COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

GRADO EN INGENIERÍA EN TECNOLOGÍAS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

PHENOMENOLOGICAL MULTIVARIATE ANALYSIS OF MENTAL HEALTH ISSUES USING PREDICTIVE MODELLING & ARTIFICIAL INTELLIGENCE ON VOICE DATA

Autor: Carlos Sánchez-Cabezudo Pinto

Director: David Martín-Corral Calvo

Madrid

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

Phenomenological Multivariate Analysis of Mental Health Issues using Predictive Modelling & 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.: Carlos Sánchez-Cabezudo Pinto Fecha: 01/07/2025

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

MARTIN CORRAL CALVO DAVID - 03909768K

Digitally signed by MARTIN CORRAL CALVO DAVID - 03909768K
Date: 2025.07.02 09:35:00 +02'00'

Fdo.: David Martín Corral Calvo Fecha: 01/07/2025

GRADO EN INGENIERÍA EN TECNOLOGÍAS DE TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

ANÁLISIS FENOMENOLÓGICO MULTIVARIANTE DE SALUD MENTAL MEDIANTE MODELOS PREDICTIVOS E INTELIGENCIA ARTIFICIAL APLICADOS A DATOS DE VOZ.

Autor: Carlos Sánchez-Cabezudo Pinto

Director: David Martín-Corral Calvo

Madrid

ANÁLISIS FENOMENOLÓGICO MULTIVARIANTE DE SALUD MENTAL MEDIANTE MODELOS PREDICTIVOS E INTELIGENCIA ARTIFICIAL APLICADOS A DATOS DE VOZ.

Autor: Sánchez-Cabezudo Pinto, Carlos.
Director: Martín Corral, David.
Entidad Colaboradora: Souly

RESUMEN DEL PROYECTO

Este trabajo desarrolla un conjunto complejo de modelos de aprendizaje automático orientados a la clasificación entre ansiedad, depresión y sujetos de control (sin patología), a partir de una base de datos de vídeos construida meticulosamente mediante un pipeline de preprocesamiento automatizado.

Palabras clave: Machine Learning; Salud mental; Pipeline de datos; Algoritmos; Souly

1. Introducción

Los trastornos de salud mental se han consolidado como una de las principales causas de discapacidad en el mundo contemporáneo, generando pérdidas económicas globales que superan el billón de dólares. En economías avanzadas, más del 6% de los adultos reciben prestaciones por incapacidad, y más de la mitad de estos casos están vinculados a condiciones psicológicas *[1]*. Sin embargo, pese a esta elevada incidencia, se estima que el 60% de los individuos con trastornos mentales no recibe ningún tipo de tratamiento.

En respuesta a este vacío, el presente trabajo propone un enfoque alternativo: transformar grabaciones espontáneas de vídeo en un instrumento de cribado mediante el uso de aprendizaje automático y análisis multimodal. Aprovechando la riqueza expresiva del rostro, la voz y el lenguaje verbal, se desarrolla una arquitectura técnica que permite inferir el estado mental a partir de testimonios grabados, acercando así la detección de ansiedad y depresión al contexto digital cotidiano, democratizando su acceso.

2. Definición del proyecto

El proyecto se articula en torno a la creación de un sistema capaz de clasificar de manera automática a los usuarios en tres categorías: ansiedad, depresión o control (sin patología). Para ello, se establecen cuatro objetivos: (1) diseñar un pipeline de preprocesamiento automatizado, desde la recopilación de datos hasta la extracción de características multimodales; (2) entrenar diversos modelos de clasificación basados en algoritmos de última generación; (3) evaluar el rendimiento de dichos modelos mediante una variedad de métricas cuantitativas y (4) formular un plan de adopción que contemple la transición del prototipo a un producto comercial viable, incluyendo un análisis de mercado.

3. Descripción del pipeline de preprocesamiento y modelos desarrollados

El sistema de preprocesamiento, mostrado en la siguiente figura, constituye una cadena automatizada que transforma contenido audiovisual no estructurado en un conjunto de datos apto para el modelado estadístico. Inicialmente, se emplean scrapers para

recolectar vídeos desde plataformas como YouTube y TikTok, seleccionando testimonios reales relacionados con salud mental mediante filtros temáticos. Estos vídeos son descargados en bloque, subidos a la nube mediante URLs prefirmadas de Amazon S3, y analizados por una API externa de alta complejidad (Souly). Los datos generados incluyen características acústicas, análisis semánticos y rasgos inferidos como autoestima o neuroticismo. Tras esto, los resultados se integran en un CSV, previa transformación de los JSON originales:

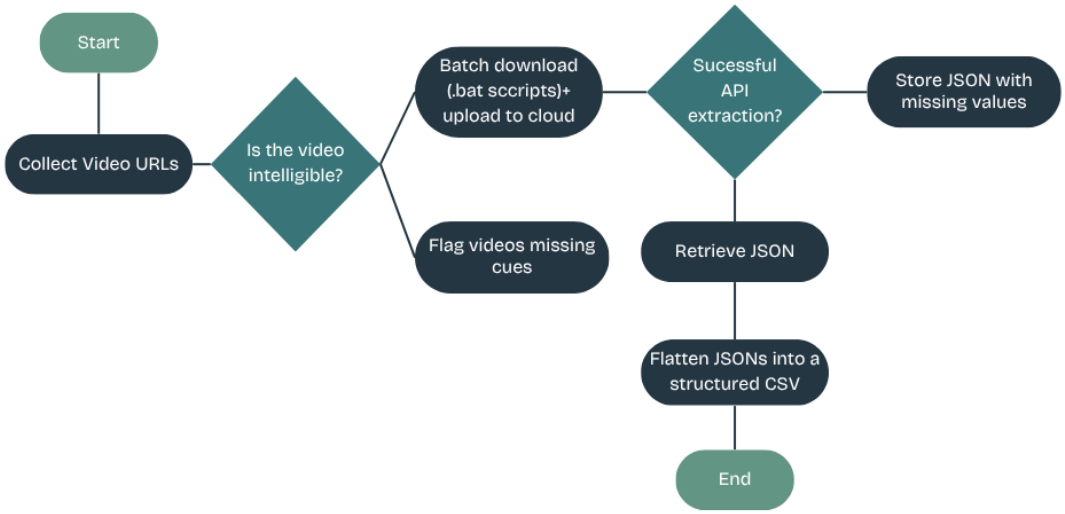

Figure 1: Preprocessing Pipeline Flowchart

Se implementan y ajustan varios modelos de aprendizaje automático como parte del proceso de modelado predictivo. Aparte de Historgram Gradient Boosting, se desarrolla una red neuronal (MLP: perceptrón multicapa), afinando hiperparámetros. Asimismo, se entrena un clasificador Support Vector Machine (SVM) con un núcleo de base radial (RBF). Finalmente, se construye y perfecciona un clasificador de Random Forest mediante técnicas de poda y selección de atributos.

Paralelamente, se lleva a cabo un análisis explicativo y exploratorio del conjunto de datos utilizando modelos de regresión logística. Esta fase permite evaluar desempeño tanto en escenarios multiclase como binarios (presencia de alteración psicológica frente a condición saludable), detallando además el impacto específico de distintos subconjuntos de variables.

4. Resultados

Los resultados confirman la eficacia del enfoque propuesto. El modelo HGB se posicionó como el más robusto, alcanzando una exactitud del 90% y mostrando una distribución de errores bien calibrada con gran facilidad de uso.

Los modelos explicativos de regresión logística también demostraron un rendimiento notable, particularmente en la clasificación binaria (salud mental alterada vs. control), con un F1-score superior a 0.92. La combinación de variables textuales, acústicas y faciales fue clave para alcanzar estos resultados, y los experimentos con subconjuntos de características evidenciaron que la voz, por sí sola, ya ofrece un importante poder discriminativo, mientras que los rasgos emocionales y de personalidad refuerzan el rendimiento cuando se integran en el modelo completo.

5. Conclusiones

Más allá de los resultados cuantitativos, esta investigación pone en evidencia el potencial transformador del aprendizaje automático aplicado a la salud mental. Souly no solo demuestra ser capaz de detectar signos de ansiedad y depresión a partir de un simple vídeo, sino que también propone un nuevo paradigma para la evaluación psicológica: uno que se basa en datos accesibles, análisis objetivo y procesamiento no invasivo, lo que abre una vía hacia un diagnóstico equitativo y proactivo.

Asimismo, se ha diseñado una estrategia de adopción realista orientada al mercado español, con vistas a su expansión internacional. El modelo de negocio planteado, tipo Software as a Service (SaaS), contempla su integración inicial en entornos corporativos, clínicos y educativos, con despliegues piloto que permitan demostrar su impacto y retorno de inversión.

6. Referencias

[1] World Health Organization, “Depression,” *WHO Fact Sheets*, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression | https://www.who.int/news-room/fact-sheets/detail/mental-health-at-work

PHENOMENOLOGICAL MULTIVARIATE ANALYSIS OF MENTAL HEALTH ISSUES USING PREDICTIVE MODELLING & ARTIFICIAL INTELLIGENCE ON VOICE DATA

Author: Sánchez-Cabezudo Pinto, Carlos.
Supervisor: Martín Corral, David
Collaborating Entity: Souly

ABSTRACT

This study develops a complex set of machine learning models to distinguish between anxiety, depression and control (healthy) subjects based on a meticulously gathered sample of video data through an automated preprocessing pipeline.

Keywords: Machine Learning; Mental Health; Processing Pipeline; Algorithms; Souly.

1. Introduction

Mental health disorders such as anxiety and depression are increasingly prevalent worldwide, and early detection is crucial for effective intervention. Globally, depression and anxiety alone lead to productivity losses of over $1 trillion each year, with 3% of adults receiving disability benefits suffering from anxiety or depression *[1]*.

Machine learning (ML) techniques, like the ones used in this thesis, can assist clinicians by automating the identification of mental health conditions from patient data. In this thesis, Souly's API is used to analyze raw video data (obtained through data scraping) to create a structured dataset with a variety of features based on voice, facial and textual cues obtained from said videos. This dataset is then used to train a variety of cutting-edge algorithms to compose a real-word useful predictive app.

2. Project Definition and Objectives

The goal of this project is to build and evaluate a system that classifies individuals into three categories: anxiety, depression, or control (no clinical condition). The specific objectives include: (1) designing a robust preprocessing pipeline that cleans raw data, extracts relevant features, and normalizes them for modeling; (2) developing and tuning classification models using different algorithms, (3) evaluating model performance using metrics such as accuracy, confusion matrices, and ROC curves, and (4) creating a detailed, figure-based, methodical plan for the path towards market adoption of the solution developed in this thesis. This systematic approach aims to ensure transparency and reproducibility of each step in the process.

3. Preprocessing Pipeline and Models Developed

The data preprocessing pipeline (as shown below) begins constitutes a fully automated, modular system designed to transform raw, user-generated video content into a structured dataset suitable for machine learning analysis. The process begins with the selection of mental health-related videos from YouTube and TikTok using scrapers and keyword filters, followed by their batch downloading via customized .bat scripts. Videos are uploaded to the cloud through secure Amazon S3 URLs, triggering multimodal analysis via Souly's external AI-powered API. This API extracts a wide range of

features—including acoustic properties, vocal emotions, facial expressions, semantic content, and inferred psychological traits—returning them as structured JSON files. These outputs are systematically parsed and converted into a unified CSV dataset, with comprehensive quality checks to ensure consistency and model-readiness.

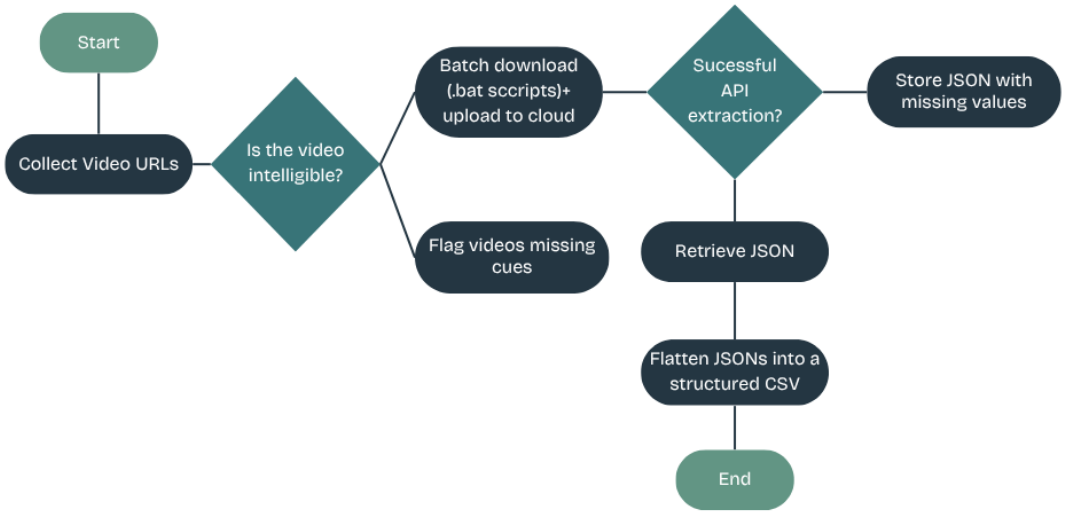

Figure 2: Preprocessing Pipeline Flowchart

Several machine learning models were implemented and tuned. Additionally, to Histogram Gradient Boosting, a neural network (multilayer perceptron) was also constructed, tuning hyperparameters such as neuron number and learning rates to optimize accuracy. A Support Vector Machine (SVM) classifier with a radial basis function kernel was also trained. Finally, a Random Forest classifier was trained and perfected through pruning.

An explanatory and exploratory analysis of the dataset was also conducted using logistic regression models, studying performance both multiclass and binarily (health issue vs. healthy) and particularizing for specific subsets of data groups and features.

