909768K
Date: 2025.06.15 22:36:55 +02'00'

Fdo.: David Marín-Corral Calvo Fecha: 13/06/2025

COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

GRADO EN INGENIERÍA EN TECNOLOGÍAS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

PHENOMENOLOGICAL ANALYSIS OF NEURODIVERGENT PROFILES USING MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE ON VOICE DATA

Autor: Enrique Sanz Tur

Director: David Martín-Corral Calvo

Madrid

ANÁLISIS FENOMENOLÓGICO DE PERFILES NEURODIVERGENTES USANDO MACHINE LEARNING E INTELIGENCIA ARITIFICIAL EN DATOS DE VOZ

Autor: Sanz Tur, Enrique.
Director: Martín-Corral Calvo, David.
Entidad Colaboradora: ICAI – Universidad Pontificia Comillas

RESUMEN DEL PROYECTO

El presente Trabajo de Fin de Grado propone un enfoque basado en el análisis de voz y lenguaje mediante técnicas de inteligencia artificial y machine learning para la detección de perfiles neurodivergentes. Este proyecto, en colaboración con la start-up *Souly* creada por David Martín-Corral, utiliza las herramientas desarrolladas por la empresa para explorar la posibilidad de identificar patrones emocionales, lingüísticos y de personalidad en dichos perfiles. Este enfoque busca no solo avanzar en el diagnóstico asistido, sino también contribuir al desarrollo de tecnologías más inclusivas, accesibles y basadas en datos objetivos.

Palabras clave: Neurodivergencia, Inteligencia Artificial, Machine Learning, Análisis, Diagnóstico.

1. Introducción

La creciente visibilidad de la neurodivergencia en la sociedad ha impulsado nuevas formas de comprender y apoyar a personas con perfiles neurodivergentes. No obstante, la mayoría de métodos diagnósticos actuales se basan en entrevistas clínicas, cuestionarios o juicios subjetivos. Esto limita su objetividad, escalabilidad y aplicabilidad en contextos reales.

Frente a esta situación, surge la oportunidad de emplear la voz y el lenguaje como marcadores objetivos del estado emocional, rasgos de personalidad y posibles patrones neurodivergentes. El análisis de voz ha demostrado ser una fuente rica en información no verbal que puede reflejar matices emocionales, patrones de estrés o dificultades comunicativas, sin necesidad de intervención directa.

Este Trabajo de Fin de Grado propone un enfoque innovador que se apoya en herramientas de inteligencia artificial y algoritmos de machine learning para analizar grabaciones de voz y vídeo reales, extraídas principalmente de redes sociales como YouTube, TikTok, o Instagram. Estas grabaciones son procesadas a través de APIs avanzadas proporcionadas por la start-up Souly, con el objetivo de extraer características acústicas y lingüísticas que sirvan como base para modelos de clasificación automatizados.

El proyecto plantea una alternativa tecnológica, no invasiva y potencialmente accesible para la evaluación de perfiles neurodivergentes. Si bien no se pretende sustituir el diagnóstico clínico, se busca proporcionar un sistema que permita orientar mejor al profesional médico, reducir los tiempos de evaluación y facilitar la identificación temprana de casos, mejorando así la experiencia tanto del paciente como del especialista.

2. Definición del proyecto

El objetivo principal de este proyecto es identificar perfiles neurodivergentes, entrándose específicamente en el autismo, el TDAH y la dislexia, a partir del análisis automatizado de grabaciones reales de voz y vídeo. Estas grabaciones, obtenidas de plataformas como YouTube, TikTok, o Instagram, recogen a personas hablando directamente a cámara en contextos naturales, permitiendo así trabajar con datos espontáneos, accesibles y no intrusivos.

A partir de cada muestra, el sistema extrae una gran cantidad de parámetros provenientes de tres fuentes principales: voz (frecuencias, tono, ritmo, intensidad), expresión facial (emociones básicas como tristeza, sorpresa, enfado, etc.) y lenguaje (transcripción del discurso, polaridad, subjetividad, entidades, temas conversacionales). Esta información se obtiene mediante las APIs de Souly, proporcionadas por Souly, que permiten obtener para cada vídeo en un conjunto estructurado de métricas emocionales, lingüísticas y de personalidad.

Una vez recogidos todos estos datos, se lleva a cabo un análisis estadístico y computacional mediante algoritmos de machine learning. El objetivo no es solo clasificar perfiles, sino también determinar qué parámetros concretos son más relevantes o significativos para identificar neurodivergencia. Así, el proyecto busca generar un modelo explicativo que permita orientar procesos diagnósticos desde una perspectiva objetiva y basada en datos.

3. Descripción del modelo/sistema/herramienta

El sistema utilizado se apoya en una arquitectura basada en APIs que permite automatizar el análisis de datos de voz y texto y expresiones faciales. El sistema se basa principalmente en las APIs de Souly, que permiten analizar automáticamente grabaciones en vídeo para obtener datos sobre emociones, lenguaje y rasgos de personalidad. La figura siguiente resume visualmente este flujo de procesamiento distribuido y modular:

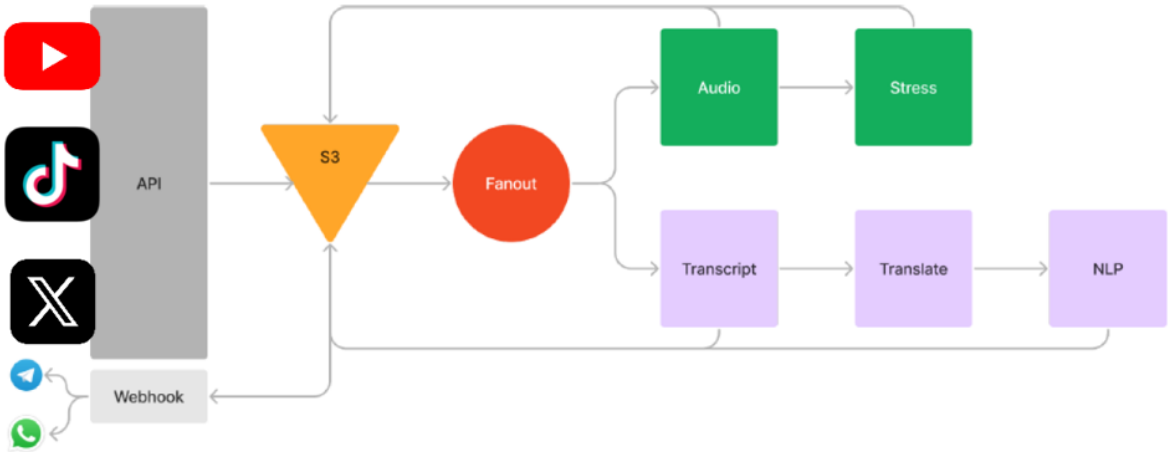


Ilustración 1 - Esquema de la arquitectura de la API de Souly para el procesamiento de datos

El proceso comienza con la subida de archivos .mp4, .wav o .ogg a un bucket en Amazon S3, que es un entorno de almacenamiento en la nube. Desde el bucket de S3, el contenido se distribuye a través de un módulo denominado fanout, que actúa como enrutador hacia distintas fases de análisis especializadas.

En primer lugar, el audio es sometido a procesos de limpieza y extracción de características acústicas (frecuencia media, desviación, tono o ritmo por ejemplo). Posteriormente, se calcula el nivel de estrés mediante modelos entrenados con etiquetas clínicas, y se realiza una transcripción del contenido verbal. Esta transcripción puede ser traducida si es necesario, y después analizada mediante modelos de procesamiento del lenguaje natural (NLP), que identifican entidades, emociones y temas de conversación.

Estos modelos NLP permiten extraer las características mencionadas, emociones presentes en el discurso, polaridad, subjetividad, temas tratados y tiempos verbales predominantes. Además, combinando las características de la voz y del texto, el sistema estima rasgos de personalidad como la autoestima, la compasión, la imaginación o la conciencia.

4. Resultados

Para evaluar el rendimiento de los modelos desarrollados, se realizaron diversos experimentos de clasificación y análisis explicativo. En la parte predictiva, se entrenaron y compararon distintos algoritmos supervisados como Random Forest, SVM y un perceptrón multicapa (MLP) utilizando como variable objetivo la condición neurodivergente (juntando los tres perfiles específicos de TDAH, autismo y dislexia) o de control.

	Precisión	Recall	F1 – Score	Support
Autism	0.90	0.77	0.83	419
Control	0.85	0.89	0.87	407
Dyslexia	0.89	0.92	0.90	405
ADHD	0.84	0.91	0.87	406
Accuracy		0.87		1637
Macro avg	0.87	0.87	0.87	1637
Weighted avg	0.87	0.87	0.87	1637

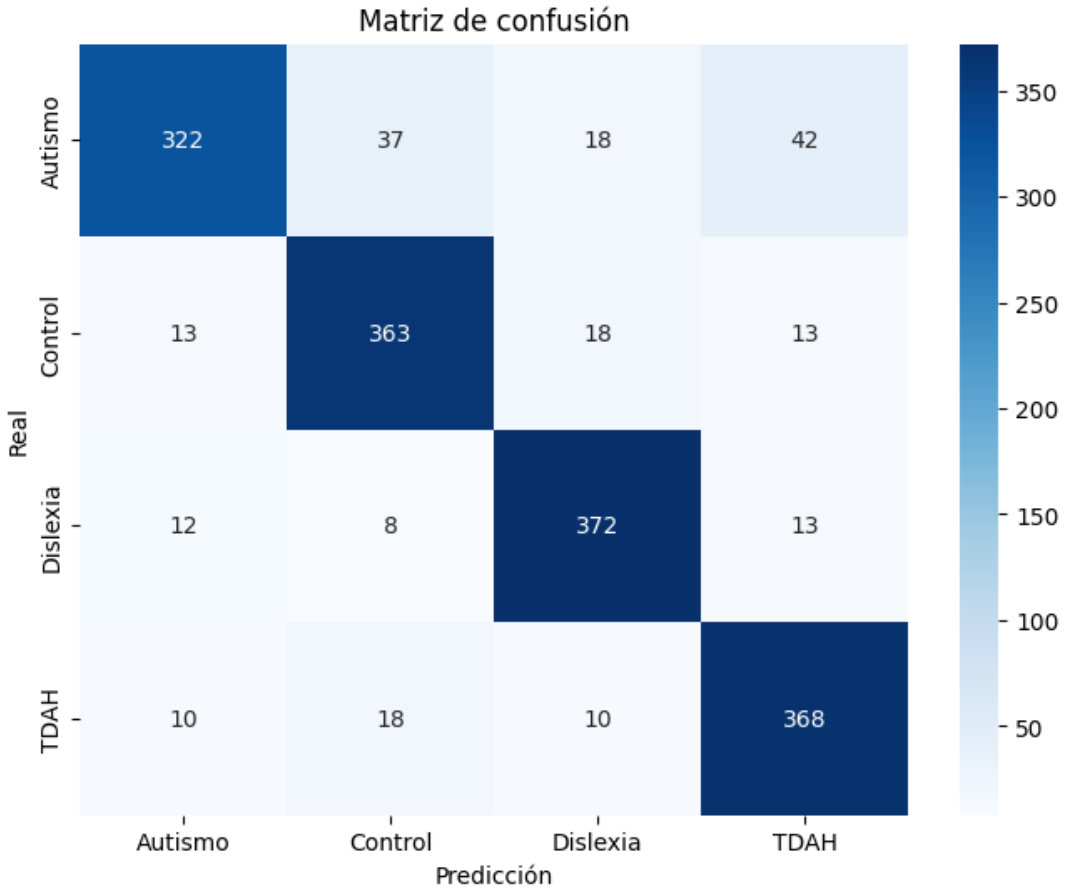


Ilustración 2 – Métricas y matriz de confusión del algoritmo de Random Forest.

Desde una perspectiva explicativa, se empleó una regresión logística utilizando tanto las variables numéricas como el contenido textual transcrito (representado con TF-IDF).

	Precisión	Recall	F1 - Score	Support
0 (control)	0.82	0.86	0.84	1532
1 (neurodiv.)	0.86	0.81	0.83	1549
Accuracy		0.84		3081
Macro avg	0.84	0.84	0.84	3081
Weighted avg	0.84	0.84	0.84	3081

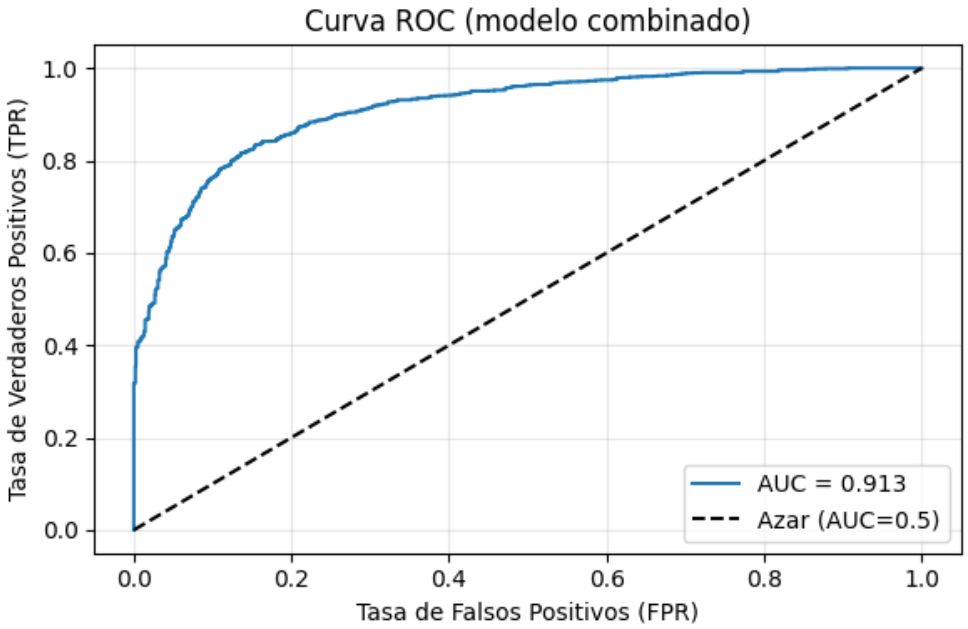


Ilustración 3 – Métricas y curva ROC del modelo combinado (regresión logística con datos numéricos + texto).

Esta combinación ofreció buenos resultados en clasificación binaria (control vs neurodivergente), con una AUC de 0.913 (ver Figura 3). Esta métrica indica un excelente poder discriminativo del modelo combinado.

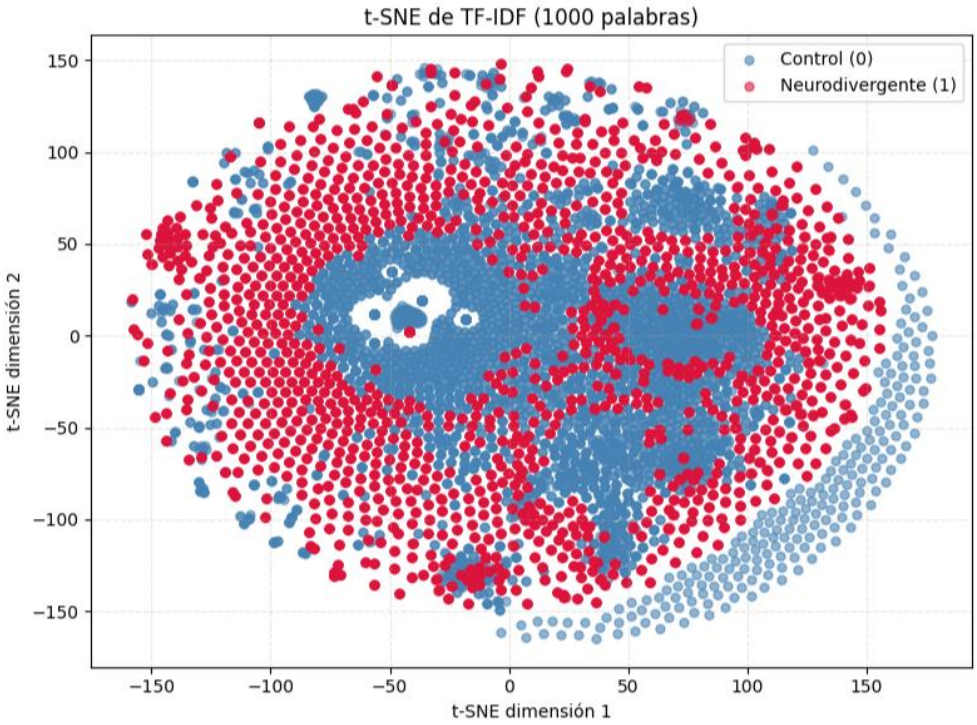


Ilustración 4 – Visualización t-SNE de las transcripciones usando TF-IDF (usando las 1000 palabras más relevantes).

Además, se aplicó la técnica de reducción de dimensionalidad t-SNE para visualizar la distribución espacial de los individuos a partir de las 1000 palabras más significativas del corpus. La Figura 4 muestra una clara tendencia a la agrupación entre los perfiles neurodivergentes y el grupo de control, lo que sugiere diferencias lingüísticas consistentes entre ambos colectivos.

5. Conclusiones

Este proyecto pretende poner a prueba el potencial del análisis automatizado de voz, lenguaje y expresión facial como herramienta objetiva para el estudio de perfiles neurodivergentes. A partir de vídeos reales procesados con APIs especializadas, se ha logrado analizar un conjunto de variables suficientemente ricas para entrenar modelos de machine learning con altos niveles de precisión.

Los resultados obtenidos con clasificadores como Random Forest, SVM o redes neuronales multicapa reflejan un rendimiento sólido, tal y como se ha comentado con anterioridad, con accuracies superiores al 85%. Esto valida el enfoque tanto en tareas de clasificación multiclase como binarias. Paralelamente, el uso de regresión logística ha permitido interpretar el peso específico de diferentes grupos de variables (acústicas, lingüísticas y emocionales), lo cual aporta gran valor desde una perspectiva explicativa para ver qué factores son los que más hubiera que tener en cuenta para detectar neurodivergencias.

Este tipo de tecnología podría integrarse fácilmente en entornos médicos o educativos como herramienta de apoyo al diagnóstico preliminar o al seguimiento de pacientes, mejorando la eficiencia del proceso clínico y facilitando la identificación temprana de casos. Además, el enfoque es compatible con soluciones digitales como la plataforma web de Souly (https://www.mysouly.com/en/), donde estos modelos podrían implementarse en aplicaciones accesibles, escalables y no invasivas para usuarios finales.

6. Referencias

[1] Malgaroli, M., Hull, T. D., Zech, J. M., & Althoff, T. “Natural language processing for mental health interventions: A systematic review and research framework”, Translational Psychiatry, 13(1), 309, 2023. https://doi.org/10.1038/s41398-023-02592-2

[2] Higuchi, M., Nakamura, M., Shinohara, S., Omiya, Y., Takano, T., Mitsuyoshi, S., & Tokuno, S. “Effectiveness of a Voice-Based Mental Health Evaluation System for Mobile Devices: Prospective Study”, JMIR Formative Research, 4(7), e16455, 2020. https://doi.org/10.2196/16455

[3] Souly API Documentation. “Análisis de voz y personalidad mediante procesamiento de lenguaje natural”, Souly, 2024. https://www.mysouly.com/en/

PHENOMENOLOGICAL ANALYSIS OF NEURODIVERGENT PROFILES USING MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE ON VOICE DATA

Author: Sanz Tur, Enrique.
Supervisor: Martín-Corral Calvo, David.
Collaborating Entity: ICAI – Universidad Pontificia Comillas

ABSTRACT

This Final Degree Project proposes an approach based on voice and language analysis using artificial intelligence and machine learning techniques to detect neurodivergent profiles. Developed in collaboration with the start-up Souly, founded by David Martín-Corral, the project leverages the company's proprietary tools to explore the identification of emotional, linguistic, and personality-related patterns within these profiles. The approach aims not only to advance assisted diagnostic processes but also to contribute to the development of more inclusive, accessible, and data-driven technologies.

Keywords: Neurodivergence, Artificial Intelligence, Machine Learning, Analysis, Diagnosis.

1. Introduction

The growing visibility of neurodivergence in society has fostered new ways of understanding and supporting individuals with neurodivergent profiles. However, most current diagnostic methods rely on clinical interviews, questionnaires, or subjective judgments, which limits their objectivity, scalability, and applicability in real-world contexts.

In light of this, there is a valuable opportunity to use voice and language as objective markers of emotional state, personality traits, and potential neurodivergent patterns. Voice analysis has proven to be a rich source of non-verbal information capable of reflecting emotional nuances, stress patterns, or communication difficulties—without requiring direct intervention.

This Final Degree Project proposes an innovative approach based on artificial intelligence tools and machine learning algorithms to analyze real voice and video recordings, primarily collected from social media platforms such as YouTube, TikTok, and Instagram. These recordings are processed through advanced APIs provided by the start-up Souly, with the goal of extracting acoustic and linguistic features that serve as the foundation for automated classification models.

The project presents a technological, non-invasive, and potentially accessible alternative for assessing neurodivergent profiles. While it is not intended to replace clinical diagnosis, the system aims to better support medical professionals, reduce evaluation times, and facilitate early identification of cases, ultimately improving the experience for both patients and specialists.

2. Definition of the project

The main objective of this project is to identify neurodivergent profiles, specifically focusing on autism, ADHD, and dyslexia, through the automated analysis of real voice and video recordings. These recordings, sourced from platforms such as YouTube, TikTok, and Instagram, feature individuals speaking directly to the camera in natural settings, thus providing spontaneous, accessible, and non-intrusive data.

From each sample, the system extracts a wide range of parameters from three primary sources: voice (frequency, tone, rhythm, intensity), facial expression (basic emotions such as sadness, surprise, anger, etc.), and language (speech transcription, polarity, subjectivity, named entities, and conversational topics). This information is obtained using Souly's APIs, provided by Souly, which process each video into a structured set of emotional, linguistic, and personality-related metrics.

Once all this data is collected, a statistical and computational analysis is performed using machine learning algorithms. The goal is not only to classify profiles, but also to determine which specific parameters are the most relevant or significant for identifying neurodivergence. In this way, the project aims to build an explanatory model that can support diagnostic processes from an objective, data-driven perspective.

3. Description of the model/system/tool

The system is built upon an API-based architecture that enables the automated analysis of voice, text, and facial expression data. It relies primarily on Souly's APIs, which are used to process video recordings to extract information related to emotions, language, and personality traits. The following figure provides a visual summary of this modular and distributed processing workflow:

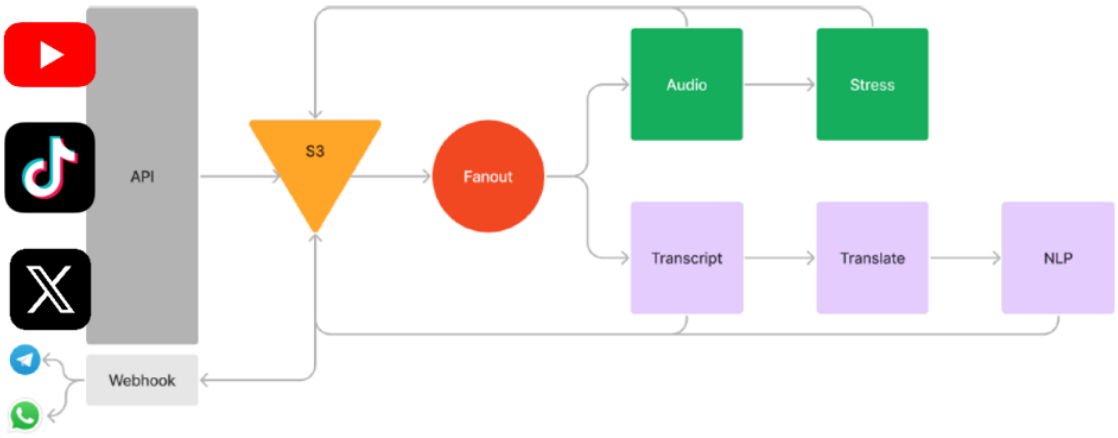


Figure 2 – Diagram of Souly's API architecture for data processing

The process begins with the upload of .mp4, .wav, or .ogg files to an Amazon S3 bucket, which serves as a cloud storage environment. From the S3 bucket, the content is routed through a fanout module, which distributes the input across various specialized analysis stages.

First, the audio undergoes cleaning and acoustic feature extraction (such as mean frequency, standard deviation, tone, or rhythm). Then, stress levels are calculated using models trained with clinically labeled data, and the spoken content is transcribed. If needed, the transcription can be translated and subsequently processed by Natural Language Processing (NLP) models that identify entities, emotions, and conversational topics.

These NLP models extract the previously mentioned features, including the emotional tone of the discourse, polarity, subjectivity, main topics, and predominant verb tenses. In addition, by combining vocal and textual features, the system estimates personality traits such as self-esteem, compassion, imagination, and awareness.

4. Results

To evaluate the performance of the developed models, several classification and explanatory experiments were conducted. On the predictive side, various supervised algorithms, such as Random Forest, SVM, and a Multilayer Perceptron (MLP) were trained and compared, using as the target variable the neurodivergent condition (combining the three specific profiles: ADHD, autism, and dyslexia) versus the control group.

	Precision	Recall	F1 – Score	Support
Autism	0.90	0.77	0.83	419
Control	0.85	0.89	0.87	407
Dyslexia	0.89	0.92	0.90	405
ADHD	0.84	0.91	0.87	406
Accuracy		0.87		1637
Macro avg	0.87	0.87	0.87	1637
Weighted avg	0.87	0.87	0.87	1637

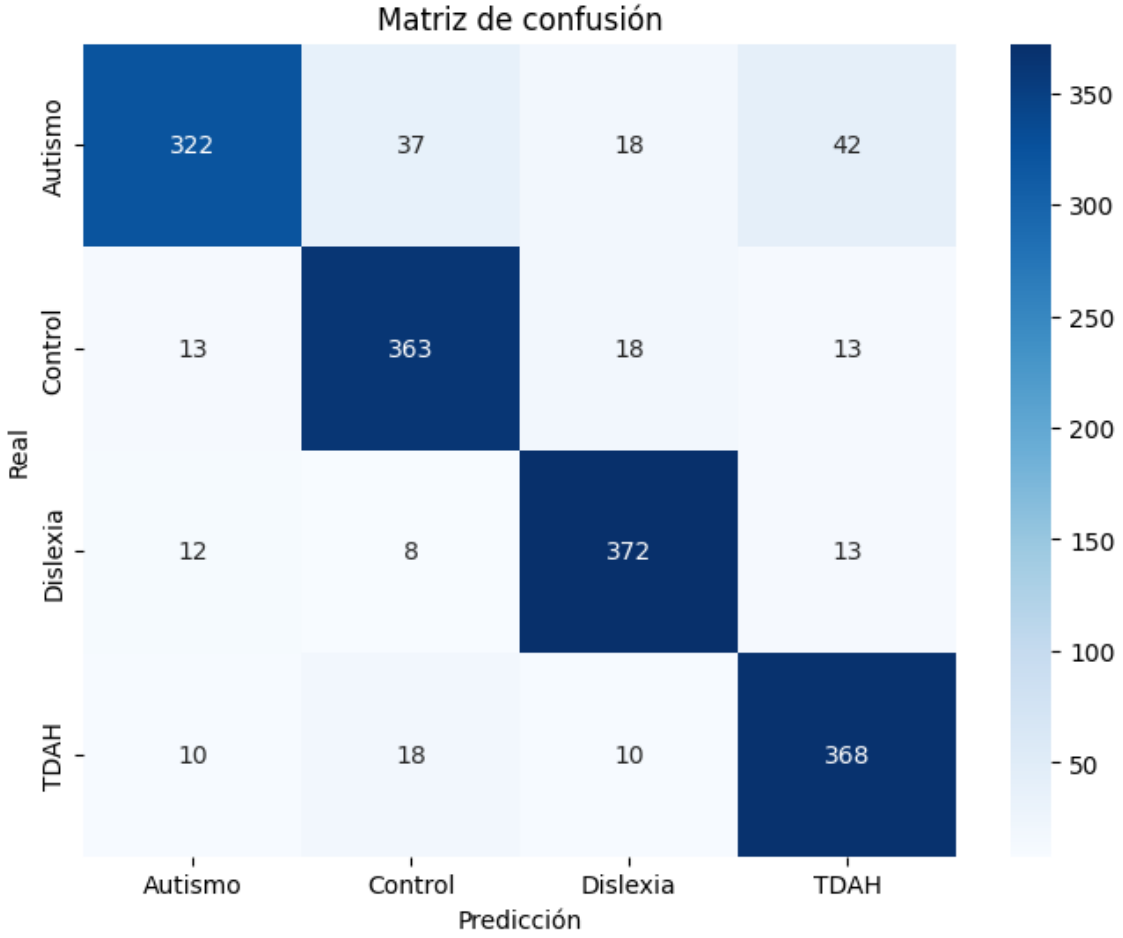


Figure 2 – Confusion matrix and performance metrics of the Random Forest model.

All trained models showed strong performance, with accuracy scores above 85% and recall generally exceeding 80% across all evaluated classes. In particular, the Random Forest algorithm achieved a global accuracy of 87%, as well as high macro and weighted averages. Figure 2 presents the confusion matrix along with detailed per-class performance metrics. The model showed balanced results across the profiles of autism, ADHD, dyslexia, and control, with F1-scores ranging from 0.83 to 0.90, confirming its reliability for multiclass classification tasks in this context.

From an explanatory perspective, a logistic regression model was used, incorporating both numerical variables and the transcribed textual content (represented using TF-IDF).

	Precision	Recall	F1 - Score	Support
0 (control)	0.80	0.85	0.82	1532
1 (neurodiv.)	0.84	0.79	0.82	1549

Accuracy	0.82			3081
Macro avg	0.82	0.82	0.82	3081
Weighted avg	0.82	0.82	0.82	3081

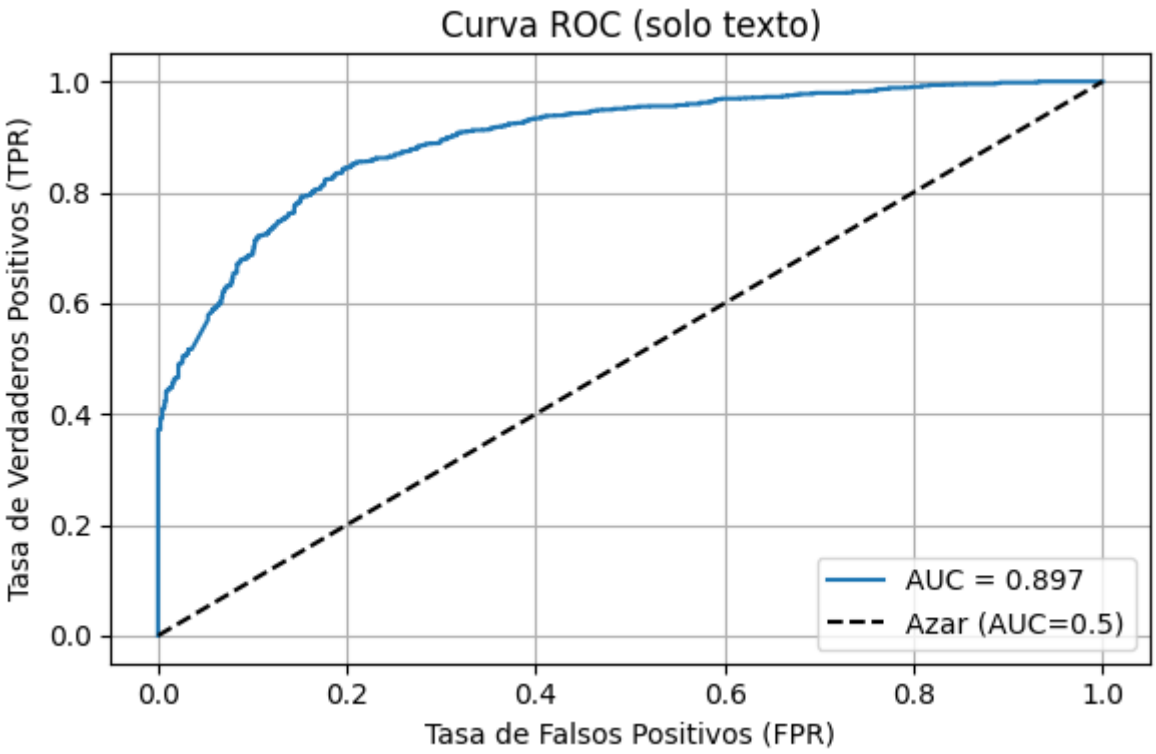


Figure 3 – Metrics and ROC curve of the combined model (logistic regression with numerical data + text).

This combined approach yielded strong results for binary classification (control vs. neurodivergent), with an AUC of 0.913 (see Figure 3), indicating excellent discriminative power.

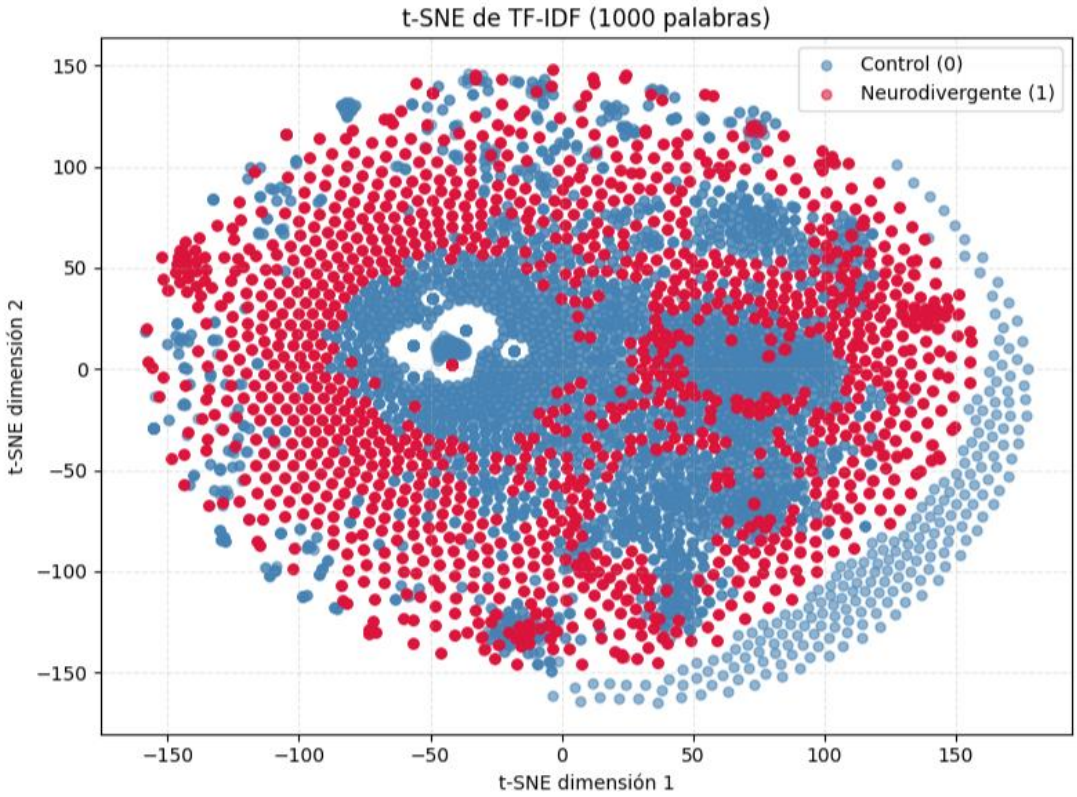


Figure 4 – t-SNE visualization of transcriptions using TF-IDF (based on the 1,000 most relevant words).