4. Results

The most effective algorithm in the multiclass scenario is the ensemble decision tree method, particularly the HGB due to its superior combination of accuracy (90%), stability, and ease of use. However, every algorithm examined demonstrated excellent capability, reinforcing confidence in the use of machine learning for mental health classification, strengthening the validity of the findings: regardless of the chosen algorithm, the data signal is strong enough to allow high-fidelity classification. This convergent result is a positive indication for real-world applicability – it means

conclusions drawn are not contingent on one model's idiosyncrasies but reflect a genuine structure in the data that multiple algorithms can learn.

The logistic regression models also demonstrated considerable accuracy, both in binary classification (89% accuracy, 0.92 F1-score for detecting anxiety/depression) and multiclass (achieving an accuracy of 88% and macro F1-scores above 0.85 for each condition), using the best-performing models that combined both textual and numeric features. Subset experiments revealed that acoustic features were the most impactful among the numeric modalities, while movement and biometric features contributed less individually but enhanced performance when included in the full model.

5. Conclusions

The research conducted in this thesis confirms the viability and effectiveness of using multimodal machine learning techniques for the detection and differentiation of anxiety and depression. However, beyond algorithmic performance, the development of the Souly system represents a broader contribution: it demonstrates how accessible digital recordings, such as short video testimonies, can be transformed into powerful diagnostic aids when appropriately processed and interpreted.

Crucially, this work outlines a realistic and structured path to real-world implementation. Market adoption of Souly is envisioned in phased stages, beginning with pilot deployments in selected segments with high unmet need and strong institutional incentive. A Software-as-a-Service (SaaS) model, coupled with clear ROI metrics and privacy safeguards, could accelerate initial uptake, further enhanced with the near-future planned integration with mobile devices.

6. References

[1] World Health Organization, "Depression," *WHO Fact Sheets*, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression | https://www.who.int/news-room/fact-sheets/detail/mental-health-at-work

Memoir Index

Figure Index

Table Index

Capítulo 1. INTRODUCTION

This first chapter provides a detailed introduction that aims to bring out the interest of the reader regarding this project and what it entails.

1.1 RELEVANCE OF THE STUDY

Mental health disorders such as depression and anxiety are widespread and pose a major public health challenge. Globally, depression and anxiety alone lead to productivity losses of over $1 trillion each year. Mental health conditions are a leading cause of disability, accounting for over one-third of disability claims in major economies like the U.S. and across Europe. Around 6% of working-age adults receive disability benefits, over 50% of which are directly related to mental health issues or linked to psychosomatic symptoms exacerbated by stress and anxiety *[1]*.

Despite the high incidence of these conditions, they often go undiagnosed and untreated. It is estimated that roughly 60% of individuals with a mental health condition do not receive any treatment, and primary care providers correctly identify mental health issues in only about 47% of cases *[2]*. Stigma and lack of objective screening tools contribute to this treatment gap. There is a clear need for more accessible, efficient and objective methods to detect and monitor mental health issues as early as possible.

Such rising prevalence of mental health issues globally, exacerbated by factors such as social media exposure, socio-economic pressures, and the COVID-19 pandemic, has led to an urgent need for innovative, effective methods to identify and manage these conditions. In recent years, advances in digital health and artificial intelligence (AI) have opened new avenues for mental health assessment. The human voice, in particular, has emerged as a rich medium for potential health biomarkers. Changes in speech patterns, tone, pitch, and other vocal features can correlate with psychological states. Research indicates that disruptions in vocal characteristics can be analyzed and used as diagnostic cues, essentially treating voice

as a "biomarker" of health *[3]*. This is especially relevant for depression and anxiety, which often manifest in subtle changes in speech (such as slower speech, flatter intonation, or jitter in the voice). Leveraging voice data for mental health assessment could enable non-invasive, quick screenings that integrate seamlessly into everyday technology like smartphones or telehealth calls. Such voice-based tools promise to augment traditional mental health screening (which typically relies on subjective questionnaires) with objective, quantitative measures.

In addition to vocal dynamics, facial expressions and the semantic content of speech have also demonstrated strong potential as complementary biomarkers—facial micro-expressions can reflect emotional dysregulation, while linguistic patters and sentiment in the spoken content (from the tense in which the speaker communicates to the arrangement and structure of the response) provide further context to assess those cognitive and affective states.

The scalability and non-invasive nature of this platform make it especially suitable for widespread adoption, allowing companies to embed mental health monitoring seamlessly into daily workflows. This innovative approach promises not only to save costs but also to foster a more supportive and stigma-free workplace culture. Ultimately, this project is positioned to bridge a major diagnostic gap in the mental health sector, offering a solution that is as cost-effective as it is expansive in reach.

1.2 SCOPE AND OBJECTIVES

1.2.1 SCOPE

This project focuses on developing and evaluating an AI-driven system for analyzing voice data to detect indicators of mental health issues, specifically depression and anxiety. The scope includes gathering a suitable dataset of voice recordings with mental health labels, extracting multidimensional acoustic features, and applying multivariate predictive modeling to discern patterns associated with depression and anxiety. The analysis is termed "phenomenological" and "multivariate" to emphasize that it will explore a wide range of

vocal features (pitch, tone, cadence, etc.) and their relationships to mental health states, rather than focusing on a single parameter.

Key objectives include:

a) Identifying which vocal, text and facial features are most indicative of depression and anxiety.
b) Building machine learning explanatory and predictive models that work efficiently and effectively using those features.
c) Validate the performance of these models and assess the implementation and commercialization of the solution.

The project also entails, through the automatized uploading of videos and results-verification, testing and developing an API-based prototype application that can process voice videos and output a mental health diagnosis, demonstrating how the research can be translated into a fully functional tool.

It could be argued that the end goal is to answer the research question: Can we reliably detect and quantify depression or anxiety levels from a person's voice using AI models, and how might such a tool be implemented in practice?

1.2.2 SPECIFIC OBJECTIVES

To achieve the general objective, several specific objectives have been established, each corresponding to a key aspect of the project:

1. Automated Dataset Construction and Feature Extraction: Collect and construct a labeled dataset of voice and video recordings relevant to mental health, extracting meaningful features from these recordings through the previously explained APIs.

2. Machine Learning Models Development: Develop a set of models that take the extracted features and predict mental-health related (anxiety and depression) outcomes. This is done both through a regression and a variety of predictive algorithms, studying both binary and multivariate capabilities of the app.

3. Validation and Evaluation: Rigorously evaluate the performance of the Souly platform and its underlying model. This includes quantitative evaluation of the model's accuracy, sensitivity and specificity in detecting mental health conditions as well as qualitative feedback from result interpretation.

4. Develop a Deployment Strategy and Ethical/Regulatory Considerations: Formulate a clear path to deploy the platform in real-world conditions, including strategies for user adoption in the Spanish market and expansion beyond. This objective involves identifying any regulatory requirements and developing an adoption strategy.

1.3 RATIONALE BEHING THE PROJECT

The proposed platform, *Souly*, addresses a pressing intersection of technological capability and societal need in mental health screening. This section articulates the justification for undertaking this work from technical, social and commercial perspectives. Recent advances in artificial intelligence (AI) and signal processing have opened new avenues for non-invasive health monitoring using everyday devices. Meanwhile, mental health problems have escalated in prevalence and impact, especially in the post-2020 era, creating urgency for innovative solutions. Souly's focus is on analyzing voice and video inputs to screen for mental health indicators (such as stress, anxiety or depression) in a way that is accessible, scalable and privacy conscious. The rationale behind developing Souly is multifaceted, as detailed below.

1.3.1 TECHNICAL MOTIVATION

From a technical standpoint, it is now feasible to extract meaningful biomarkers of mental health from voice and facial video data using AI. Research in the past few years, as previously commented, has demonstrated that machine learning models can detect patterns in speech and facial expressions that correlate with psychological conditions. Such results validate the premise that voice alone can serve as a reliable indicator of mental state.

Similarly, advances have been made in video-based emotion recognition: AI models using facial-image processing can detect depression onset with promising accuracy. Dartmouth College researchers in 2024 developed a smartphone app that passively monitors facial cues; it detected early depressive symptoms with about 75% accuracy in clinical subjects *[13]*. These technical breakthroughs justify Souly's approach – by combining voice and video analysis, we leverage a multi-modal AI strategy in hopes of achieving higher screening accuracy and robustness.

Another technical motivation is the non-invasive nature of this approach. Traditional mental health screenings often rely on self-report questionnaires or clinical interviews, which can be time-consuming and require active patient participation. In contrast, voice/video analysis can be done passively or in the background of regular interactions, thus lowering barriers to screening. The fact that these AI models can run online at any time and even provide explainable outputs (highlighting which speech or facial features influenced the result) is a strong technical justification for their adoption. Moreover, high smartphone penetration and improvements in device sensors mean that the technical infrastructure needed for massive adoption of the platform created and presented in this project is largely in place.

1.3.2 SOCIAL MOTIVATION

Socially, the need for improved mental health screening and early intervention is acute. In Spain and globally, mental health disorders have become highly prevalent and are sometimes referred to as the "epidemic of our century" *[14]*. According to Spain's National Health System report in 2023, 34% of the Spanish population had suffered from a mental illness at some point – a figure significantly higher than the often-cited global statistic of 25%. Common conditions such as anxiety disorders affect about 10% of Spaniards (14% of women and 7% of men) each year. Depression and anxiety particularly surged amid and after the COVID-19 pandemic, with younger demographics showing alarming trends, and this increase has remained prevalent the last couple of years. Early detection is crucial: conditions like depression are the leading cause of disability in adults under 40, yet many cases go untreated and even undetected until they reach a crisis stage.

However, significant barriers prevent people from accessing timely mental health care. Stigma remains a major issue, but there still are structural barriers, such as shortage of mental health professionals and long wait times. These gaps mean that a large segment of those suffering do not receive support early enough, if at all. Because of this, this platform is socially justified as a tool to help bridge this gap. By providing quick, private screening through voice and video, it lowers the threshold for individuals (especially those in stigmatized or underserved groups) to check their mental well-being. For example, an employee or student could use Souly on a regular basis to monitor their anxiety and depression levels. This aligns with the public health goal of prevention and early detection (as emphasized in Spain's Mental Health Strategy 2022–2026 *[15]*.

The social rationale also ties to improving well-being in key community settings. For instance, in workplaces, especially high-pressure environments like could be Big Four consultancy firms or banks, chronic anxiety and depression embedded in burnout have become prevalent. These not only harm employees' quality of life but also affect teamwork and productivity. A non-intrusive screening platform like this one can help an organization proactively offer support or adjustments for employees showing signs of depression or anxiety (with appropriate privacy safeguards). In addition, in clinical settings it could assist general practitioners or mental health triage nurses in quickly assessing patients. Given a primary care doctor often has limited time, having an automated voice/video screening before or during a consultation could flag mental health risks that the patient might not volunteer. Moreover, in educational institutions, where counseling resources are often stretched, automated screening could help identify students in need of support. This project provides a scalable solution in these contexts, potentially being able to screen entire classes or company departments in real time at a reduced cost, something not feasible with limited human counselors.

1.3.3 COMMERCIAL AND ECONOMICAL MOTIVATION

From a commercial and economic standpoint, mental health represents a significant market and an area of substantial cost savings if addressed correctly. Spain is no exception; if

anything, the pandemic exacerbated costs related to absenteeism and disability. A recent 2025 study highlighted that in Spain, sickness absences (all causes) in 2023 accounted for 5.4% of GDP, translating to €81.6 billion lost in economic output *[16]*. Notably, mental health disorders were among the leading causes of these work absences – in 2023 the social security system recorded nearly 600,000 new cases of temporary work incapacity due to mental health issues, a figure that had more than doubled (+110%) since 2018.

Therefore, the commercial justification for this platform is that it offers a solution in this high-impact area. By enabling early detection and intervention, companies and health systems could reduce the severity and duration of mental health-related absences, thereby saving substantial costs. Even a modest reduction in stress-related sick leave in a large corporation could yield meaningful financial returns when aggregated.

Investing in mental health not only avoids costs but also has a positive return on investment (ROI) for employers. Studies by Deloitte in the UK have quantified this ROI, finding that on average every £1 invested in workplace mental health programs returns about £5.30 in improved productivity and reduced absenteeism *[17]*. Similar analyses in other countries have consistently found ROI in the range of 4:1 to 6:1 for proactive mental wellness spending. These figures make a compelling economic case to any enterprise: tools that help employees stay mentally healthy effectively pay for themselves multiple times over. In this context, Souly can be positioned as the core of a company's employee wellness toolkit – an innovative, AI-driven program that helps identify who might benefit from counseling, stress management resources, or workload adjustments before problems worsen. The commercial opportunity here is twofold: large organizations (corporate and public sector) have incentives to adopt such solutions to reduce costs, and there is willingness to pay for effective platforms that demonstrate results. Souly could be offered as a Software-as-a-Service (SaaS) to companies, generating revenue while delivering cost savings to the client, thus creating a win-win scenario.

The mental health technology market itself has grown rapidly in recent years, indicating a strong commercial momentum. Globally, the mental health tech market was estimated at $38

billion in 2024, comprising about 20% of the entire digital health market *[18]*. In Spain, the digital mental health market size was about $420 million in 2024 and is projected to reach $2.7 billion by 2035, implying an aggressive compound annual growth of ~18.5% *[19]*. This growth is driven by increasing societal awareness and government support for integrating digital solutions into healthcare. The Spanish mental health app segment specifically is also expanding, with one market report noting revenue of $140 million in 2024 for Spain, expected to more than double by 2030 [20]. These trends illustrate a ripe market for a platform like Souly. Early entrants that can demonstrate clinical validity and user engagement stand to capture significant market share.

Furthermore, governments are injecting funding into mental health initiatives which can indirectly support commercial adoption. The Spanish government launched a Mental Health Action Plan (2022–2024) with €100 million dedicated to improving mental health services and outcomes *[15]*. Such funding streams and policy support can create opportunities for public-private partnerships, pilot programs or grants that a platform like Souly can benefit from, accelerating its path to market.

1.4 METHODOLOGY

The project methodology encompasses the data sources and preparation, the model development process and the approach to evaluation.

1.4.1 DATA COLLECTION AND PREPROCESSING

The first step involved assembling a dataset suitable for training and testing the mental health screening model. Given the dual-modal nature of Souly, the dataset contains voice recordings and video footage with associated mental health labels or scores. The dataset construction was based on the automated recollection of Youtube and Tik Tok videos through sophisticated .bat archives. Having done this, the videos were then uploaded to the API in an automated way, using a variety of Python codes that will be described in the next section of this thesis. The result of this stage, as will be explained in the future in greater

detail, was a csv file with key voice, facial and text features, which provided a proper dataset to then use for the multivariate phenomenological explanatory and predictive modelling.

1.4.2 MODEL DEVELOPMENT

With feature vectors obtained, the next step was building the models. The explanations of the models developed, and the results obtained are included in the sixth section (for the explanatory model) and seventh section (for a series of predictive algorithms). Models developed were based on a variety of techniques and algorithms that were compared between themselves to tailor the best results for the use case of this project. Different variables importance in the model results, text vs non-text analysis, binary vs multivariate approaches and, I general, an in-depth study was conducted as will be detailed in those sections. The result is a fully functioning and perfected explanatory model curated through a logistic regression as well as the obtention of the best working algorithm to predict anxiety and depression with the Souly deep learning platform.

1.4.3 RESULTS AND DEPLOYMENT STRATEGY

The final part of the project involved evaluating the results and setting a future approach towards incorporating the models developed in the current architecture and creating a scalable deployment strategy for the Souly app, considering the target users and sectors previously mentioned in this section. This will be commented on in the conclusions section of this thesis in more detail.

1.5 TIMELINE AND KEY MILESTONES

The development of this TFG was carried out to a structured timeline spanning six months (from December through May) and one month left for redacting and evaluating the thesis (June) with clearly defined milestones each month. This structured scheduling was crucial given the multifaceted nature of the project, ensuring steady progress on data, modeling and deployment tasks. Below is an overview of the timeline and key milestones achieved, broken down by month and step:

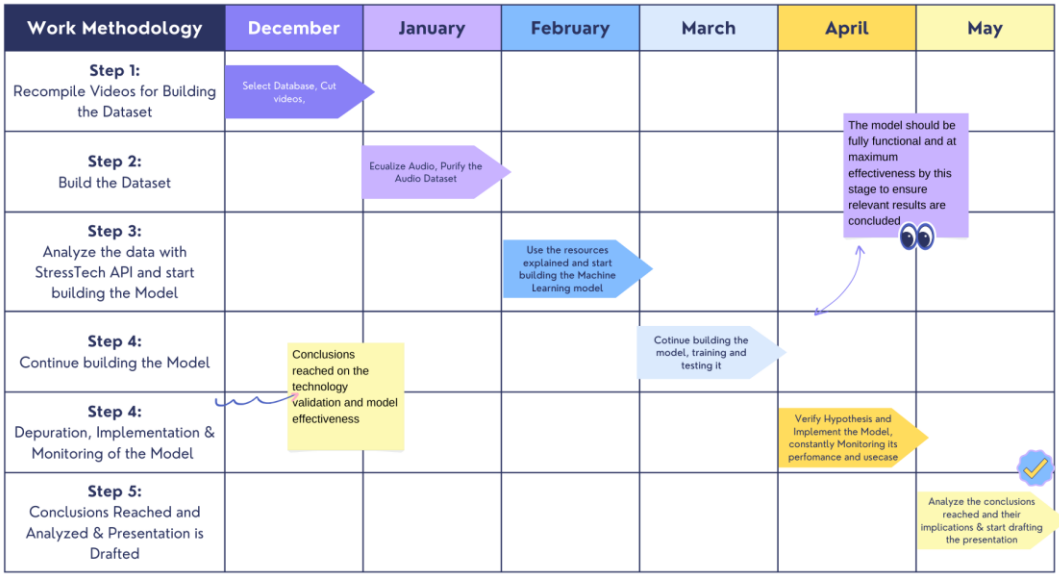

Figure 3: Timeline and Key Milestones

1. December 2024: During this month the project began with recompiling the videos to be used in the dataset. The automation channel was initiated with .bat files, unique for Youtube video-scraping and Tik Tok scraping, using a variety of platforms (such as Tik Tok Scraper). Manual review of videos was, at times, necessary to be performed to check the validity of the selected videos.
2. January 2025: The dataset construction was further enhanced until reaching approximately a dataset of 7000 videos. The first API calls were tried to ensure compatibility and that the API responded as desired.
3. February 2025: A series of Python codes were meticulously programmed to fulfill the complete pre-processing pipeline. These codes enabled automated API calling for the dataset analysis and recollection of features for future model creation. The API was stressed out with hundreds of simultaneous calls, and it responded swiftly and effectively, with no errors. Once the pipeline was fully executed, the overall design of which models to use and how to develop the algorithms was initiated.

4. March 2025: The ML explanatory and predictive models were carefully programmed using Python in the Jupyter environment through a series of ypnb files that will be later discussed.
5. April 2025: The models were finalized and depurated, ensuring no errors remained in any of the files and useful conclusions had been reached in all realms of the analysis.
6. May 2025: The models were meticulously analyzed and compared to obtain the best working parameters and algorithms. Conclusions were reached and further studied. Initial Drafting of the written thesis and presentation begins.

Capítulo 2. TECHNOLOGY OVERVIEW

2.1 OVERVIEW OF SOULY'S API

The technology underpinning this project revolves around a cloud-based API which performs advanced multimodal analysis of human voice (and video) data for mental health insights. These platforms employ state-of-the-art speech processing and natural language processing (NLP) techniques to transform raw audio recordings into detailed information about the speaker's emotional state and personality traits. In essence, a user's voice recording (or a video clip containing voice) is analyzed in near real-time, yielding objective indicators of stress levels, mood/emotions, and personality characteristics. This API extends this capability by processing both audio and video for emotion analysis, incorporating facial expression data alongside voice data for a more comprehensive assessment (i.e. multi-modal analysis). This system is designed to provide rapid feedback on mental well-being metrics, making it suitable for interactive applications in health and wellness.

At a high level, it ingests a video file from a user and return a structured analysis. The output is returned as a JSON document containing numerous fields that quantify emotional and linguistic features extracted from the input. These APIs thus serve as the core analytical engine for the project, handling tasks from voice signal processing and speech transcription to emotion detection and personality profiling.

One of the main characteristics of these APIs is that they provide real-time and interactive solutions, therefore being designed to be integrated into applications that provide immediate feedback. Indeed, Souly's API processes data with fast latency (low delay) and at economical cost per analysis. This means a user can upload a recording through an interface and receive results almost instantly (typically within seconds for a short recording), a critical requirement for interactive mental health applications. This system, therefore, offers a cloud service that can analyze user-contributed media in real-time to infer a rich set of psychological and emotional indicators, forming the technological foundation for this thesis.

2.2 CAPABILITIES AND FEATURES OF THE ANALYSIS PLATFORM

Souly's platform provides a broad range of analysis modules that collectively extract information about the speaker's emotional well-being, speaking patterns, and linguistic content. These capabilities can be grouped into several categories:

Figure 4: Core Capabilites of Souly's API

- Emotional Health Metrics (Voice-Based): The API evaluates the speaker's stress level and categorizes it into three tiers (high, medium, or low stress). It also gauges signs of depression and self-efficacy in the voice, similarly rating these on a scale. Additionally, the system captures indicators of helplessness and self-compassion, which are nuanced emotional states relevant to mental well-being.
- Personality and Psychological Traits: The API provides assessments of qualities such as creativity, self-esteem, compassion and communication style, which are high-level inferences drawn from both vocal characteristics and linguistic content.
- Voice Signal Analysis: Statistics regarding the speaker's pitch or frequency are also computed, mapping voice frequencies to musical notes and analyzing the

contribution of notes in the speech. This associates vocal tones with emotional "centers" and personality traits.

- Real-Time Emotional Tone Detection: Building on the acoustic features, it can classify the emotional tone of the speaker's voice in real time, analyzing vocal qualities to detect emotions such as calm, sadness, anger, fear, surprise, happiness, or neutrality in the speaker's tone. This is essentially speech emotion recognition from audio. The system can indicate which emotion (or mix of emotions) the speaker is expressing.
- Speech Content Analysis (Transcription and NLP): The API converts the spoken audio into text via an integrated speech-to-text transcription engine, leveraging OpenAI's Whisper model, automatically transcribing the audio to text while identifying the language of the speech, applying a suite of NLP analyses to the text content:
    - Named Entity Recognition (NER): The system runs an entity recognition model to extract names of people, places, organizations, and other entities mentioned in the transcribed speech, detecting a variety of entity types. Identified entities are returned in the output with their type labels, providing context on what or who the speaker is talking about.
    - Topic Classification: In addition to specific entities, the API determines the conversational topics being discussed. A multi-label text classification model analyzes the transcript and classifies it into one or more broad topic categories. This model is based on a fine-tuned RoBERTa-large transformer, enabling it to recognize multiple relevant topics in the text, generating as output a list of the detected topic labels.
    - Emotion Recognition from Text: Separate from voice tone analysis, the system also performs text-based emotion analysis. This NLP module detects the presence of various emotion categories expressed in the linguistic content of speech. It uses a multi-label classifier (also based on a RoBERTa-large model) to tag the transcript with these fine-grained emotions. The purpose is

to capture the emotional context or sentiment explicitly expressed in language, which complements the vocal tone analysis.

- Sentiment Analysis: The API performs sentiment analysis on the transcribed text, evaluating the overall polarity (positive, negative, or neutral tone) as well as the degree of subjectivity in the speaker's statements. This functionality identifies whether the speaker's utterances are predominantly positive/optimistic or negative, and whether they are factual/objective or emotional/subjective.
- Grammatical and Tense Analysis: The dominant verb tense used by the speaker is also extracted. The system employs a part-of-speech tagging or rule-based mechanism to determine if the speaker mostly talks about past events, current situations, or future plans, which is included in the output as well.

- Facial Expression Analysis: The speaker's facial expressions are examined to detect emotions (such as smiling, frowning, eye gaze, etc). The API result includes a flag indicating if facial analysis was performed and would incorporate any identified facial emotion data under its output, providing a more robust understanding of the user's emotional state.

2.3 ARCHITECTURE AND DESIGN OF THE SOLUTION

Internally, Souly's API is architected as a pipeline of specialized AI/ML models and processing modules, each handling a particular aspect of the analysis. This multi-layer architecture ensures that different types of data (audio signals, linguistic content, visual data) are processed by the most appropriate techniques, and intermediate results feed into subsequent analyses.

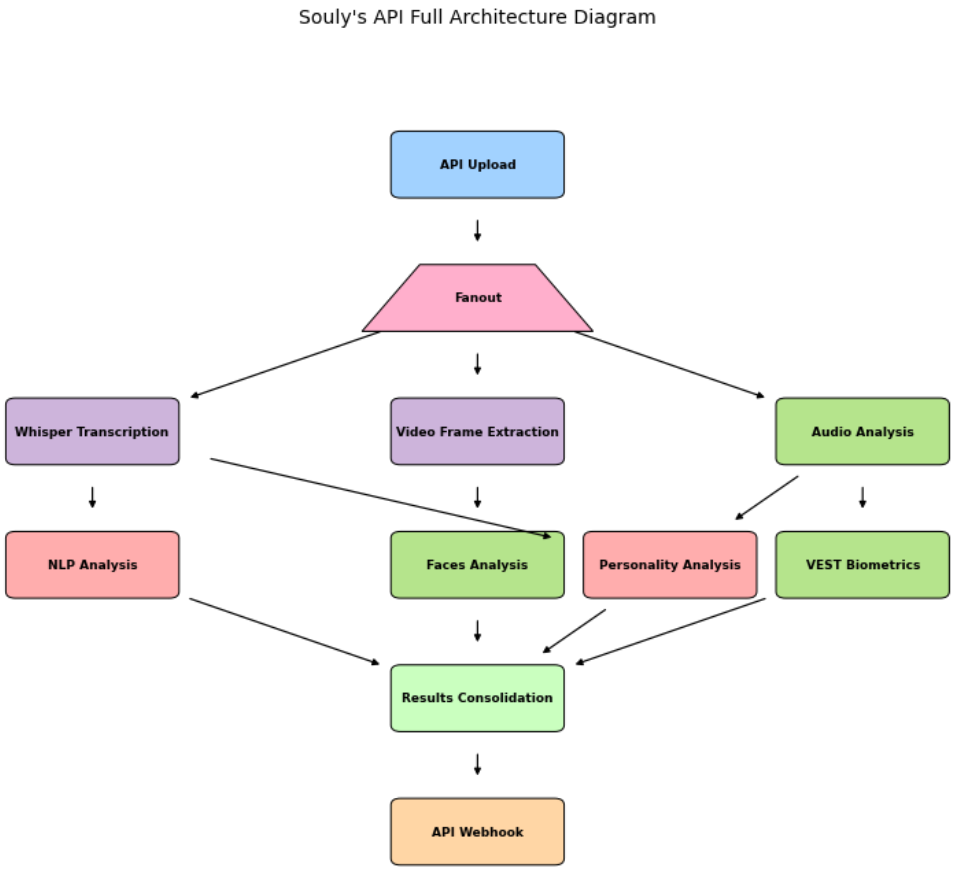

Figure 5: Souly API's Architecture Pipeline

1. Signal Processing Layer: The first layer deals with the raw data file, which undergoes preprocessing: noise reduction, voice activity detection (to focus on spoken segments), and conversion into standard formats. The system then performs feature extraction on the signal, computing voice features such as fundamental frequency (pitch) over time, intensity (loudness), and spectral features (via a spectrogram). This may involve using digital signal processing libraries and ensures that the audio is cleaned and transformed into a form suitable for machine learning models. The spectrogram or other features might be passed to downstream models as input. In parallel, if a video is provided, the video processing subsystem would extract the

audio track (to feed into the same pipeline) and also use computer vision to track the speaker's face across frames, preparing images for facial analysis.