Additionally, the t-SNE dimensionality reduction technique was applied to visualize the spatial distribution of individuals based on the 1,000 most significant words from the corpus. Figure 4 shows a clear clustering tendency between neurodivergent profiles and the control group, suggesting consistent linguistic differences between the two populations.

5. Conclusions and insights

This project aims to explore the potential of automated analysis of voice, language, and facial expressions as an objective tool for studying neurodivergent profiles. By processing real videos through specialized APIs, it has been possible to analyze a sufficiently rich set of variables to train machine learning models with high levels of accuracy.

The results obtained with classifiers such as Random Forest, SVM, and multilayer neural networks reflect strong performance, as previously discussed, with accuracies above 85%. This validates the approach for both multiclass and binary classification tasks. In parallel, the use of logistic regression has enabled the interpretation of the relative importance of different variable groups (acoustic, linguistic, and emotional), which adds significant value from an explanatory perspective, highlighting which factors are most relevant for detecting neurodivergence.

This type of technology could be easily integrated into medical or educational environments as a support tool for preliminary diagnosis or patient monitoring, improving the efficiency of clinical workflows and facilitating early case identification. Moreover, the approach is fully compatible with digital solutions such as the Souly web platform (https://www.mysouly.com/en/), where these models could be implemented in accessible, scalable, and non-invasive applications for end users.

6. References

[1] Malgaroli, M., Hull, T. D., Zech, J. M., & Althoff, T. "Natural language processing for mental health interventions: A systematic review and research framework", Translational Psychiatry, 13(1), 309, 2023. https://doi.org/10.1038/s41398-023-02592-2

[2] Higuchi, M., Nakamura, M., Shinohara, S., Omiya, Y., Takano, T., Mitsuyoshi, S., & Tokuno, S. "Effectiveness of a Voice-Based Mental Health Evaluation System for Mobile Devices: Prospective Study", JMIR Formative Research, 4(7), e16455, 2020. https://doi.org/10.2196/16455

[3] Souly API Documentation. "Análisis de voz y personalidad mediante procesamiento de lenguaje natural", Souly, 2024. https://www.mysouly.com/en/

Índice de la memoria

Índice de figuras

Índice de tablas

Capítulo 1. INTRODUCCIÓN

1.1 CONTEXTO Y MOTIVACIÓN DEL PROYECTO

En los últimos años, el término neurodivergencia ha adquirido una mayor presencia en los ámbitos científico, clínico y social. Este concepto hace referencia a formas diversas de procesar la información, percibir el entorno o interactuar con los demás. La neurodivergencia engloba perfiles como el Trastorno del Espectro Autista (TEA), el Trastorno por Déficit de Atención e Hiperactividad (TDAH) y la dislexia. En España, por ejemplo, el número de diagnósticos de TEA se ha cuadruplicado en la última década [3], pasando de 19.023 alumnos en el curso 2011-2012 a más de 78.000 en el curso 2022-2023, según datos del Ministerio de Educación y la Confederación Autismo España. De forma similar, los diagnósticos de TDAH han aumentado un 30% en los últimos seis años [2], [3], reflejando una mayor concienciación social y avances en los procesos de detección.

No obstante, este aumento en el número de diagnósticos ha puesto en evidencia varias limitaciones del sistema actual. La falta de personal especializado y la elevada carga asistencial provocan tiempos de espera que pueden superar un año en la sanidad pública. Además, los métodos de evaluación clínica empleados (principalmente entrevistas, cuestionarios y observaciones estructuradas), si bien son herramientas validadas, dependen en gran medida del juicio subjetivo del profesional. Esto introduce posibles sesgos y limita su escalabilidad, accesibilidad y objetividad en determinados contextos [1], [2].

Este proyecto surge precisamente como respuesta a esa necesidad. Busca explorar nuevas formas de apoyo diagnóstico que permitan reducir tiempos de espera, aumentar la precisión y facilitar la intervención temprana desde una perspectiva objetiva basada en datos.

Mi interés en este campo se debe tanto a razones académicas como personales. Durante la búsqueda de un tema para mi Trabajo de Fin de Grado, conocí al profesor David Martín-Corral, fundador de la startup Souly, que se especializa en el uso de inteligencia artificial

para el análisis de la salud mental. Su proyecto me pareció especialmente relevante y alineado con mis intereses, no solo por el enfoque técnico, sino también por mi vínculo familiar con el ámbito de la psiquiatría. Mis dos padres se dedican profesionalmente a ello, y esto me ha permitido entender desde cerca las dificultades reales del diagnóstico en salud mental. Desde ese punto de partida, vi la posibilidad de desarrollar un proyecto con impacto social, combinando el potencial de la inteligencia artificial con una problemática actual.

1.2 ENFOQUE Y PLANTEAMIENTO GENERAL

El proyecto se basa en el análisis automatizado de voz, lenguaje y expresión facial, con el objetivo de identificar patrones característicos de perfiles neurodivergentes a partir de vídeos de personas hablando a cámara. Estas grabaciones, obtenidas de plataformas como YouTube, TikTok o Instagram, se procesan mediante herramientas proporcionadas por Souly, como las propias APIs de Souly [5], que permiten transformar cada archivo de vídeo en un conjunto estructurado de variables emocionales, lingüísticas y acústicas.

El sistema extrae métricas como nivel de estrés, frecuencia vocal, emociones predominantes, polaridad del discurso, subjetividad, temas conversacionales y rasgos de personalidad. Esta información se organiza en datasets que sirven como base para entrenar modelos supervisados de clasificación, entre los que se han utilizado algoritmos como Random Forest, SVM, redes neuronales multicapa y técnicas explicativas como regresión logística o visualización con t-SNE. Los resultados obtenidos superan el 85% de accuracy y han demostrado ser consistentes y robustos en la clasificación entre perfiles neurodivergentes y control.

1.3 JUSTIFICACIÓN CLÍNICA, TECNOLÓGICA Y EDUCATIVA

Desde una perspectiva clínica, esta propuesta presenta una posible solución complementaria para sistemas de salud saturados, ofreciendo herramientas objetivas para mejorar la precisión diagnóstica en fases tempranas. Permite reducir la dependencia de criterios subjetivos y avanzar hacia procesos más rápidos, medibles y escalables. Su integración en entornos

clínicos o educativos puede mejorar la calidad asistencial y facilitar la intervención en etapas cruciales del desarrollo.

Tecnológicamente, el proyecto demuestra la aplicabilidad de técnicas de machine learning e inteligencia artificial a una problemática compleja, ética y socialmente sensible. La plataforma web desarrollada por Souly ya permite el análisis automático de grabaciones y plantea un modelo escalable, no invasivo y accesible que puede extenderse tanto a centros clínicos como a usuarios individuales.

Finalmente, el enfoque educativo del sistema no debe subestimarse. Los datos generados y el análisis automatizado pueden servir como herramienta formativa para futuros profesionales, estudiantes de psicología, medicina o educación, y contribuir a una mayor comprensión, al igual que visibilizar la diversidad cognitiva como parte de plataformas interactivas en cursos de salud mental o neuropsicología.

Capítulo 2. DESCRIPCIÓN DE LAS TECNOLOGÍAS

En este capítulo se describirán las tecnologías principales utilizadas en el desarrollo del sistema de análisis automatizado de voz y texto aplicado a perfiles neurodivergentes. El trabajo se apoya en dos APIs desarrolladas por Souly. A continuación, se detalla la funcionalidad, arquitectura y papel que desempeña cada una dentro del sistema:

2.1 API PARA EL PROCESAMIENTO DE ARCHIVOS

La primera API de Souly constituye el núcleo del sistema de análisis. Esta herramienta transforma grabaciones de audio y vídeo en un conjunto estructurado de métricas que describen el estado emocional, lingüístico y de personalidad del hablante. Su diseño permite una evaluación objetiva de variables psicológicas a partir de la voz.

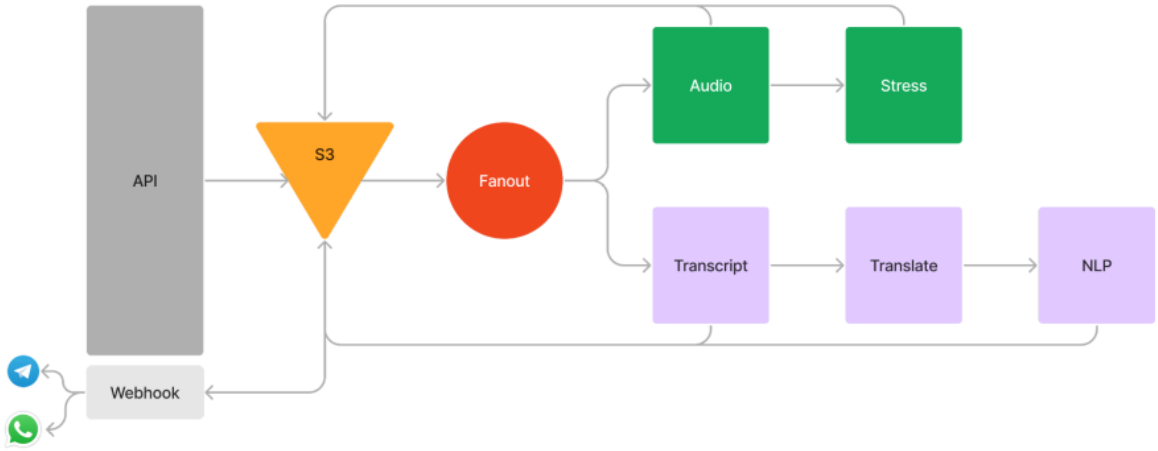


Figura 1 - Esquema de la arquitectura de la API de Souly para el procesamiento de archivos

2.1.1 ARQUITECTURA TÉCNICA

El funcionamiento del sistema sigue una arquitectura modular que automatiza la cadena completa de análisis:

1. **Subida del archivo:** El archivo de entrada (en formato .mp4, .ogg o .wav) se almacena en un "bucket" de Amazon S3. Amazon S3 es un servicio de almacenamiento en la nube proporcionado por Amazon Web Services (AWS) , y un "bucket de Amazon S3 actúa como un contenedor lógico donde se organizan los archivos y permite su acceso posterior por parte de otros servicios o aplicaciones.
2. **Webhook y fanout:** Una vez almacenado el archivo, se activa un webhook, que es una notificación automática enviada por el sistema para indicar que un evento (en este caso, la subida del archivo) ha sucedido. Esta notificación desencadena el inicio del procesamiento. Posteriormente, entra en juego el componente fanout, que se encarga de distribuir la información del archivo hacia diferentes módulos especializados de análisis, como el de voz o el de lenguaje, permitiendo así un procesamiento en paralelo más eficiente.
3. **Procesamiento distribuido:** El audio es limpiado y procesado por modelos de tratamiento de señal que permiten mejorar la calidad de los inputs mediante normalización, detección de ruido y análisis del silencio. Posteriormente, se extraen características vocales como la frecuencia fundamental (F_0), ritmo, intensidad, y patrones de entonación. Esta información se utiliza como entrada para modelos de procesamiento de voz, los cuales detectan marcadores acústicos relacionados con el estrés, la carga emocional y la variabilidad prosódica (es decir, los cambios en la entonación, el ritmo y la intensidad de la voz a lo largo del habla, que permiten identificar diferencias expresivas o patrones atípicos).

Una vez procesado el audio, se realiza la transcripción automática del contenido hablado. Esta transcripción puede ser traducida al inglés si el idioma original es otro, y es posteriormente analizada por modelos de procesamiento del lenguaje natural (NLP). Estos

modelos permiten detectar la polaridad emocional, subjetividad, temas conversacionales, y métricas de coherencia semántica.

2.1.1.1 Métricas y análisis

El sistema proporciona métricas como:

- Nivel de estrés, depresión y autoeficacia (evaluados mediante modelos clínicos).
- Emociones detectadas tanto en la voz como el texto.
- Análisis de polaridad y subjetividad del discurso.
- Identificación de entidades y temas conversacionales.
- Estimación de rasgos de personalidad: autoestima, compasión, creatividad, etc.

Todos estos datos son devueltos mediante el output API, que incluye además información sobre latencia, duración de las muestras, formato del audio, idioma detectado y probabilidades de silencio o entropía textual.

2.2 API PARA LA SUBIDA DE ARCHIVOS

La segunda herramienta empleada es otra API proporcionada por Souly, cuyo propósito principal es permitir la subida automatizada de archivos de vídeo a la plataforma de análisis. Esta funcionalidad es esencial en el sistema, ya que permite alimentar el sistema con grandes cantidades de datos de redes sociales, siendo las principales fuentes YouTube, TikTok o Instagram.


```
C:\Users\Enrique Sanz Tur\Desktop\TFG>curl -X GET "https://heratropic-main-c6ba0ae.d2.zuplo.dev/v1/result/394ea0bd-1455-4c65-bf55-c740fe8ec82f" -H "Authoriz
ation: Bearer zpka_ed56a7576f47465095a2f3ee1d08c2ca_56c05972"
{"status":"success","response":{"created_at":1742669583,"aid":"394ea0bd-1455-4c65-bf55-c740fe8ec82f","result_url":"/v1/result/394ea0bd-1455-4c65-bf55-c740fe
8ec82f","original_file":{"extension":".mp4","format":"video","duration":169},"status":{"FILE_STORED":true,"FACIAL_ANALYSED":true,"VOICE_ANALYSED":true,"VOIC
E_TRANSCRIBED":false,"BIOMETRICS_EXTRACTED":true,"SPEECH_ANALYSED":false,"PERSONALITY_ANALYSED":true,"FACES_EXTRACTED":true},"external_vars":{"id":"1"},"dat
a":{"facial":{"average_emotions":{"angry":0.0324,"disgust":0.0011,"fear":0.0001,"happy":0.0009,"sad":0.0,"surprise":0.9654,"neutral":0.0001},"most_frequent_
dominant_emotion":"surprise","dominant_emotion_counts":{"surprise":1},"average_face_confidence":0.0093},"traits":{"extraversion":0.4043,"neuroticism":0.4664
,"agreeableness":0.4678,"conscientiousness":0.4561,"openness":0.4743,"survival":0.28380000591278076,"creativity":0.18369999527931213,"self_esteem":0.1134999
9904632568,"compassion":0.20579999685287476,"communication":0.32580000162124634,"imagination":0.2953000068664551,"awareness":0.22939999401569366,"stress":{"
high":0.7538054585456848,"medium":0.27349480986595154,"low":0.0891864076256752},"helplessness":{"high":0.7327077388763428,"medium":0.24481886625289917,"low"
:0.1151895746588707},"self_efficacy":{"medium":0.48075833916664124,"low":0.2958986163139343,"high":0.22542314231395721},"depression":{"high":0.7900807261466
98,"medium":0.19586016237735748,"low":0.07875201851129532}},"voice":{"frequencies":{"mean":2551,"sd":3076,"median":1205,"mode":554,"Q25":494,"Q75":3534,"IQR
":3041,"skewness":7.190000057220459,"kurtosis":82.9000015258789,"mean_note":"D#","median_note":"D","mode_note":"C#","Q25_note":"B","Q75_note":"A","rmse":0.0
4919999837875366},"pitch":512,"tone":1959,"emotions":{"sad":0.3802557587623596,"disgust":0.33326825499534607,"fearful":0.08544372767210007,"neutral":0.07614
624500274658,"happy":0.04327226057648659,"angry":0.034562282264232635,"calm":0.024014070630073547,"surprised":0.023037387058138847}}}}}
```


Figura 2 - Ejemplo del método GET para obtener los resultados de la subida de un vídeo

2.2.1 MÉTODO DE OPERACIÓN

La API proporciona dos endpoints para la subida de archivos:

- Para archivos de menos de 5MB, se utiliza una URL directa.
- Para archivos mayores de 5MB, se solicita una URL preasignada mediante un primer método POST, tras lo cual se realiza la carga del fichero al bucket en S3 mediante un método PUT.

En ambos casos, el sistema devuelve un identificador único del archivo y una URL para consultar los resultados una vez procesado.

2.2.2 AUTENTICACIÓN Y METADATOS

El acceso está protegido mediante API Keys, y el sistema permite enviar también variables externas como ID del canal o metadatos adicionales. Esto facilita la organización y trazabilidad de los datos en estudios posteriores.

Aunque esta API no realiza el análisis directamente, constituye un componente clave para la integración fluida entre los datos recogidos y la posterior inferencia realizada por la otra API.

Ambas APIs, utilizadas de forma complementaria, permiten ayudar en la automatización de principio a fin del proceso de recogida, análisis y explotación de datos, ofreciendo una base técnica robusta para desarrollar modelos diagnósticos objetivos más adelante.

Capítulo 3. ESTADO DE LA CUESTIÓN

En los últimos años, la inteligencia artificial y el procesamiento del lenguaje natural (NLP) se han consolidado como una herramienta prometedora en el ámbito de la salud mental. Gracias al crecimiento masivo en el uso de dispositivos móviles y la mejora en algoritmos de aprendizaje automático, se ha abierto la posibilidad de analizar patrones de voz, lenguaje y expresión emocional para obtener indicadores de estados psicológicos.

En este contexto, se ha impulsado una nueva generación de soluciones tecnológicas orientadas a la monitorización y evaluación no invasiva del bienestar mental, tanto desde la esfera clínica como desde la investigación académica.

3.1 SOLUCIONES EXISTENTES EN EL MERCADO

3.1.1 VOCALIS HEALTH

Vocalis Health es una empresa centrada en el análisis de biomarcadores vocales con fines diagnósticos, especialmente patologías físicas. Durante la pandemia de COVID-19, esta tecnología fue utilizada para detectar patrones respiratorios anómalos en la voz. Sin embargo, su enfoque se aleja de la evaluación psicológica profunda. Vocalis Health no proporciona análisis semántico, ni detecta emociones o indicadores de personalidad. Su sistema está orientado a la detección de condiciones como la fibrosis pulmonar o insuficiencia cardíaca, basándose principalmente en parámetros físicos de la voz [11].

3.1.2 KINTSUGI

Kintsugi es una startup estadounidense que utiliza inteligencia artificial para detectar niveles de ansiedad y depresión mediante el análisis del tono y el contenido del habla. Su aplicación móvil permite realizar grabaciones de voz cortas, que son analizadas en tiempo real para estimar el estado emocional del usuario.

Si bien representa un avance importante, su campo de acción está limitado a dos ámbitos diagnósticos, la ansiedad y la depresión. No contempla la detección de perfiles neurodivergentes como el TDAH, el autismo o la dislexia, ni ofrece análisis multimodal que integre voz, texto y emociones [9].

3.1.3 CANARY SPEECH

Canary Speech también emplea algoritmos de análisis vocal para evaluar estados como el estrés, la fatiga cognitiva o el deterioro neurológico. La empresa ha desarrollado modelos que capturan características vocales relacionadas con la carga mental, y ha establecido colaboraciones con instituciones médicas para validar sus resultados. Al igual que Vocalis Health y Kintsugi, se centra principalmente en la voz como canal único de entrada, sin explotar el contenido lingüístico ni otras señales como la expresión facial o la estructura discursiva [10].

3.1.4 ELLIPSIS HEALTH

Ellipsis Health es una plataforma que utiliza el análisis de voz para identificar signos de ansiedad y depresión en tiempo real. Su tecnología está diseñada para integrarse con llamadas telefónicas o videollamadas, evaluando aspectos como la entonación, la velocidad del habla y las pausas, sin necesidad de preguntas directas o formularios. Aunque su enfoque es útil en el ámbito de la salud mental, se limita a detectar estados emocionales generales y no aborda perfiles neurodivergentes específicos, ni utiliza información textual o visual como apoyo [8].

3.1.5 WINTERLIGHT LABS

Winterlight Labs ha desarrollado una herramienta orientada al diagnóstico de deterioro cognitivo leve, demencia, y enfermedades neurodegenerativas como el Alzheimer. A partir de una muestra de voz y lenguaje de menos de un minuto, el sistema analiza más de 400 características lingüísticas y acústicas para generar métricas. A diferencia e otras soluciones centradas en la depresión o el estrés, Winterlight adopta un enfoque clínico más estructurado, pero su ámbito de aplicación no incluye perfiles neurodivergentes [7].

Plataforma	Modalidad de entrada	Enfoque principal	Análisis de texto	Emociones	Personalidad	Perfiles neurodivergentes
Vocalis Health	Voz	Salud física (respiratoria)	No	No	No	No
Kintsugi	Voz	Ansiedad y depresión	Parcial	Sí	No	No
Canary Speech	Voz	Estrés y carga cognitiva	No	Sí	No	No
Ellipsis Health	Voz	Estado de ánimo	No	Sí	No	No
Winterlight Labs	Voz y texto	Enfermedades cognitivas	Sí	Parcial	No	No
Souly	Voz + texto + video	Salud mental y neurodivergencia	Sí	Sí	Sí	Sí

Tabla 1 – Comparativa de plataformas del sector y sus características

3.2 INVESTIGACIONES ACADÉMICAS RELEVANTES

Numerosos estudios han explorado el uso de modelos de inteligencia artificial para evaluar la salud mental mediante la voz o el lenguaje. A continuación, se destacan algunos trabajos representativos:

3.2.1 MALGAROLI ET AL. (2023)

En este estudio se llevó a cabo una revisión sistemática en la que se analizaron más de 80 artículos centrados en el uso de NLP para intervenciones en salud mental. Uno de los

hallazgos clave fue que, aunque muchas herramientas muestran el potencial diagnóstico o terapéutico, la mayoría carece de validación externa y presentan limitaciones metodológicas que dificultan su aplicación en entornos clínicos reales.

La revisión enfatiza la necesidad de marcos más rigurosos y de evaluar la equidad en modelos entrenados con datos sesgados.

3.2.2 HIGUCHI ET AL. (2020)

En este estudio se desarrolló y probó un sistema de evaluación psicológica basado en voz llamado MIMOSYS, orientado a dispositivos móviles. El estudio, con más de 180 usuarios, encontró una correlación significativa entre las características vocales extraídas (entonación, ritmo y pausas), y los resultados de escalas psicológicas como el CES-D (Center for Epidemiologic Studies Depression Scale). El sistema demostró eficacia como herramienta de selección no invasiva para evaluar la estabilidad mental diaria.

3.2.3 GHOSH ET AL. (2022)

En este estudio se propuso un modelo de aprendizaje profundo multimodal para la detección de autismo que combina características de audio y texto. Utilizando un conjunto de datos de entrevistas clínicas, el modelo alcanzó una precisión de hasta el 88% en tareas de clasificación binarias (autista vs no autista).

Además, se realizó un análisis de importancia de características, lo que aportó interpretabilidad a los resultados y permitió identificar señales específicas del lenguaje y la prosodia asociadas al espectro autista.

Estos trabajos confirman la validez de las señales vocales y lingüísticas como marcadores de salud mental. No obstante, la mayoría aún presenta limitaciones en cuanto a generalización, explicatividad o análisis multimodal, aspectos que el presente proyecto busca abordar integrando múltiples fuentes de información (voz, texto y expresión facial) y aplicando algoritmos tanto predictivos como interpretables.

3.3 LIMITACIONES DE LOS ENFOQUES ACTUALES

A pesar del avance en las tecnologías descritas, persisten importantes limitaciones:

- **Monocanalidad:** La mayoría de las plataformas analizan exclusivamente la voz, sin considerar el lenguaje textual ni la expresión facial.
- **Alcance limitado:** Se centran en estados comunes como ansiedad o depresión, dejando de lado otros perfiles psicológicos complejos.
- **Escasa explicatividad:** Muchos modelos funcionan como si fueran "cajas negras" y no ofrecen interpretabilidad sobre qué variables influyen en el diagnóstico.
- **Poca adaptabilidad:** Las soluciones actuales presentan dificultades para generalizar a distintos idiomas, acentos, o contextos culturales.

3.4 CONTRIBUCIÓN DIFERENCIAL DEL PROYECTO

El enfoque desarrollado en este trabajo supera muchas de las barreras mencionadas anteriormente. Al integrar voz, texto y expresión facial, el sistema ofrece una evaluación multimodal que permite una detección más precisa y contextualizada de perfiles neurodivergentes. Además, combina modelos predictivos con modelos explicativos para facilitar la interpretación de los resultados.

El uso de datos reales, obtenidos de entornos naturales como YouTube o TikTok, aporta realismo al modelo y mejora su potencial de generalización. Finalmente, su compatibilidad con plataformas web como Souly facilita su aplicación en entornos clínicos o educativos, democratizando el acceso a herramientas avanzadas de apoyo al diagnóstico.

Souly, además, no se limita al análisis de perfiles neurodivergentes, sino que también trabaja con problemas como el estrés, la ansiedad u otros trastornos del bienestar emocional, ampliando el espectro e utilidad de sus tecnologías.

Capítulo 4. DEFINICIÓN DEL TRABAJO

4.1 JUSTIFICACIÓN

Tal y como se abordó en la introducción, la identificación de perfiles neurodivergentes sigue enfrentando múltiples desafíos en términos de accesibilidad, precisión y objetividad. Este proyecto surge como respuesta a esta problemática, proponiendo un sistema automatizado basado en el análisis multimodal de datos subjetivos y percibidos obtenidos de fuentes abiertas y no clínicas como YouTube, TikTok o Instagram. El enfoque busca facilitar un primer cribado no profesional, que oriente tanto a usuarios como a profesionales.

Esta justificación se estructura en tres dimensiones complementarias: la técnica, la investigadora y la social.

4.1.1 JUSTIFICACIÓN TÉCNICA

Los sistemas actuales de apoyo al diagnóstico en salud mental suelen basarse en entrevistas clínicas o test psicométricos cuya interpretación depende en gran medida del juicio del profesional. Esto, aunque obviamente válido, puede generar limitaciones en términos de escalabilidad y acceso. Por tanto, se hace necesaria la integración de soluciones tecnológicas que permitan estructurar, automatizar y ampliar la capacidad diagnóstica de forma no invasiva.

En este sentido, el proyecto propone el uso de las dos APIs desarrolladas por Souly. Estas permiten transformar grabaciones reales en métricas acústicas, emocionales, lingüísticas, y de personalidad, permitiendo generar un conjunto de datos estructurados listos para ser procesados por modelos de machine learning. La combinación de voz, texto y expresión facial convierte esta propuesta en un sistema de análisis multimodal, mucho más robusto que los enfoques convencionales centrados únicamente en la voz [5][8].

Además, la infraestructura técnica implementada se basa en scripts automáticos para subir vídeos, procesarlos en paralelo y estructurar los resultados en CSV de forma rápida y escalable. Durante el desarrollo del proyecto se han procesado más de 20.000 llamadas a la API, lo que permite estimar su viabilidad operativa y económica. El coste aproximado por procesamiento ha sido de 0,01 €/vídeo, lo que valida el sistema como una opción técnica sólida y económicamente sostenible.

4.1.2 JUSTIFICACIÓN INVESTIGADORA

Desde el punto de vista académico, este trabajo busca cubrir varios vacíos existentes en la literatura. En primer lugar, la mayoría de estudios aplicados al análisis de voz para salud mental se centran exclusivamente en emociones básicas, como ansiedad o depresión [1][2]. Muy pocos abordan perfiles neurodivergentes, y menos desde un enfoque multimodal que combine texto, prosodia y expresión facial, y menos aún utilizando datos percibidos, generados de forma espontánea y en contextos naturales.

Además, gran parte de los modelos desarrollados en este ámbito presentan una baja transparencia, lo que dificulta comprender qué variables o factores están influyendo realmente en el diagnóstico. Por esta razón, en este proyecto se implementan no solo modelos predictivos (como Random Forest o SVM, por ejemplo), si no también modelos explicativos, permitiendo identificar que grupos de variables (lingüísticas, acústicas o emocionales) son más representativas para cada perfil evaluado [6].

El enfoque fenomenológico adoptado también introduce un componente de individualización (esto es, analizar de manera subjetiva como se manifiesta cada caso a través de métricas objetivas) que contrasta con enfoques estadísticos tradicionales. En conjunto, este proyecto pretende conseguir un marco reproducible y transparente que pueda ser replicado o ampliado en un futuro.

4.1.3 JUSTIFICACIÓN SOCIAL Y DE ACCESIBILIDAD

El diagnóstico de perfiles neurodivergentes en España presenta importantes barreras estructurales. Según eldiario.es [3], las listas de espera para una evaluación diagnóstica en

salud mental infantil superan el año en algunas comunidades autónomas. A esto se suma la falta de especialistas en zonas rurales o de difícil acceso, lo que genera desigualdad territorial en la calidad del diagnóstico [4].

Frente a esta situación, este trabajo propone una herramienta accesible, escalable y objetiva, basada en hechos reales obtenidos de plataformas abiertas como YouTube o TikTok. Esta aproximación tiene el potencial de aplicarse en entornos clínicos, pero también educativos o incluso domésticos, permitiendo realizar un primer cribado automatizado no profesional que oriente a pacientes y familias.

Además, al tratarse de una solución de bajo coste operativo, gratuita en su versión básica y fácilmente integrable en plataformas digitales, puede ser especialmente útil en contextos con barreras económicas, culturales o geográficas. Esto refuerza su utilidad tanto en el ámbito doméstico como en entornos escolares, sanitarios o comunitarios. Su impacto potencial en términos de equidad, accesibilidad y visibilización de la neurodivergencia es, por tanto, significativo.

4.2 OBJETIVOS

4.2.1 OBJETIVO GENERAL

Desarrollar un sistema automatizado basado en inteligencia artificial capaz de analizar voz, texto y vídeo para identificar perfiles neurodivergentes, con el fin de asistir en procesos diagnósticos preliminares de forma objetiva, y multimodal.

Este sistema responde a la necesidad de soluciones tempranas y de bajo impacto que permitan reducir las barreras actuales en la detección de condiciones como el TDAH, la dislexia o el autismo. A diferencia de enfoques tradicionales, el proyecto integra señales perceptuales obtenidas de contextos naturales ("*on-the-wild*"), generando así un marco más representativo de la variabilidad humana.

4.2.2 OBJETIVOS ESPECÍFICOS

- Recopilar y estructurar datos reales (vídeos) de individuos con perfiles neurodivergentes (autismo, dislexia y TDAH) y de control.
- Automatizar el proceso de subida, análisis y extracción de métricas mediante las APIs mencionadas anteriormente.
- Obtener variables acústicas emocionales, lingüísticos y de personalidad a partir de cada muestra.
- Entrenar modelos de clasificación supervisada como Random Forest o SVM, por ejemplo.
- Implementar un modelo explicativo que permita interpretar la relevancia de cada grupo de variables.
- Visualizar los resultados con herramientas (t-SNE, curvas ROC, matrices de confusión).
- Evaluar la aplicabilidad clínica y educativa de los resultados.
- Evaluar el rendimiento operativo de la API y estimar su viabilidad técnica y económica en escenarios de uso real.

4.3 METODOLOGÍA

La metodología del proyecto se ha estructurado en las siguientes fases:

4.3.1.1 Adquisición de datos

Los datos se obtuvieron mediante scrapers disponibles en la plataforma Apify [12], configurados para recopilar vídeos desde YouTube, TikTok e Instagram. Las búsquedas se personalizaron mediante combinaciones de palabras clave y hashtags como #autismo, #tdah, #podcast, con el fin de localizar grabaciones en las que los usuarios relataran en primera persona su experiencia con alguna condición neurodivergente.

Se recogieron más de 13.000 vídeos, distribuidos del siguiente modo: 6.000 de control y 2.000 por cada perfil neurodivergente (TDAH, autismo y dislexia). Esta estrategia ha permitido generar un dataset balanceados, tanto para modelos binarios como multiclase.

4.3.1.2 Procesamiento de datos

a) **Preprocesamiento:** Cada lote de vídeos fue tratado de forma automatizada para su etiquetado y almacenamiento en la nube, permitiendo su posterior análisis. Una vez almacenados, se procesaron los resultados generados por el sistema de análisis, que fueron estructurados en un único conjunto de datos tabular. Este proceso incluyó tanto las métricas cuantitativas derivadas del análisis de voz, texto y expresión facial, como la integración de las transcripciones asociadas a cada muestra.

b) **Posprocesamiento:** Una vez obtenido el dataset consolidado, se depuró eliminando columnas vacías o redundantes, normalizando las variables numéricas y codificando las categóricas mediante librerías de *scikit-learn* [29]. Finalmente, se generaron dos versiones del dataset: una multiclase y otra binaria (neurodivergente vs. control), ambas equilibradas.

4.3.1.3 Modelado predictivo

Para evaluar la capacidad de clasificación del sistema, se han implementado modelos supervisados como Random Forest, Support Vector Machines (SVM) y redes neuronales multicapa (MLP). Las métricas utilizadas para valorar su rendimiento han sido precisión (accuracy), recall y F1-score, permitiendo una evaluación completa tanto del rendimiento general como del comportamiento por clase. Este enfoque ha permitido comparar el potencial predictivo de distintos algoritmos detectando perfiles neurodivergentes.

4.3.1.4 Modelado explicativo

Con el objetivo de interpretar qué variables tienen mayor relevancia en la detección de perfiles neurodivergentes, se ha desarrollado un modelo explicativo basado en regresión logística. A partir del conjunto completo de métricas extraídas (emocionales, lingüísticas, acústicas y de personalidad), se ha realizado un análisis por grupos de variables, tanto de forma individual como combinada. El modelo también incorpora el contenido textual de los

vídeos procesado con técnicas de vectorización, permitiendo identificar patrones relevantes en el lenguaje. Este enfoque complementa el análisis predictivo y aporta transparencia y explicabilidad al sistema.

4.3.1.5 Visualización y evaluación

Se han empleado técnicas de visualización para representar de forma intuitiva la distribución espacial de los perfiles analizados. En concreto, se ha aplicado t-SNE sobre el espacio vectorial del texto para explorar la agrupación natural de clases, revelando estructuras presentes en el discurso. Además, se han generado curvas ROC, reportes de clasificación y matrices de confusión para evaluar cuantitativamente la calidad de los modelos entrenados, tanto en escenarios multiclase como binarios.