2. Voice Analysis Layer: The next layer uses machine learning models to analyze the processed audio signal for psychological indicators:
    a. The Voice Emotional Stress Transformer outputs predictions for stress-related measures, essentially learning audio patterns that correlate with high stress or low self-efficacy, enabling it to infer those states from new speakers.
    b. The Voice Depression Transformer is another audio-based model that specifically predicts the likelihood of depression from voice, outputting categorical levels (e.g., "high" vs "low" risk) or probability scores for each category, which the API then translates into the high/medium/low labels seen in the output.
    c. The Voice Emotional Transformer handles the multi-class emotion recognition from audio (anger, joy, sadness, etc.), outputting probabilities for each emotion category, including the highest probability emotions in the results.
3. Speech-to-Text and NLP Layer: Simultaneously, or immediately after the voice models, the audio is passed to the Whisper speech recognition model (which could be considered part of the voice layer or as the start of the NLP layer). Whisper transcribes the speech to text and identifies the language spoken. The generated transcript is then fed into the NLP analysis sub-modules previously explained.
4. Output Synthesis: After all these processing layers, the system compiles the results into a unified JSON output, which contains sections for each category of analysis. Additionally, a timestamp (or ID) is included to identify each session by a unique ID (aid) for tracking.

From an engineering perspective, the API is implemented as a set of cloud services (microservices) behind a unified RESTful API. The files are stored on Amazon S3, and processing is done via AWS Lambda functions or similar, orchestrated by API Gateway. This approach allows the system to scale on demand – each analysis request triggers a chain of serverless functions that run in parallel (for example, one for audio ML, one for NLP) and

then aggregate results. The benefit is that even if many users use the service concurrently, AWS will scale out the necessary resources, and charges only incur per execution.

Capítulo 3. STATE OF THE ART

3.1 STATE OF THE ART

Beyond commercial products, a body of academic research in recent years has explored using AI and machine learning to detect mental health conditions from voice and, in some cases, facial data. This section highlights a selection of representative studies, emphasizing technical approaches and findings relevant to using these biomarkers for effectively diagnosing mental health issues, especially anxiety and depression.

3.1.1 VOICE BIOMARKERS FOR DEPRESSION – EFFICACY AND LIMITS

Donaghy et al. (2024), conducted a comprehensive review of 19 studies (2019–2022) that used machine learning to identify vocal biomarkers of depression *[8]*. Most studies attempted to classify patients with depression versus healthy controls based on acoustic features (such as pitch variability, formant frequencies, speech rate, etc.) extracted from recorded speech. The review found that, on average, ML classifiers achieved moderate performance (mean sensitivity ≈ 0.78 and F1-score ≈ 0.76) in detecting depression from voice. These accuracy levels, while promising, were generally below the accuracy of standard clinical screening tools like the Patient Health Questionnaire (PHQ-9).

However, *Donaghy et al*. note a large variability in methodologies and evaluation metrics across studies, which makes direct comparison difficult. For example, some works reported high accuracy on small datasets, but lacked validation on independent samples – highlighting generalizability as a challenge. The review concluded that voice-based depression detection is feasible but not yet as reliable as existing clinical instruments, calling for larger datasets, consistent benchmarking and that integration of voice analysis with other data was paramount to improve robustness.

3.1.2 ADVANCED SPEECH MODELLING – DEEP LEARNING APPROACHES

Recent studies have leveraged state-of-the-art deep learning models and large datasets to improve voice-based mental health assessment. *Huang et al. (2024)* proposed using a pre-trained speech representation model (wav2vec 2.0) to automatically extract high-quality features from raw audio for depression detection *[9]*. By fine-tuning this model on the DAIC-WOZ depression corpus, the authors achieved remarkably high performance: ~95% accuracy in classifying depressed vs. non-depressed individuals, and similarly strong results for multi-class depression severity prediction. This approach outperformed earlier methods that relied on hand-crafted acoustic features, demonstrating the benefit of transfer learning from large speech datasets. The use of wav2vec (which had been trained on thousands of hours of general audio) provided robust feature extraction even with limited psychiatric speech data, which further showcased the relevance of said study.

The study also underscores the importance of sufficient data, addressing the data scarcity issue by harnessing a model which already was based on general speech patterns. The impressive accuracy obtained suggests that deep neural networks, combined with self-supervised pre-training are able to capture subtle vocal cues of depression that traditional feature-engineering might miss.

3.1.3 MULTIMODAL VOICE-FACIAL ANALYSIS

Some of the latest research integrates facial expressions with voice to improve mental health assessments. An example is *Jin et al. (2025),* who developed a deep learning framework fusing facial video and audio data to diagnose depression *[10]*. In their approach, a spatiotemporal attention CNN was applied to facial expression sequences while a graph convolutional network (GCN) and LSTM analyzed vocal features, and the two modalities were combined in a joint model which was then tested on the E-DAIC (Extended Distress Analysis Interview Corpus). The Mean Absolute Error (MAE) of 3.51 in predicting patients' PHQ-8 depression scores, which is substantially lower (that is, the results obtained were considerably better) than previous benchmarks on that dataset, indicating the model's accuracy in gauging depression severity. The authors reported that their fusion approach

outperformed single-modal models, confirming that voice and facial cues together yield more robust depression indicators. For instance, a depressed individual might exhibit a flat affect (detected via facial features) and slow, monotonic speech (detected via audio features) simultaneously; the model's attention mechanism can capture such correlated patterns.

This study, therefore, highlighted that multimodal systems are suitable for early evaluation of depression, as they can pick up complementary signals - especially in scenarios like telemedicine where both audio and video may be available. This research line opens opportunities to also include other modalities (e.g. text transcripts or physiological signals) for even more accurate mental health assessments.

3.1.4 SCREENING FOR ANXIETY AND STRESS

While depression has been the primary focus of voice biomarker research, some studies have targeted anxiety and stress detection. *Espinola et al. (2022)* conducted an exploratory study using vocal acoustic analysis to distinguish multiple mental conditions: major depression, bipolar disorder, schizophrenia, generalized anxiety and controls *[11]*. Using recordings from 78 individuals across those five groups, their machine learning model (a Random Forest classifier) achieved about 75% overall accuracy in multi-class classification, with ~75% sensitivity and 93.8% specificity for identifying any disorder vs. healthy control. Notably, the vocal features that differed for anxiety tended to relate to speech rate and tension in the voice, whereas depressive speech was characterized more by prosodic flattening. The high specificity in that study suggests that when the model predicts a person has a mental health condition based on voice, it is often correct, though sensitivity varies by disorder.

In related work on stress, researchers have identified certain vocal changes due to acute stress (such as higher pitch and speaking intensity, or irregular phonation) and used deep learning classifiers to detect high-stress vs. low-stress states from voice samples. For example, one 2023 study developed vocal stress biomarkers for Korean speakers using a convolutional neural network, integrating these into a mobile health app for convenient stress monitoring. Likewise, a 2025 study by *Sharma et al*. combined speech analysis with behavioral data for early detection of various mental illnesses, reporting over 99% accuracy distinguishing

"normal" vs. "pathological" conditions in their dataset (such an exceptionally high accuracy likely reflects a controlled experimental setting; real-world performance is expected to be lower).

3.2 EXISTING SOLUTIONS IN THE MARKET

Several companies have developed voice-based AI solutions aimed at mental health screening and monitoring. These platforms leverage vocal biomarkers – acoustic or linguistic patterns in speech correlated with health conditions – to detect signs of depression, anxiety, and other disorders. Below, we review four prominent solutions and their approaches.

Comparison of Solutions

Criteria	Vocalis Health	Canary Speech	Kintsugi	SOULY
Target Conditions	Depression, COVID	Anxiety, Depression	Anxiety, Depression	Multiple Mental Health Issues
Input Modality	Audio	Audio	Audio	Audio + Text + Video
Analysis Type	Acoustic ML	Acoustic ML	Acoustic ML	Acoustic + NLP + Multivariate
Language Support	English Only	Language Agnostic	Language Agnostic	Multilingual
Validation Status	CE Mark	Internal Validation	Peer-Reviewed	Research-Grade
Integration Options	Phone APP	API	API, EHR, Telehealth	Modular API + CLI Uploads + Phone APP
Real-Time Capacity	No	Yes	Yes	Yes
Differentiators	Respiratory Focus	Short Audio, Pharma Use	Large Voice Dataset	Multimodal, Non-Invasive

Figure 6: Comparison Table of Solutions, including Souly.

3.2.1 VOCALIS HEALTH

Vocalis Health (formed in 2019 from the merger of Beyond Verbal and Healthymize) was an early pioneer in vocal biomarkers *[5]*. The company's platform analyzes voice recordings (e.g. via phone apps) to screen for a range of conditions, from chronic respiratory and cardiac ailments to mental health issues like depression. Vocalis's technology uses machine learning models to extract vocal features that may indicate cognitive or mood disorders, in addition to other diseases. Notably, voice indicators linked to depression (such as reduced pitch variation or slower speech) are part of the targeted features. Vocalis gained regulatory attention by achieving CE mark approval for a voice-based COVID-19 screening tool, demonstrating >80% accuracy in detecting infection via voice. This success underscored the broader potential of Vocalis's voice analysis platform for health monitoring. While the company's offerings have emphasized physical health and respiratory conditions, its inclusion of depression in the list of target ailments signals an application of its vocal analysis technology to mental health as well.

While Vocalis has established a strong presence in physiological health assessment, its applications in mental health are less comprehensive. Although their technology offers a model for physiological analysis, its use in detecting mental health conditions is limited, indicating an opportunity for a more specialized solution that addresses the specific vocal characteristics associated with mental health disorders.

3.2.2 CANARY SPEECH

Canary Speech is a U.S.-based company that applies patented voice biomarker technology for early detection of neurological and mood disorders *[6]*. Canary's platform can screen for mood states such as depression and anxiety, as well as cognitive conditions like dementia, using just short samples of speech. In practice, their solution requires only about 20 seconds of a person speaking to generate an assessment. The system analyzes numerous acoustic features and compares them to known patterns associated with various conditions. According to the company, this approach enables clinical-grade screening for mental health conditions and diseases using machine learning models trained on large voice datasets. The

technology is largely language-agnostic, focusing on audio features rather than specific words, which could allow broader use.

However, Canary Speech primarily focuses on medical conditions with strong cognitive or neurological components, which, while valuable, offers limited flexibility for broader applications in mental wellness or everyday emotional health monitoring. The emphasis on cognitive decline and clinical disorders highlights the need for a platform that can detect a wider range of emotional and psychological states, particularly as they manifest in less clinically overt ways, such as workplace stress or social anxiety.

3.2.3 KINTSUGI

Kintsugi is a startup founded in 2019 that specializes in AI-based voice analysis for mental health. The company's core product ("Kintsugi Voice") uses machine learning algorithms to detect signs of clinical depression and anxiety from short clips of free-form speech *[7]*. Kintsugi's approach is notable for not requiring any specific prompted script – users can speak naturally, and the system evaluates how they speak (vocal qualities) rather than what they say. The platform provides a real-time mental wellness score or assessment, enabling it to function as an "emotional vital sign" in primary care or telehealth settings.

Kintsugi's models have been trained on what is claimed to be one of the largest voice datasets for mental health, collected via a consumer wellness app across many cities and languages. The system can be integrated via API into call centers, electronic health records, or telemedicine platforms, alerting providers to patients who may need mental health support.

In 2022, Kintsugi received recognition for its innovation – such as a Frost & Sullivan technology leadership award – highlighting that it was one of the first voice biomarker technologies to provide real-time mental health assessments with as little as 20 seconds of audio, which further emphasized Kintusgi's value proposition of enabling objective, quick screening for depression and anxiety.

While Kintsugi's technology has demonstrated strong results in identifying emotional distress in speech, its application often leans toward therapeutic support rather than

proactive, preventive measures within corporate or non-clinical environments. Moreover, its diagnostic capabilities are generally limited to specific conditions like depression and anxiety, without the multivariate approach necessary to capture a more nuanced spectrum of mental health markers.

3.3 CHALLENGES IN CURRENT APPROACHES

Despite the progress in voice-based mental health technology, there remain significant challenges limiting widespread adoption and clinical impact of current approaches.

Achieving high accuracy in diverse, real-world settings is difficult. Many models that perform well in research trials or with certain demographics can fail on new population.. Furthermore, these types of models tend to struggle capturing the full variability of real-world speech. The complex, multifaceted nature of mental health disorders means that a voice model might not capture every relevant signal, which leads to produce, as it could be observed in most of the previously studied scientific publications, moderate sensitivity and specificity in practice.

Collecting clinical-grade voice data with ground truth (e.g. clinician diagnoses or standardized assessment scores) involves extensive effort and ethical considerations. This scarcity of data hinders model training and validation. Moreover, there is a lack of standardization in how studies evaluate models, with some studies reporting only accuracy, others report AUC, F1-score, etc. while also thresholds for "detection" hugely vary. This makes it hard to compare systems or set industry benchmarks, which is further magnified by the fact that few voice AI systems have undergone rigorous clinical trials. Without external validation, it is challenging for clinicians to trust these tools or for regulators to approve them, which, consequently, implies that most voice-based depression and anxiety detectors remain in protype or pilot phases.

The fact that voice and video recordings are sensitive personal health and intimate data also raises privacy and ethical concerns around data storage, consent and potential misuse of the

information provided, not yet fully contemplated by the solutions in the market and research performed. Researchers have highlighted the need to address these challenges: ensuring secure encryption and anonymization of voice data and being clear about how the analysis is used in care decisions. Additionally, the previously mentioned bias in algorithms may not only be a technical issue but an ethical one – if a model works less accurately for certain groups, it could exacerbate healthcare disparities. Misclassification is another risk: a false positive could lead to unnecessary anxiety or intervention, while a false negative might leave someone's condition unnoticed.

Even if the underlying technology is sound and verified, there still are multiple integration and adoption barriers for the current solutions in the present landscape of mental health diagnosis. Many primary care clinics have limited time and may not easily add a new screening procedure unless it clearly reduces workload or improves outcomes. Furthermore, some clinicians may be skeptical of AI "black box" tools – lacking explainability, they might be reluctant to act on a machine-generated mental health risk score without understanding the basis, which highlights the need for an easily interpretable, visually understandable and user friendly front-end. Workflow integration is another issue, for instance for solutions like Kintsugi, trying to embed into EHR systems, as it takes a considerable effort for healthcare IT departments to implement these and train staff.

3.4 UNMET NEEDS AND OPPORTUNITIES

Video-based tools could provide an early detection mechanism for those currently underserved in the realm of mental health diagnosis. Mental health conditions often go undiagnosed or untreated for long periods. It is estimated that about 60% of people suffering from mental health issues never receive any treatment *[12]*. One reason is the lack of accessible, routine screening. For instance, despite recommendations, only roughly 4% of primary care patients are actually screened for depression in practice due to time constraints and stigma. This represents a huge opportunity for quick, private voice-based tools which could enable quick, ubiquitous mental health check-ups. It should also be noted that early intervention greatly improves outcomes in depression and anxiety, so identifying those at

risk earlier is a critical need. In addition, solutions that can be deployed at scale and automatically alert care providers to concerning signs could dramatically increase screening coverage. Moreover, such tools can serve populations with limited access to mental health professionals, particularly in rural areas, to obtain a self-diagnosis of at least a preliminary assessment of their potential conditions remotely. The opportunity here is to democratize mental health screening using everyday technology (phones, apps), catching issues before they escalate.

The technologies studied could serve to objectively monitor the evolution of pathologies and provide continuity of care. Mental health status can fluctuate over time, and clinicians currently rely on patient self-reporting at appointments to gauge progress. There is an unmet need for objective, continuous monitoring between visits. Voice analysis could fill this gap by tracking a patient's mood trends through regular voice samples. For example, a patient with depression might use an app to speak a short journal entry each day; subtle changes in vocal tone or energy could indicate improvement or relapse. Opportunities exist to create "mental health vital signs" that are measured as frequently as physical vital signs, making it possible to quantitatively track depression and anxiety over weeks or months, providing immediate feedback to users. Importantly, this kind of monitoring could improve treatment of chronic conditions and prevent crises by noticing early warning signs.

Another need in the field is for tools that clinicians can trust and easily use. This involves making AI decisions explainable, because many current models act as black boxes that output a score without rationale. An opportunity exists to research and develop interpretable ML models for mental health with transparency to increase clinician buy-in and also provide insight to patients.

3.5 DIFFERENTIATING VALUE PROPOSITION OF THIS PROJECT

This project aims to contribute a unique value proposition within the landscape described above. Several aspects differentiate this work from the existing solutions and research previously covered:

1. Phenomenological and Multivariate Focus: Unlike commercial products that function as black-box predictors, this project places an emphasis on phenomenological analysis – that is, understanding and interpreting how various voice features relate to the subjective experience of mental health issues. By conducting a multivariate analysis, this project will examine multiple acoustic, facial and linguistic variables in tandem, rather than outputting a single opaque score. In doing so, the project provides richer insight into the "how and why" of voice changes in mental illness, not just a yes/no detection, which is especially valuable for clinicians and researchers seeking to ground AI findings in clinical reality.
2. Integrating Predictive Modeling with Clinical Interpretation: The outcome of the project is not limited to predicting a mental health state from voice bio markers, but also an analysis of the patterns observed. For instance, the project aims to link certain vocal, facial and textual specific attributes to specific occurrence of symptomatology (depression and anxiety), providing not only a predictive tool but also interpretive analysis.
3. Research-Oriented Innovation: As a TFG academic project, the work is positioned not only to compete in the market, but also with a focus to contribute to the scientific community. This scope has allowed the project to entail novel techniques without constraints, such as using cutting edge algorithms.
4. Alignment with Unmet Needs: Finally, it is important to note that this project is initiated and defines its goals in light of the gaps previously identified. For instance, this project entails evaluating the model on different subsets of data, aiding with the current struggle of generalizability. The value proposition here is that the project is not merely reimplementing an existing idea, but thoughtfully addressing specific weaknesses in current approaches. For example, an improved method of feature selection is proposed to handle the multicollinearity of many voice, facial and text features, yielding a simpler yet effective model.

Capítulo 4. DATA COLLECTION AND PROCESSING

This chapter details the end-to-end data acquisition and preprocessing pipeline developed for the study. It covers the identification of raw video sources, automated download procedures, cloud-based processing via an external API, and the assembly of a structured dataset ready for modeling.

No data anonymization or filtering was applied during these steps, as preserving the integrity of the raw multimodal signals (voice, facial expressions, and speech content) was crucial for subsequent analysis. By the end of this pipeline, a comprehensive multi-feature dataset (csv_finalCSVMulti.csv) was obtained, encapsulating all extracted features for use in predictive modeling.

An illustration of the preprocessing pipeline has been provided for rapid, visual analysis of the process developed, and can be observed in the next page. Also, in addition to this depiction, a link to the Github where all the thesis architecture codes are uploaded. This Github entails a more code specific explanation of the files than the one rendered during this chapter, whose aim is solely to make the reader understand the preprocessing pipeline and the reasoning behind every step taken along its design and creation:

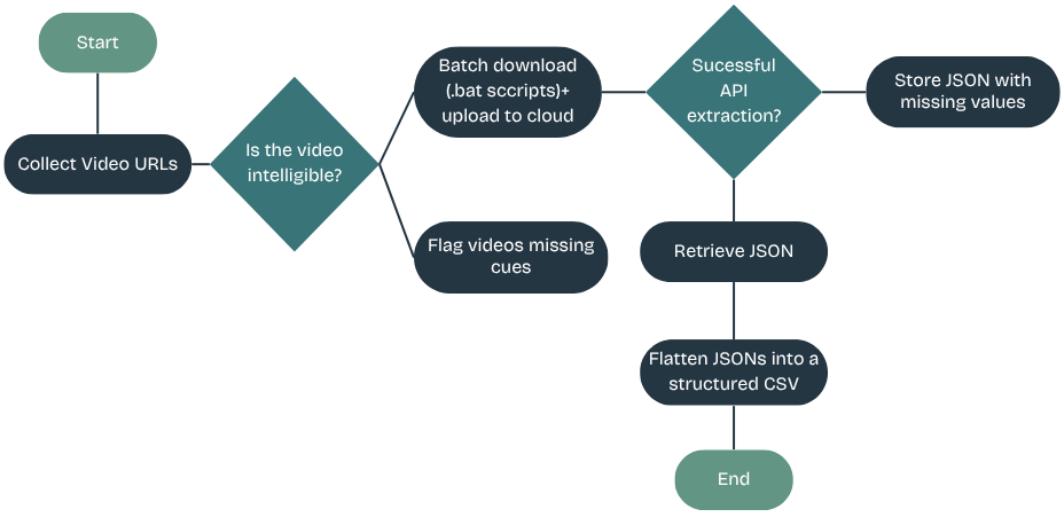

Figure 7: Preprocessing Pipeline Flowchart

4.1 IDENTIFICATION AND SELECTION OF VIDEO SOURCES

The first step in the pipeline was to identify and select appropriate video data sources that could provide rich audio (voice) and visual (facial expressions) content relevant to mental health issues, specifically focusing on phenomenological accounts of conditions like anxiety. Given the nature of the research, user-generated videos from social platforms were chosen as they often feature personal narratives, testimonials, or discussions about mental health challenges in an authentic voice. Two platforms were targeted in particular: YouTube and TikTok. These platforms were selected due to their popularity and the abundance of content wherein individuals openly talk about their mental health experiences, providing natural speech data in varying contexts.

A systematic search strategy was employed on each platform using relevant keywords and hashtags as well as selecting specific medical channels which focus on testimonies from patients with either anxiety or depression. This helped surface videos in which speakers discuss anxiety and related mental health topics. To automate this process, several Tik Tok and Youtube scrapers were used like Apify, to capture hundreds of links to videos in an effective and efficient manner *[21]*.

The selection criteria for videos were guided by the need for clear voice recordings and relevant content: videos had to contain discernible spoken language (preferably monologues or dialogues about personal mental health experiences) and sufficient length to extract meaningful vocal features (which, due to the accuracy of the APIs, was only seconds). Another important criterion was that the videos depicted the user talking to the camera, so that their facial expressions were observable and, therefore, could be captured by the subsequent API calls. Notably, the quality of both video and speech ought to be intelligible so that no errors were encountered because of faulty data collection. Furthermore, the diversity of sources selected aimed to ensure that the voice data reflected different demographics and speaking conditions (formal vlogs, casual phone camera recordings, etc.), enriching the phenomenological dataset.

For each platform, a link aggregator text file was created to compile the URLs of all selected videos. Specifically, a file urls.txt listed the YouTube video links, and a separate file urls_tiktok.txt listed the TikTok video links. These plain text files served as manifest lists of raw data sources for automated processing. Each URL in these lists corresponds to a video identified as relevant to the study's theme. Using text files for link aggregation ensured reproducibility and convenience: the entire set of source videos could be fed into automated tools without manual intervention. It is important to note that at this data collection stage, no content was removed or anonymized – the videos were taken as-is from the public domain, preserving all spoken content and any visible personal information in the footage. This decision was made to maximize the fidelity of the dataset's raw features, which would later play a critical role in the multivariate analysis.

4.2 DOWNLOAD AUTOMATION

Having compiled the lists of target video URLs, the next stage involved the automated bulk download of videos from the respective platforms. Manual downloading of dozens of videos would be time-consuming and error-prone, so a programmatic approach was adopted for efficiency and consistency. The tool of choice was yt-dlp, a widely used command-line program for downloading videos from the internet. Yt-dlp supports both YouTube and TikTok (among many platforms) and allows batch processing of multiple links.

Two platform-specific batch scripts (.bat files) were developed to drive the download process, one for YouTube and one for TikTok. The YouTube downloader script (video_downloading.bat) uses yt-dlp with the aggregator file urls.txt as input. This script initiates a mass download by feeding the entire list of YouTube URLs to yt-dlp's -a option (which accepts a filename containing multiple URLs). All videos are thus downloaded in succession without manual input. The script ensures videos are saved in a designated output directory, in this case VideosDescargados, by using yt-dlp's output parameter -P. Additionally, to enforce a uniform format, the --recode-video mp4 flag was used, converting any video to MP4 if it was in a different format. This guarantees that subsequent processing (which expects .mp4 files) can handle all files consistently, which are now stored in a folder that constitutes a local repository of raw video data.

A similar approach was used for TikTok videos, with necessary adjustments for that platform's specifics. The TikTok downloader script (tiktok_downloading.bat) also leverages yt-dlp but points to the urls_tiktok.txt list and includes extra parameters to handle TikTok's environment. An interesting difference between the two approaches was that during this one, it was necessary so specify a custom user agent string (in this case Mozilla/5.0) through a command because TikTok's servers may restrict automated downloads. By doing this, the code effectively mimics a standard web browser user agent and the script avoids potential download blocks while ensuring a smooth retrieval of content. The TikTok script also directed downloads to a chosen folder (in this case, a unified Videos directory within the project workspace).

Another implementation detail for TikTok downloads was the standardization of filenames. TikTok video filenames as saved by yt-dlp can be verbose or include unique identifiers that are not meaningful for analysis. To impose a consistent naming convention, the script performed a post-download renaming loop. Each downloaded TikTok video file was systematically renamed to a simpler schema (for example, ansiedad1.mp4, ansiedad2.mp4, and so forth, using a sequential counter). This naming scheme not only ensures that filenames are succinct, but also implicitly encodes the content category ("ansiedad", Spanish for "anxiety") which was the thematic label for all these clips. The use of sequential numbering avoids any ambiguity or duplication in file identifiers.

By the end of this stage, all targeted videos from both YouTube and TikTok had been downloaded locally in MP4 format. The videos were now stored in a consolidated directory (hereafter referred to as the Videos folder), ready for the next phase of processing. While tedious to code and implement, this automated downloading stage greatly enhanced the data collection capabilities of this project, yielding a raw dataset of audio-visual files with minimal manual manipulation and with consistency across sources.

4.3 AMAZON S3 BUCKET INPUT

After obtaining the raw video files, the pipeline's focus shifted to uploading these files to a cloud-based analysis service. The volume and size of video data, as well as the need for advanced processing (e.g., voice feature extraction using AI models), made it practical to leverage a cloud infrastructure for analysis. The chosen approach was to utilize an external API service that accepts video input and returns a rich set of analysis results. This service uses Amazon S3 cloud storage as an intermediary for data ingestion: instead of sending large video files directly through API calls, the pipeline obtains a pre-signed S3 URL for each video. Uploading via a pre-signed URL allows the video file to be transferred directly into the cloud storage bucket managed by the API, offloading the data payload from the API's interface while still associating it with a specific analysis job.

To perform this function, the pipeline implemented next a Python automation script: API_uploading.py. The script iterates through each video in the local Videos folder and performs a two-step operation for each file:

1. Request an Upload URL (and a corresponding result endpoint) from the API
2. Uploads the video file to the provided URL.