4.4 PLANIFICACIÓN TEMPORAL

La planificación del proyecto se ha estructurado siguiendo una distribución temporal comprendida entre diciembre y mayo, con tareas escalonadas y parcialmente superpuestas. A continuación se describe cada fase:

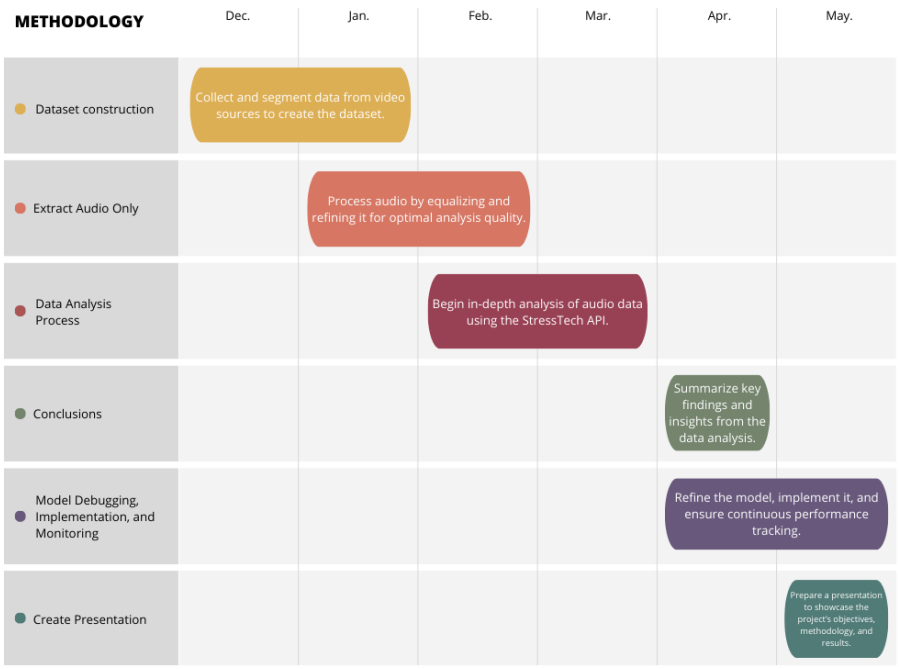


*Figura 3 - **¡Error! No se encuentra el origen de la referencia.***

- **Construcción del dataset (Dic. – Ene.):** recopilación y segmentación de vídeos de fuentes abiertas para construir el dataset inicial.
- **Extracción de datos (Ene. – Feb.):** extracción de los canales de audio y texto, normalización y ecualización del sonido con el objetivo de conseguir una calidad óptima de la señal para su análisis posterior.
- **Análisis de datos (Feb. – Mar.):** análisis detallado de los audios mediante la primera API de Souly, obteniendo métricas emocionales, lingüísticas y de personalidad para cada una de todas las muestras procesadas.
- **Conclusiones (Abr.):** síntesis de hallazgos principales y elaboración de insights extraídos de los análisis cuantitativos y visuales realizados.
- **Depuración, implementación, y monitorización del modelo (Abr. – May.):** ajuste fino de modelos, implementación definitiva de la arquitectura del sistema y verificación de su rendimiento en distintos escenarios.
- **Creación de la presentación (May. – Jun.):** elaboración del material visual y expositivo que resume los objetivos, la metodología y los resultados clave del proyecto.

4.5 POTENCIAL DE COMERCIALIZACIÓN Y ESCALABILIDAD

Este proyecto no solo tiene valor académico, sino que también presenta un alto potencial de aplicación real en entornos clínicos, educativos y comerciales. Durante su desarrollo, el sistema ha sido puesto a prueba con unas 20.000 llamadas a la API, lo que ha permitido validar su rendimiento operativo y estimar con precisión su viabilidad técnica y económica. El coste de procesamiento por vídeo se ha mantenido constante en torno a los 0,01 € por llamada, mientras que el coste mensual total ha oscilado entre 15 y 45 euros, dependiendo de la carga puntual del sistema. Esta estructura de costes sostenida y predecible confirma que la solución es escalable y económicamente eficiente.

Desde un punto de vista comercial, la arquitectura del sistema permite su integración en plataformas ya existentes o su despliegue como producto autónomo en formato web o móvil.

Dada su capacidad de procesamiento masivo a bajo coste, el sistema podría generar un margen económico positivo si se comercializa a escala. Por ejemplo, incluso con una tarifa por uso simbólica (por ejemplo, 0,10 € por evaluación), el margen bruto cubriría ampliamente los costes operativos actuales. Esto abre la puerta a distintos modelos de negocio, desde licencias SaaS (software como servicio) para clínicas y centros educativos, hasta soluciones *freemium* con módulos básicos gratuitos y análisis avanzados mediante suscripción.

El sistema, por tanto, no solo es viable desde el punto de vista técnico, sino que también presenta argumentos sólidos para su sostenibilidad financiera y expansión futura.

Capítulo 5. GENERACIÓN DEL DATASET

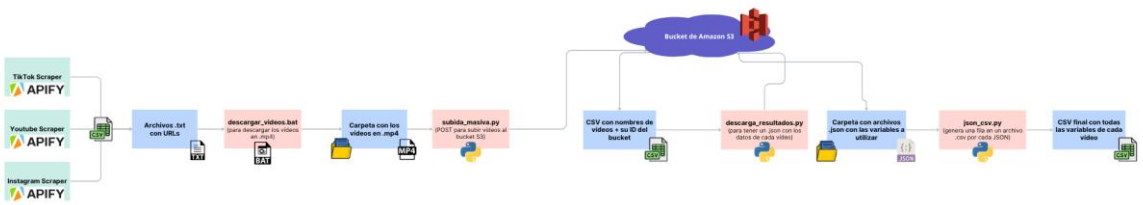


Figura 4 - Diagrama de bloques del proceso de generación del dataset

La elaboración del conjunto de datos utilizado en este trabajo ha requerido la integración de herramientas de extracción, procesamiento y estructuración de contenido procedente de redes sociales. El objetivo ha sido construir un dataset robusto de más de 10.000 muestras que permita entrenar modelos de machine learning orientados al reconocimiento de perfiles neurodivergentes. A continuación se detalla el proceso, dividido en fases operativas consecutivas.

5.1 OBTENCIÓN DE URLS DESDE PLATAFORMAS ONLINE

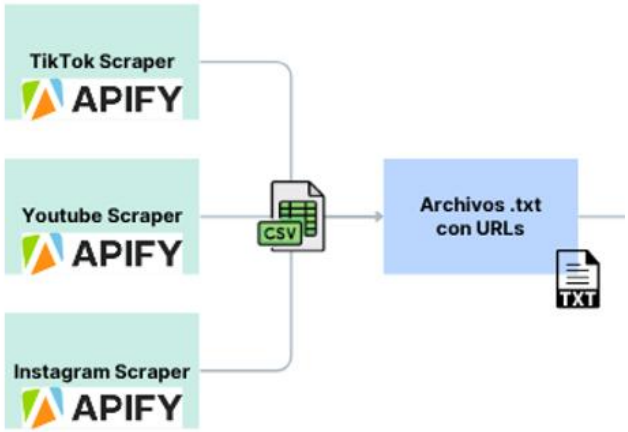


Figura 5 - Parte del diagrama de bloques que ilustra la obtención de URLs

Para la recopilación inicial de muestras, se utilizó la plataforma Apify [12], que proporciona scrapers configurables para redes sociales como TikTok, YouTube e Instagram. Estos scrapers permiten automatizar búsquedas en función de hashtags, términos clave o descripciones de contenido.

El criterio de selección se ha centrado en identificar vídeos en los que personas hablaran directamente a cámara, en primera persona, relatando su experiencia con alguno de los perfiles que se quieren analizar (TDAH, autismo / TEA, dislexia, grupo de control).

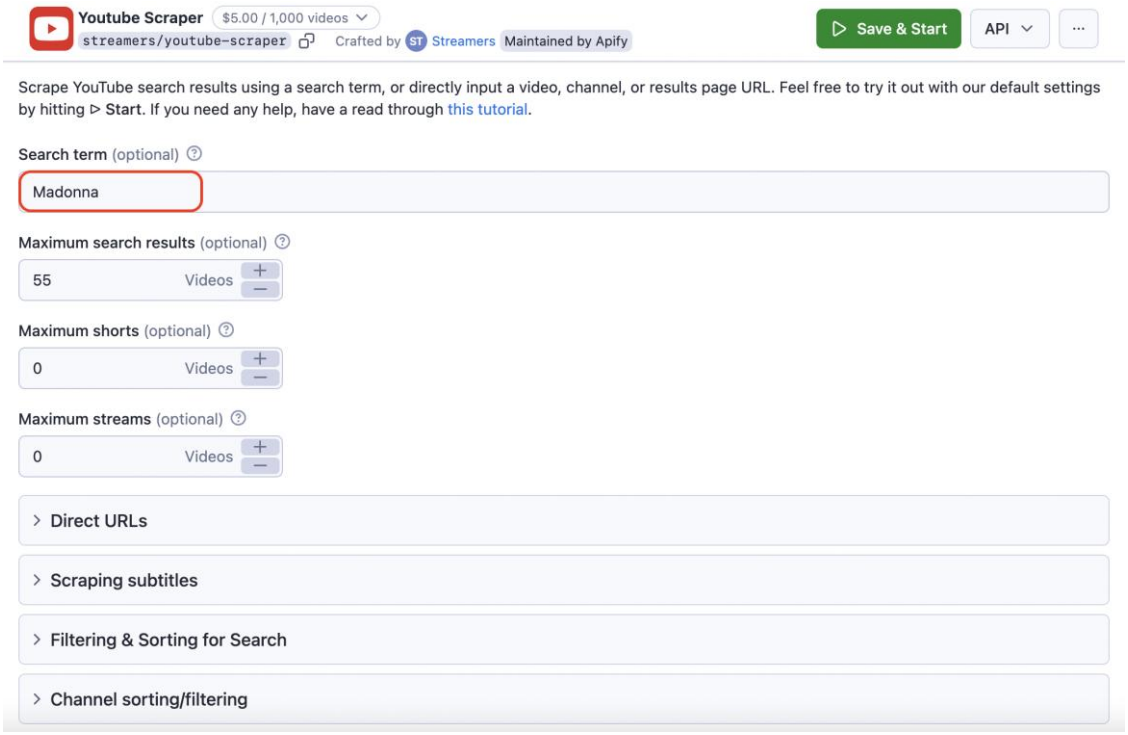


Figura 6 - Ejemplo de la interfaz del scraper de YouTube de Apify para configurar la búsqueda de vídeos masiva

Se realizaron múltiples consultas en Apify por variable, en los distintos scrapers disponibles en la plataforma, para obtener una mayor variedad en los datos. En cada una de estas

consultas, se han utilizado combinaciones de términos como "mi experiencia con dislexia", "viviendo con autismo" o "testimonio TDAH". También, se ha filtrado mediante hashtags relevantes como "#autismo" o "#TDAH", y se han incorporado canales personales de creadores con neurodivergencia (por ejemplo, el canal de Pau Brunet [13], un joven diagnosticado con autismo que ha aparecido en diversos medios en España para dar visibilidad al TEA infantil).

Para el caso de los vídeos de control, se han seleccionado podcasts generalistas que traten de cualquier tema ajeno a la neurodivergencia, por la adecuación del formato de este tipo de programas y vídeos a lo que se busca para analizar.

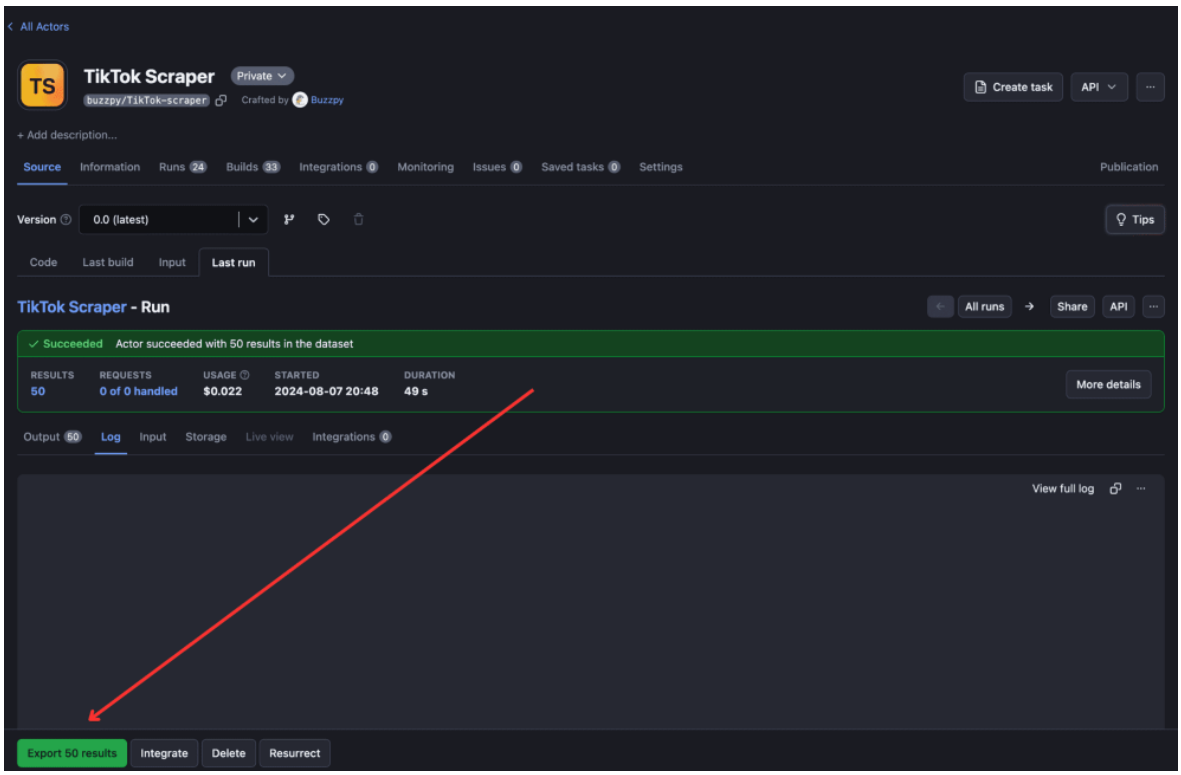


Figura 7 - Ejemplo de la exportación de 50 vídeos haciendo uso del scraper de TikTok de Apify

Cada consulta generaba un archivo .csv con entre 800 y 1000 URLs de vídeos, que se almacenaban por separado. Para construir el dataset completo de más de 10.000 muestras,

ha sido necesario hacer varias consultas en Apify, como se comentaba al inicio del apartado, y hacer todo el proceso mostrado en el diagrama para cada consulta.

Con el fin de construir dos datasets distintos, uno para clasificación multiclase balanceada (con 2.000 vídeos por clase), y otro para regresión logística binaria (6.000 vídeos de neurodivergentes y 6.000 vídeos de control), se ha programado a lo largo del proyecto la extracción escalonada de aproximadamente 13.000 vídeos.

5.2 DESCARGA DE VÍDEOS DESDE URLS

Figura 8 - Parte del diagrama de bloques que ilustra el proceso de descarga de vídeos en .mp4 a partir de URLs

Una vez obtenido el archivo .csv con las URLs de cada tanda, se copian y pegan sin ningún tipo de manipulación en un archivo llamado "urls_[plataforma].txt", teniendo también una URL por vídeo. A partir de este documento .txt se procedió a la descarga de estos en formato .mp4.

Para este apartado, un reto encontrado a la hora de construir el dataset era la descarga de vídeos en formato .mp4, ya que el bucket de Amazon S3 no almacena URLs, sino los vídeos como tal, y a priori, descargar más de 10.000 vídeos en formato .mp4 uno a uno es un proceso tedioso que bien puede durar varias semanas.

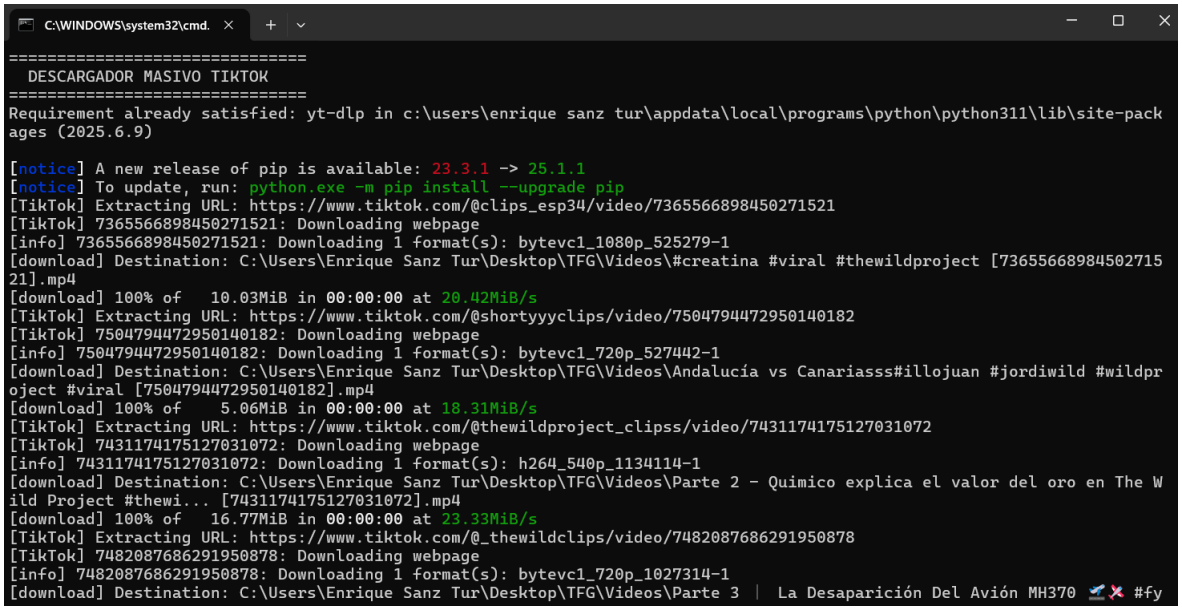


Figura 9 - Ejecución del archivo "descargar_videos.bat", que va mostrando la descarga de vídeos secuencial

Para esto, se utilizó un script Batch (descargar_videos.bat), que apoyándose en el documento .txt, recorre secuencialmente cada una de las URLs para ir descargando los vídeos uno a uno en una carpeta específica, asignando un nombre único con prefijo de variable, en función de la consulta que se hubiera hecho anteriormente (Autismo1.mp4, Autismo2.mp4, …)

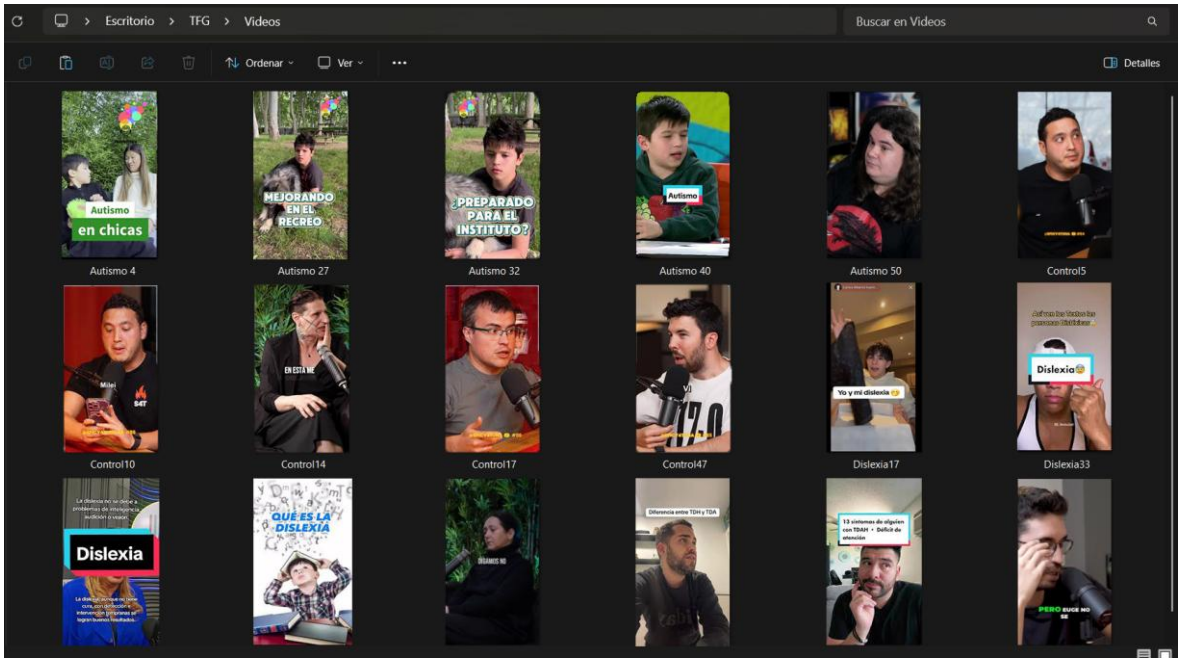


Ilustración 1

Este paso garantiza una organización de las muestras, y permite mantener la trazabilidad a lo largo de todo el proceso entre el contenido y su clase correspondiente.

5.3 SUBIDA DE VÍDEOS AL BUCKET DE AMAZON S3

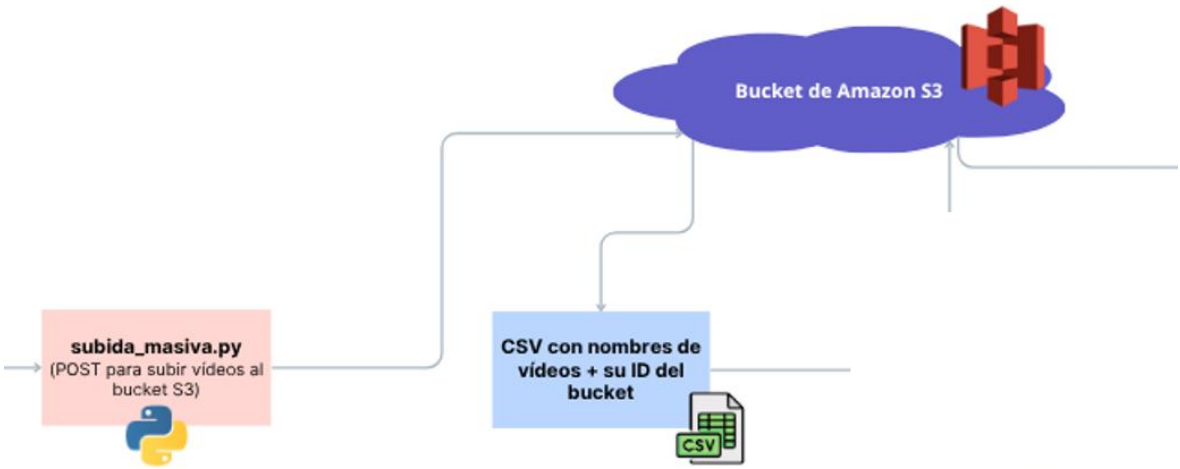
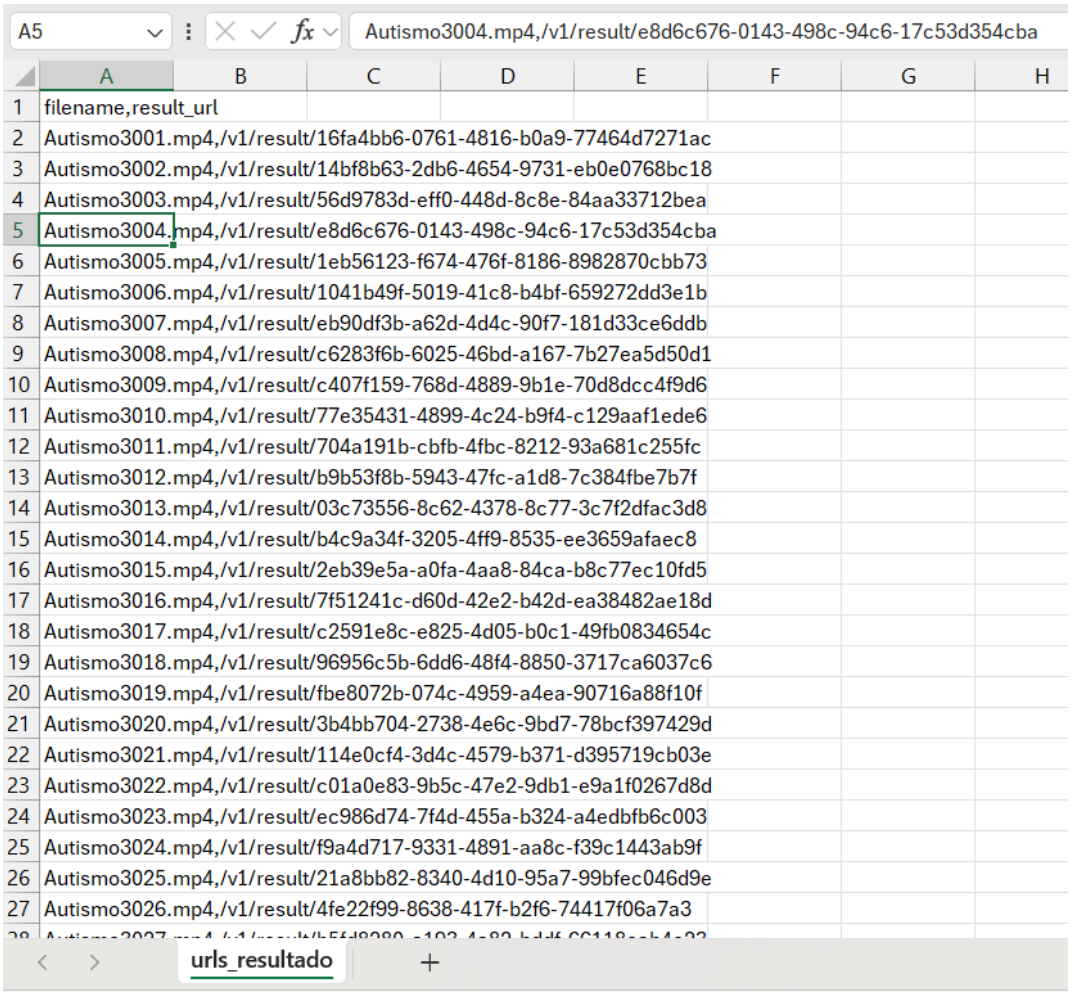


Figura 10 - Parte del diagrama de bloques que ilustra el proceso de subida de vídeos al bucket S3

Antes de ser procesados por la plataforma de análisis de Souly, hay que hacer un último paso con los vídeos descargados en formato .mp4.

Se ha desarrollado un script de Python (subida_masiva.py) que recorre la carpeta de vídeos que contiene los previamente descargados, y sube cada archivo al bucket de S3 de forma automática mediante peticiones HTTP tipo POST.



A5 | Autismo3004.mp4,/v1/result/e8d6c676-0143-498c-94c6-17c53d354cba

	A	B	C	D	E	F	G	H
1	filename,result_url							
2	Autismo3001.mp4,/v1/result/16fa4bb6-0761-4816-b0a9-77464d7271ac							
3	Autismo3002.mp4,/v1/result/14bf8b63-2db6-4654-9731-eb0e0768bc18							
4	Autismo3003.mp4,/v1/result/56d9783d-eff0-448d-8c8e-84aa33712bea							
5	Autismo3004.mp4,/v1/result/e8d6c676-0143-498c-94c6-17c53d354cba							
6	Autismo3005.mp4,/v1/result/1eb56123-f674-476f-8186-8982870cbb73							
7	Autismo3006.mp4,/v1/result/1041b49f-5019-41c8-b4bf-659272dd3e1b							
8	Autismo3007.mp4,/v1/result/eb90df3b-a62d-4d4c-90f7-181d33ce6ddb							
9	Autismo3008.mp4,/v1/result/c6283f6b-6025-46bd-a167-7b27ea5d50d1							
10	Autismo3009.mp4,/v1/result/c407f159-768d-4889-9b1e-70d8dcc4f9d6							
11	Autismo3010.mp4,/v1/result/77e35431-4899-4c24-b9f4-c129aaf1ede6							
12	Autismo3011.mp4,/v1/result/704a191b-cbfb-4fbc-8212-93a681c255fc							
13	Autismo3012.mp4,/v1/result/b9b53f8b-5943-47fc-a1d8-7c384fbe7b7f							
14	Autismo3013.mp4,/v1/result/03c73556-8c62-4378-8c77-3c7f2dfac3d8							
15	Autismo3014.mp4,/v1/result/b4c9a34f-3205-4ff9-8535-ee3659afaec8							
16	Autismo3015.mp4,/v1/result/2eb39e5a-a0fa-4aa8-84ca-b8c77ec10fd5							
17	Autismo3016.mp4,/v1/result/7f51241c-d60d-42e2-b42d-ea38482ae18d							
18	Autismo3017.mp4,/v1/result/c2591e8c-e825-4d05-b0c1-49fb0834654c							
19	Autismo3018.mp4,/v1/result/96956c5b-6dd6-48f4-8850-3717ca6037c6							
20	Autismo3019.mp4,/v1/result/fbe8072b-074c-4959-a4ea-90716a88f10f							
21	Autismo3020.mp4,/v1/result/3b4bb704-2738-4e6c-9bd7-78bcf397429d							
22	Autismo3021.mp4,/v1/result/114e0cf4-3d4c-4579-b371-d395719cb03e							
23	Autismo3022.mp4,/v1/result/c01a0e83-9b5c-47e2-9db1-e9a1f0267d8d							
24	Autismo3023.mp4,/v1/result/ec986d74-7f4d-455a-b324-a4edbfb6c003							
25	Autismo3024.mp4,/v1/result/f9a4d717-9331-4891-aa8c-f39c1443ab9f							
26	Autismo3025.mp4,/v1/result/21a8bb82-8340-4d10-95a7-99bfec046d9e							
27	Autismo3026.mp4,/v1/result/4fe22f99-8638-417f-b2f6-74417f06a7a3							

urls_resultado

Figura 11 - Pantallazo del csv resultante después de ejecutar el script de Python para la subida masiva de vídeos al bucket S3

Una vez finalizada cada tanda, se genera un archivo .csv que contiene el nombre del vídeo local y el identificador único asignado por el bucket. Este archivo será necesario posteriormente para vincular resultados a los vídeos originales.

Dado que el sistema de backend de Souly requiere cierto tiempo para procesar completamente los vídeos, se ha establecido un periodo de espera de unas pocas horas (entre 2 y 4 horas) antes de lanzar los siguientes scripts, permitiendo que la plataforma pueda procesar de manera completa todos los vídeos subidos a la vez, y pueda recolectar sus datos correctamente.

5.4 DESCARGA DE RESULTADOS PROCESADOS

Figura 12 - Parte del diagrama de bloques que ilustra el proceso de descarga de los resultados obtenidos por el bucket S3

Una vez transcurrido el tiempo de procesamiento estimado, se ejecuta el script "descarga_resultados.py". Este script accede al archivo .csv con los identificadores del bucket y descarga para cada vídeo un archivo .json con los resultados del análisis, almacenándolos todos en una carpeta por organización.

Los campos que se encuentran en el JSON resultado utilizado para hacer el dataset se pueden observar en la siguiente tabla:

Campo	Descripción
`status`	Indica el estado de la operación, normalmente "success" si el análisis se ha completado correctamente.
`created_at`	Marca temporal de creación del análisis (en formato Unix).
`aid`	Identificador único del análisis realizado sobre el archivo.
`original_file`	Información sobre el archivo original: tipo (.mp4, .wav, etc.) y duración en segundos.
`status`	Lista de etapas completadas durante el análisis (voz, expresión facial, transcripción, biometría, etc.).
`external_vars`	Variables externas asociadas al archivo, como un identificador propio.
`data.facial.average_emotions`	Promedio de emociones básicas detectadas en las expresiones faciales (e.g., tristeza, alegría, enfado).
`data.facial.dominant_emotion_counts`	Número de veces que una emoción fue dominante en distintos fotogramas.

`data.facial.average_face_confidence`	Confianza media del sistema en la detección de rostros durante el análisis facial.
`data.traits`	Métricas de personalidad y rasgos psicológicos inferidos (Big Five, autoestima, compasión, estrés, depresión, etc.).
`data.traits.stress, helplessness, self_efficacy, depression`	Subgrupos que dividen cada rasgo en niveles de intensidad: bajo, medio y alto.
`data.voice.frequencies`	Parámetros estadísticos de la voz (frecuencia media, desviación típica, curtosis, notas musicales predominantes, etc.).
`data.voice.pitch`	Altura tonal media de la voz en Hz.
`data.voice.emotions`	Distribución de emociones detectadas a partir del tono de voz (e.g., enfado, tristeza, felicidad).
`data.speech.language`	Idioma principal detectado en la transcripción del audio.
`data.speech.text`	Transcripción automática del discurso completo.
`data.speech.tense`	Tiempos verbales predominantes en el discurso (presente, pasado, futuro).
`data.speech.sentiment`	Análisis de sentimiento: polaridad (positivo/negativo) y subjetividad.
`data.speech.emotions`	Emociones inferidas a partir del contenido del texto transcrito.
`data.speech.entities`	Entidades conversacionales detectadas y clasificadas por tipo (personas, lugares, actividades, etc.).

`data.speech.topics`	Temas de conversación más probables inferidos por modelos de lenguaje entrenados.
`data.translation`	Traducción automática del texto original en caso de que estuviera en otro idioma distinto al inglés.
`latency`	Tiempo que ha tardado el análisis completo en ejecutarse, expresado en segundos.

Tabla 2 – Campos del JSON resultado, y su significado

Estos JSON contienen métricas extraídas por las APIs de Souly, incluyendo información sobre voz (frecuencia, tono, ritmo), lenguaje (polaridad, subjetividad, entidades, temas), texto transcrito (text, translation), emociones detectadas y estimaciones de rasgos de personalidad.

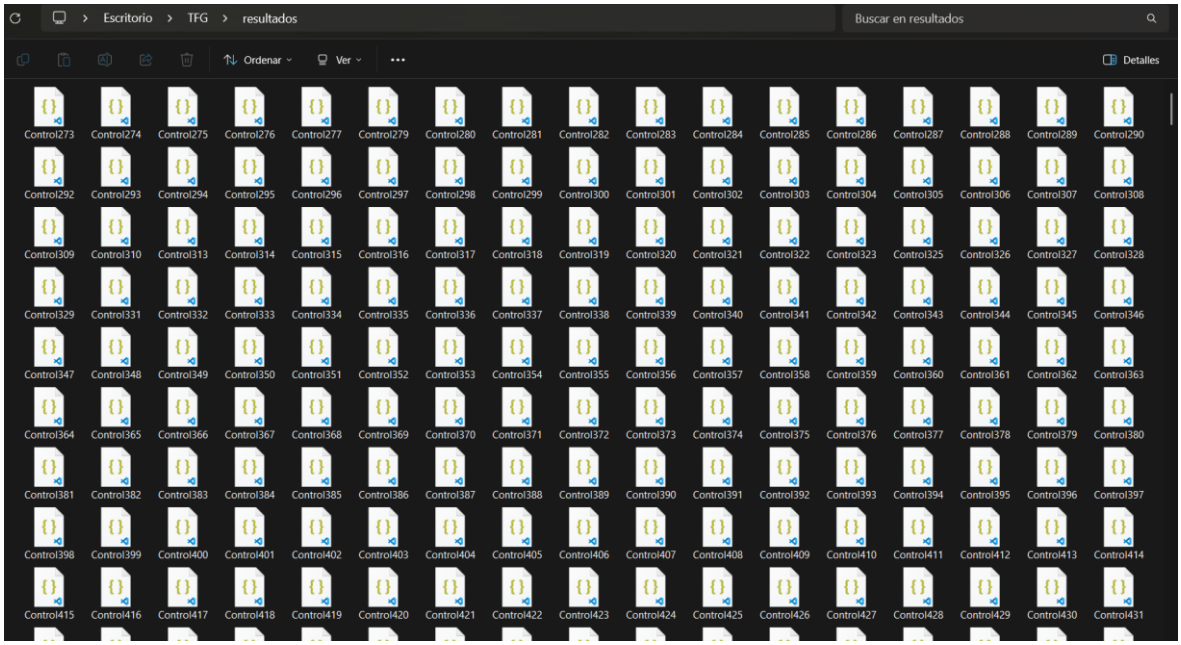


Figura 13 - Carpeta resultante que contiene todos los JSONs con los datos de cada vídeo extraído del bucket S3

Cada tanda genera una carpeta con el volumen de la consulta inicial a los scrapers de Apify, de unos 800-1000 JSONs, uno por vídeo. Así, se consigue mantener el orden y la etiqueta de clase correspondiente.

5.5 *UNIFICACIÓN DE RESULTADOS EN CSV ESTRUCTURADO*

Figura 14 - Parte del diagrama de bloques que ilustra la parte del proceso de unificación de resultados en un archivo .csv

Para poder trabajar con los datos en herramientas de análisis y modelado, será necesario transformar todos los archivos JSON en un formato estructurado, como CSV. Para conseguirlo, se ha desarrollado el script "json_csv.py", que recorre todos los JSONs de la carpeta en la que están contenidos, extrae las variables y construye un archivo .csv donde cada fila representa un vídeo. Un ejemplo de fila de este archivo final se muestra a continuación:

```
1747823367,3b56a8d2-532d-44b1-97af-
915c0ce059d4,.mp4,video,129,True,True,True,True,True,True,True,1,0.0063,0.00
01,0.1956,0.2278,0.2384,0.0,0.3318,neutral,null,0.9533,0.4534,0.4699,0.4779,0.484
5,0.5912,0.25529998540878296,0.22120000422000885,0.5095000267028809,0.28780001401
901245,0.4083999991416931,0.28610000014305115,0.1873999983072281,0.05131414532661
438,0.8721030354499817,0.21103714406490326,0.1005195900797844,0.8312807083129883,
0.15121488273143768,0.34965744614601135,0.06189623847603798,0.7879638671875,0.123
60988557338715,0.9044023752212524,0.035425979644060135,2512,3210,731,392,352,3805
,3453,13.869999885559082,668.0999755859375,Dâ™¯,Fâ™¯,G,F,Aâ™¯,0.11779999732971191
,349,1505,0.018845342099666595,0.598314642906189,0.006962723098695278,0.231140390
0384903,0.032178692519664764,0.06420004367828369,0.011754118837416172,es,0.036604
09897565842,0.11768902838230133,4.3146,0.25,0.75,0.0,-0.0034,0.3764,Control
```

Este archivo final contiene más de 70 columnas, correspondientes a las variables acústicas, lingüísticas y de personalidad. También, el script de Python utilizado se ha encargado de añadir una columna al final, basándose en el nombre del vídeo original en .mp4 (que se utiliza también para nombrar su archivo JSON correspondiente) para añadir la variable asociada, y tener el identificador de cada vídeo de cara a su futuro análisis y tratamiento.

5.6 *CONSOLIDACIÓN DEL DATASET FINAL*

El proceso anterior se ha repetido por separado para cada clase del estudio, manteniendo la separación por variables (no se mezclaban vídeos entre clases en una misma tanda, aunque de lo contrario no hubiera habido ningún problema). Tras finalizar todas las tandas, los distintos archivos .csv generados fueron consolidados manualmente en un único archivo maestro.

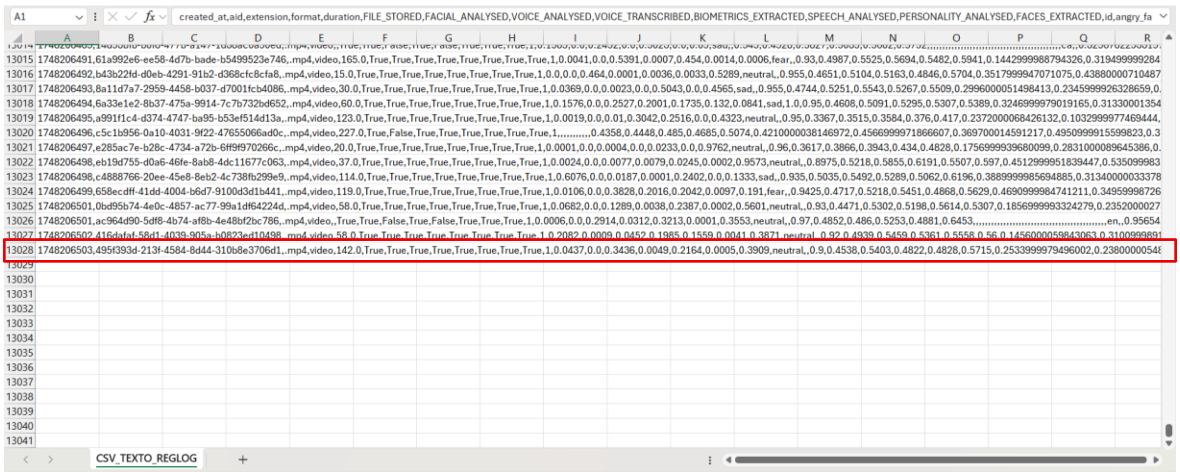


Figura 15 - Pantallazo del archivo final, que muestra los 13.000 vídeos procesados en total

Este CSV final contiene más de 13.000 filas, una por vídeo, y sirve como base para los modelos predictivos y explicativos desarrollados en este trabajo. Se ha verificado que todos los registros están correctamente etiquetados, completos y en un formato homogéneo.

5.7 INTEGRACIÓN DE TRANSCRIPCIONES PARA EL MODELO EXPLICATIVO

Aunque la construcción inicial del dataset se centró en variables numéricas extraídas por las APIs de Souly, en fases posteriores del análisis se consideró relevante incorporar el contenido textual de cada vídeo. Esta decisión respondió a la necesidad de explorar, en el modelo explicativo de regresión logística, si el lenguaje utilizado por los participantes podía aportar información significativa sobre su perfil neurodivergente.

Para ello, se ha desarrollado un script de Python (Add_textoCSV.py) que recorrió el dataset consolidado de más de 13.000 vídeos y añadió dos nuevas columnas: una con el texto original ("text"), y otra con la traducción automática al inglés ("translation"). Ambas variables se extraen directamente de los archivos .json generados durante el procesamiento inicial. Un ejemplo de línea final del .csv con las transcripciones se muestra a continuación:

```
1744825037,5b442dc9-a39a-47e2-ba05-
b0fe880ee44b,.mp4,video,135.0,True,True,True,True,True,True,True,1,0.0315,0.
0001,0.0204,0.4411,0.2166,0.0009,0.2895,neutral,,0.9167,0.5171,0.5625,0.5803,0.55
75,0.6256,0.2748000025749206,0.1768999993801117,0.2211000025272369,0.235300004482
2693,0.2242999970912933,0.3068999946117401,0.3048000037670135,0.342178076505661,0
.5435426831245422,0.0908569023013114,0.396858662366867,0.5517717003822327,0.08131
40422105789,0.5691758394241333,0.171218529343605,0.2396633923053741,0.81402820348
73962,0.2560266852378845,0.0534319654107093,1830.0,2609.0,828.0,173.0,419.0,1772.
0,1353.0,6.130000114440918,52.61000061035156,AÃ¢â„¢Â¯,GÃ¢â„¢Â¯,F,GÃ¢â„¢Â¯,A,0.057
7000007033348,367.0,1483.0,0.0182029567658901,0.0428850576281547,0.12466012686491
01,0.6944096684455872,0.0300897695124149,0.0222449786961078,0.020281421020627,es,
0.0472260043025016,0.0554920881986618,4.244,0.0833,0.9167,0.0,0.217,0.5825,Autism
o,"Hola, soy Federico GarcÃƒÂ-a, tengo 14 aÃƒÂ±os y tengo Asperger. La verdad de
un lado, el escente con Asperger es normal, bastante tranquila, como la de
cualquier otro. Pero puede llegar a ser muy solitaria, sin la orientaciÃƒÂ³n
apropiada. Soy diferente, soy como tÃƒÂº, es como el lema de nuestra
fundaciÃƒÂ³n, la FundaciÃƒÂ³n Federico GarcÃƒÂ-a y Llegas. Nase, pues, porque
Federico me decÃƒÂ-a, mamÃƒÂ-, yo no quiero que otras personas sientan lo que yo
```

he sentido. Lo que nosotros buscamos es acompaÃƒÂ±ar a familias que encuentran que alguno de sus hijos o alguna persona puede dentro del nÃƒÂ°cleo familiar. EstÃƒÂ¡ o puede estar dentro del espectro del autismo, hacerles ese acompaÃƒÂ±amiento, esa orientaciÃƒÂ³n, lo que necesiten para poder atravesar, digamos, este camino de la manera mÃƒÂ¡s transida. Yo trabajo de la mano con los especialistas de la FundaciÃƒÂ³n Palla de Lili, porque es un equipo maravilloso y un equipo que conocen, el diagnÃƒÂ³stico conocen la condiciÃƒÂ³n, entonces, pues pueden darnos apoyo en todos los aspectos, hacemos un trabajo integral. Soy diferente, soy como tÃƒÂº, originalmente, a una campaÃƒÂ±a que puede desarrollando y fue evolucionando hasta ser mÃƒÂ¡s una iniciativa, un lema personal o llamado a la diversidad. Seamos muy respetuosos, muy tolerantes, con la diferencia, cualquiera que sea, todos merecemos respeto para poder convivir todos en el mundo tan complicado como este, pero de manera armoniosa. Dos mensajes para las personas que son como yo. El primero, no usen su condiciÃƒÂ³n para como excusa para hacer cosas que no deberÃƒÂ-an. Y dos, nunca dejen que nadie les diga que no solamente por su condiciÃƒÂ³n. Soy diferente, soy como tÃƒÂº."," Hello, I'm Federico Garcia, I'm 14 years old and I'm a hospital. What's your favorite subject? I've had a lesson with the Pregueres. It's normal, quite calm, like any other. But it can be very solitary without the own orientation. I'm different, I'm like you, it's like the issue of our foundation, the Federico Garcia and you arrive. I don't know why Federico said, but I don't want other people to feel what I've felt. What we're looking for is to accompany the families who find out that some of their children or some person can be within the family's core, or can be within the spectrum of the autism, to make them accompany that orientation, what they need to be able to move through this path in a more calm way possible. I work with the specialists of the Lili Foundation, because they're a wonderful team and they know the diagnosis, they know the condition, so they can give us support in all aspects, we do an integral work. I'm different, I'm like you, I originally had a campaign that can develop and evolve, to make a more initiative, a personal issue or a call for diversity. We are very respectful, very tolerant, with the difference, anyone who is, we all respect, to be able to coexist in a complicated world like this, but in a harmonious way. I'm very grateful to the people who are like me. The first thing is not in their condition, for excuse, for doing things they would never do. And two, never let anyone say they don't just do their condition. I'm different, I'm like you. Thank you."

El procedimiento del script consistió en localizar el archivo .json correspondiente a cada fila del CSV a partir de su identificador único ("aid"), acceder a la carpeta donde se almacenaron todos los archivos JSON, y en caso de coincidencia, añadir los campos mencionados a la fila correspondiente. Tras todo este procedimiento, se guardó un archivo CSV con la estructura original y las nuevas variables incluidas.

Esta extensión del dataset permitirá realizar el análisis de los datos combinando texto y variables numéricas, así como estudiar con mayor profundidad el contenido semántico y expresivo de los discursos de los vídeos.

Capítulo 6. MODELO EXPLICATIVO

Este capítulo está enfocado en el análisis explicativo del conjunto de datos generado, con el objetivo de comprender qué variables contribuyen de manera más significativa a la identificación de perfiles neurodivergentes. Para ello se ha empleado un modelo de regresión logística, tanto en su versión numérica como textual y combinada, como una herramienta de interpretabilidad y explicatividad del modelo.

6.1 PREPROCESAMIENTO DE LOS DATOS

El dataset utilizado se compone de más de 13.000 muestras, cada una correspondiente a un vídeo previamente etiquetado como perteneciente a una de las clases objetivo (Autismo, TDAH, Dislexia y Control). La distribución de vídeos se muestra a continuación:

Control	6155
Autismo	2113
TDAH	2030
Dislexia	2026
1 (Neurodivergente)	6199
0 (Control)	6155

Tabla 3 – Distribución de vídeos para el modelo explicativo

Para la regresión logística, se consideró una variable binaria donde la clase de control toma el valor 0, y las clases de TDAH, dislexia y Autismo se agrupan bajo el valor 1 (neurodivergente). El dataset fue dividido en conjuntos de entrenamiento (train) y de prueba (test) en una proporción de 75/25, respectivamente.

```
🟦 Columna categórica 'most_frequent_dominant_emotion' rellenada con su moda
🟦 Columna categórica 'voice_mean_note' rellenada con su moda
🟦 Columna categórica 'voice_median_note' rellenada con su moda
🟦 Columna categórica 'voice_mode_note' rellenada con su moda
🟦 Columna categórica 'voice_Q25_note' rellenada con su moda
🟦 Columna categórica 'voice_Q75_note' rellenada con su moda
🟦 Columna categórica 'language' rellenada con su moda
🟦 Columna categórica 'text' rellenada con su moda
🟦 Columna categórica 'translation' rellenada con su moda
✅ Nulos restantes tras limpieza: 0
```

Figura 16 - Pantallazo del proceso de rellenado de valores con su moda, y comprobación de nulos

Como siguiente paso en el proceso de preprocesado de los datos, se trató primero con las variables categóricas. Para evitar posibles errores con el contenido de estas, se procedió a identificar cualquier valor no existente que pudiera haber en el dataset, y se procedió a rellenar dicho valor con la moda de su variable, para todas aquellas muestras que tuvieran igual variable objetivo. Por ejemplo, si en una muestra la columna 'voice_mean_note' no tuviera valor, y esta muestra fuera de un vídeo de autismo, el valor de dicha columna se rellenaría con el valor más repetido de la columna 'voice_mean_note' de todas las 2113 muestras de vídeos de autismo. De esta manera, se evita la mezcla con otras variables objetivo, evitando también que se interfiera en la eficacia del modelo y sus resultados.

Posteriormente, se eliminaron las variables de identificación y aquellas que fueran redundantes o irrelevantes. Las variables fueron normalizadas mediante la herramienta StandardScaler [14], y las categóricas codificadas si fuera necesario. También, se aseguró la ausencia de valores nulos en el conjunto final.

```
✅ Nuevas dimensiones de X tras eliminar texto: (12324, 61)
```

Figura 17 - Pantallazo con el recuento de las dimensiones del dataset

6.2 MODELADO GLOBAL

6.2.1 MODELO BASADO SOLO EN VARIABLES NUMÉRICAS

En la primera fase, se entrenó una regresión logística clásica sobre las variables numéricas generadas por la API de Souly, que trataban mayormente de emociones, personalidad o polaridad.

	Precision	Recall	F1 - Score	Support
0 (control)	0.69	0.72	0.70	1532
1 (neurodiv.)	0.71	0.68	0.69	1549
Accuracy		0.70		3081
Macro avg	0.70	0.70	0.70	3081
Weighted avg	0.70	0.70	0.70	3081

Tabla 4 Métricas del modelo numérico

En términos métricos, el modelo alcanza una precisión de 0.69 para la clase de control y de 0.71 para la clase neurodivergente. Los valores de recall (métrica que mide la frecuencia con la que se miden verdaderos positivos de todas las muestras positivas reales) se mantienen relativamente equilibrados en ambas clases, con 0.72 para la clase 0 y 0.68 para la clase 1, lo que indica que el modelo tiene una capacidad razonable para identificar correctamente tanto a individuos neurotípicos como neurodivergentes. El F1-score, que combina precisión y recall en una única medida, se sitúa en 0.70 para ambas clases, mostrando simetría en el comportamiento predictivo del modelo. El accuracy general alcanza un 70%, lo que significa que el modelo acierta aproximadamente 7 de cada 10 casos.

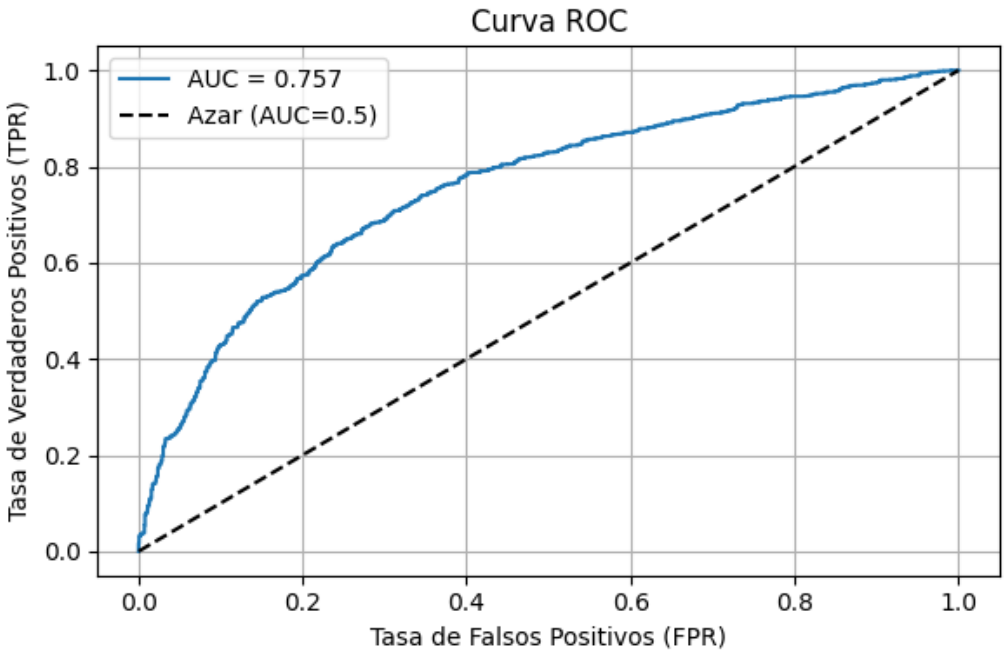


Figura 18 - Curva ROC del modelo numérico

Por su parte, la curva ROC muestra una separación clara respecto a la diagonal del azar (o clasificador aleatorio, con un AUC / área bajo la curva = 0.5), lo que se traduce en un AUC de 0.757. Este valor refleja una capacidad discriminativa moderadamente buena, el modelo logra distinguir entre las clases mejor que el azar, aunque sin llegar a ser un clasificador altamente eficaz. La forma de la curva, ligeramente arqueada sobre la diagonal, refuerza esta interpretación, dejando ver que si bien las variables numéricas contienen información útil, su capacidad por si solas para diferenciar claramente entre perfiles no es ideal.

6.2.2 MODELO BASADO EN SOLO TEXTO (TF-IDF)

Después, se procedió a entrar un modelo utilizando exclusivamente el texto transcrito de los vídeos (usando solo la variable 'translation'), aplicando para ello una representación mediante TF-IDF [15] que extrae las 1000 palabras más relevantes del corpus tras eliminar stopwords (palabras vacías) y caracteres irrelevantes. Esta transformación permite convertir los fragmentos textuales en vectores numéricos con los que alimentar un clasificador de regresión logística binaria. El objetivo ha sido evaluar el valor predictivo que puede tener el lenguaje utilizado por los individuos a la hora de diferenciar entre perfiles neurodivergentes y no neurodivergentes.

	Precision	Recall	F1 – Score	Support
0 (control)	0.80	0.85	0.82	1532
1 (neurodiv.)	0.84	0.79	0.82	1549
Accuracy		0.82		3081
Macro avg	0.82	0.82	0.82	3081
Weighted avg	0.82	0.82	0.82	3081

Tabla 5 – Métricas del modelo textual

En cuanto a resultados obtenidos, el modelo alcanza una precisión de 0.80 para la clase de control y de 0.84 para la clase neurodivergente, lo que representa una mejor respecto al modelo anterior. También, se observan valores sólidos de recall, obteniendo un 0.85 para la clase 0 y 0.79 para la clase 1. En conjunto, estas métricas se reflejan en un F1-Score equilibrado de 0.82 en ambas clases, lo que indica un buen balance entre precisión y

cobertura en la predicción. La exactitud global del modelo es del 82%, lo que implica clasificar correctamente algo más de 8 de cada 10 casos de media.

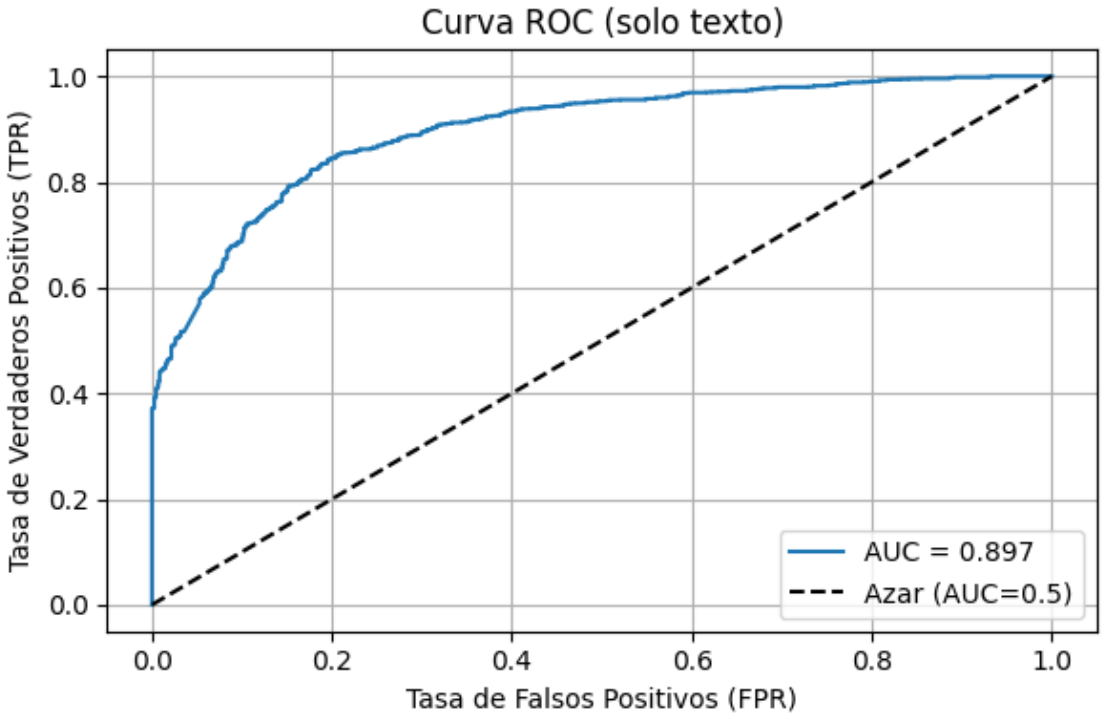


Figura 19 - Curva ROC del modelo usando solo el texto de las muestras

El rendimiento se confirma con una curva ROC significativamente más separada de la diagonal de azar. El valor del área bajo la curva (AUC) alcanza un 89.7%, muy superior al obtenido con los datos numéricos. Esto indica que el contenido lingüístico encierra una señal altamente informativa y que aporta gran valor para la tarea de clasificación. La forma de la curva, con un ascenso inicial pronunciado, muestra que el modelo logra mantener una tasa de verdaderos positivos elevada incluso con tasas bajas de falsos positivos.

Todo esto evidencia que las palabras que emplean los sujetos, extraídas y ponderadas adecuadamente mediante TF-IDF, permiten al modelo identificar patrones lingüísticos distintivos entre personas con neurodivergencia y personas sin neurodivergencia con un nivel de confianza elevado.

6.2.3 MODELO COMBINADO (TEXTO + VARIABLES NUMÉRICAS)

Por último, se ha construido un modelo conjunto que integra los dos anteriores. Cuenta tanto con variables numéricas (rasgos faciales, vocales y de personalidad), como con las características textuales extraídas mediante TF-IDF con las 1000 palabras más significativas. Esta combinación permite aprovechar la complementariedad entre las señales cuantificadas a nivel biométrico y el contenido semántico del discurso, ofreciendo una visión más completa del comportamiento verbal y no verbal de los sujetos.

	Precision	Recall	F1 – Score	Support
0 (control)	0.88	0.89	0.89	1532
1 (neurodiv.)	0.89	0.88	0.88	1549
Accuracy		0.89		3081
Macro avg	0.89	0.89	0.89	3081
Weighted avg	0.89	0.89	0.89	3081

Tabla 6 – Métricas del modelo combinado de datos numéricos y textuales

Los resultados reflejan un rendimiento excelente. El modelo alcanza una precisión del 88% para la clase de control y del 89% para la clase neurodivergente. Además, tanto el recall como el F1-Score se sitúan en torno al 89% para ambas clases, lo que indica que el modelo no solo acierta con frecuencia, sino que lo hace de forma equilibrada entre las dos categorías. La accuracy también es del 89%, muy superior a la de los modelos que se entrenaron con una única fuente de datos.

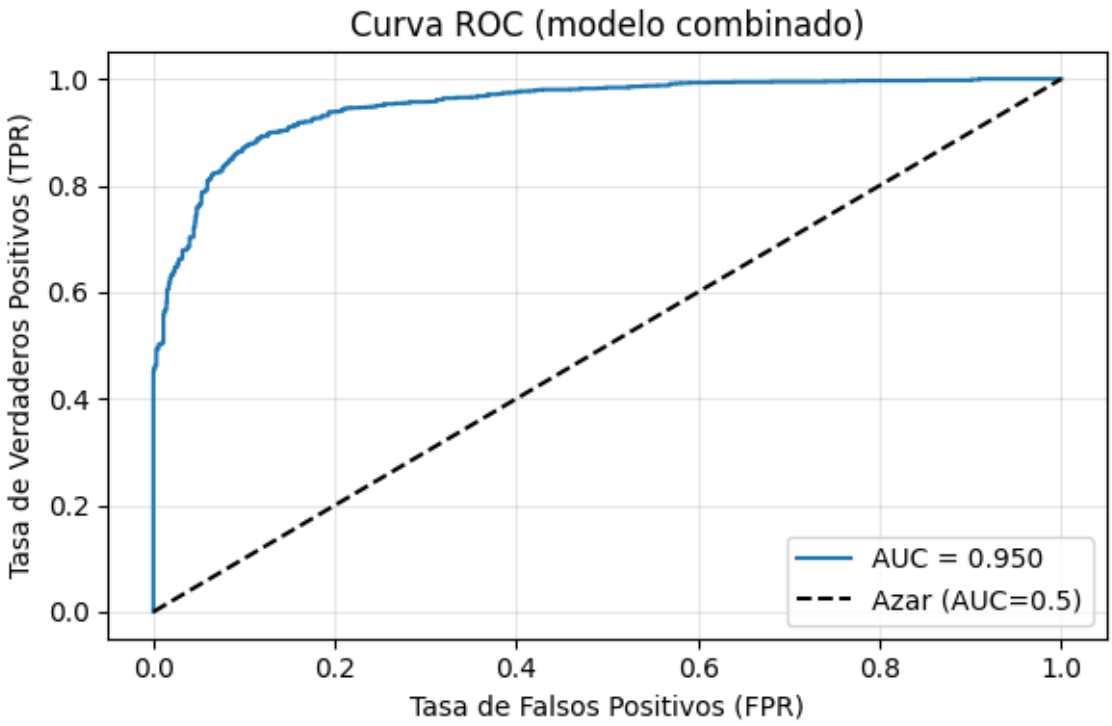


Figura 20 - Curva ROC del modelo combinado con datos numéricos y textuales

Estos resultados se ven respaldados visualmente por la curva ROC, que muestra una separación muy clara respecto a la diagonal aleatoria. El área bajo la curva es del 95%, lo que implica una capacidad predictiva casi perfecta. La curva asciende rápidamente hacia el vértice superior izquierdo, lo que indica una alta tasa de verdaderos positivos aún con bajos niveles de falsos positivos.

Esto refuerza la hipótesis de que la combinación de texto y datos numéricos ofrece una mayor riqueza informativa que permite al modelo aprender patrones complejos de forma más precisa y robusta.

En definitiva, este modelo combinado no solo mejora sensiblemente el rendimiento con respecto a sus versiones por separado, sino que demuestra el potencial de abordar tareas de clasificación compleja como esta, desde una perspectiva multimodal, integrando señales verbales y no verbales en un mismo proceso de decisión.

6.3 ANÁLISIS SEMÁNTICO Y REDUCCIÓN DE LA DIMENSIONALIDAD

6.3.1 VISUALIZACIÓN CON T-SNE

Con el objetivo de explorar visualmente la capacidad de las transcripciones textuales para distinguir entre individuos neurodivergentes y no neurodivergentes, se aplicó la técnica t-SNE (t-distributed Stochastic Neighbour Embedding) [16] sobre los vectores TF-IDF generados a partir del texto.

Esta técnica de reducción de dimensionalidad no lineal es ampliamente utilizada en el ámbito del aprendizaje automático para proyectar datos de alta dimensión en espacios de dos o tres dimensiones, conservando de forma local la estructura de los datos y facilitando la interpretación visual de posibles agrupamientos y/o patrones.

En este caso, los vectores TF-IDF, que codifican la relevancia de las palabras más frecuentes en las transcripciones, constituyen una representación semántica rica (como se ha analizado anteriormente) que encapsula aspectos del contenido verbal de cada participante. Al aplicar t-SNE sobre estos vectores, se busca observar hasta que punto los patrones de habla se agrupan de forma natural según la etiqueta de clase binaria que se está utilizando en este modelo (0 – control, 1 – neurodivergente), y podría sugerir que la información textual contiene rasgos diferenciadores detectables incluso en un espacio reducido.

En los siguientes gráficos, cada punto representará una muestra del conjunto de datos. Cada punto estará coloreado en azul si pertenece a la clase de control, y en rojo si se corresponde a la clase neurodivergente.

6.3.1.1 Con las 1000 palabras más relevantes

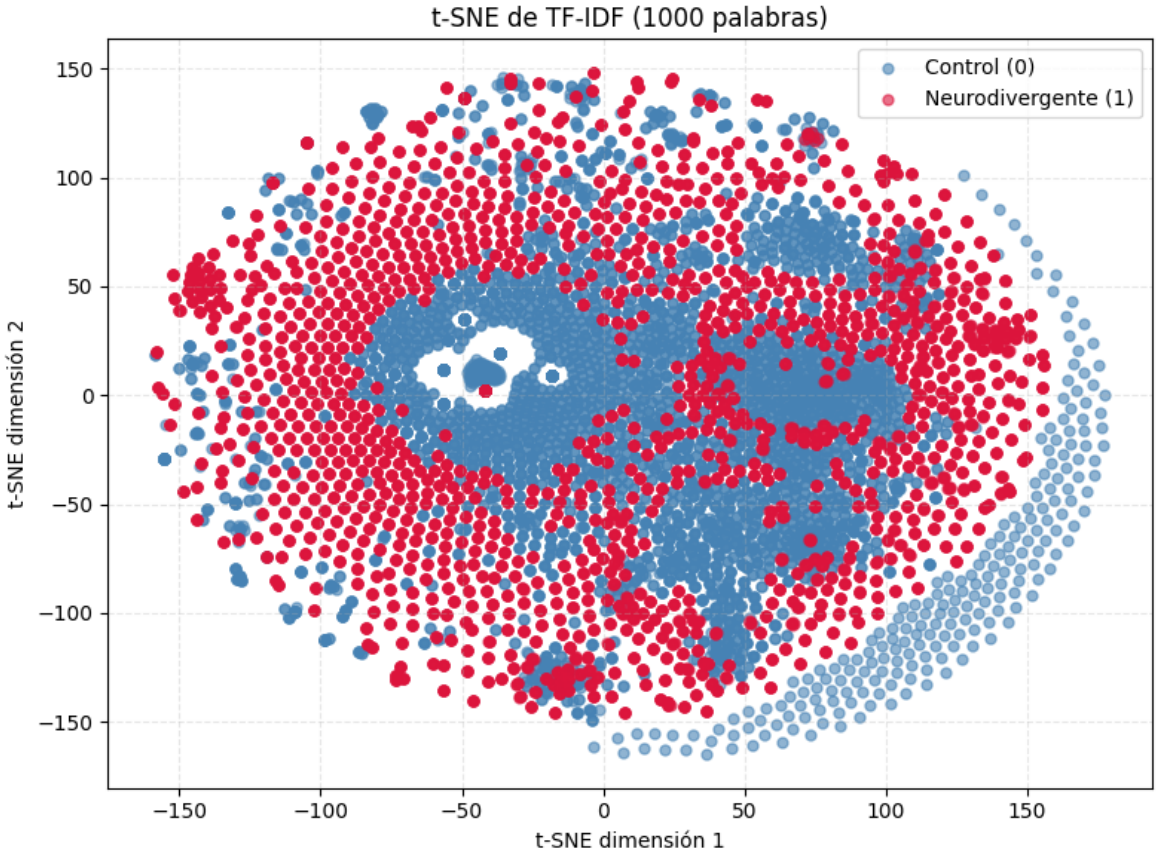


Figura 21 - Visualización con t-SNE usando las 1000 palabras más relevantes

Para este caso, aplicando t-SNE sobre los vectores TF-IDF construidos con las 1000 palabras más frecuentes, se puede apreciar que, a pesar del solapamiento entre ambos grupos, existen regiones claramente dominadas por cada clase, lo que sugiere que el contenido semántico del texto transcrito contiene rasgos informativos distintivos entre ambos perfiles. Por ejemplo, hay una zona densa y bien delimitada en el centro derecha del gráfico, donde predominan los puntos azules, así como áreas periféricas superiores e izquierdas con alta concentración de puntos rojos. Esto muestra que t-SNE, al operar sobre una representación textual suficientemente rica, logra identificar y proyectar en el espacio reducido ciertas estructuras y patrones distintivos empleados por cada grupo.