More specifically, for a given video filename, the script issues a POST request to the API's upload endpoint (/v1/upload/large) to obtain a pair of URLs. The API responds with a JSON containing an upload URL (a pre-signed URL to an S3 bucket location) and a result URL (an endpoint to later retrieve the analysis results for that file). Upon successfully receiving these URLs, the script then opens the video file in binary mode and performs an HTTP PUT request to the provided S3 upload URL. This effectively streams the video data to the cloud. The use of a pre-signed URL ensures that the file is uploaded securely to the correct storage location tied to the analysis service, without requiring direct manual interaction with S3 consoles or, in general, any further intervention apart from code execution.

Robustness and automation were key considerations in this stage. The upload script incorporated error handling and retry logic to handle transient network issues or server unavailability. If an upload attempt failed, for example, due to a network timeout, the script logs the incident and optionally retries a limited number of times with exponential backoff. This design choice was important to avoid data loss or omissions: all videos had to be successfully uploaded for analysis to ensure the completeness of the dataset. The script recorded each upload outcome in real time, printing statuses to the console (that is, confirming a successful upload or flagging a failure) and writing to a log file for any errors encountered during the previously described process.

Critically, as each video was uploaded, its associated result URL was saved into a CSV file (e.g., resulting_urls.csv) along with the original filename. This CSV became a manifest linking each input video to its future analysis results on the cloud. In essence, it is a table with two columns: the local video filename and the API-provided result endpoint for that

video. This mapping was crucial for the next stage of the pipeline, as it allowed the system to know where to get the results for each video once analysis was complete.

By leveraging Amazon S3 via pre-signed URLs, the pipeline achieved a scalable data input mechanism: large video files were offloaded to cloud storage efficiently, and the analysis service could pick them up from there for processing. The use of cloud infrastructure at this stage also means that the compute-intensive feature extraction would happen server-side, which is more effective than on the local machine (which is a laptop with no specialized hardware or software).

The outcome of this stage of the pipeline was that all videos resided in the cloud (S3) associated with unique result identifiers, and the system had a local record (CSV) of these identifiers to query for analysis outputs.

4.4 PROCESSING AND FEATURE EXTRACTION

Once the videos were uploaded to the cloud via Souly's API, the processing and feature extraction phase commenced. In this stage, the heavy analytical work was performed by the remote AI service: each video underwent a comprehensive multi-modal analysis to extract features relevant to our predictive modeling of mental health indicators. The purpose of this stage was to transform raw audiovisual data into a rich set of quantitative and qualitative features that characterize each video along several dimensions (voice acoustics, facial expressions, speech linguistics, etc.).

Using an already established, well designed API, which included complex deep-learning algorithms, ensured that the future machine learning models to be applied (for example to emotion detection or to the speech transcription), will be applied in a consistent and validated manner with useful data, rather than implemented those complex models from scratch.

This automated analysis pipeline part is, therefore, designed to process (through theAPI) an uploaded video and produce a JSON-formatted result encapsulating numerous features. The

analysis can be thought of as comprising several parallel sub-processes, each targeting a different modality or aspect of the data:

- Facial Expression Analysis: If the video contains a visible face of the speaker, the API performs facial emotion recognition across the frames of the video. It detects expressions such as happiness, sadness, anger, fear, disgust, surprise, and neutrality. The output includes metrics like the average intensity of each facial emotion over the video's duration, as well as measures derived such as the most frequent dominant emotion observed and the confidence levels of the face detections. For example, the API might report that a video's facial expressions were 60% neutral on average with a certain percentage of frames showing a happy expression as the dominant emotion, etc. These features provide insight into the non-verbal emotional cues present in the content.
- Voice Acoustic Features: The audio track of the video is analyzed to extract statistical and signal features from the speaker's voice. The API computes fundamental frequency-related features (pitch) and various descriptive statistics of the audio frequency spectrum. This includes measures such as mean and standard deviation of the voice frequency (pitch), median and mode frequency, frequency quartiles (Q25, Q75), interquartile range (IQR), as well as higher-order moments like skewness and kurtosis of the frequency distribution. Additionally, it may compute the root mean square energy (volume indicator) and other voice characteristics. These acoustic features quantitatively characterize the speaker's voice, which can be correlated with emotional or mental states (for instance, a flatter prosody or low variation in pitch might relate to certain affective conditions).
- Vocal Emotion Recognition: Beyond raw acoustic measures, the pipeline also applies models to estimate the probability of various emotions conveyed through the voice (often termed tonal or vocal emotions). The result JSON for each video includes a set of features like happy_voice, sad_voice, angry_voice, fearful_voice, calm_voice, disgust_voice, surprised_voice, etc., each representing a confidence score of that emotional tone being present in the speaker's voice with a determined

frequency. For example, a video might have a high sad_voice score if the speaker's tone and prosody align with sadness. These features provide another layer of understanding of the emotional content, complementing the facial analysis.

- Speech Content Analysis (Transcription and NLP): The spoken words in the audio are transcribed and analyzed for linguistic features. The API first detects the language spoken (in our dataset, predominantly Spanish, which the API identifies accordingly). It then likely generates a transcript of the speech. Given that the videos involved personal testimonies, the transcript captures the verbal content of their messages.
    - One aspect analyzed in this stage is, for instance, tense analysis, which the API examines providing proportions of past, present and future usage. This can be phenomenologically interesting, as, for example, a person speaking about their mental health might dwell on past experiences or focus on present feelings depending on their health state.
    - Another aspect is sentiment analysis of the speech text, yielding a sentiment polarity score and a subjectivity score, which help quantify the emotional tone of the spoken content itself (for instance, a very negative polarity could indicate a distressed narrative)
    - The API's result also included a "no_speech_probability" and an "entropy" measure for the speech segment, which together indicate the clarity and information density of the speech.
    - Finally, a text translation component is included. Since the original language was Spanish, the API (or a subsequent step in our pipeline) provided an English translation of the transcript. This translation is included in the results so that further text-based analyses or comparisons can be done in English if needed (and it serves as a normalization step, given many NLP tools and lexicons are geared towards English). The presence of a translated transcript greatly improves the potential international use cases and acceptance of this work.

- Psychological Trait and State Inference: Uniquely, the employed API also generated higher-level inferred features that attempt to quantify psychological traits or states of the speaker, based on the combination of vocal, verbal, and facial cues. For example, the results include five big personality trait scores as numerical values: extraversion, agreeableness, openness, conscientiousness, neuroticism. Additionally, the API provides estimations related to mental health constructs such as stress, helplessness, self-efficacy, and depression, each typically broken down into probabilities or indicators of high, medium, or low levels. For example, the output might have fields like stress_high, stress_medium, stress_low indicating the model's assessment of the likelihood that the speaker is experiencing high stress, moderate stress, or low stress respectively. Similar tripartite indicators are given for helplessness and depression, and for self-efficacy. Other trait-like measures include constructs such as self-esteem, compassion, communication, imagination, and awareness. These high-level features are particularly relevant to phenomenological analysis, as they aim to bridge the gap between observable signals and psychological states.

All these analyses occur remotely on the cloud server once the video is uploaded. The pipeline is asynchronous: the upload request triggers processing, and the system must then retrieve results when they are ready. When implemented, after a short waiting period (typically a few seconds per video, although the exact timing depends on video length and server processing speed) a result retrieval mechanism was executed.

A Python script (output_API_download.py) automated the collection of the analysis outputs by querying each stored result URL from the earlier CSV. This script read the list of result endpoint paths (e.g., /v1/result/xyz123...) from the CSV (urls_resultado.csv) and issued an HTTP GET request to each, authenticated with the API token, to fetch the JSON results. For reliability and simplicity, the implementation used a command-line HTTP client (curl) invoked via a subprocess call to perform these GET requests. As each response was received, the script parsed the JSON and saved it to a local file in a "resultados" directory, using the original video's filename (but with a json extension) as the identifier. This ensured a one-to-one correspondence between each video and its result file.

By the end of this step, for every video that was uploaded, there existed a local JSON file containing all the extracted features and analysis outcomes for that video.

4.5 DATASET STRUCTURING AND CSV INTEGRATION

The collection of JSON files produced in the previous stage, while rich in information, needed to be transformed and consolidated into a single dataset for ease of analysis and modeling. JSON is a flexible format that can represent nested data, but for statistical modeling a tabular CSV (Comma-Separated Values) format is more practical.

This stage of the pipeline focused on structuring the data: flattening each video's JSON analysis output into a fixed set of features (columns) and integrating all videos into one coherent table. The goal was to create a multi-feature dataset where each row corresponds to one video (one data sample) and each column corresponds to a specific feature or attribute extracted from that video.

A custom Python script (csv_generator_json_source.py) was written to automate this JSON-to-CSV transformation. The script defined a comprehensive schema of features that we expected to extract from each JSON – essentially a header for the CSV. This schema was derived from the previous knowledge of the API's output structure and content and was meticulously designed so that all relevant fields needed for the future multivariate analysis were included.

A summary of the defined header structure is included to provide greater insight into the csv dataset structure that will be used in the following chapter's models.

- Metadata and Status Flags: Fields such as created_at (timestamp of analysis), aid (analysis ID), and original file properties like extension, format, duration are captured. Additionally, boolean flags indicating whether each analysis component was successfully executed (e.g., file_stored, facial_analysed voice_analysed, voice_transcribed, biometrics_extracted, speech_analysed, personality_analysed, faces_extracted) are included.

- Facial Emotion Averages: For each of the core emotions (angry, disgust, fear, happy, sad, surprise, neutral), a column stores the average intensity value from the facial analysis. There is also a column for the most_frequent_dominant_emotion identified by the facial analysis, and one for the count of how often "surprise" was the dominant emotion (as a representative example of dominant emotion counts).
- Personality and Trait Scores: columns for the previously commented personalities: (extraversion, neuroticism, agreeableness, conscientiousness, openness) are included, alongside columns for the additional traits such as survival, creativity, self_esteem, compassion, communication, imagination, awareness. Each of these holds the numerical score output by the API's trait analysis for the video. Similarly, the triadic mental state metrics are each expanded into three columns: for instance, stress_high, stress_medium, stress_low capture the probabilities of the subject having high, medium, or low stress levels.
- Voice Frequency Features: A set of columns store the statistical descriptors of the voice frequency distribution: voice_mean, voice_sd (standard deviation), voice_median, voice_mode, voice_Q25, voice_Q75, voice_IQR, voice_skewness, voice_kurtosis. Additionally, fields like voice_mean_note, voice_median_note, voice_mode_note, voice_Q25_note, voice_Q75_note are included (these are the nearest musical note corresponding to those frequency values). The voice_rmse (root mean square energy) is also recorded as a measure of voice loudness dynamics. Two other voice-related fields, pitch and tone, were captured if present.
- Voice Emotion Probabilities: Columns for each vocal emotion category (sad_voice, disgust_voice, fearful_voice, neutral_voice, happy_voice, angry_voice, calm_voice, surprised_voice) store the scores from the voice-based emotion model, typically with values between 0 and 1 indicating the strength of each emotion in the speaker's tone.
- Speech and Language Features: The language detected in the video's speech is recorded, along with the no_speech_prob (indicates segments of silence or non-speech) and entropy (reflects information density). Three columns tense_past, tense_present, tense_future capture the proportions of verb tenses in the transcript, as explained earlier. Two columns sentiment_polarity and sentiment_subjectivity

contain the sentiment analysis results, quantifying the emotional valence and objectivity of the speech content. Finally, a translation column contains the English translated transcript of the spoken content, providing the full translated speech from the video in a common language for reference.

- Target/Variable Label: Since this is a supervised learning context focusing on a specific mental health condition, a column was reserved for the phenomenological variable of interest. In this case, all videos selected were either related to anxiety and depression or were control videos to enrich the dataset and improve the quality of future multivariate analysis.

With the schema in place, the script proceeded to read each JSON file in the results directory. For each file, it parsed the JSON structure, navigated to each required field as per the schema, and extracted the value. If a particular field was missing in the JSON, due to, for example the face going undetected or the video having no face (which would be due to a sampling error in the video selection because, as previously described, the overarching objective of the first step in the pipeline was to collect quality videos with both intelligible voice and facial expressions) the script was designed to insert a "null" as a default placeholder so that the structure of the resulting csv file was not altered. This way, every row had the same number of columns, and values were aligned under the correct headers.

The script also took care to convert data types appropriately (numbers remained numeric, booleans as true/false or 1/0, and text as strings in quotes in the CSV) and to handle nested JSON objects via helper functions.

The output of this JSON integration script was a CSV file, which we can refer to as the "final results CSV" (in practice named resultados_final.csv or a similar moniker). Each line in this CSV represents one video sample, starting with identification information and followed by all the extracted feature values in their respective columns. This single file thus neatly organizes the multi-modal features (facial, vocal, linguistic) for all videos in the study, providing a very high-dimensional structured dataset, also reflecting the rich analyses performed by the API.

Basically, throughout this stage of the pipeline, the unstructured analysis outputs were successfully transformed into a machine-learning friendly format. The resultant CSV integrated all relevant information, enabling direct import into data analysis libraries or tools. This structured dataset is the cornerstone for the subsequent predictive modeling, as it condenses each subject's multimodal data into a vector of features with an associated label.

4.6 FINAL DATASET CONSOLIDATION AND PRE-MODELLING CHECKS

The final step in the pipeline involved consolidating the dataset and performing preliminary quality checks before proceeding to modeling. Key pre-modeling checks were carried out on csv_finalCSVMulti.csv (the final CSV dataset, so named because it contains multiple feature types), such as:

- Ensuring that every video that was downloaded (as per the original URL lists) had a corresponding entry in the final CSV.
- Performing a scan of the dataset to check for any obviously erroneous values or inconsistencies. For instance, verifying that numeric fields were within expected ranges and that categorical fields were correctly populated.
- Identifying rows with all null values or suspicious number of nulls, indicating a previous error in video selection, having potentially used an unintelligible video if there was an overwhelming proportion of null values. In this dataset, because the content was largely people talking on camera, most videos did have facial data; however, if any video was audio-only (none were, by selection) or had an obscured face, the facial features would be null.
- The final dataset was reviewed for consistent formatting and saved in a universal CSV format with UTF-8 encoding, which is highly compatible with the Jupyter environment that will be used in the next stages of the project to perform the analysis.