6.3.1.2 Con las 300 palabras más relevantes

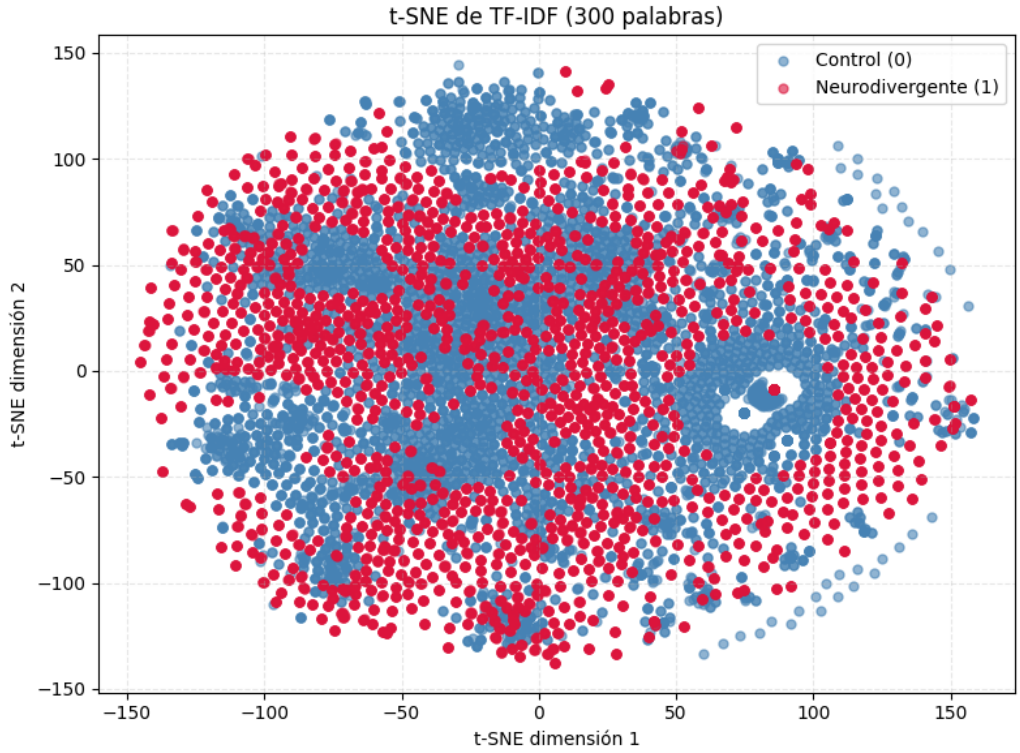


Figura 22 - Visualización con t-SNE usando las 300 palabras más relevantes

En este segundo análisis, se ha reducido el vocabulario TF-IDF utilizado para generar los vectores textuales de entrada a las 300 palabras más frecuentes, con el objetivo de analizar el impacto de esta reducción sobre la estructura proyectada por t-SNE. A pesar de esta simplificación, se sigue observando una organización espacial clara en la representación visual.

Aunque la separabilidad entre clases es menos marcada que en el caso anterior, continúan apareciendo agrupamientos evidentes. Por ejemplo, se pueden distinguir zonas densamente ocupadas por una de las dos clases, como regiones internas dominadas por las muestras de control (puntos azules), o áreas más dispersas con predominancia de muestras neurodivergentes (puntos rojos). Sin embargo, la cantidad de regiones solapadas ha aumentado, implicando una mayor mezcla de muestras entre ambas clases.

Esto sugiere que, aunque el modelo con 300 palabras sigue capturando el contenido semántico distintivo entre grupos, la reducción del vocabulario conlleva cierta pérdida de información relevante para la discriminación. Aun así, el gráfico revela que incluso con un número más limitado de características, el contenido textual contiene señales suficientes para que t-SNE extraiga patrones espaciales útiles que reflejan diferencias entre las clases.

6.3.1.3 Con las 50 palabras más relevantes

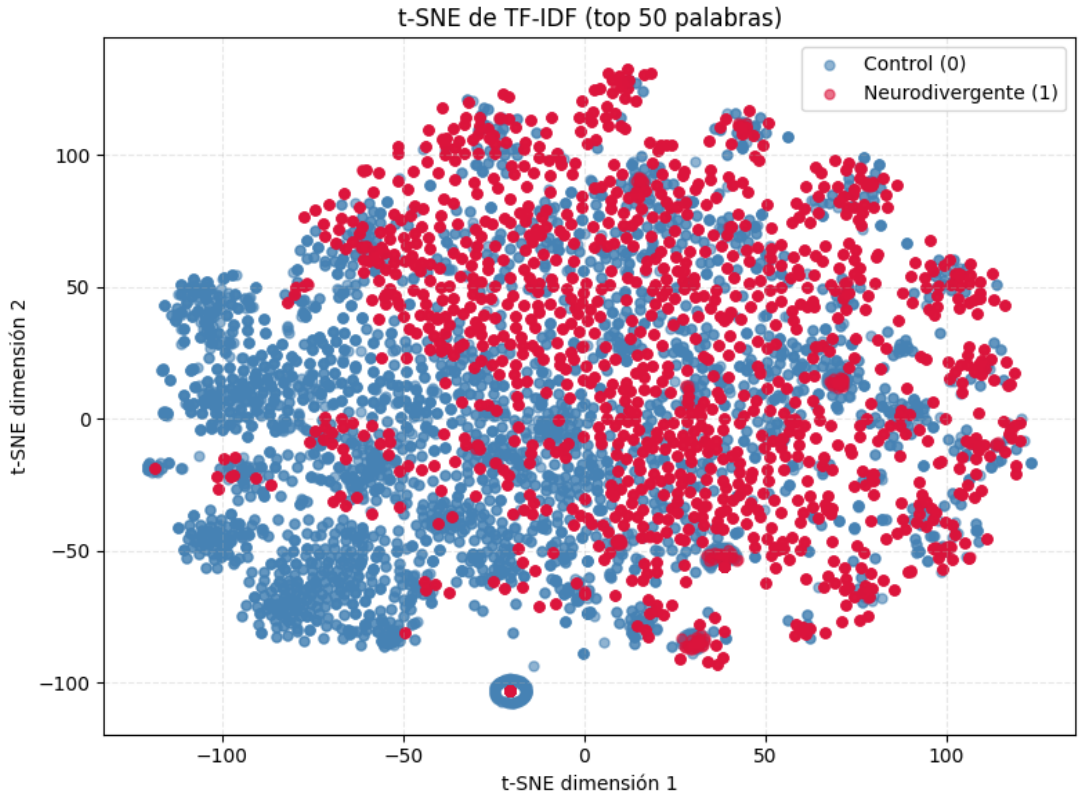


Figura 23 - Visualización con t-SNE usando las 50 palabras más relevantes

En este tercer caso, el modelo se ha restringido a un vocabulario compuesto únicamente por las 50 palabras más frecuentes del conjunto de transcripciones, con el objetivo de evaluar hasta qué punto una representación altamente comprimida conserva la capacidad de reflejar diferencias semánticas entre clases.

La visualización muestra un escenario en el que la separación entre clases es ciertamente más difusa que en los casos anteriores. A diferencia de en dichos casos, en esta vez predominan las regiones de mezcla. Los puntos de clase de control y los de clase neurodivergente se encuentran más entrelazados, lo que dificulta la identificación de agrupamientos consistentes por clase.

A pesar de esta pérdida de definición, pueden observarse aún ciertos indicios de estructura. Se mantiene un núcleo compacto de puntos azules en la parte izquierda del gráfico, así como una ligera concentración de puntos rojos hacia la mitad derecha. Sin embargo, estas agrupaciones son más débiles, y el solapamiento entre clases es considerable.

Esto refleja cómo la reducción drástica del vocabulario elimina parte de la riqueza semántica necesaria para distinguir con precisión ambos perfiles. El t-SNE proyecta con mayor dispersión, y las estructuras visuales que reflejan clases distintas tienden a diluirse a medida que el número de características decrece.

6.3.1.4 Con las 10 palabras más relevantes

Las 10 palabras más relevantes, por valor absoluto de su coeficiente, se muestran a continuación:

Palabra	Coeficiente en el modelo
Autism	10.919077
Dyslexia	7.238995
Autistic	5.888901
Read	5.026503
Dyslexic	4.934355
Things	4.703112

Brain	4.460925
Attention	4.392295
Child	4.358873
Diagnosis	3.972264

Tabla 7 – 10 palabras más relevantes por coeficiente (en valor absoluto)

Para este último caso, se ha construido una representación utilizando exclusivamente las 10 palabras más relevantes para la predicción. Estas palabras incluyen términos altamente significativos, como se puede apreciar en el ranking, con palabras como *autism, dyslexia, autistic, read, dyslexic* o *attention*. Estos términos reflejan directamente el contenido característico de los perfiles analizados.

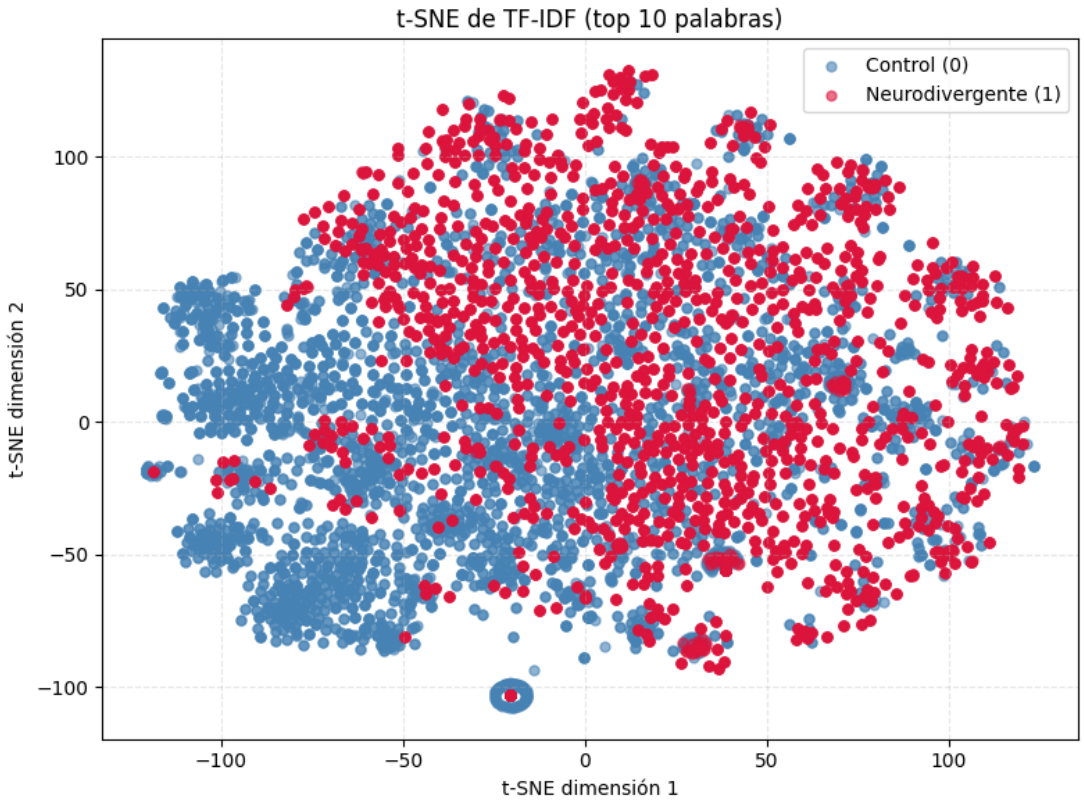


Figura 24 - Visualización con t-SNE usando las 10 palabras más relevantes

La visualización t-SNE generada con esta configuración muestra un patrón mucho más difuso que en los casos anteriores con mayor número de palabras. A pesar de que se pueden observar algunos núcleos con predominancia de puntos azules o rojos, la superposición entre clases es considerable y hay bastantes regiones que aparecen ocupadas por ambos grupos de forma indistinta. Esto era previsible, ya que al reducirse tan drásticamente la dimensionalidad del vector TF-IDF, se pierde gran parte de la riqueza semántica que permitía a t-SNE representar la estructura local con mayor fidelidad.

Sin embargo, lo más llamativo es que, a pesar de esta pérdida visual de separabilidad, el modelo entrenado únicamente con estas 10 palabras consigue unos resultados notablemente competitivos.

	Precision	Recall	F1 – Score	Support
0 (control)	0.66	0.94	0.77	1532
1 (neurodiv.)	0.89	0.52	0.66	1549
Accuracy		0.73		3081
Macro avg	0.78	0.73	0.72	3081
Weighted avg	0.78	0.73	0.72	3081

Tabla 8 – Métricas del modelo con únicamente las 10 palabras más relevantes

El modelo lingüístico entrenado exclusivamente con estas 10 palabras tiene una exactitud global (accuracy) del 73%, con unos valores F1-Score de 0.77 para la clase de control (0) y de 0.66 para la clase neurodivergente (1). Aunque el rendimiento global es inferior al modelo de 1000 palabras (que obtenía una accuracy del 82% y F1-Scores por encima del 80% en

ambas clases), los resultados siguen siendo sólidos, especialmente si se considera que el modelo está operando con un conjunto de variables 100 veces inferior.

La pérdida de rendimiento se observa principalmente en la clase 1, que ha experimentado una caída significativa en el recall (del 79% al 52% en este modelo, 27 puntos porcentuales), lo que significa que el modelo con solo 10 palabras tiende a fallar más al identificar correctamente a individuos neurodivergentes. Sin embargo, su precisión en esta clase sigue siendo elevada (de 0.89), lo que sugiere que cuando predice neurodivergencia, lo hace con alta fiabilidad. Esto, en resumen, indica que el modelo consigue mantener una alta capacidad de discriminación aún con un conjunto mínimo de términos (altamente representativos).

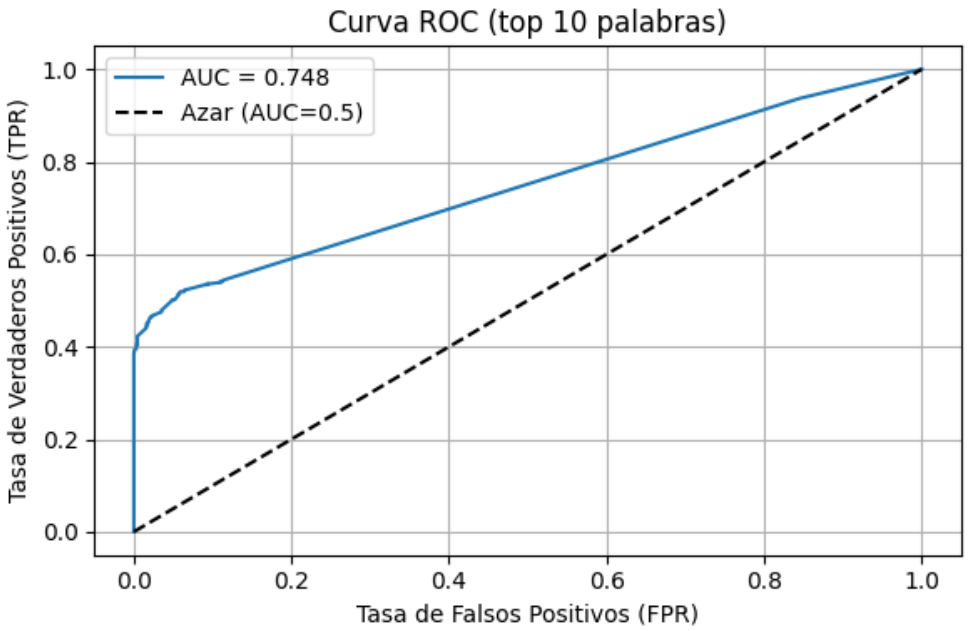


Figura 25 - Curva ROC del modelo entrenado con las 10 palabras más relevantes

La curva ROC asociada a este modelo, con un AUC del 74.8%, confirma visualmente lo que se ha observado en las métricas. A pesar de una caída respecto al modelo completo (con un AUC del 89.7%), el área bajo la curva sigue indicando un rendimiento bastante superior al azar y refleja una capacidad razonable de separación entre clases. El hecho de que

únicamente 10 palabras, seleccionadas por su alta contribución al modelo, sean capaces de generar un clasificador con esta AUC refuerza su peso informativo.

Al compararlo con el modelo de 1000 palabras, puede afirmarse que, aunque la riqueza semántica global aporta valor añadido, una fracción sustancial de la capacidad predictiva ya está contenida en estas 10 palabras más significativas. Esta observación es altamente relevante en la práctica, ya que permite que se puedan plantear modelos más simples, rápidos de entrenar y más fácilmente interpretables, sin renunciar por completo al rendimiento predictivo.

6.3.2 INTERPRETABILIDAD MEDIANTE REGRESIÓN LOGÍSTICA REDUCIDA

Con el fin de favorecer la interpretabilidad del modelo, se ha entrenado una regresión logística reducida con las 50 palabras más relevantes (según sus coeficientes). Este proceso busca determinar cuáles de estos términos tienen una contribución estadísticamente relevante a la predicción de la clase objetivo, permitiendo identificar el peso específico de cada palabra y aportar al análisis una perspectiva explicativa complementaria.

En la tabla de resultados, se presentan los coeficientes estimados junto con sus errores estándar, p-valores y los intervalos de confianza al 95%. Este desglose permite observar tanto la dirección como la significancia estadística del efecto de cada término.

Como puede apreciarse, los términos con coeficientes positivos y significativos (esto es, con un p-valor < 0.05) están asociados a una mayor probabilidad de pertenecer al grupo neurodivergente. Por ejemplo, palabras como *reading, read, diagnosis, brain, attention, learn, difficult* o *help* presentan tanto valores positivos elevados como p-valores muy bajos, indicando que tienen un peso predictivo claro. Estas palabras remiten a conceptos relacionados con el entorno clínico, lo cognitivo, el aprendizaje y la experiencia subjetiva de dificultades, lo que puede indicar una fuerte carga semántica ligada a discursos típicos de personas neurodivergentes.

Por otro lado, varios términos aparecen con coeficientes negativos significativos, como *make, money, guy, called,* o *10*. Estos términos parecen responder patrones conversacionales más genéricos o triviales, más comunes en la clase de control (tiene sentido también por la naturaleza de los podcasts que específicamente se indicaron en Apify para extraer los vídeos). El hecho de que estas palabras sean significativas y tengan peso negativo indica que su presencia en la transcripción reduce la probabilidad de que esta pertenezca al grupo neurodivergente.

Además del análisis de los coeficientes, se visualizó el intervalo de confianza para cada término en un gráfico, que se muestra a continuación:

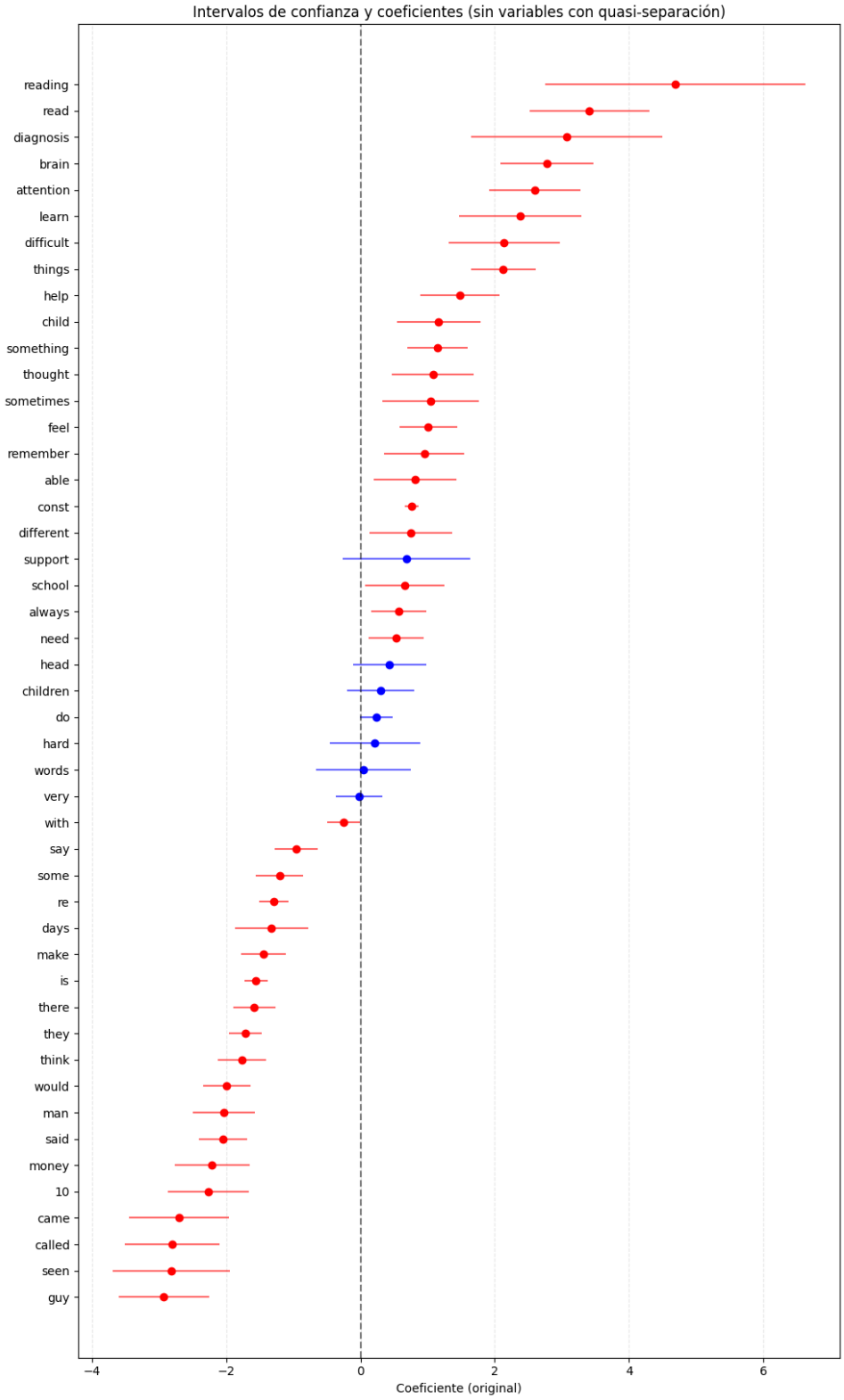


Figura 26 - Gráfico de las 50 variables más significativas y sus intervalos de confianza

En él, cada punto representa el coeficiente de una palabra, mientras que las líneas indican el intervalo de confianza del 95%. El color rojo se asigna a variables con un p-valor menor a 0.05 (las estadísticamente significativas), mientras que el azul identifica a aquellas sin evidencia estadística suficiente. La mayoría de coeficientes significativos se ubican en los extremos del gráfico, indicando su mayor impacto predictivo.

Este tipo de análisis permite identificar con una mayor precisión cuáles son los aspectos de las experiencias comunicadas por los participantes que están contribuyendo en mayor medida a la clasificación. Obteniendo así más información para el estudio del lenguaje como marcador de perfiles neurodivergentes.

6.4 EVALUACIÓN POR GRUPOS DE VARIABLES

En este siguiente punto, con el fin de valorar el peso relativo de diferentes bloques temáticos, se entrenaron modelos separados con un solo grupo de variables.

6.4.1 VARIABLES EMOCIONALES (FACIALES)

Este grupo se compone de siete variables que representan la probabilidad media de expresión de emociones básicas detectadas en el rostro, y son *angry_facial, disgust_facial, fear_facial, happy_facial. sad_facial, surprise_facial* y *neutral_facial*. Este grupo busca encontrar patrones emocionales no verbales que puedan estar vinculados a diferencias en la expresión facial entre individuos de ambas clases.

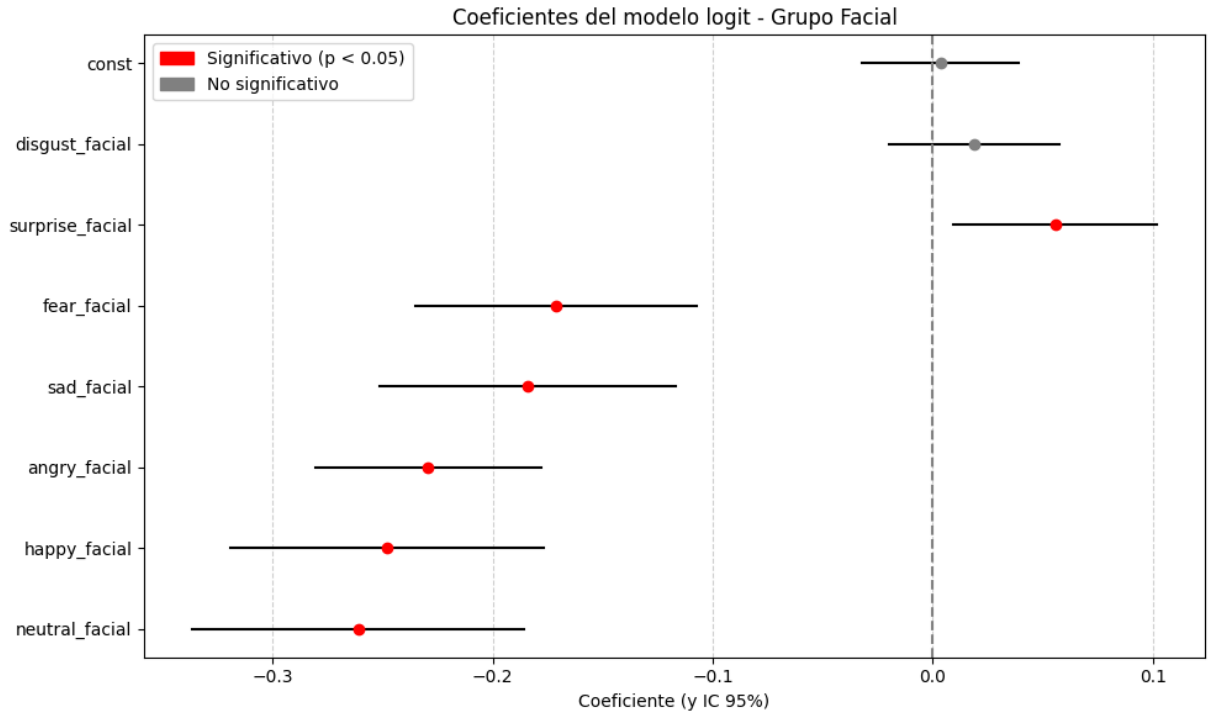


Figura 27 - Representación por intervalos de confianza y coeficientes del grupo de variables emocionales del rostro

Desde un punto de vista estadístico, el modelo Logit ajustado solo con estas variables muestra como casi todas las variables son estadísticamente significativas (con un p-valor menor a 0.05), destacando las expresiones de enfado, miedo, felicidad, tristeza y neutralidad. Se observa como también todas las expresiones menos las de disgusto y sorpresa tienen coeficientes negativos (menor probabilidad de neurodivergencia), por lo que, el hecho de aparecer disgustado o con cara de sorpresa indica una mayor probabilidad de estar asociado con el grupo de neurodivergencia. Por otro lado, sugiere que el enfado, la felicidad o la tristeza, al igual que otras emociones más básicas parecen tener una distribución más marcada en el grupo de control.

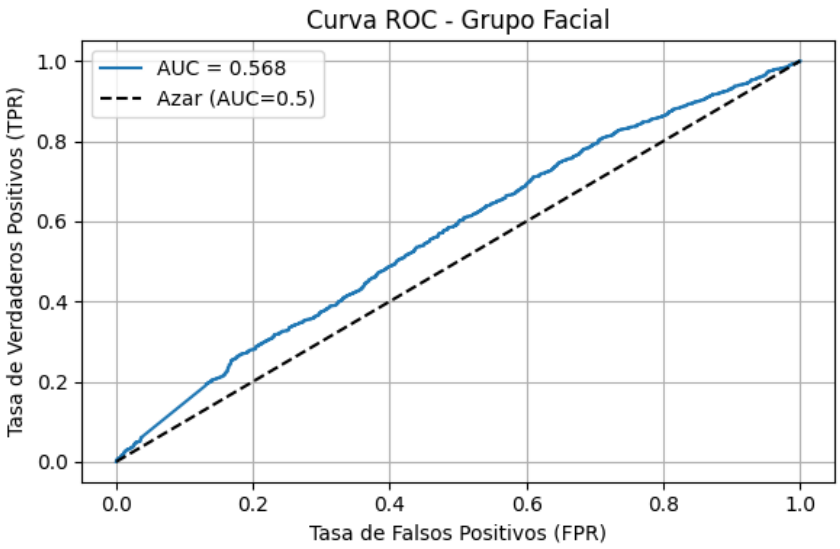


Figura 28 - Curva ROC del modelo con las variables emocionales del rostro

En cuanto al modelo predictivo, el modelo basado exclusivamente en estas variables emocionales obtiene un valor de AUC del 56.8%, muy cercano a la línea de azar. Esto indica que, aunque existen diferencias estadísticamente significativas entre grupos en algunas de las emociones faciales, estas diferencias no son lo suficientemente fuertes como para construir un clasificador robusto por sí solo con estas variables. En otras palabras, la capacidad de separar eficazmente a los sujetos únicamente utilizando las emociones faciales es limitada.

6.4.1.1 Impacto del texto sobre el grupo de variables

Ahora, se procede a combinar las variables con el texto vectorizado mediante TF-IDF de 1000 términos, obteniendo los resultados que se muestran a continuación.

	Precision	Recall	F1 – Score	Support
0 (control)	0.86	0.89	0.87	1532
1 (neurodiv.)	0.89	0.86	0.87	1549
Accuracy		0.87		3081
Macro avg	0.87	0.87	0.87	3081
Weighted avg	0.87	0.87	0.87	3081

Tabla 9 – Métricas del modelo del grupo facial con texto

La matriz de clasificación resultante muestra una notable mejora en todos los indicadores, situándose todos entre el 86% - 89%, frente al pobre rendimiento anterior. Esto sugiere que, si bien las emociones pueden ofrecer ciertas pistas, es la combinación con el lenguaje lo que permite capturar patrones más complejos y relevantes para la detección de perfiles neurodivergentes.

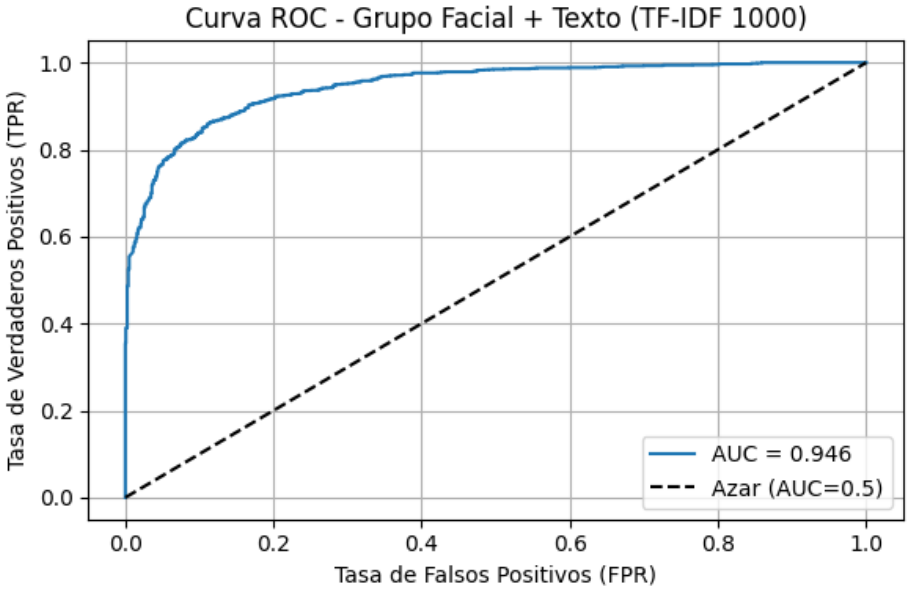


Figura 29 - Curva ROC del modelo de variables emocionales del rostro, con métricas numéricas y textuales

Se puede observar en la curva ROC como el rendimiento del modelo aumenta drásticamente, alcanzando un AUC de 0.946. Este salto representa una ganancia significativa en discriminación, reforzando la idea de que el contenido verbal aporta información fundamental que no se refleja de forma tan clara en las expresiones faciales.

6.4.2 VARIABLES DE PERSONALIDAD

El grupo de variables de personalidad incluye rasgos derivados de modelos ampliamente aceptados como el Big Five [17] y otras métricas complementarias relacionadas con aspectos intrapersonales. Concretamente, se consideran variables como extraversión, neuroticismo, amabilidad (*agreeableness*), responsabilidad (*conscientiousness*), apertura mental (*opennes*), autoestima, creatividad, imaginación, supervivencia, comunicación, conciencia y compasión. Estas dimensiones psicológicas han sido teóricamente asociadas con patrones de comportamiento, cognición y lenguaje que podrían ser útiles para distinguir entre perfiles neurodivergentes y no neurodivergentes.

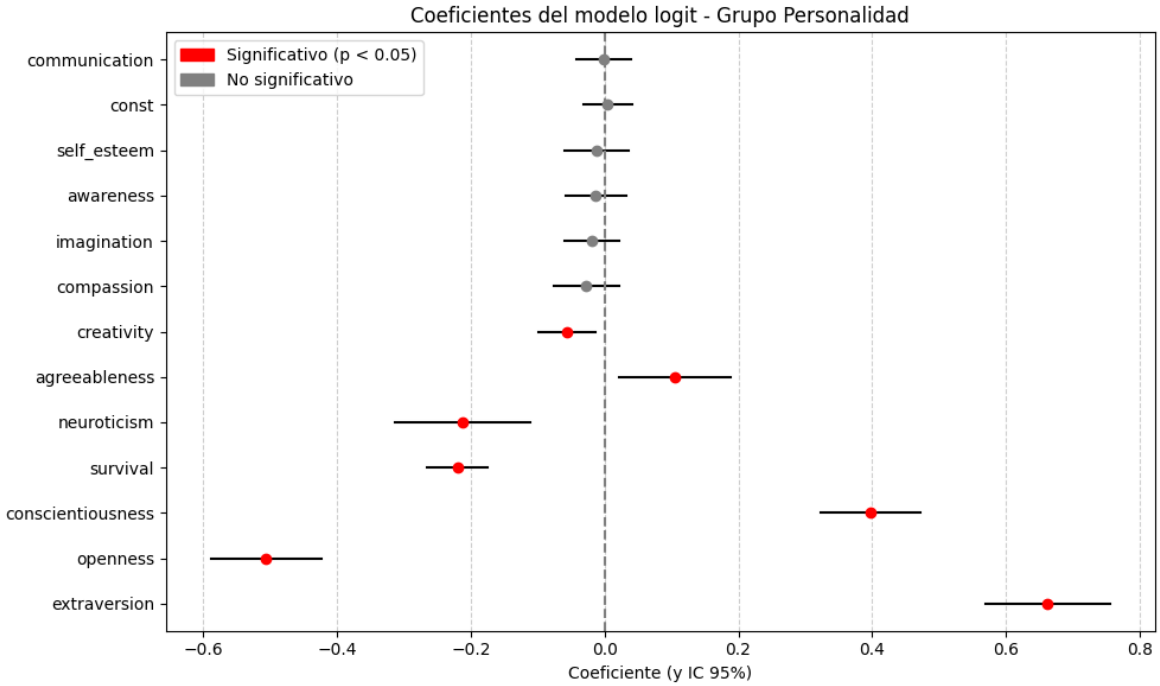


Figura 30 - Representación por grupo de confianza y coeficientes del grupo de variables de personalidad

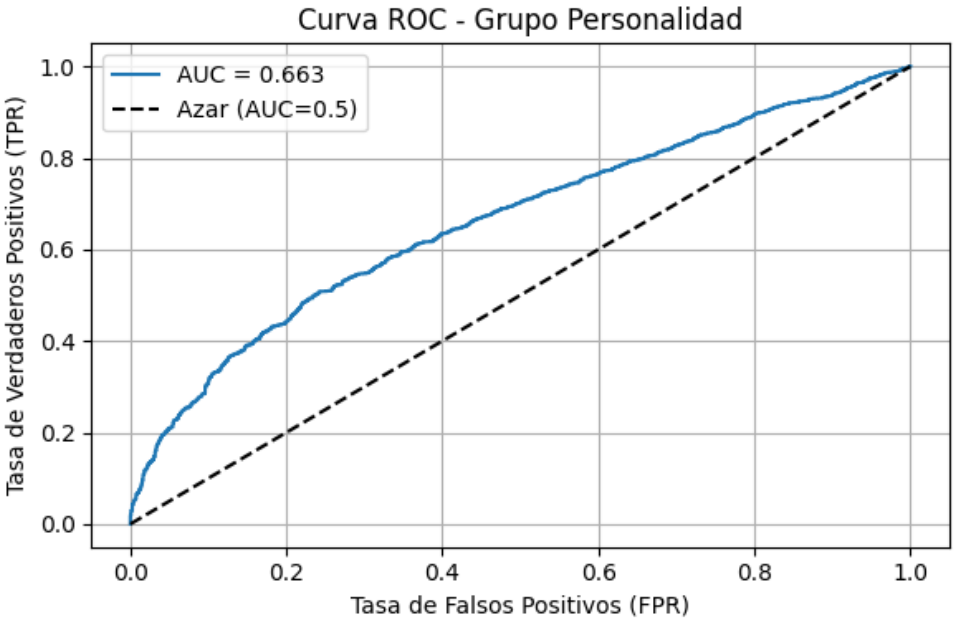


Figura 31 - Curva ROC del modelo del grupo de variables de personalidad

El modelo logit entrenado exclusivamente con estas variables de personalidad muestra un desempeño moderado, con un AUC de 0.663, que si bien está por encima del umbral de aleatoriedad, refleja una capacidad limitada de estas variables por si solas para discriminar de forma robusta entre clases.

Desde el punto de vista estadístico, varias variables muestran coeficientes significativos. En concreto, destacan:

- **Neuroticismo y *opennes*:** presentan coeficientes negativos significativos. Esto sugiere que puntuaciones más altas en estas dimensiones tienden a asociarse con una menor probabilidad de clasificación como neurodivergente. Esto puede interpretarse para cada caso:
    - Una baja puntuación en *opennes* podría reflejar una menor flexibilidad o preferencia por rutinas más estructuradas, aspectos que se suelen asociar con algunos perfiles neurodivergentes.
    - Una baja puntuación en neuroticismo (es decir, mayor estabilidad emocional) también podría ser menos común de perfiles neurodivergentes, como por

ejemplo personas autistas o con TDAH, que con frecuencia reportan niveles más altos de ansiedad o reactividad emocional.

- ***Conscientiousness* y *agreeableness*:** presentan coeficientes positivos y altamente significativos.
    - En el caso de *conscientiousness*, un mayor nivel puede reflejar la hiper-responsabilidad, necesidad de control o adherencia a normas rígidas, que son conductas observadas en perfiles neurodivergentes (TOC, autismo).
    - Altos niveles de agreeableness podrían estar relacionados con una mayor tendencia a la cooperación y a la sensibilidad interpersonal, y con un mayor esfuerzo consciente por adaptarse socialmente, algo que algunos individuos neurodivergentes desarrollan como estrategia compensatoria.

6.4.2.1 Impacto del texto sobre el grupo de variables

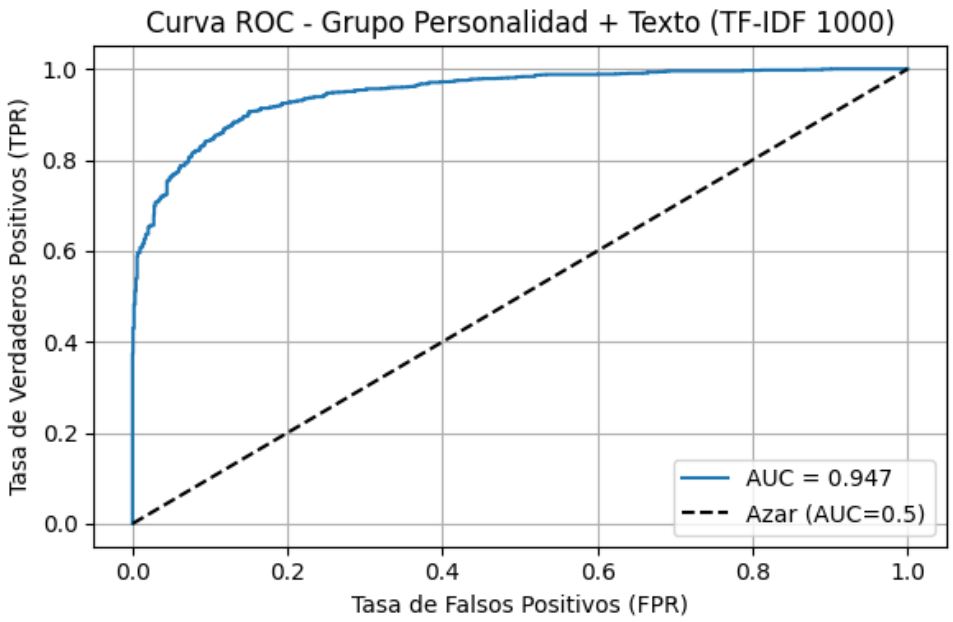


Figura 32 - Curva ROC del modelo de variables de personalidad, combinado con datos textuales

Cuando se combinan estas variables con los vectores de texto generados a partir del análisis TF-IDF de las 1000 palabras más frecuentes, el rendimiento mejora de forma drástica, alcanzando un AUC de 0.947. Esta ganancia sustancial refleja que la información verbal contenida en las transcripciones textuales refuerza y complementa significativamente los patrones detectables a partir de los rasgos de personalidad.

	Precision	Recall	F1 – Score	Support
0 (control)	0.85	0.89	0.87	1532
1 (neurodiv.)	0.89	0.85	0.87	1549
Accuracy		0.87		3081
Macro avg	0.87	0.87	0.87	3081
Weighted avg	0.87	0.87	0.87	3081

Tabla 10 – Métricas del grupo de personalidad combinado con texto

La matriz de clasificación confirma esta mejora, mostrando métricas de accuracy, precisión, recall y F1-Score de entorno al 87% en ambas clases. La combinación entre la riqueza semántica del lenguaje utilizado y las características individuales de personalidad permite al modelo capturar de forma mucho más precisa las diferencias entre perfiles.

6.4.3 VARIABLES DE ESTADO EMOCIONAL / PSICOLÓGICO

Este grupo está compuesto por variables que reflejan estados internos de tipo afectivo y psicológico. Estos son el estrés, la indefensión, la autoeficacia o la depresión. Se incluyen versiones codificadas en niveles (bajo, medio, alto) para cada variable, lo que permite capturar matices en la intensidad de estas experiencias. Estas variables se consideran especialmente relevantes para este análisis, ya que numerosos estudios han evidenciado que los perfiles neurodivergentes pueden experimentar patrones emocionales y psicológicos diferenciados, tanto en intensidad como en regulación de los mismos.

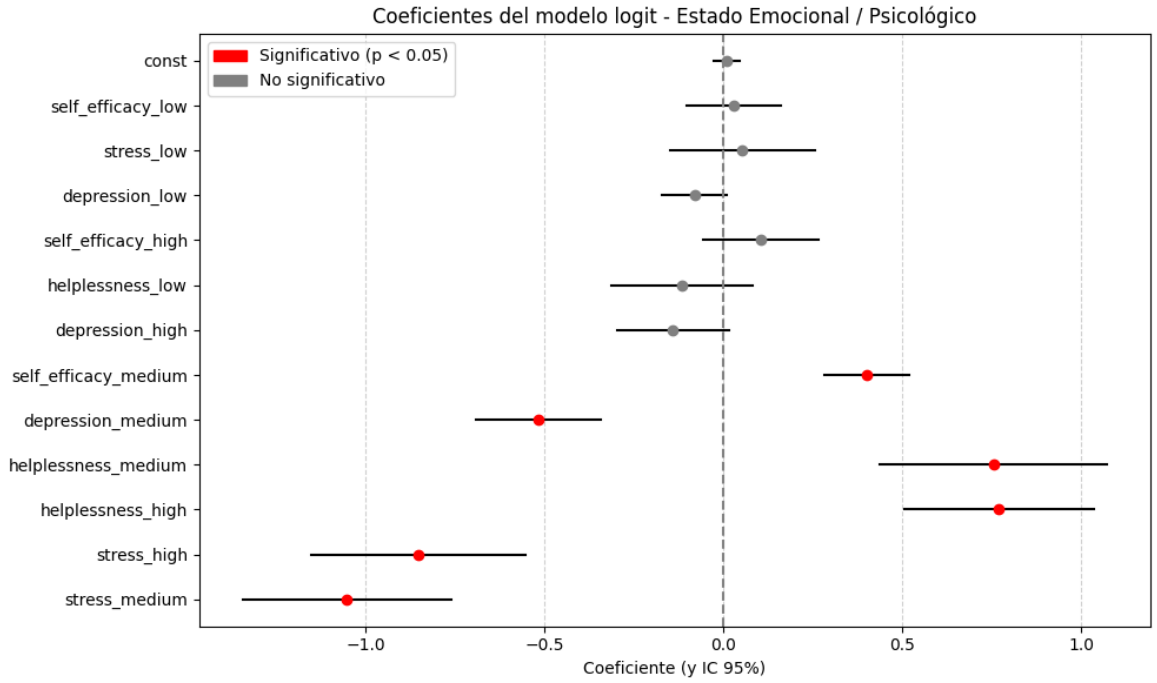


Figura 33 - Representación por grupo de confianza y coeficientes del grupo de variables de estado emocional / psicológico

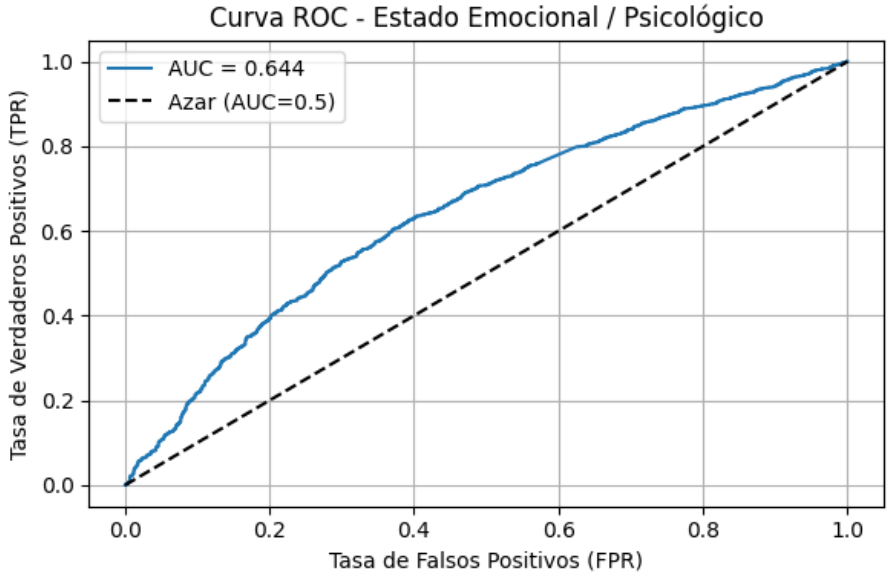


Figura 34 - Curva ROC del grupo de variables de estado emocional / psicológico

El modelo logit ajustado únicamente con las variables de este grupo muestra un rendimiento discreto, con un AUC del 64.4%. Aunque este valor se encuentra por encima del azar, indica que las variables psicológicas por sí solas no permiten una separación robusta entre perfiles neurodivergentes y no neurodivergentes.

Desde un punto de vista estadístico, varias variables muestran coeficientes significativos. En particular:

- **Estrés:** los niveles medio y alto de estrés presentan coeficientes negativos estadísticamente significativos ($p < 0.001$). Esto sugiere que a medida que aumenta el estrés, disminuye la probabilidad de clasificación como neurodivergente. Este resultado, aunque puede sonar contraintuitivo, podría explicarse por un posible efecto de desensibilización o adaptación al estrés en algunos perfiles no neurodivergentes, o por la presencia de estrategias de afrontamiento más visibles o explícitas en estos casos.

- **Indefensión (*helplesness*):** los niveles medio y alto muestran coeficientes positivos y significativos, indicando que a mayor sensación de indefensión, mayor probabilidad de ser clasificado como neurodivergente. Este patrón tiene sentido, ya que sentimientos de falta de control o desesperanza son comúnmente reportados por personas con TDAH, autismo u otras condiciones neurodivergentes, especialmente en contextos sociales o estructurales poco adaptativos.
- **Depresión media (*depression_medium*):** significativo y con un coeficiente negativo, lo que indica que puntuaciones moderadas en síntomas depresivos se asocian con menor probabilidad de ser clasificado como neurodivergente. Esto puede deberse a que, en ciertos casos, la sintomatología depresiva moderada se da también en perfiles no neurodivergentes, y no resulta un rasgo tan específico como otros estados emocionales más extremos.

6.4.3.1 Impacto del texto sobre el grupo de variables

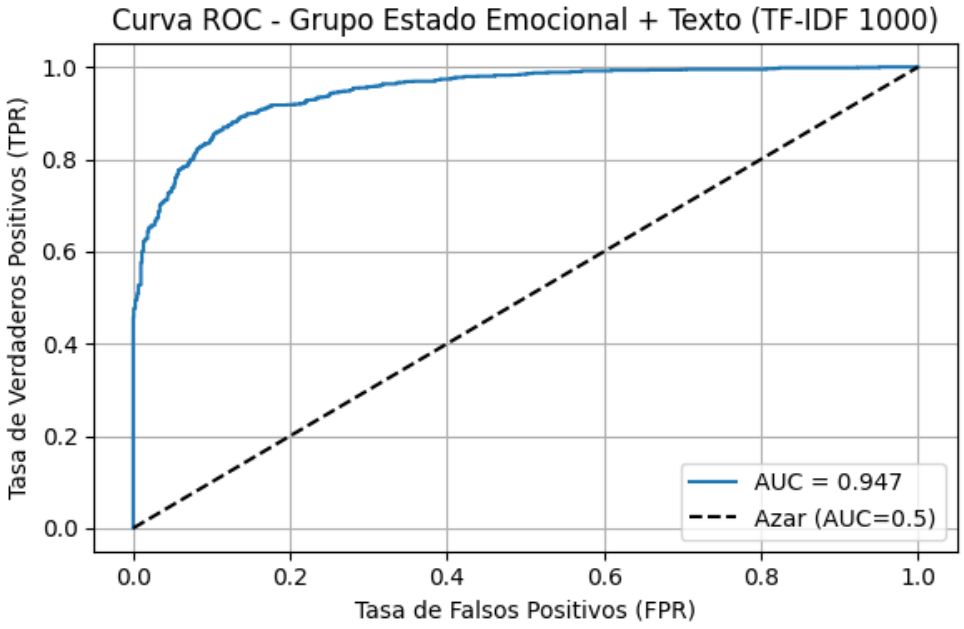


Figura 35 - Métricas y curva ROC del modelo con las variables del grupo de estado emocional, con texto añadido

La incorporación de las variables textuales mediante TF-IDF con las 1000 palabras más frecuentes mejora drásticamente el rendimiento del modelo, alcanzando un AUC del 94.7%. Esta diferencia de más de 30 puntos porcentuales con respecto al modelo basado únicamente en variables numéricas refuerza la idea de que el contenido verbal expresa de forma mucho más extensa y matizada estos estados internos, captando detalles implícitos en el lenguaje y otros rasgos lingüísticos altamente predictivos, que no se podrían capturar solo con las variables puramente numéricas.

	Precision	Recall	F1 – Score	Support
0 (control)	0.86	0.89	0.88	1532
1 (neurodiv.)	0.89	0.86	0.87	1549
Accuracy		0.88		3081
Macro avg	0.88	0.88	0.88	3081
Weighted avg	0.88	0.88	0.88	3081

Tabla 11 – Métricas del grupo de estado emocional con texto

Las métricas obtenidas (precisión, recall y F1-score en torno al 88%) evidencian que la combinación con el lenguaje permite capturar perfiles emocionales de manera mucho más efectiva y diferenciada. El texto no solo complementa, sino que amplifica la sensibilidad del modelo a las diferencias latentes entre ambos grupos.

6.4.4 VARIABLES VOCALES

Este grupo recoge características cuantitativas relacionadas con la voz de los participantes, como la media, mediana, desviación típica, percentiles (Q25, Q75), rango intercuartílico (IQR), curtosis, sesgo, tono, pitch y modo de la señal. Estas métricas capturan propiedades estadísticas del audio, como la variabilidad, la energía y la distribución espectral de la voz, que podrían estar relacionadas con el estado emocional o patrones de habla específicos de perfiles neurodivergentes.

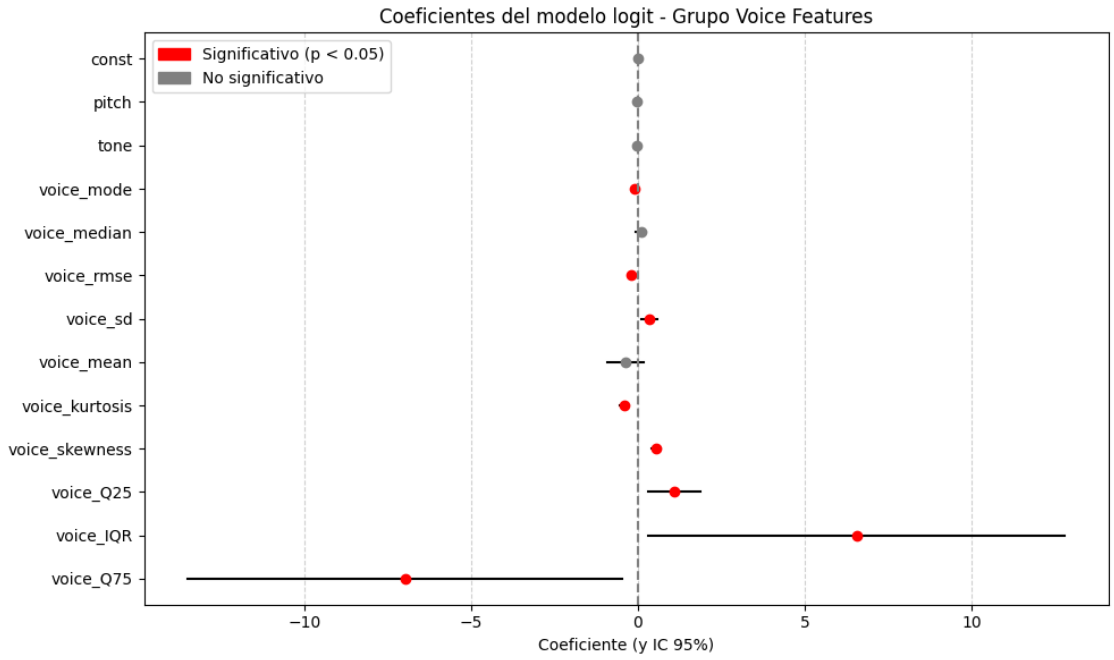


Figura 36 - Representación por intervalo de confianza y coeficientes usando el grupo de variables vocales

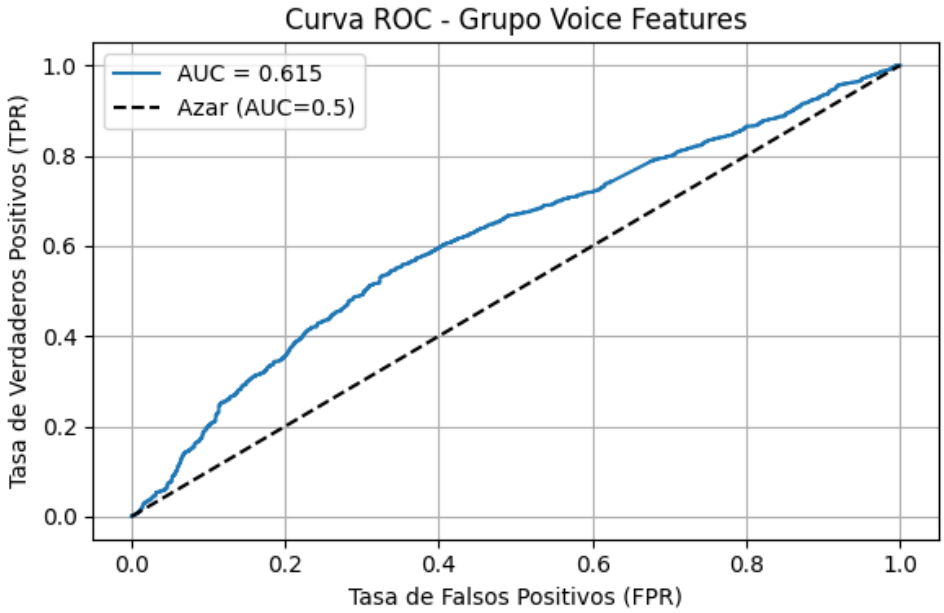


Figura 37 - Curva ROC del grupo de variables vocales

El modelo logit entrenado exclusivamente con estas variables vocales obtuvo un AUC de 0.615. Un valor modesto, aunque superior al azar. Esto indica que, aunque las variables vocales contienen información útil, no son suficientemente potentes por sí solas para una discriminación entre clases.

Desde el punto de vista estadístico, varias variables resultaron significativas:

- **Percentiles *Q25* y *Q75*, y *IQR*:** se muestran como variables significativas, por lo que ciertos rangos de intensidad o dispersión de la voz pueden ser más frecuentes en una de las dos clases. Por ejemplo, un valor alto en *Q75* o *IQR* puede indicar mayor variabilidad o energía en la voz, lo cual se suele relacionar con perfiles más expresivos o impulsivos, presentes normalmente en personas con TDAH.
- Variables como ***voice_sd*** y ***voice_mode*** también resultan significativas por su pequeño p-valor, mostrando su peso en el desempeño del modelo.

Por otro lado, variables como *tone*, *pitch* o *voice_mean* no resultaron significativas estadísticamente. Esto sugiere que ciertas medidas convencionales de tono e intensidad podrían no ser suficientemente distintivas entre perfiles.

6.4.4.1 Impacto del texto sobre el grupo de variables

	Precision	Recall	F1 - Score	Support
0 (control)	0.86	0.89	0.88	1532
1 (neurodiv.)	0.88	0.86	0.87	1549
Accuracy		0.87		3081
Macro avg	0.87	0.87	0.87	3081
Weighted avg	0.87	0.87	0.87	3081

Tabla 12 – Métricas del grupo de voice features con texto

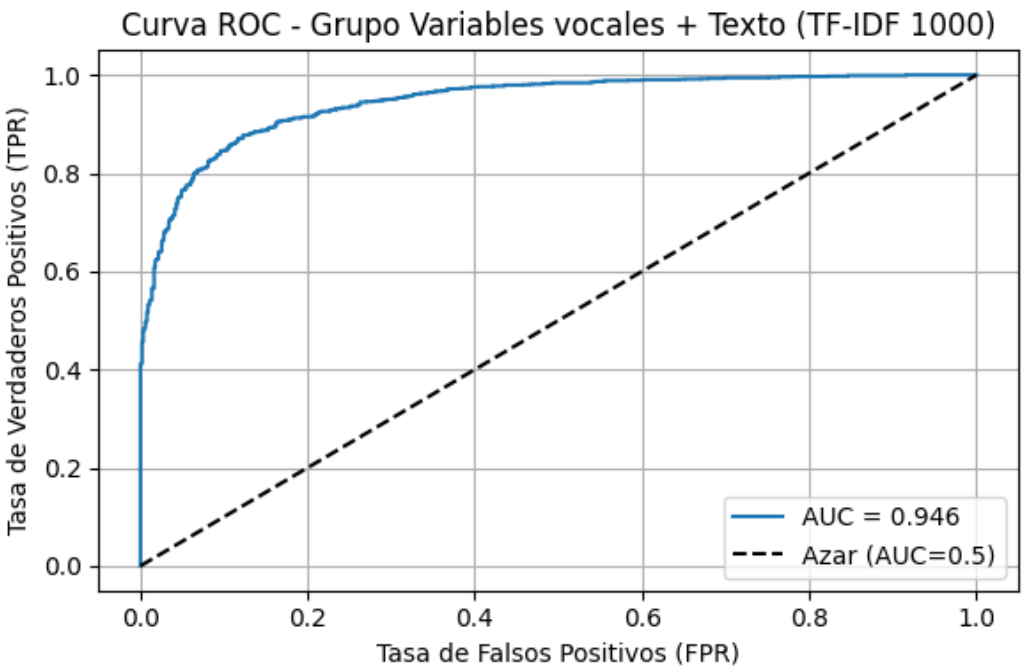


Figura 38 - Métricas y curva ROC del grupo de variables vocales, con texto añadido

Cuando se combinan las variables vocales con el modelo TF-IDF de las 1000 palabras más relevantes en las transcripciones, el rendimiento mejora drásticamente, alcanzando un AUC de 0.946, comparable al resto de modelos combinados. Además, se observan métricas equilibradas para las dos clases, siendo la precisión, el recall y el F1-score cercanas al 88%.

Esta mejora tan significativa sugiere que, aunque los indicadores vocales por sí solos no sean altamente predictivos, sí aportan información complementaria al contenido verbal, reforzando la capacidad del modelo para discriminar entre perfiles. Posiblemente, la entonación, la modulación y otros matices de la voz permiten matizar o enfatizar elementos del discurso que ayudan a revelar patrones asociados a la neurodivergencia.

6.4.5 VARIABLES EMOCIONALES DE LA VOZ

Este conjunto de variables busca encontrar las emociones directamente del tono de voz de los participantes, identificando emociones como tristeza, felicidad, miedo, enfado, sorpresa o calma, entre otras. El objetivo de incluir estas variables es comprobar si la expresión emocional de la voz proporciona información útil para distinguir entre perfiles neurodivergentes y no neurodivergentes, más allá de lo puramente textual o visual.

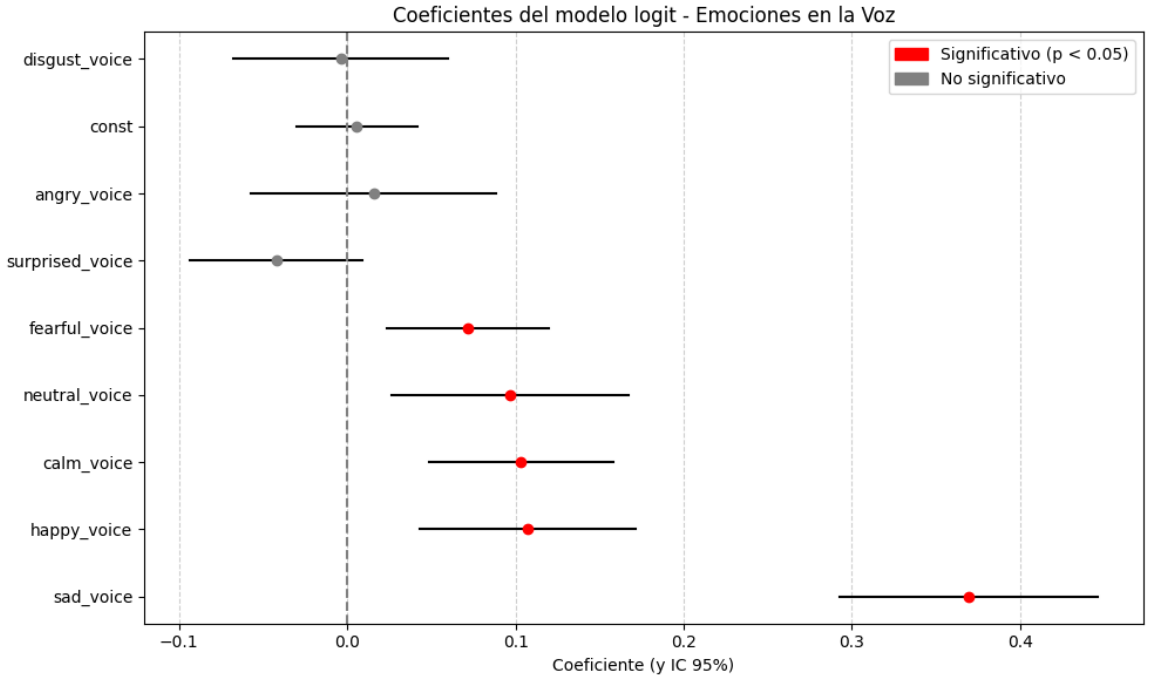


Figura 39 – Representación de las variables del grupo por intervalos de confianza y coeficientes

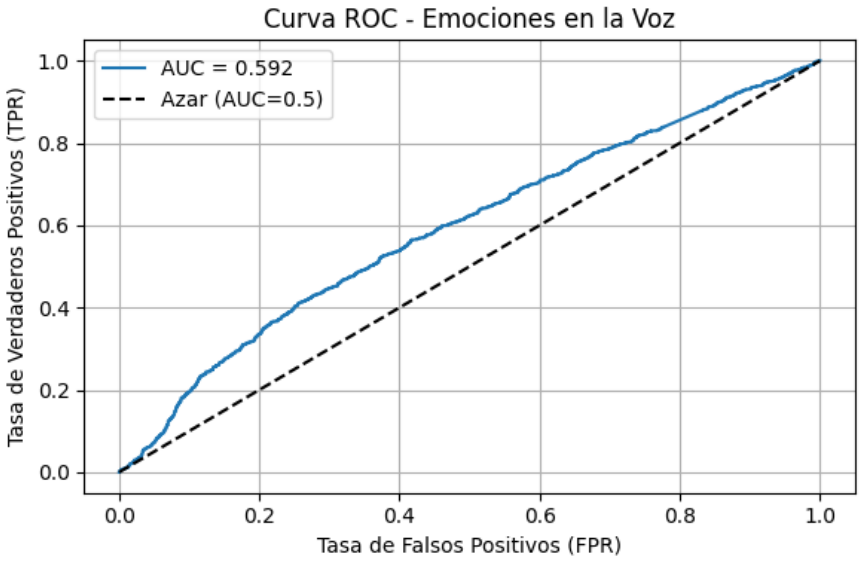


Figura 40 - Curva ROC del grupo de variables emocionales de la voz

El modelo logístico entrenado exclusivamente con las variables de emociones vocales obtiene un resultado limitado, con un AUC del 59.2%. Esta cifra está apenas por encima del umbral de aleatoriedad (AUC = 0.5), lo que sugiere que, por sí solas, estas variables tienen una capacidad baja para discriminar entre clases.

No obstante, algunas variables individuales muestran coeficientes estadísticamente significativos ($p < 0.05$), como:

- ***Sad_voice:*** coeficiente positivo y altamente significativo. Una puntuación alta en tristeza vocal se asocia con mayor probabilidad de pertenecer al grupo neurodivergente, lo cual es coherente con la teoría, que vincula expresiones de afecto negativo con ciertos perfiles neurodivergentes.
- ***Calm_voice* y *happy_voice*:** también tienen coeficientes positivos significativos, lo que puede resultar algo más ambiguo desde un punto de vista interpretativo. Una posible explicación es que algunos perfiles neurodivergentes pueden manifestar expresividad emocional intensa o desajustada en distintos contextos.

Por el contrario, variables como *fearful_voice* o *surprised_voice* no presentan significancia estadística clara, lo que puede sugerir que no aportan información relevante para clasificar en este modelo aislado.

6.4.5.1 Impacto del texto sobre el grupo de variables

	Precision	Recall	F1 - Score	Support
0 (control)	0.86	0.88	0.88	1532
1 (neurodiv.)	0.88	0.86	0.87	1549
Accuracy		0.87		3081
Macro avg	0.87	0.87	0.87	3081
Weighted avg	0.87	0.87	0.87	3081

Tabla 13 – Métricas del grupo de variables de emociones de la voz con texto

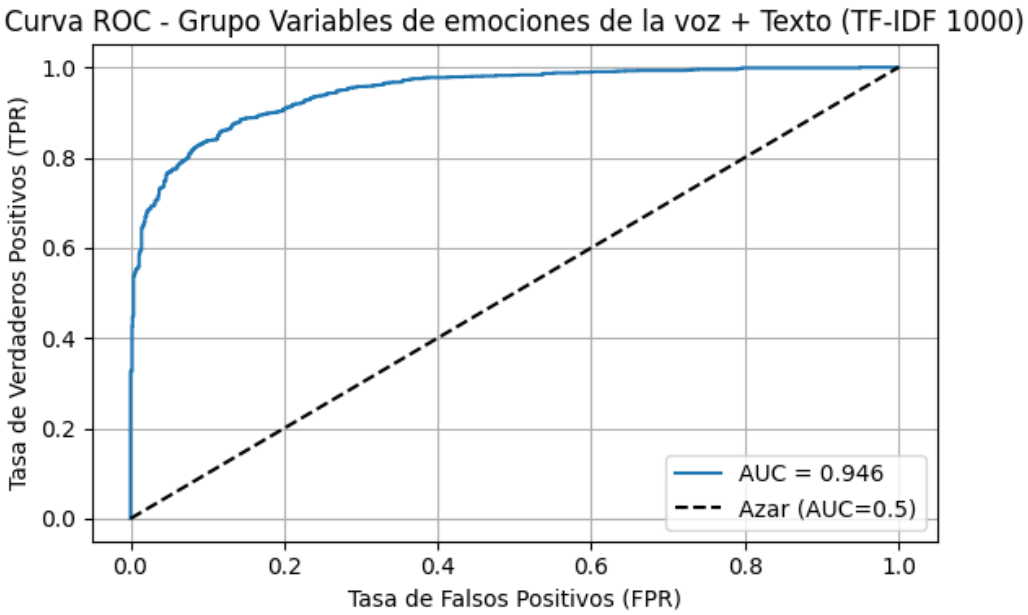


Figura 41 - Curva ROC del modelo de variables emocionales de la voz con texto

Cuando se combinan las variables de emociones vocales con los vectores TF-IDF del texto, usando las 1000 palabras más frecuentes, el rendimiento mejora de forma considerable también. El modelo combinado alcanza un AUC de 0.946, frente al 0.592 del modelo anterior. Esta diferencia evidencia que el texto transcrito aporta información muy valiosa, que no sólo complementa sino que multiplica la capacidad predictiva de estas variables.