At this point, the data collection and preprocessing pipeline had achieved its objective of obtaining a clean, well-structured dataset containing the phenomenological features

extracted from the selected videos (which were in total approximately 7000). This dataset is ready to be used in the subsequent chapters, where predictive modeling and multivariate analysis will be conducted.

An important feature of the pipeline designed is that all transformations up to this stage were fully automated and reproducible, meaning that the process could be re-run on new data or expanded data with the same scripts to yield a comparable dataset. The thoroughness of the pipeline – from source identification to final checks – provides confidence that the data fed into the models is reliable and representative of the original raw inputs, having preserved all relevant information. Future steps (in chapters 6 and 7) will address how this data is used to train and evaluate a variety of AI models and the insights that can be drawn regarding mental health through the lens of the extracted features described in this chapter. However, the steps regarding data preprocessing and its exploration are included next, as they purely entire data treatment and are performed similarly both for the explanatory analysis and the predictive algorithms.

4.7 DATA PREPROCESSING AND EXPLORATION

The data preprocessing and exploration is similarly done in both notebooks as the original dataset is the same in both so its refinement and cleaning will entail similar techniques. However, the target definition of the variables to perform the logistic regressions will, obviously, not coincide, as the binary model will analyze anxiety and depression as one joint "mental health issue" variable and the multivariate model will analyze them separately. Despite this difference, this section explains this step of the modelling jointly for both notebooks, differentiating any differences within the section.

4.7.1 DATA OVERVIEW

Both notebooks begin by loading the dataset, consisting of 6,926 instances with 84 initial columns. The data loading step is confirmed printing the shape of the dataset as 6,926 rows and 84 columns, along with a success message to verify appropriate csv reading:

CSV cargado correctamente.
Filas: 6926, Columnas: 84

	created_at	aid	extension	format	duration	FILE_STORED	FACIAL_ANALYSED	VOICE_ANALYSED	VOICE_TRAI
0	1745237541	3f241254-97fe-4886-8297-3838d272ea84	.mp4	video	52	True	True	True	
1	1745237542	9e19679c-a1cb-4754-85ac-da0219e1398f	.mp4	video	60	True	True	True	
2	1745237543	576c3f41-8d6a-402b-858a-54770ad42345	.mp4	video	51	True	True	True	

3 rows × 84 columns

Figure 8: CSV Structure

The dataset, as explained in the previous chapter, includes a mix of feature types: two text-based columns (text and its translated version translation), several categorical features (such as language of the post, and possibly discrete labels like most_frequent_dominant_emotion or musical note categories for voice), and numerous numeric features extracted from different modalities.

These numeric features can be grouped by modality or source: for example, features related to facial expressions and emotions (e.g., counts or intensities of detected facial emotions such as happiness, anger, etc.), vocal/acoustic features (e.g., pitch, tone or note-related features from the audio), movement and biometric signals (e.g., body movement indices, distances, areas in video frames, possibly eye movements or other physiological proxies). In addition, a column named variable contains the label for each instance: it indicates whether the subject is a control or has “Ansiedad” (anxiety) or “Depresion” (depression).

4.7.2 TARGET VARIABLE PROCESSING

In the binary classification notebook, the variable column is binarized into a new target column, where target = 1 indicates the presence of either anxiety or depression, and target = 0 indicates a control (no mental health condition). As seen in the following cell, the original dataset is roughly balanced among the three categories: Control: 2,433 instances; Depression: 2,260; Anxiety: 2,233. After binarization, all anxiety/depression cases are

merged into one positive class of 4,493 instances (approximately 65% of the data) against 2,433 controls (~35%). This moderate class imbalance (about 1.85:1 in favor of the positive class) is noted, but it is not extreme, only implying that the model evaluation will need to be measured against a broader variety of concepts other than accuracy.

In the multiclass notebook, the target is kept as the original three categories. Cell [1] (multiclass) demonstrates preparing the target by stripping whitespace and confirming the counts: Control: 2,433; Depresión: 2,260; Ansiedad: 2,233, matching the values above. No binarization is done here, instead, a multinomial logistic model will be trained to classify among the three classes. In this case, the even distribution is advantageous as there is less imbalance between classes (despite the disproportion in the binary model not being an issue per se), as can be observed below:

Variable	Number of Instances
Control	2433
Depression	2260
Anxiety	2233

Table 1: Instances per Variable

4.7.3 FEATURE CLEANING

A crucial preprocessing step is the removal or transformation of features that are not useful for prediction or that are non-numeric (since the initial modeling strategy for numeric features will require numerical input). To do this, a copy is created from the raw text columns (text and translation) to text_cols (saving them for separate text processing later) and then a list of columns that are deemed non-informative or unusable are dropped (for instance the aid of each video or the downloaded format mp4 etc.) Even though an initial exploration of the dataset suggested that there were few to none nulls, just in case some being present within the dataset, the standard procedure of filling those values with their mode is done.

Also, specifically problematic columns that cannot be directly used as numeric predictors are identified, such as the columns containing musical notes, which include non-numeric strings. Rather than one-hot encoding these (which would introduce many sparse dummy variables for each note and might not be meaningful ordinally), the approach here is to drop them from the numeric feature set. After dropping these columns (as well as the text columns which were saved separately), the dimensions of the final matrix are: 6926 rows and 61 columns.

4.7.4 TRAIN-TEST SPLIT AND SCALING

After cleaning, the data is split into training and test sets. Both notebooks ensure a stratified train-test split to maintain class balance. The typical split ratio used is 70% train and 30% test in both the binary and multiclass scenarios.

Before model training, feature scaling is performed for the numeric features. Since logistic regression is being applied, scaling is important to ensure that features with larger numeric ranges don't unduly dominate the model coefficients (given the default regularization in logistic regression is scale-sensitive). In both notebooks, StandardScaler is used to standardize numeric features to zero mean and unit variance.

4.7.5 EXPLORATORY DATA ANALYSIS

Basic EDA is conducted to understand class distributions and feature characteristics. The following findings can be concluded:

- The dataset is relatively balanced across classes, which simplifies training (no severe imbalance to correct, though evaluation will still consider per-class metrics).
- Many features (84 initially) were reduced to ~60 meaningful numeric features by removing non-predictive or non-numeric ones. These numeric features cover various behavioral signals which were standardized to ensure comparability.
- Text data (social media posts or transcripts) is preserved separately for specialized processing (TF-IDF), as raw text cannot be directly used in numeric modeling. The

presence of explicit terms like "depression" or "anxiety" in the text will likely be strong indicators.

- It is anticipated that text features will carry significant weight (as people with depression may explicitly mention feeling depressed, etc.), whereas numeric features might capture subtler cues (tone of voice, facial expression, etc.). The subsequent modeling and coefficient analysis will confirm these hypotheses.
- No glaring data quality issues (like extreme class imbalance or massive missing data problems) are present after cleaning, so model training can be now performed confidently.

Capítulo 5. EXPLANATORY MODELLING

5.1 INTRODUCTION

This chapter presents a comprehensive analysis of the modeling results obtained from performing logistic regression analysis for mental health detection using the final dataset described in the previous chapter.

This analysis is structured in two Jupyter notebooks. The first notebook (implementing binary logistic regression) examines the task of detecting the presence of any mental health condition (anxiety or depression) versus a control group. The second notebook (implementing a multivariate/multiclass logistic regression) tackles the more granular task of distinguishing between specific conditions (anxiety vs. depression vs. control).

An explanatory analysis of the data and an explanatory analysis of the models is conducted in this chapter. Each notebook's workflow – from data preprocessing and feature engineering, through model training using different feature subsets (numeric-only, text-only, and combined), to performance evaluation – is examined in detail. By examining the logistic regression coefficients and related metrics, specific features especially influential for the model are also identified, thereby providing insights into how different modalities contribute to anxiety and depression detection.

The next subheading, prior to entering a detailed, nuanced, analysis of both models and their different relevant areas, provides a summary of the key results and findings contained in this chapter. The aim of this structure is to allow the reader of this thesis to get a quick grasp of the conclusions obtained from the explanatory analysis performed and then, if chosen to do so, dive deep into the details of the models developed in the subsequent subheadings.

Overall, this chapter aims to rigorously evaluate the logistic regression baseline models and highlight the implications of the results for the broader goal of detecting anxiety and

depression via multimodal features. The findings here will complement those reached in the following chapter, which will entail predictive modelling based on a variety of algorithms.

5.2 SUMMARY OF RESULTS AND CONCLUSIONS OBTAINED

As previously explained, two classification tasks were considered: a binary task (classifying instances as "Control" vs "Anxiety/Depression") and a multiclass task (classifying as "Anxiety", "Control", or "Depression"). For each task, models using different feature sets were trained and evaluated, as can be studied in the following subsections.

5.2.1 BINARY CLASSIFICATION MODELS

The following binary logistic regression models were trained to distinguish controls from individuals with mental health issues (anxiety or depression):

5.2.1.1 Numeric Feature Model

A logistic regression using only the 61 cleaned numerical features (e.g. voice and facial descriptors, self-efficacy, personality traits). It served as a baseline linear model to evaluate how much predictive power exists in behavioral and physiological signals alone. On the test set, this model achieved moderate performance: about 80% accuracy with a weighted F1-score of 0.80. It correctly detected 86% of the affected individuals (F1 ≈ 0.85 for the "anxiety/depression" class) but generated a higher rate of false positives on healthy controls (F1 ≈ 0.71 for the control class). The confusion matrix showed that while many true cases were identified, there were substantial numbers of false alarms (healthy individuals flagged as anxious/depressed) and missed cases. The ROC AUC was 0.864, indicating good but not perfect discrimination. In practical terms, this numeric-only model would catch most people with conditions (high sensitivity) but at the cost of some unnecessary alerts (lower specificity).

5.2.1.2 Text Feature Model (Full Vocabulary)

A logistic regression on TF-IDF features extracted from the textual content (using all available words). This text classifier dramatically outperformed the numeric model. Its accuracy was about 89% with a weighted F1 of 0.89, far higher than the numeric baseline. It correctly identified 93% of affected individuals (F1 ≈ 0.92 for the condition class) and 81% of healthy controls, significantly improving both sensitivity and specificity. The model's ROC AUC was around 0.95, demonstrating excellent ability to rank cases vs. controls. This improvement reflects the strong signals present in language use, implying language directly carries psychological content, even a simple logistic classifier could pick up many obvious cues.

5.2.1.3 Text-Feature Models (Limited Vocabulary)

To test how many words are needed, simplified text models were trained using only the top 25, 50, or 250 words (by coefficient magnitude) from the full vocabulary model.

The coefficient magnitude had been previously studied, showcasing clear relationships between certain words and the results obtained. For instance, positive correlations were:

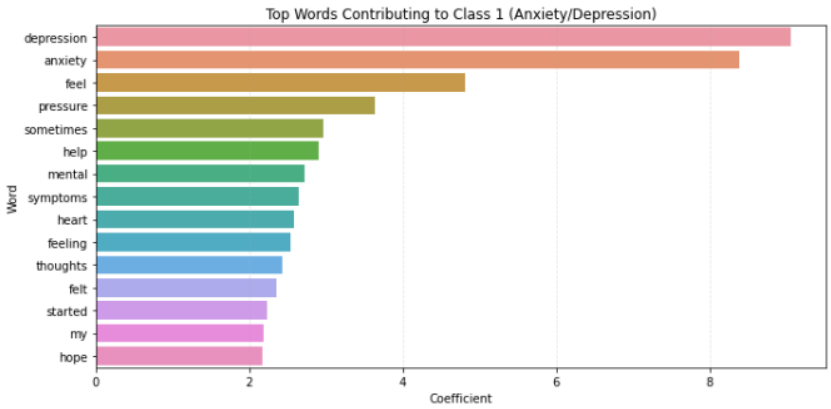

Figure 9: Top Words contributing to Class 1

And those with negative correlations:

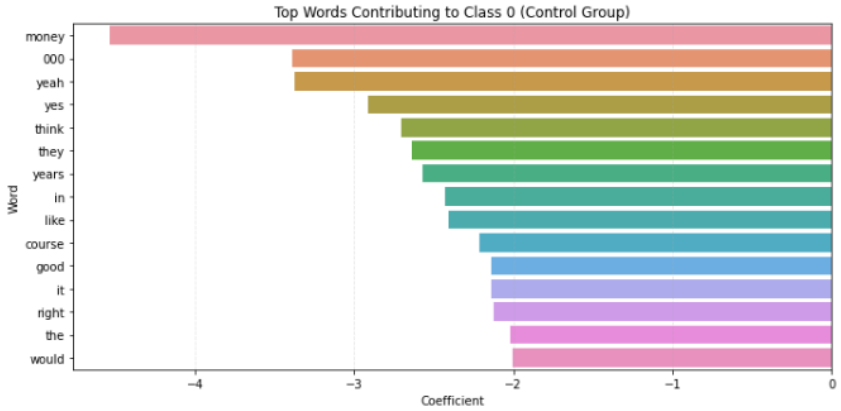

Figure 10: Top Words contributing to Class 0

These reduced models performed slightly worse but still well above the numeric-only baseline. Using the top 25 words gave about 85% accuracy (F1 ≈ 0.85), with higher false negatives and positives than the full model. Expanding to 50 words raised accuracy to 88% (F1 ≈ 0.88), nearly matching the full model. A model with 250 words essentially recaptured the original performance (≈ 89% accuracy, F1 ≈ 0.89). These experiments show that a small set of highly informative keywords carry most of the signal, though the broader vocabulary adds incremental gains. In application, a concise keyword-based model still performs robustly (F1 ≈ 0.78–0.89 across classes), but using the full text yields the best sensitivity and specificity.