Además, se observa un aumento significativo en todas las métricas de clasificación. Precisión, recall y F1-score en ambas clases alcanzan el 87–88%, lo que refleja una mejora robusta y consistente en la discriminación entre grupos. Esto permite afirmar que, si bien las emociones vocales por sí solas son limitadas, su interacción con el contenido verbal puede descubrir patrones relevantes no detectables mediante un único tipo de señal.

6.4.6 OTRAS VARIABLES

Este grupo recoge características más heterogéneas y complementarias respecto al resto de dimensiones analizadas. En concreto, incluye métricas como la probabilidad de no habla (*no_speech_prob*), entropía (*entropy*), orientación temporal de las expresiones verbales (*tense_past, tense_present, tense_future*) y medidas de polaridad y subjetividad del texto (*sentiment_polarity* y *sentiment_subjectivity*). A pesar de no formar un bloque homogéneo, estas variables pueden aportar información indirecta sobre el estilo de comunicación, la carga emocional del discurso o la forma de estructurar el relato temporalmente.

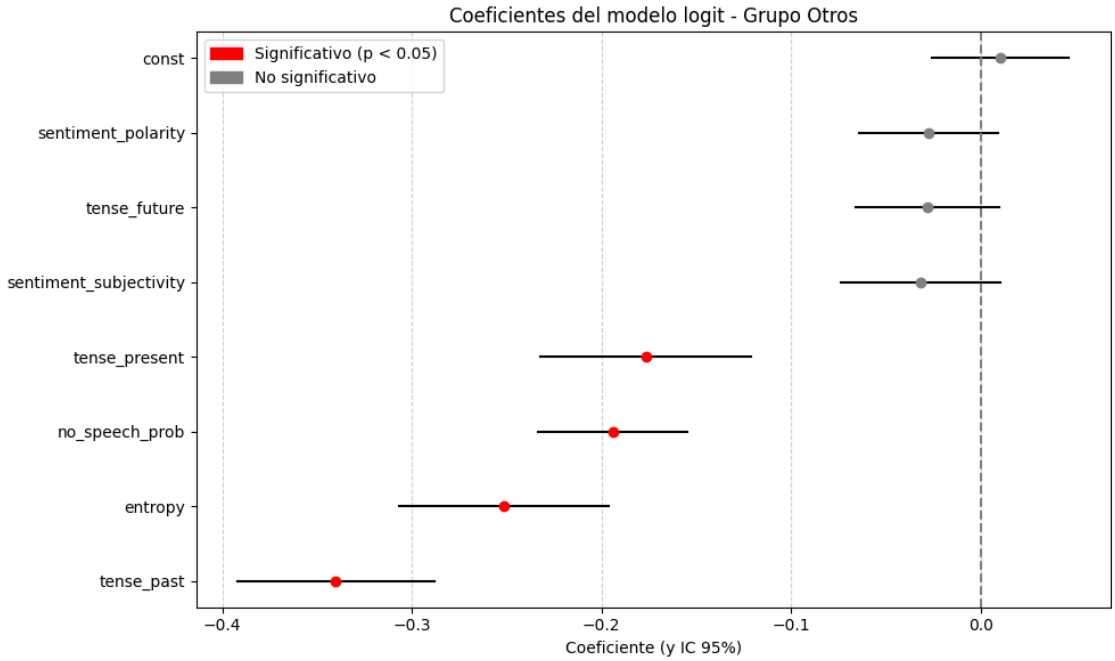


Figura 42 - Representación de las variables del grupo de otras variables por intervalos de confianza y coeficientes

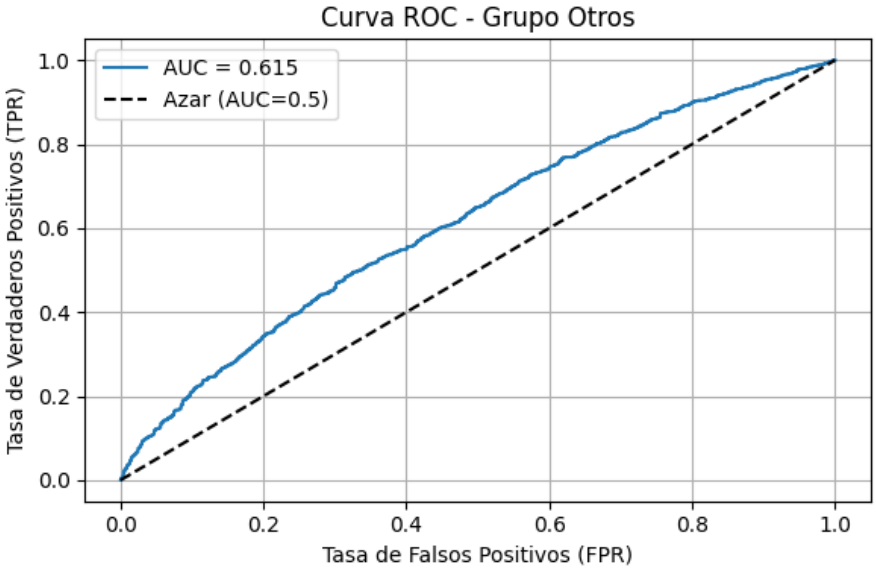


Figura 43 - Curva ROC del modelo de otras variables

El modelo logit entrenado exclusivamente con estas variables muestra un rendimiento moderado, con un AUC de 0.615. Aunque este valor se sitúa ligeramente por encima del umbral de aleatoriedad, indica una capacidad discriminativa limitada cuando se utilizan estas variables por sí solas.

Desde un punto de vista estadístico, destacan principalmente tres variables con coeficientes negativos y p-valores estadísticamente significativos:

- **No Speech probability:** Un mayor valor en esta variable (es decir, menos habla detectada) se asocia con una mayor probabilidad de pertenecer al grupo neurodivergente. Esto podría reflejar pausas más largas, dificultades de fluidez o reticencia al discurso, características que pueden darse en perfiles autistas o con ansiedad social.
- **Entropy:** Un menor valor de entropía (que sugiere menor variabilidad en el discurso), también se relaciona con neurodivergencia. Esto puede deberse a un lenguaje más repetitivo, literal o concreto, que limita la diversidad expresiva.
- **Past tense:** Una menor presencia de tiempos pasados se asocia con la clase neurodivergente. Esto podría interpretarse como una menor tendencia a narrar

experiencias pasadas o dificultades para estructurar relatos temporales complejos, observadas en algunos perfiles neurodivergentes.

6.4.6.1 Impacto del texto sobre el grupo de variables

	Precision	Recall	F1 - Score	Support
0 (control)	0.85	0.90	0.87	1532
1 (neurodiv.)	0.89	0.84	0.87	1549
Accuracy		0.87		3081
Macro avg	0.87	0.87	0.87	3081
Weighted avg	0.87	0.87	0.87	3081

Tabla 14 – Métricas del grupo de otras variables con texto

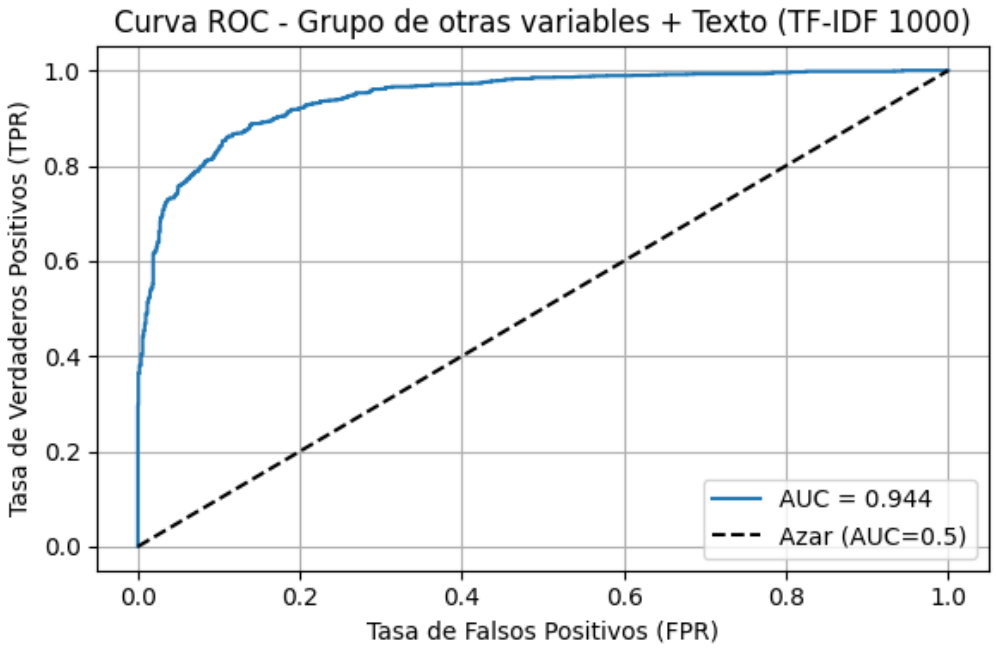


Figura 44 - Curva ROC del modelo con las otras variables, mas texto

Cuando estas variables se combinan con las 1000 palabras más relevantes extraídas mediante TF-IDF, el rendimiento mejora notablemente. El AUC alcanza un valor de 0.944, muy similar al observado en otros grupos más estructurados. Esta mejora sustancial pone de manifiesto que, si bien estas variables por sí solas aportan poco valor predictivo, su combinación con las características lingüísticas permite reforzar el modelo y capturar patrones de comportamiento comunicativo útiles para la clasificación.

Las métricas del modelo combinado muestran un rendimiento robusto y equilibrado. Un accuracy del 87%, acompañado de precisiones y recalls similares en ambas clases, que confirman que la complementariedad entre indicadores de estilo discursivo y contenido semántico enriquece la capacidad discriminativa del sistema.

6.5 COMPARACIÓN GLOBAL Y CONCLUSIONES DEL MODELO EXPLICATIVO

Con el objetivo de evaluar de forma integrada la capacidad predictiva de los diferentes grupos de variables analizados, se ha construido un modelo que combina todas las fuentes numéricas, excluyendo inicialmente el componente textual. Este modelo conjunto alcanza un AUC de 0.88, lo que representa una mejora considerable respecto a cualquiera de los bloques por separado. Esta mejora pone de manifiesto el carácter complementario de los distintos tipos de información (emocional, vocal, lingüística o de personalidad) y refuerza la idea de que una aproximación multimodal es clave para ver mejor la complejidad lo que se estudia.

La siguiente tabla resume los resultados obtenidos en términos de AUC para cada grupo de variables, tanto de forma aislada como en combinación con las 1000 palabras más relevantes del texto procesado con TF-IDF:

Grupo de variables	AUC (numérico)	AUC (numérico + texto)
Emociones faciales	0.568	0.946
Personalidad	0.663	0.947
Estado emocional / psicológico	0.644	0.947
Características vocales	0.615	0.946
Emociones en la voz	0.592	0.946
Otras variables	0.615	0.944
Modelo global (sin texto)	0.880	-

Tabla 15 – Comparativa global de los resultados de los grupos de variables

Como se puede observar, los modelos basados exclusivamente en variables numéricas tienden a tener un rendimiento más limitado (con AUCs en torno a 0.6), mientras que al incorporar la dimensión semántica del texto, todos los grupos mejoran de forma notable su capacidad de clasificación, alcanzando valores superiores al 94% en la mayoría de los casos. Esto evidencia que el contenido verbal, extraído de las transcripciones, aporta información rica y complementaria que potencia el valor diagnóstico de los modelos.

Desde una perspectiva metodológica, el uso de regresión logística ha demostrado ser especialmente valioso no solo por su rendimiento, sino también por su interpretabilidad. Este modelo ha permitido identificar qué variables concretas tienen un mayor peso en la diferenciación entre perfiles neurodivergentes y no neurodivergentes, facilitando la comprensión científica del fenómeno y aportando transparencia en contextos aplicados. Esto lo convierte en una herramienta prometedora para el desarrollo de sistemas de ayuda al diagnóstico que sean comprensibles, auditables y alineados con principios éticos.

Capítulo 7. MODELOS PREDICTIVOS

Este capítulo está dedicado al estudio de diversos modelos de clasificación supervisada con el objetivo de predecir, de forma automática, el perfil neurodivergente de los individuos analizados. A diferencia del modelo explicativo de regresión logística desarrollado previamente, en este caso se busca maximizar el rendimiento predictivo mediante algoritmos de aprendizaje más complejos y no necesariamente interpretables.

El dataset utilizado contiene un total de 8000 registros, balanceado entre las cuatro clases objetivo: TDAH, dislexia, autismo y control (2000 observaciones por clase). Este equilibrio permite comparar el rendimiento de los distintos modelos sin introducir sesgos debidos al desbalance de clases. Los algoritmos evaluados incluyen un perceptrón multicapa (MLP), Random Forest y Support Vector Machines (SVM) en sus versiones binaria y multiclase.

7.1 PREPROCESAMIENTO DEL DATASET

El preprocesado del dataset constituye una fase fundamental para garantizar la calidad de los modelos. Por ello, antes de aplicar cualquier modelo, se ha realizado un proceso de limpieza, transformación y preparación del dataset original, con el objetivo de asegurar la calidad de los datos y su idoneidad para el entrenamiento supervisado.

7.1.1 CARGA Y LIMPIEZA INICIAL

Se parte de un fichero que cuenta con 82 columnas. De estas columnas, o variables, no solo hay columnas numéricas, sino que también hay variables categóricas, y metadatos adicionales sobre los vídeos originales.

	created_at	aid	extension	format	duration	FILE_STORED	FACIAL_ANALYSED	VOICE_ANALYSED	VOICE_TRANSC
0	1744824974	63612104-cc5c-4d54-b136-1f5880dece96	.mp4	video	NaN	True	True	False	
1	1744825004	1afe2c00-8488-40f2-b3e1-5bd90fd57ad8	.mp4	video	213.0	True	True	True	
2	1744825017	6c987224-499c-469a-b908-dffef38c48b4	.mp4	video	115.0	True	True	True	
3	1744825025	e34d8573-7371-486a-900c-32a6ee78fabe	.mp4	video	202.0	True	True	True	
4	1744825037	5b442dc9-a39a-47e2-ba05-b0fe880ee44b	.mp4	video	135.0	True	True	True	

5 rows × 82 columns

Figura 45 - DataFrame inicial con las 82 columnas

```
Número de columnas tras la limpieza: 68
```

	angry_facial	disgust_facial	fear_facial	happy_facial	sad_facial	surprise_facial	neutral_facial	most_frequent_dominant_emotion
0	0.2098	0.0008	0.4117	0.1561	0.2053	0.0063	0.0099	sad
1	0.0271	0.0000	0.1126	0.0005	0.6216	0.0001	0.2381	sad
2	0.1140	0.0000	0.1678	0.0062	0.4205	0.0020	0.2895	sad
3	0.0931	0.0771	0.1322	0.0408	0.3593	0.0007	0.2969	sad
4	0.0315	0.0001	0.0204	0.4411	0.2166	0.0009	0.2895	neutral

5 rows × 68 columns

Figura 46 - DataFrame después de la limpieza de columnas con metadatos

En primer lugar, se eliminan aquellas columnas que no aportan información relevante para el modelado o que constituyen identificadores directos, como *aid, created_at, status*, entre otras. Tras este filtrado, se conservan 68 variables predictoras potencialmente útiles.

7.1.2 ELIMINACIÓN DE LA CLASE MINORITARIA

Para los algoritmos usados para predecir, en un inicio se pretendía clasificar entre 5 clases, las 4 mencionadas más la dispraxia. Esta clase se ha tenido que eliminar posteriormente debido a la falta de muestras suficientes, ya que su presencia desbalanceada podía introducir ruido y dificultar el aprendizaje. Por ello, se ha optado por eliminar esta clase y centrarse en un problema de clasificación multiclase con las 4 restantes.

La distribución de las clases tras la eliminación queda de la siguiente manera:

Control	2033
Autismo	2113
TDAH	2030
Dislexia	2026

Tabla 16 – Distribución de clases para los modelos predictivos

Tras esta depuración, la distribución final del dataset queda aproximadamente equilibrada, con unas 2000 muestras por clase, y un total de 8182 observaciones.

7.1.3 TRATAMIENTO DE VALORES NULOS

Variables	Número de nulos
dominant_emotion_counts_surprise	7184
neutral_facial, disgust_facial, average_face_confidence, most_frequent_dominant_emotion, angry_facial, surprise_facial, sad_facial, happy_facial, fear_facial	988
voice_kurtosis, voice_median, voice_mode, voice_Q25, voice_Q75, voice_IQR, voice_skewness, voice_Q75_note, voice_mean_note, voice_median_note	351

Tabla 17 – Distribución de las variables según su número de valores nulos

Durante el análisis exploratorio se detectó una presencia de valores nulos en varias variables, especialmente entre aquellas extraídas del análisis facial y vocal (por ejemplo, *neutral_facial, voice_median_note*, etc.). Haciendo un recuento de los nulos por cada variable, se puede observar una en concreto, *dominant_emotion_counts_surprise*, que cuenta con un 87.8% de valores nulos (7184 de 8182 muestras totales), por lo que se procedió a su eliminación al no considerarse su aportación relevante. El resto de variables, al no tener más de un 13% de valores nulos no se consideró su eliminación del dataset.

```
i  Columna categórica 'most_frequent_dominant_emotion' rellenada con su moda
i  Columna categórica 'voice_mean_note' rellenada con su moda
i  Columna categórica 'voice_median_note' rellenada con su moda
i  Columna categórica 'voice_mode_note' rellenada con su moda
i  Columna categórica 'voice_Q25_note' rellenada con su moda
i  Columna categórica 'voice_Q75_note' rellenada con su moda
i  Columna categórica 'language' rellenada con su moda

✅ Nulos restantes tras limpieza: 0
```

Figura 47 - Comprobación de que la imputación ha sido exitosa y no quedan nulos

Para el resto de variables:

- En las numéricas, los valores nulos se imputaron mediante su mediana condicionada por clase. Se escogió frente a la media por considerarse más robusta frente a posibles valores extremos.
- En las categóricas, la imputación se realiza con la moda, también condicionada por clase. Es decir, para cada muestra con un valor categórico nulo, se rellena utilizando el valor más frecuente dentro de su misma clase objetivo. Esta estrategia permite preservar las diferencias entre clases objetivo y evitar sesgos derivados de una imputación global. Por ejemplo, si en la variable *most_frequent_dominant_emotion* hay valores nulos en algunas muestras etiquetadas como "Autismo", se imputan utilizando la moda de esa variable solo entre los ejemplos que también pertenecen a la clase "Autismo".

```
Tamaño del dataset después (para comprobar que no se eliminan filas): (8182, 67)
```

Figura 48 - Comprobación de las dimensiones finales del dataset

7.1.4 TRANSFORMACIÓN DE VARIABLES CATEGÓRICAS Y ESCALADO DE VARIABLES

```
Columnas categóricas a codificar: ['most_frequent_dominant_emotion', 'voice_mean_note', 'voice_median_note', 'vo
ice_mode_note', 'voice_Q25_note', 'voice_Q75_note', 'language']
Tamaño X_train: (6545, 179)
Tamaño X_test: (1637, 179)
Distribución de clases en y_train:
 variable
Autismo     1674
Control     1626
TDAH        1624
Dislexia    1621
Name: count, dtype: int64
```

Figura 49 - Comprobación final del dataset ya preprocesado antes de aplicar algoritmos

Se identifican un total de 7 columnas con valores categóricos, como *language, gender*, o *most_frequent_dominant_emotion*, entre otras. Estas variables se transforman mediante codificación one-hot [18], eliminando una de las categorías mediante `drop_first=True` para evitar multicolinealidad. Esto da lugar a un conjunto de variables binarias indicadoras.

Posteriormente, se aplica una normalización estándar (StandardScaler) [14] sobre todas las variables numéricas, incluyendo tanto las variables originales como las generadas a partir del one-hot encoding. Esta transformación permite centrar las variables en media cero y varianza unitaria, requisito especialmente relevante para algunos algoritmos como redes neuronales o SVM.

Finalmente, se divide el dataset en un conjunto de entrenamiento (80%) y uno de test (20%) asegurando que la proporción de clases se mantenga en ambas particiones. El conjunto de entrenamiento final consta de 6545 muestras y 179 columnas, ya completamente preprocesadas y listas para alimentar los modelos supervisados que se describen en los siguientes apartados.

7.2 RED NEURONAL TIPO PERCEPTRÓN MULTICAPA (MLP)

El primer algoritmo supervisado aplicado para la clasificación de perfiles neurodivergentes es una red neuronal tipo perceptrón multicapa (MLP, Multi Layer Perceptron). Este tipo de modelo es especialmente útil para capturar relaciones no lineales entre las variables predictoras, gracias a su estructura basada en capas y neuronas artificiales.

7.2.1 ARQUITECTURA Y CONFIGURACIÓN DEL MODELO

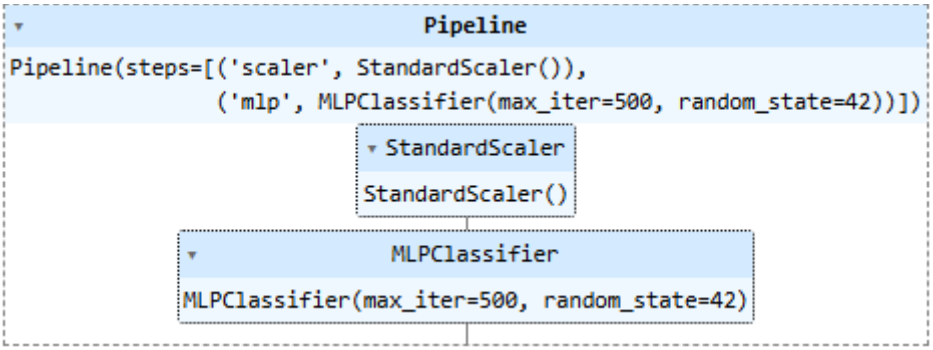


Figura 50 - Visualización del pipeline utilizado para entrenar el modelo

El modelo se entrena a través de un pipeline que incluye una normalización previa con StandardScaler, con el objetivo de escalar todas las variables a media cero y varianza unitaria antes de alimentar la red neuronal.

También, este modelo ha sido implementado con la clase MLPClassifier [19], empleando su configuración por defecto, especificando el número máximo de iteraciones (max_iter=500) y la fijación de la semilla aleatoria (random_state=42) para asegurar la reproducibilidad. Esta configuración implica que el modelo contiene una única capa oculta con 100 neuronas, utiliza la función de activación ReLU y el optimizador Adam (todo esto para hacer que los modelos aprendan de forma más eficiente, ajustando continuamente la tasa de aprendizaje de cada parámetro).

7.2.2 RESULTADOS DEL MODELO

	Precision	Recall	F1 – Score	Support
Autismo	0.86	0.76	0.81	419
Control	0.87	0.84	0.85	407
Dislexia	0.86	0.92	0.89	405
TDAH	0.81	0.87	0.84	406
Accuracy		0.85		1637
Macro avg	0.85	0.85	0.85	1637
Weighted avg	0.85	0.85	0.85	1637

Tabla 18 – Métricas del modelo MLP

Los resultados del modelo en el conjunto de test muestran una accuracy global del 85%, con un rendimiento equilibrado entre clases. En concreto, destaca la clase “Dislexia”, con el mayor valor de F1-score (0.89) y “Autismo" como la clase más desafiante, con un recall algo más bajo (0.76), indicando mayor dificultad del modelo para identificar correctamente perfiles autistas.

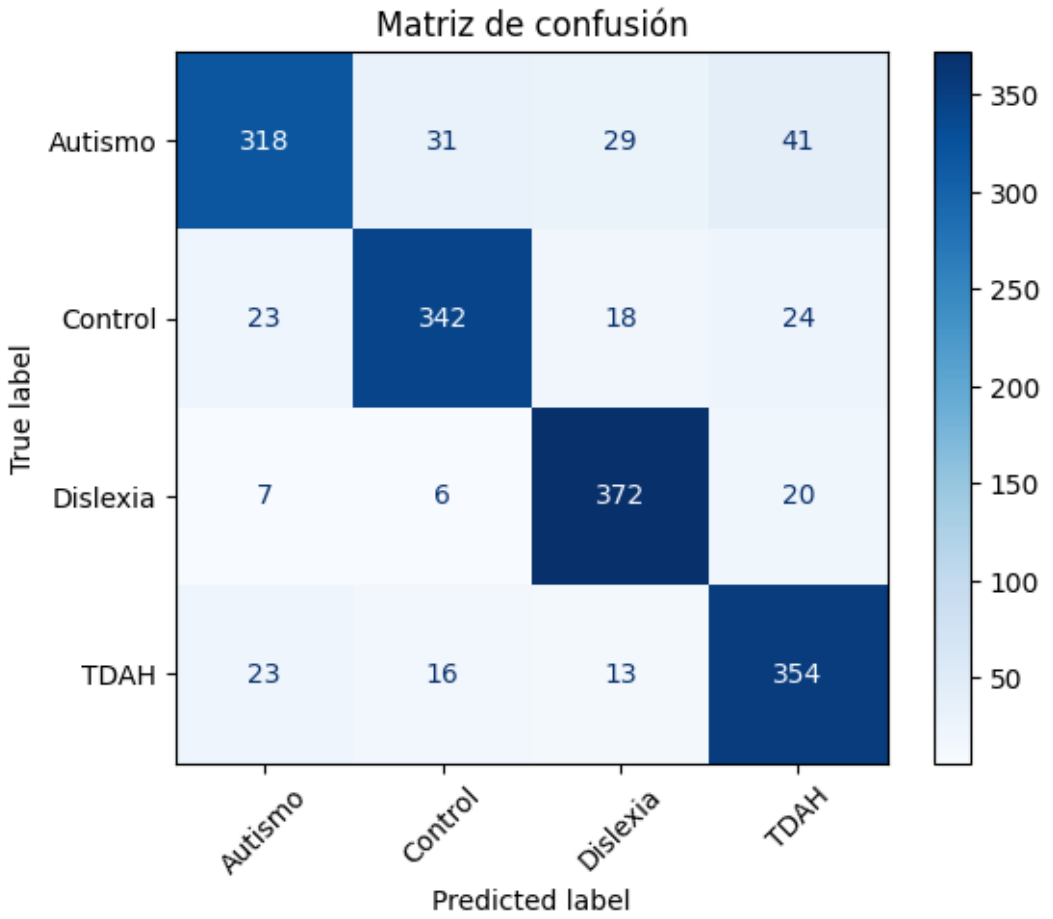


Figura 51 - Matriz de confusión por clase del modelo MLP

La matriz de confusión muestra patrones claros de clasificación correcta, reflejando la capacidad del modelo para captar patrones diferenciales entre perfiles. Si bien existe cierta superposición entre clases clínicamente similares, como autismo y TDAH, esta situación también resalta la sensibilidad del modelo para capturar matices entre categorías cercanas, mostrando un rendimiento sólido y con buen equilibro, a pesar de tratarse de un modelo con una configuración relativamente sencilla.

7.3 RANDOM FOREST

El segundo modelo aplicado ha sido el algoritmo Random Forest, una técnica de ensamblado basada en árboles de decisión que agrega los resultados de múltiples clasificadores para mejorar la precisión y reducir el riesgo de sobreajuste.

El modelo se ha entrenado utilizando la clase RandomForestClassifier [20], manteniendo los hiperparámetros por defecto para la primera evaluación. Esta configuración incluye un número de árboles (n_estimators=100), sin restricción explícita de profundidad máxima (max_depth=None), utilizando el criterio de Gini [21] para medir la calidad de las divisiones.

7.3.1 RESULTADOS DEL MODELO

	Precision	Recall	F1 - Score	Support
Autismo	0.90	0.77	0.83	419
Control	0.85	0.89	0.87	407
Dislexia	0.89	0.92	0.90	405
TDAH	0.84	0.91	0.87	406
Accuracy		0.87		1637
Macro avg	0.87	0.87	0.87	1637
Weighted avg	0.87	0.87	0.87	1637

Tabla 19 – Métricas del modelo Random Forest

Los resultados obtenidos en el conjunto de test reflejan una accuracy global del 87%, superando al modelo MLP. En particular, destaca la clase "Dislexia" con un F1-score de

0.90, lo que muestra su capacidad de detección. La clase "Autismo" sigue mostrando un rendimiento más moderado con un recall del 0.77, aunque mejora respecto al modelo anterior.

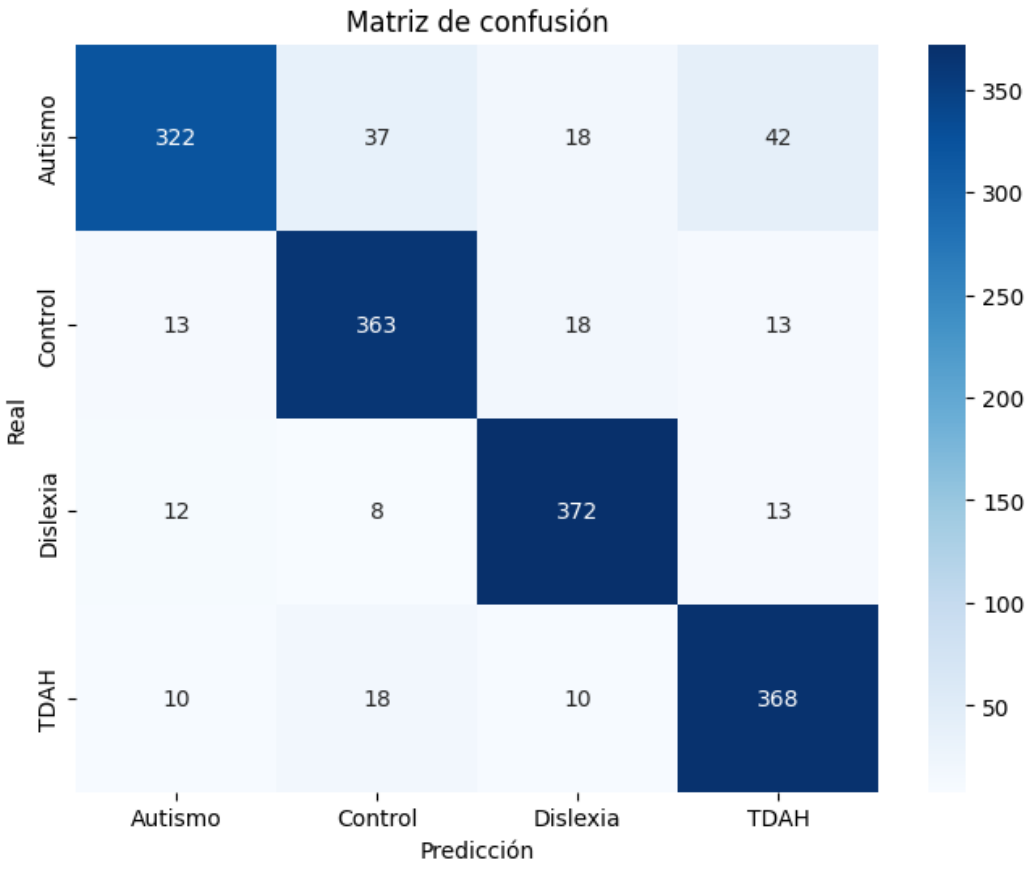


Figura 52 - Métricas y matriz de confusión del modelo Random Forest

La matriz de confusión muestra un mejor equilibrio en la clasificación entre clases, con reducciones visibles en los errores de clasificación cruzada entre perfiles similares. El modelo logra mejorar la identificación del perfil "Control" y mantiene buenos resultados en TDAH.

En resumen, Random Forest ha mostrado un rendimiento notable, con mejoras respecto al modelo MLP tanto en accuracy global como en métricas por clase. Además, su naturaleza

robusta ante datos ruidosos y su capacidad de interpretar la importancia relativa de las variables lo convierten en una opción aceptable para tareas de clasificación multiclase. Aunque no es tan fácilmente interpretable como un modelo lineal, ofrece un equilibrio entre rendimiento, estabilidad y facilidad de implementación que lo posiciona de buena manera dentro del conjunto de modelos evaluados.

7.4 SVM MULTICLASE – ONE-VS-REST

El tercer modelo implementado ha sido una Máquina de Vectores de Soporte (SVM) con estrategia multiclase basada en One-vs-Rest (OvR). Este enfoque permite construir un clasificador independiente para cada clase, enfrentándola al resto, lo que resulta especialmente útil en problemas de clasificación multiclase como el presente.

7.4.1 ARQUITECTURA Y CONFIGURACIÓN DEL MODELO

```
Fitting 3 folds for each of 9 candidates, totalling 27 fits
🏆 Mejor combinación: {'svc__C': 10, 'svc__gamma': 0.01, 'svc__kernel': 'rbf'}
🔍 Mejor accuracy en validación: 0.746
🎯 Accuracy en test con mejores hiperparámetros: 0.866
```

Figura 53 - Pantallazo con los resultados del ajuste de hiperparámetros con GridSearchCV

El modelo se ha entrenado con la clase SVC de scikit-learn [22], utilizando un kernel RBF (función de base radial) [23] como núcleo no lineal. Para mejorar su rendimiento, se ha aplicado un procedimiento de ajuste de hiperparámetros mediante GridSearchCV [24], probando distintas combinaciones de los parámetros C (complejidad) y gamma (coeficiente del kernel). La mejor combinación encontrada fue C=10 y gamma=0.01, lo que permitió maximizar la precisión sin caer en sobreajuste, como refleja la coherencia entre la validación cruzada (0.746) y el test final (0.866).

7.4.2 RESULTADOS DEL MODELO

	Precision	Recall	F1 - Score	Support
Autismo	0.90	0.78	0.84	419
Control	0.83	0.89	0.86	407
Dislexia	0.91	0.92	0.92	405
TDAH	0.83	0.88	0.85	406
Accuracy		0.87		1637
Macro avg	0.87	0.87	0.87	1637
Weighted avg	0.87	0.87	0.87	1637

Tabla 20 – Métrica del modelo SVM multiclase

Los resultados del modelo afinado muestran una accuracy global del 87%, situándose a la par con Random Forest y por encima del MLP. Se muestra un rendimiento equilibrado entre clases, con la "Dislexia" alcanzando nuevamente el mejor F1-score (0.92), y la clase "Autismo" mejorando significativamente hasta un F1 de 0.84.

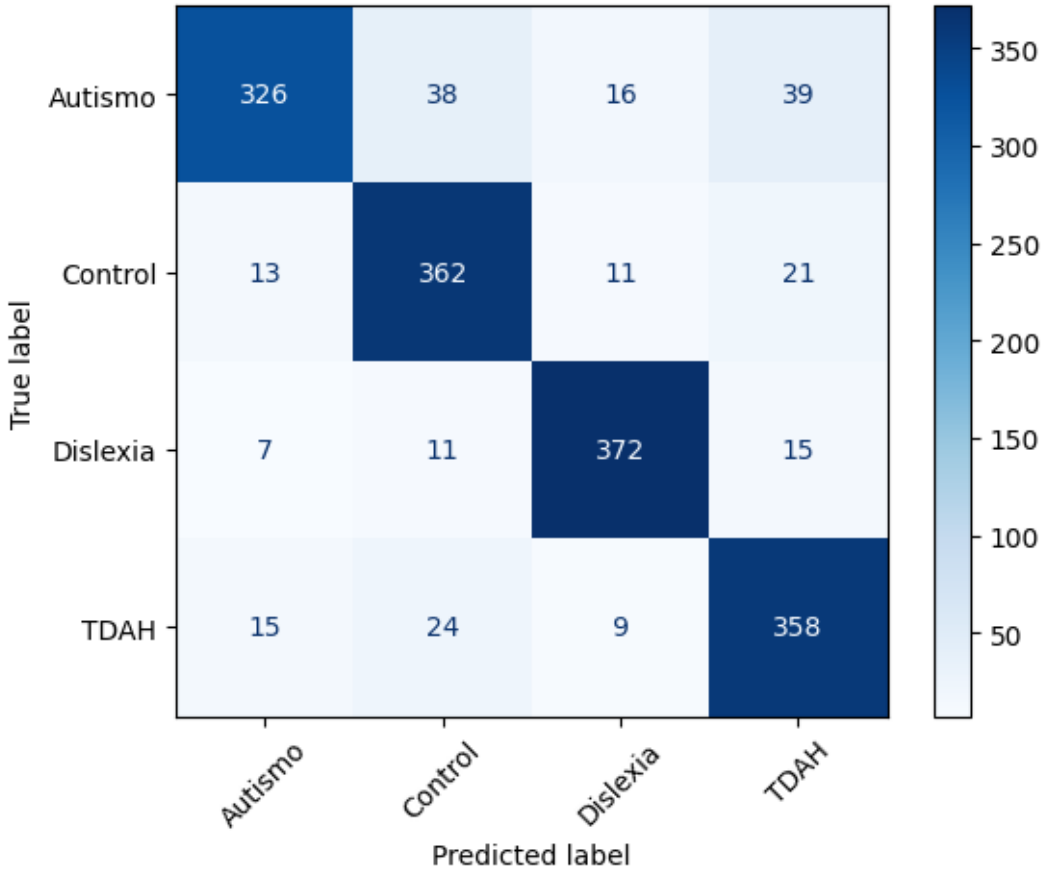


Figura 54 - Matriz de confusión del modelo SVM multiclase tras utilizar GridSearchCV

La matriz de confusión muestra un rendimiento notable y equilibrado, con pocas confusiones entre clases similares. Se observa, por ejemplo, una mejora en la distinción entre TDAH y Autismo, y una clasificación más precisa de la clase Control. Estos resultados reflejan la eficacia del ajuste de hiperparámetros mediante GridSearchCV, optimizando la generalización del modelo sin incurrir en sobreajuste.

El modelo SVM multiclase afinado destaca por su capacidad para clasificar perfiles con alta precisión, alcanzando un rendimiento competitivo respecto al resto de modelos evaluados, también gracias a su capacidad de ajuste mediante hiperparámetros.

7.5 SVM BINARIO

El último modelo implementado ha sido una Máquina SVM con enfoque binario, diseñado para distinguir entre perfiles neurodivergentes y no neurodivergentes, al igual que con la regresión logística. Este enfoque simplifica el problema de clasificación y permite centrarse en la capacidad del modelo para diferenciar entre estos dos grandes grupos.

7.5.1 ARQUITECTURA Y CONFIGURACIÓN DEL MODELO

El modelo, al igual que el multiclase, se ha implementado con la clase SVC de scikit-learn [22], utilizando un kernel RBF [23] para capturar relaciones no lineales entre las variables. El preprocesamiento previo al entrenamiento incluye la separación entre variables numéricas y categóricas, aplicando una normalización estándar (StandardScaler [14]) sobre las primeras y una codificación one-hot (OneHotEncoder [18]) sobre las segundas. Esta transformación combinada se lleva a cabo a través de un ColumnTransformer [25] dentro de un Pipeline [26], asegurando así una correcta integración de los distintos tipos de datos.

Para evitar sesgos en el modelo, se ha dividido el dataset en conjunto de entrenamiento y test con un random_state=42. El modelo se ha entrenado sin ajuste de hiperparámetros adicional, utilizando los valores por defecto salvo la elección del kernel RBF.

7.5.2 RESULTADOS DEL MODELO

	Precision	Recall	F1 - Score	Support
0 (control)	0.87	0.95	0.91	1231
1 (neurodiv.)	0.95	0.86	0.90	1234
Accuracy		0.91		2465
Macro avg	0.91	0.91	0.91	2465
Weighted avg	0.91	0.91	0.91	2465

Tabla 21 – Resultados del modelo SVM binario

Los resultados del modelo binario muestran un rendimiento muy sólido, con una accuracy del 91% sobre el conjunto de test. La clase neurodivergente alcanza un F1-Score de 0.90, mientras que la clase de control obtiene un 0.91, lo que sugiere un rendimiento equilibrado y eficaz en la detección de ambos perfiles.

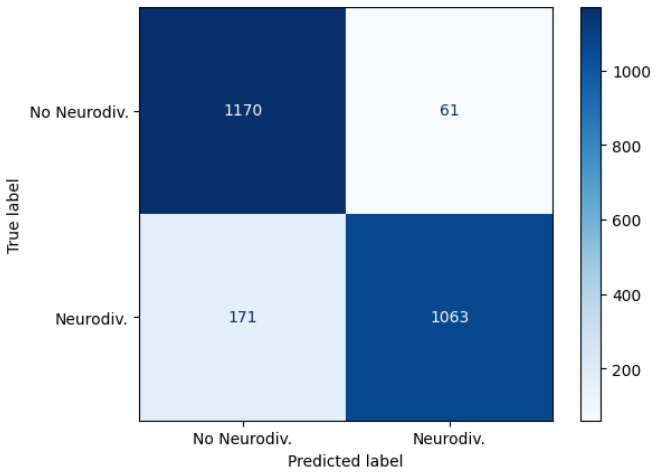


Figura 55 - Matriz de confusión del modelo SVM binario

La matriz de confusión muestra una clara separación entre ambas clases, con un número reducido de falsos positivos y negativos (entre un 5%-15% para ambas clases). Esta capacidad de distinguir entre perfiles neurodivergentes y no neurodivergentes sugiere que el modelo ha aprendido patrones generales consistentes y representativos, sin depender de ajustes complejos o estrategias de afinado adicionales. El modelo SVM binario, gracias a su simplicidad y precisión, se muestra como una herramienta altamente eficaz para la detección de neurodivergencia. Su rendimiento, junto con la facilidad de implementación, lo convierte en una opción especialmente útil en contextos donde se requiere una clasificación binaria fiable y eficiente.

7.6 COMPARATIVA GLOBAL DE LOS MODELOS

A continuación se presenta una comparativa general entre los modelos desarrollados, tomando como referencia las métricas de accuracy, macro-average y tiempo de entrenamiento aproximado:

Modelo	Tipo de clasificación	Accuracy (%)	Macro Avg	Tiempo de entrenamiento
MLP	Multiclase	87.0	0.86	Medio
Random Forest	Multiclase	87.2	0.87	Bajo
SVM	Binario	91.3	0.91	Alto
SVM	Multiclase	84.1	0.84	Muy alto

Tabla 22 – Comparativa global de los modelos de predicción utilizado

Como se observa en la tabla, el modelo SVM binario destaca por su capacidad discriminativa en la separación entre perfiles neurodivergentes y no neurodivergentes. Sin embargo, su elevado coste computacional lo hace menos viable para implementaciones a gran escala o

en tiempo real. Por otro lado, Random Forest combina un alto rendimiento con una velocidad de entrenamiento significativamente menor, lo que lo convierte en una opción equilibrada y adecuada para despliegues operativos.

El perceptrón multicapa ofrece resultados muy similares en cuanto a exactitud, con un tiempo de entrenamiento intermedio. Finalmente, el SVM multiclase, a pesar de su robustez teórica, muestra una eficiencia algo inferior en este contexto específico, tanto en precisión como en coste computacional.

Esta comparativa refuerza la idea de que, más allá de la métrica de rendimiento puro, la selección de modelos debe contemplar también su coste computacional, interpretabilidad y escalabilidad en entornos reales.

Capítulo 8. CONCLUSIONES Y TRABAJOS FUTUROS

8.1 RESUMEN DE LO REALIZADO

A lo largo de este Trabajo de Fin de Grado se ha desarrollado un sistema completo para el análisis automatizado de perfiles neurodivergentes a partir de vídeos obtenidos en redes sociales. El proyecto ha abarcado todas las fases de una solución basada en inteligencia artificial, desde la recopilación de datos hasta la validación de modelos predictivos y explicativos.

En primer lugar, se ha construido un dataset de más de 13.000 muestras de vídeo, representando cuatro perfiles distintos, como el TDAH, autismo, dislexia, y una clase de control con vídeos de gente con un perfil no neurodivergente. Este dataset se generó de forma progresiva mediante scrapers personalizados, con un enfoque que priorizaba la diversidad y naturalidad del contenido.

Posteriormente, se procesaron todos los vídeos utilizando una API avanzada desarrollada por Souly. A través de este sistema se extrajeron más de un centenar de métricas por vídeo, cubriendo aspectos acústicos, emocionales, lingüísticos y de personalidad. Esta riqueza de variables permitió realizar un análisis profundo tanto desde una perspectiva explicativa como predictiva.

En la parte explicativa, se desarrolló un modelo de regresión logística que permitió identificar qué grupos de variables resultaban más determinantes para clasificar perfiles neurodivergentes. El modelo obtuvo un AUC de 0.913, demostrando su solidez incluso en tareas de clasificación binaria (control vs neurodivergente). Además, se realizaron visualizaciones mediante t-SNE que evidenciaron diferencias lingüísticas consistentes entre clases.

En la vertiente predictiva, se implementaron distintos algoritmos supervisados, incluyendo Random Forest, perceptrones multicapa (MLP) y máquinas de vectores soporte (SVM), alcanzando accuracies superiores al 85% en todos los casos, con valores de F1 equilibrados por clase. Estos resultados validan la viabilidad de este enfoque para tareas de clasificación multiclase en contextos reales.

8.2 APORTACIONES DEL TRABAJO

Este proyecto tiene capacidad para aportar en el ámbito del análisis automático de perfiles neurodivergentes mediante técnicas de machine learning aplicadas a datos percibidos, espontáneos y en formato vídeo. Su novedad no reside únicamente en la combinación de voz, texto y señales faciales, sino también en la escala del dataset, el enfoque multimodal y la capacidad de explicar los resultados con modelos interpretables.

En comparación con estudios previos, como el de *Hu et al. (2024)* [27], donde se emplearon características acústicas extraídas de diálogos clínicos estructurados para detectar autismo con un accuracy del 87,8%, este trabajo ofrece un enfoque más amplio y generalizable. En lugar de limitarse a contextos clínicos, se ha trabajado con contenido informal de redes sociales, lo que implica una mayor variabilidad, pero también refleja mejor las condiciones reales de uso. Además, se han incorporado métricas lingüísticas, emocionales y de personalidad, que no están presentes en la mayoría de trabajos previos.

Por otro lado, el estudio de *Ramesh y Assaf (2021)* [28] se centra únicamente en el análisis de transcripciones de voz, alcanzando un rendimiento máximo cercano al 75% en clasificación binaria. En este proyecto, el análisis textual, basado en vectores TF-IDF sobre más de 13.000 muestras, alcanza una AUC de 0.87 sin necesidad de usar señales acústicas. Este resultado muestra que el lenguaje por sí solo ya contiene patrones suficientemente informativos, y que su combinación con variables numéricas puede mejorar sustancialmente la capacidad predictiva, como se observa en el modelo combinado (AUC = 0.913).

En resumen, este trabajo se aumenta la robustez del análisis al incorporar fuentes de datos variadas y no clínicas, se mejora la explicabilidad mediante modelos como la regresión logística, frente a "cajas negras" habituales, y se introduce una estructura escalable que puede integrarse en plataformas reales como Souly.

8.3 LIMITACIONES DEL ESTUDIO

A pesar de los resultados obtenidos y del valor práctico del sistema desarrollado, es importante señalar algunas limitaciones que pueden haber condicionado el alcance o la generalización de los modelos entrenados.

1. **Naturaleza no controlada de los datos:** El dataset está compuesto por vídeos extraídos de plataformas sociales como YouTube, TikTok e Instagram. Aunque este enfoque aporta diversidad y cercanía a contextos reales, también introduce una importante variabilidad, al tratarse de datos subjetivos y no controlados, es decir, grabaciones "*on-the-wild*". Esta heterogeneidad afecta a factores como la calidad del audio, el entorno de grabación, el idioma o el tipo de discurso. La falta de estandarización puede poner en duda la consistencia de las métricas extraídas y, por lo tanto, influir en el rendimiento de los modelos.
2. **Ausencia de validación clínica:** La categorización de los perfiles neurodivergentes se ha realizado en base al contenido de los vídeos y su etiquetado semántico (títulos, hashtags, contexto). No se dispone de diagnósticos clínicos validados, por lo que no es posible garantizar que todos los vídeos etiquetados como, por ejemplo, "autismo" correspondan a un diagnóstico profesional confirmado. Esto introduce un posible sesgo en la veracidad de las clases objetivo.
3. **Simplificación de la clasificación:** Aunque se han incluido tres perfiles neurodivergentes (TDAH, autismo y dislexia), en el análisis binario estos se agrupan bajo una sola etiqueta. Este agrupamiento, si bien útil desde un punto de vista explicativo, puede ocultar diferencias significativas entre los perfiles. Del mismo modo, no se han contemplado otros perfiles como dispraxia, debido a la escasa disponibilidad de datos.

4. **Modelos estáticos y sin validación externa:** Todo el desarrollo se ha llevado a cabo sobre un único conjunto de datos. No se ha validado el modelo en nuevos contextos (por ejemplo, con otros idiomas o acentos), ni se ha integrado en ningún entorno real para evaluar su rendimiento en condiciones operativas.

8.4 PROPUESTA DE MEJORA Y TRABAJOS FUTUROS

Este proyecto ha demostrado el potencial del análisis automatizado de voz, texto y expresión facial como herramienta de apoyo en la detección de perfiles neurodivergentes. No obstante, existe un amplio margen de mejora y diversas líneas de trabajo que podrían reforzar la utilidad clínica, técnica y social del sistema.

1. **Validación clínica estructurada:** Una de las líneas prioritarias de mejora podría ser la réplica de este análisis con muestras diagnósticas verificadas, recogidas en entornos clínicos o educativos. Esto permitiría validar los modelos frente a etiquetas confirmadas por especialistas, eliminando el sesgo implícito en el uso de datos subjetivos extraídos de redes sociales. También facilitaría adaptar los algoritmos a contextos normativos, como la atención primaria o la orientación psicopedagógica.
2. **Ampliación del espectro neurodivergente:** Como se ha comentado anteriormente, el sistema actual se centra solo en tres perfiles. Sería relevante ampliar el abanico de condiciones consideradas, incluyendo dispraxia, ansiedad social, alexitimia o incluso perfiles combinados. Esto permitiría construir modelos de mayor granularidad y avanzar hacia una clasificación más rica y representativa de la diversidad cognitiva.
3. **Desarrollo de modelos multimodales avanzados:** Los modelos utilizados en este trabajo se basan en pipelines convencionales (Random Forest, MLP, SVM). Futuras investigaciones podrían explorar arquitecturas más complejas y dinámicas, como redes neuronales recurrentes (RNN), transformers para audio-texto o modelos de atención multimodal. Este enfoque permitiría explotar mejor las relaciones temporales y contextuales entre voz, texto y emociones.
4. **Aplicación práctica en plataformas digitales:** El sistema tiene un alto potencial de transferencia tecnológica, especialmente en su integración en plataformas como

Souly. Como línea de trabajo, se podría perfeccionar una interfaz que permita a los usuarios subir sus vídeos, obtener un análisis automático de sus patrones de voz y lenguaje, y recibir una devolución estructurada y visual, respetando criterios éticos y de privacidad, tal y como está avanzando Souly.

5. **Utilidad educativa y de concienciación:** Finalmente, el modelo podría adaptarse como herramienta didáctica para formar a estudiantes de medicina, psicología o educación en el reconocimiento de rasgos neurodivergentes. Al trabajar con ejemplos reales y métricas objetivas, el sistema puede contribuir a una comprensión más empática y científica de la diversidad cognitiva, visibilizando perfiles que suelen pasar desapercibidos.

Capítulo 9. BIBLIOGRAFÍA

[1] Malgaroli, M., Hull, T. D., Zech, J. M., & Althoff, T. (2023). Natural language processing for mental health interventions: A systematic review and research framework. Translational Psychiatry, 13(1), 1–14. https://doi.org/10.1038/s41398-023-02592-2

[2] SEMERGEN. (2022). TDAH en el adulto: diagnóstico y tratamiento en Atención Primaria. Sociedad Española de Médicos de Atención Primaria. https://semergen.es/files/docs/grupos/salud%20mental/adultoTDAH.pdf

[3] eldiario.es. (2021). TDAH: el trastorno que no siempre se diagnostica en la infancia. https://www.eldiario.es/sociedad/tdah-trastorno-diagnostica-infancia_1_8072729.html

[4] eldiario.es. (2023). Los diagnósticos de autismo se han cuadruplicado en diez años y las familias no quieren esperar más. https://www.eldiario.es/sociedad/diagnosticos-autismo-han-cuadruplicado-diez-anos-familias-no-esperan_1_12112735.html

[5] Souly. (2024). Souly API Documentation. https://www.mysouly.com/en/

[6] Ghosh, S., Choudhury, T., & Kalantidis, Y. (2022). Multimodal detection of autism spectrum disorder using audio and text features. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2022, 3697–3706. https://ieeexplore.ieee.org/document/9761583

[7] Winterlight Labs. (2023). Winterlight: Technology for the detection of cognitive impairment. https://www.winterlightlabs.com/

[8] Ellipsis Health. (2023). Voice-based mental health analytics. https://www.ellipsishealth.com/

[9] Kintsugi. (2023). Mental Health Voice Biomarkers. https://www.kintsugihealth.com/

[10] Canary Speech. (2023). Speech analysis for cognitive and mental health. https://www.canaryspeech.com/

[11] Vocalis Health. (2021). Vocal biomarkers for remote respiratory health monitoring. https://www.vocalishealth.com/

[12] Apify. (2024). Apify: Web scraping and automation platform. https://apify.com/

[13] Brunet, P. [@PauPautista]. (n.d.). Vídeos subidos por Pau Brunet. YouTube. https://www.youtube.com/@PauPautista/videos

[14] Scikit-learn. (2024). sklearn.preprocessing.StandardScaler. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

[15] Scikit-learn. (2024). sklearn.feature_extraction.text.TfidfVectorizer. https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[16] Scikit-learn. (2024). sklearn.manifold.TSNE. https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

[17] Wikipedia. (2024). Modelo de los cinco grandes. En Wikipedia, La enciclopedia libre. https://es.wikipedia.org/wiki/Modelo_de_los_cinco_grandes

[18] Scikit-learn. (2024). sklearn.preprocessing.OneHotEncoder. https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html

[19] Scikit-learn. (2024). sklearn.neural_network.MLPClassifier. https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

[20] Scikit-learn. (2024). sklearn.ensemble.RandomForestClassifier. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[21] Wikipedia. (2024). *Coeficiente de Gini*. En Wikipedia, La enciclopedia libre. https://es.wikipedia.org/wiki/Coeficiente_de_Gini

[22] Scikit-learn. (2024). sklearn.svm.SVC. https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

[23] Scikit-learn. (2024). RBF SVM parameters. https://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html

[24] Scikit-learn. (2024). sklearn.model_selection.GridSearchCV. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[25] Scikit-learn. (2024). sklearn.compose.ColumnTransformer. https://scikit-learn.org/stable/modules/generated/sklearn.compose.ColumnTransformer.html

[26] Scikit-learn. (2024). sklearn.pipeline. https://scikit-learn.org/stable/api/sklearn.pipeline.html

[27] Hu, C., Zhou, Z., Tang, Y., Liu, D., Zhang, H., & Du, Y. (2024). Exploring Speech Pattern Disorders in Autism using Machine Learning. IEEE Access, 12, 45678–45689. https://www.researchgate.net/publication/381351509_A_Comparison_of_Deep_Learning_and_Supervised_Machine_Learning_Methods_in_RF_tracking_for_Neurodivergent_Learners

[28] Ramesh, A., & Assaf, T. (2021). Detecting Autism Spectrum Disorders with Machine Learning Models Using Speech Transcripts. Procedia Computer Science, 194, 85–92. https://doi.org/10.1016/j.procs.2021.10.010

[29] Buitrago, F. J. (2020). Interpretable machine learning with Python: Learn to build interpretable high-performance models with hands-on real-world examples. Apress. https://doi.org/10.1007/978-1-4842-5373-1

[30] Sanz Tur, E. (2025). TFG_EnriqueSanz [Repositorio GitHub]. GitHub. https://github.com/Enriquesanzzz/TFG_EnriqueSanz

ANEXO I: ALINEACIÓN DEL PROYECTO CON LOS ODS

Este Trabajo de Fin de Grado se alinea con varios de los Objetivos de Desarrollo Sostenible (ODS) propuestos por las Naciones Unidas, al fomentar el uso responsable de tecnologías emergentes para mejorar la salud mental, reducir desigualdades y promover la innovación. A continuación, se detalla la correspondencia directa con los ODS más relevantes:

ODS3: Salud y Bienestar

El objetivo principal del proyecto es contribuir a una detección más accesible, objetiva y temprana de perfiles neurodivergentes mediante el análisis automatizado de voz, lenguaje y expresiones. Esto permite mejorar la eficiencia del diagnóstico y el seguimiento de condiciones como el TDAH, la dislexia o el autismo. Al reducir los tiempos de espera y ofrecer una herramienta de apoyo al profesional clínico, el proyecto se alinea directamente con la meta 3.4 de este ODS, que establece la necesidad de mejorar la salud mental y el bienestar.

ODS8: Trabajo Decente y Crecimiento Económico

La tecnología propuesta también tiene aplicaciones potenciales en el ámbito laboral, donde podría implementarse como herramienta de screening (cribado) emocional o detección de sobrecarga cognitiva. Esto permitiría prevenir el burnout, mejorar el clima laboral y fomentar entornos más inclusivos y saludables, impactando positivamente en la productividad de las organizaciones.

ODS 9: Industria, Innovación e Infraestructura

El uso de inteligencia artificial y machine learning sobre datos de voz percibidos representa una apuesta clara por la innovación tecnológica en sectores tradicionalmente humanistas como el de la salud mental. El diseño y despliegue de APIs escalables, así como la integración de estas en plataformas accesibles como Souly, muestran una infraestructura tecnológica moderna y sostenible.

ODS 10: Reducción de las Desigualdades

La herramienta tiene un enfoque inclusivo al basarse en vídeos espontáneos y accesibles ("*on-the-wild*") de redes sociales, lo que permite analizar perfiles diversos sin necesidad de recurrir a clínicas costosas o pruebas presenciales. Esto promueve la equidad en el acceso a recursos diagnósticos y podría ser especialmente útil en zonas con menor cobertura sanitaria o educativa.

ANEXO II

CÓDIGO FUENTE, SCRIPTS Y ARCHIVOS UTILIZADOS A LO LARGO DEL TRABAJO

Toda la carpeta usada para la realización de este trabajo, con todos los scripts utilizados, la estructura de carpetas, archivos finales y configuración se puede encontrar en el siguiente repositorio de GitHub: https://github.com/Enriquesanzzz/TFG_EnriqueSanz

EJEMPLO DE UN JSON OBTENIDO PARA CONSTRUIR EL DATASET:


```
{
    "status": "success",
    "response": {
        "created_at": 1747823943,
        "aid": "631470e3-9984-4f06-b21f-1083b399f9e5",
        "result_url": "/v1/result/631470e3-9984-4f06-b21f-1083b399f9e5",
        "original_file": {
            "extension": ".mp4",
            "format": "video",
            "duration": 51
        },
        "status": {
            "FILE_STORED": true,
            "FACIAL_ANALYSED": true,
            "VOICE_ANALYSED": true,
            "VOICE_TRANSCRIBED": true,
            "BIOMETRICS_EXTRACTED": true,
            "SPEECH_ANALYSED": true,
            "PERSONALITY_ANALYSED": false,
            "FACES_EXTRACTED": true
        },
        "external_vars": {
            "id": "1"
        },
        "data": {
            "voice": {
                "frequencies": {
                    "mean": 3698,
                    "sd": 4416,
                    "median": 1184,
                    "mode": 70,
                    "Q25": 352,
```

```
                "Q75": 6875,
                "IQR": 6523,
                "skewness": 23.06999969482422,
                "kurtosis": 1566.8499755859375,
                "mean_note": "A\u266f",
                "median_note": "D",
                "mode_note": "C\u266f",
                "Q25_note": "F",
                "Q75_note": "A",
                "rmse": 0.23549999296665192
            },
            "pitch": 381,
            "tone": 1958,
            "emotions": {
                "neutral": 0.7147389650344849,
                "disgust": 0.10668990015983582,
                "surprised": 0.07564735412597656,
                "happy": 0.044156309217214584,
                "angry": 0.023150773718953133,
                "calm": 0.013163724914193153,
                "sad": 0.012112542986869812,
                "fearful": 0.01034053135663271
            }
        },
        "traits": {
            "survival": 0.26969999074935913,
            "creativity": 0.40049999952316284,
            "self_esteem": 0.1670999974012375,
            "compassion": 0.1615999937057495,
            "communication": 0.22609999775886536,
            "imagination": 0.2574000060558319,
            "awareness": 0.23510000109672546,
            "stress": {
                "high": 0.7778382897377014,
                "medium": 0.27287501096725464,
                "low": 0.06373893469572067
            },
            "helplessness": {
                "medium": 0.6754592657089233,
                "high": 0.37763720750808716,
                "low": 0.05240708589553833
            },
            "self_efficacy": {
                "low": 0.7212090492248535,
                "medium": 0.2028292715549469,
                "high": 0.11354091018438339
            },
            "depression": {
                "medium": 0.6546377539634705,
                "high": 0.34123390913009644,
                "low": 0.02922504022717476
            }
        },
```

```
            "speech": {
                "language": "en",
                "no_speech_prob": 0.07699567079544067,
                "text": "Do you know Taylor Swift's favorite number? Taylor
Swift's favorite number? I think a lot of people know it. Do you know it or no? I
said no. I don't know which is favorite number. It's 13. 13. Okay. What's Super
Bowl is this? I don't know. 58. What's 5 plus 8? 5 plus 8. Uh huh. 8, 9, 10, 11,
12. 13. Exactly. When is the Super Bowl? I don't know. Okay. It's on 2-11. What's
plus 2 plus 11? 13. Exactly. Exactly. Who are they playing in the Super Bowl? I
don't know. I don't watch NFL football. Okay. The 49ers. What's 4 plus 9? 4 plus
9. Uh huh. 13. Exactly. And what's 100 minus 13? 100 minus 13. Uh huh. 87. A
coincidence? I think not.",
                "entropy": 4.7758,
                "tense": {
                    "past": 0.0909,
                    "present": 0.9091,
                    "future": 0.0
                },
                "sentiment": {
                    "polarity": 0.325,
                    "subjectivity": 0.5321
                },
                "emotions": [
                    {
                        "label": "curiosity",
                        "score": 0.4773
                    },
                    {
                        "label": "neutral",
                        "score": 0.4559
                    },
                    {
                        "label": "confusion",
                        "score": 0.4438
                    }
                ],
                "entities": [
                    {
                        "entity": "Taylor Swift",
                        "label": "person"
                    },
                    {
                        "entity": "Taylor Swift",
                        "label": "person"
                    },
                    {
                        "entity": "Super Bowl",
                        "label": "event"
                    },
                    {
                        "entity": "Super Bowl",
                        "label": "event"
                    },
```

```
                    {
                        "entity": "Super Bowl",
                        "label": "event"
                    }
                ],
                "topics": [
                    {
                        "label": "music",
                        "score": 0.5008
                    },
                    {
                        "label": "diaries_&_daily_life",
                        "score": 0.358
                    },
                    {
                        "label": "sports",
                        "score": 0.3272
                    }
                ]
            },
            "translation": " Do you know Taylor Swift's favorite number? Taylor
Swift's favorite number? I think a lot of people know it. Do you know it or no? I
said no. I don't know which is favorite number. It's 13. 13. Okay. What's Super
Bowl is this? I don't know. 58. What's 5 plus 8? 5 plus 8. Uh huh. 8, 9, 10, 11,
12. 13. Exactly. When is the Super Bowl? I don't know. Okay. It's on 2-11. What's
plus 2 plus 11? 13. Exactly. Exactly. Who are they playing in the Super Bowl? I
don't know. I don't watch NFL football. Okay. The 49ers. What's 4 plus 9? 4 plus
9. Uh huh. 13. Exactly. And what's 100 minus 13? 100 minus 13. Uh huh. 87. A
coincidence? I think not.",
            "facial": {
                "average_emotions": {
                    "angry": 0.0002,
                    "disgust": 0.0,
                    "fear": 0.0787,
                    "happy": 0.1764,
                    "sad": 0.0196,
                    "surprise": 0.0,
                    "neutral": 0.7251
                },
                "most_frequent_dominant_emotion": "neutral",
                "dominant_emotion_counts": {
                    "happy": 1,
                    "neutral": 3
                },
                "average_face_confidence": 0.945
            }
        },
        "latency": {
            "time": 24,
            "unit": "seconds"
        }
    }
}
```

MODELO LOGIT ENTRENADO CON LAS 50 VARIABLES MÁS SIGNIFICATIVAS:

Logit Regression Results

Dep. Variable:	target	No. Observations:	12324
Model:	Logit	Df Residuals:	12273
Method:	MLE	Df Model:	50
Date:	Thu, 12 Jun 2025	Pseudo R-squ.:	0.4199
Time:	16:26:46	Log-Likelihood:	-4955.5
converged:	False	LL-Null:	-8542.3
Covariance Type:	nonrobust	LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.7618	0.052	14.718	0.000	0.660	0.863
autism	50.2304	10.378	4.840	0.000	29.891	70.570
dyslexia	126.3816	1204.336	0.105	0.916	-2234.073	2486.836
autistic	168.3838	933.598	0.180	0.857	-1661.435	1998.203
read	3.4127	0.454	7.522	0.000	2.523	4.302
dyslexic	301.1763	1.37e+05	0.002	0.998	-2.68e+05	2.69e+05
things	2.1290	0.246	8.649	0.000	1.647	2.611
brain	2.7790	0.354	7.859	0.000	2.086	3.472
attention	2.5972	0.350	7.427	0.000	1.912	3.283
child	1.1628	0.317	3.666	0.000	0.541	1.784
diagnosis	3.0738	0.729	4.215	0.000	1.645	4.503
reading	4.6873	0.989	4.738	0.000	2.748	6.626
school	0.6626	0.300	2.211	0.027	0.075	1.250
learn	2.3785	0.464	5.128	0.000	1.469	3.288
help	1.4786	0.302	4.901	0.000	0.887	2.070
difficult	2.1421	0.420	5.100	0.000	1.319	2.965
feel	1.0115	0.219	4.628	0.000	0.583	1.440
different	0.7485	0.313	2.388	0.017	0.134	1.363
remember	0.9502	0.305	3.119	0.002	0.353	1.547
words	0.0462	0.361	0.128	0.898	-0.662	0.754
10	-2.2727	0.308	-7.390	0.000	-2.876	-1.670
able	0.8114	0.315	2.576	0.010	0.194	1.429
money	-2.2112	0.283	-7.824	0.000	-2.765	-1.657
man	-2.0387	0.236	-8.630	0.000	-2.502	-1.576
thought	1.0796	0.310	3.481	0.000	0.472	1.687
something	1.1501	0.229	5.020	0.000	0.701	1.599
called	-2.8099	0.361	-7.788	0.000	-3.517	-2.103
re	-1.2917	0.114	-11.377	0.000	-1.514	-1.069
make	-1.4433	0.172	-8.381	0.000	-1.781	-1.106
said	-2.0544	0.183	-11.217	0.000	-2.413	-1.695
children	0.2976	0.254	1.173	0.241	-0.200	0.795
is	-1.5550	0.088	-17.705	0.000	-1.727	-1.383
hard	0.2138	0.345	0.620	0.535	-0.462	0.889
support	0.6827	0.484	1.410	0.159	-0.266	1.632
do	0.2372	0.122	1.942	0.052	-0.002	0.477
guy	-2.9315	0.343	-8.558	0.000	-3.603	-2.260
there	-1.5837	0.159	-9.931	0.000	-1.896	-1.271
seen	-2.8220	0.444	-6.359	0.000	-3.692	-1.952
say	-0.9622	0.163	-5.886	0.000	-1.283	-0.642
with	-0.2521	0.127	-1.989	0.047	-0.501	-0.004
need	0.5317	0.209	2.546	0.011	0.122	0.941
came	-2.7062	0.381	-7.101	0.000	-3.453	-1.959
very	-0.0215	0.174	-0.123	0.902	-0.363	0.320
they	-1.7103	0.124	-13.797	0.000	-1.953	-1.467
think	-1.7632	0.184	-9.569	0.000	-2.124	-1.402
days	-1.3265	0.280	-4.744	0.000	-1.874	-0.778
sometimes	1.0422	0.368	2.833	0.005	0.321	1.763
some	-1.2041	0.181	-6.643	0.000	-1.559	-0.849
head	0.4347	0.275	1.578	0.115	-0.105	0.975
always	0.5696	0.209	2.726	0.006	0.160	0.979
would	-1.9938	0.181	-11.028	0.000	-2.348	-1.639