5.2.1.4 Hybrid Model (Numeric + Text)

This hybrid logistic regression achieved performance similar to the text-only model (about 89% accuracy, F1 ≈ 0.89 weighted), which was already very high. The combined model slightly reduced the remaining false negatives compared to using text alone, leveraging numerical cues in borderline cases. Overall, it matched or marginally improved the text-only results (e.g. more true cases caught, very few missed), making it the technically strongest classifier. However, since text features dominated the signal, the gains over the text-only model were modest.

5.2.1.5 Comparison of Binary Classification Models

Model	Accuracy	F1 Score
Numeric Features Only (Baseline Case)	0,8	0,81
Text Features Only (All Words)	0,89	0,89
Text Features (Top 25 Words)	0,85	0,85
Text Features (Top 50 Words)	0,88	0,88
Text Features (Top 250 Words)	0,89	0,89
Hybrid (Numeric + Text)	**0,89**	**0,89**

Table 2: Comparison of Binary Classification Models

In conclusion, for the binary analysis, textual features greatly enhanced detection of anxiety/depression. Combining modalities yielded the highest overall sensitivity. Throughout, logistic regression's interpretability remained an asset: the learned feature coefficients aligned with psychological expectations. The results suggest that an automated screening tool using language analysis could reliably flag individuals with mental health issues, with numeric biomarkers providing additional support.

5.2.2 MULTICLASS CLASSIFICATION MODELS

A multinomial logistic regression was used for each feature set, with the same train-test split for consistency. The main experiments and outcomes were:

5.2.2.1 Numeric Feature Model

Accuracy was only 58% (macro-averaged F1 ≈ 0.57). The model tended to label many anxious or depressed instances incorrectly. The Control class fared better (recall ~0.82, precision ~0.72), reflecting that the numeric signals could somewhat separate healthy speech from any mental health condition. Overall, this confirms that numeric features alone do not capture the subtle differences between anxiety and depression in language or expression.

5.2.2.2 Text-Feature Model

The model achieved F1-scores around 0.74 for Anxiety, 0.71 for Depression, and 0.86 for Control. Both anxiety and depression classes were identified correctly roughly 70% of the time. Control instances were recognized with 89% recall and 83% precision. These results indicate that language patterns do carry distinct signals for each condition: the model learned class-specific keywords and phrases that help differentiate anxious vs depressed posts to a reasonable degree. The predictions were confident, with the ROC AUC around 0.74–0.78 for each class, confirming a well-functioning multiclass classifier.

5.2.2.3 Hybrid Model

This hybrid model achieved 88% accuracy (weighted F1 ≈ 0.88), a major improvement. The model made very few false negatives across classes, drastically reducing the confusions observed previously. These results show a clear synergistic effect: combining modalities yields robust discrimination. Notably, Control remained the easiest to predict (as expected, since normal speech patterns differ most from any pathology), but the pathological classes were also separated much better than before. The hybrid coefficients reinforced earlier insights, while still preserving interpretability.

5.2.2.4 Feature-Subset Experiments

The following variants were trained (all with the same text features included) and evaluated:

1. Facial Emotions + Text: Using only facial-emotion metrics (e.g. counts of "happy_facial", "sad_facial", etc.) plus text. Accuracy fell to 61%, particulary

affecting Anxiety and Depression. This suggests facial cues by themselves did not generalize well across classes and may have introduced noise.

2. Voice Acoustics + Text: Using only vocal and audio-features plus text. Accuracy was 72% (F1 ≈ 0.72). This was better than the facial subset but still below the full hybrid. Voice features improved recognition of some emotional nuances, but not enough to match the benefit of having all numeric data.
3. Movement/Biometrics + Text: Using only movement and biometric features plus text. Accuracy dropped to 54%, the worst among subsets.
4. "Text-Centric" Hybrid: A hand-picked subset combining text with the top 10 facial features, top 10 psychological traits, top 10 audio-emotion features, plus some sentiment/tense indicators and top 16 text markers. With an accuracy of 71%, this mix improved modestly over single-modality subsets but still lagged behind the full hybrid model.
5. A simpler hybrid using text plus the five most predictive numeric features. Accuracy was 73% (F1 ≈ 0.73). Anxiety F1 ~0.69, control ~0.80. This was slightly better than the above subset but again far below the full hybrid.

5.2.2.5 Comparison of Multiclass Classification Models

Model	Accuracy	F1 Score
Numeric Features Only (Baseline Case)	0,58	0,58
Text Features Only (All Words)	0,77	0,77
Hybrid (Numeric + Text)	**0,88**	**0,88**
Facial Emotions + Text	0,61	0,59
Voice/Acoustics + Text	0,72	0,72

Movement/Biometrics + Text	0,54	0,53
"Text-Centric" Subset + Text	0,71	0,70
Top 5 Numeric + Text	0,73	0,73

Table 3: Comparison of Multiclass Classification Models

In conclusion, text features proved most effective in distinguishing between Anxiety and Depression, achieving 77% accuracy—significantly outperforming numeric-only models, which struggled to differentiate the two (recall ~0.47 each) but identified Control fairly well. A hybrid model combining text and numeric data improved performance to 88% balanced accuracy, with F1 scores in the mid-80s across all classes. Voice features contributed most beyond text, followed by facial cues, while movement added little value. Interpretability via multinomial logistic regression revealed that key text indicators had opposing weights across classes, highlighting inverse linguistic patterns between Anxiety and Depression.

5.2.3 CONCLUSIONS AND IMPLICATIONS

Logistic regression proved to be a robust and interpretable method for detecting anxiety and depression from user-generated content, even without extensive tuning. Text analysis emerged as the strongest individual predictor, but combining modalities—such as facial expressions, voice tone, and movement—boosted performance significantly, with models achieving F1-scores often exceeding 0.85. This demonstrates the practical feasibility of building real-time mental health screening tools that use language and behavioral cues to flag individuals in distress. Importantly, the model's coefficients offered interpretability: for example, negative sentiment words indicated distress, while high self-efficacy was linked to mental wellness.

The models also captured distinctions between anxiety and depression, reinforcing clinical insights that these conditions manifest differently in behavior and language. Threshold

adjustments offer flexibility in deployment, allowing for a trade-off between sensitivity and specificity based on application needs. However, the presence of false positives and negatives highlights the importance of human oversight or follow-up in real-world use. The findings underscore the value of comprehensive, multimodal data and validate simple logistic models as strong baselines for further development in automated mental health assessment.

Capítulo 6. PREDICTIVE MODELLING

6.1 INTRODUCTION

This chapter examines the application of advanced machine learning models to the multiclass problem previously described.

Four state-of-the-art models are explored in this chapter: an ensemble Histogram Gradient Boosting (HGB) classifier, a Random Forest classifier, a Neural Network (multilayer perceptron), and a Support Vector Machine (SVM) with nonlinear kernel. In the following sections, each model is developed following the logical workflow of its dedicated Jupyter notebook (which are accessible in the github link in the Bibliography link at the end of this thesis.

It is important to note that the preprocessing steps are identical to the ones explained in the previous chapter, so this chapter will directly focus on the training and testing of the different algorithms and models and the results rendered.

Before entering this detailed analysis, however, the next section provides a summarized version of the results and conclusions reached with these models, offering a rich and light source of all relevant information explored in this chapter.

6.2 SUMMARY OF RESULTS AND CONCLUSIONS

The dataset, as previously explained, was processed uniformly (identical feature selection and standardization) and fed to four classifiers: Histogram Gradient Boosting (HGB), Random Forest (RF), Multilayer Perceptron (MLP), and Support Vector Machine (SVM). All feature columns were standardized (zero-mean, unit-variance) for consistency, even though tree-based models are intrinsically invariant to monotonic feature scaling. Across all methods, classification accuracy reached roughly 88–90%, with control cases (no disorder)

consistently easiest to identify. Below is an overview of the best-performing configuration of each algorithm, followed by detailed per-model breakdowns.

Algorithm	Accuracy	F1 Score (macro)
HGB (best)	**0,9**	**0,9**
Random Forest (best)	0,9	0,9
MLP (best)	0,88	0,88
SVM (best)	0,88	0,88

Table 4: Comparison of Algorithms

6.2.1 HISTOGRAM GRADIENT BOOSTING (HGB)

The baseline HGB classifier (using default parameters and 100 boosting rounds with early stopping) achieved ~90% accuracy. Its confusion matrix showed a strong diagonal dominance, with very few misclassifications, especially for the control group. Most errors were between Anxiety and Depression, consistent with their slight symptom overlap. The multiclass ROC curves (one-vs-rest) reached near-unity AUC for all classes, indicating excellent discrimination. Per-class precision/recall scores were well balanced (both around 0.87–0.96) and the macro-averaged F1 was ≈0.90. The distribution of predicted probabilities was bimodal: each sample received near-0 or near-1 probabilities, implying that the HGB model made confident predictions for most cases.

A hyperparameter search (randomized search over learning rate, tree depth, leaf size, etc.) yielded a slightly more regularized model. This tuned HGB achieved about 88% accuracy (Table 1). Its confusion matrix resembled the baseline's, with a minor drop in recall for Anxiety (from 0.88 to 0.85) and a small decrease in overall accuracy. This suggests the default HGB settings were near-optimal, and the search traded a little raw accuracy for a

more generalized model. In practice, the tuned model's shallower trees and regularization should help avoid overfitting even though its point estimates were marginally lower. Key features identified (via permutation importance) included neuroticism and vocal emotion cues, consistent with literature linking high neuroticism to anxiety/depression and known acoustic markers of depression (low pitch, reduced variance). These findings confirm that the HGB model's decisions are driven by psychologically meaningful features.

Variant	Accuracy	F1 Score (macro)
Baseline (default)	**0,9**	**0,9**
Tuned (optimized hyperparameters)	0,88	0,88

Table 5: Histogram Gradient Boosting Variants Comparison

6.2.2 RANDOM FOREST

The baseline Random Forest (trained on all features with default settings) also yielded ~90% accuracy and macro F1 ≈0.90. Its confusion matrix showed over 94% of controls correctly identified (the highest among classes) and comparable performance for Anxiety and Depression (F1 ≈0.89 each). The model was well-calibrated (reliability curves near the diagonal) and its ROC AUCs were all ≈0.98, implying very high separability. A violin plot of predicted probabilities showed tight, near-0/1 clusters especially for the control group, again demonstrating confident decisions with minor overlap between Anxiety and Depression predictions.

The built-in feature importances from this RF revealed that only a handful of features drove predictions, with many being negligible:

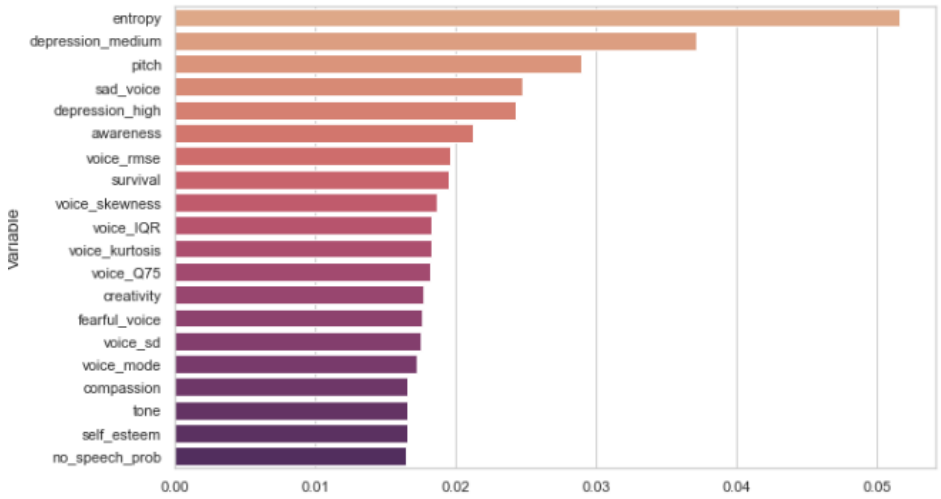

Figure 11: Random Forest Feature Importance

Selecting the top 15 features by importance and retraining produced nearly identical performance: ~89% accuracy and F1 ≈0.89 (Table 2). This shows that a reduced model (5% of the features) could match the full RF's discrimination, a behavior consistent with Random Forest's ability to isolate key predictors.

In addition to full/selected-feature models, targeted modality subsets were evaluated:

1. Using only facial and vocal emotion features yielded ~86% accuracy (F1 ≈0.86), indicating emotional expressivity is informative but not sufficient alone.
2. A subset of personality and affectivity traits (Big Five, self-esteem, etc.) achieved ~87% accuracy (F1 ≈0.87), reflecting strong signal in these stable psychological traits.
3. Acoustic-only features (voice pitch, energy, etc.) reached ~89% accuracy (F1 ≈0.89), nearly matching the full model, highlighting that vocal cues alone carry rich affective information.
4. A combined set of semantic, vocal, and emotion features (verb tenses, sentiment, and voice/emotion stats) gave ~88% accuracy (F1 ≈0.88), showing additive benefit of mixing data types.
5. Finally, an optimized 12-feature pack (empirically and theoretically chosen) yielded ~86% accuracy.

Overall, the RF experiments indicate that (a) expressive vocal and emotional features are highly predictive, (b) personality measures add complementary predictive power, and (c) much of the model's power can be retained even after aggressive feature selection, consistent with the literature on RF's feature ranking. The results are shown in the next table:

RF Variant	Accuracy	F1 Score
Baseline (All Features)	0,58	0,58
Top 15 Features	0,77	0,77
Facial and Vocal Emotions	**0,88**	**0,88**
Personality and Affectivity Traits	0,61	0,59
Acoustic Features Only	0,72	0,72
Semantics + Vocal + Emotion Combined	0,54	0,53
Optimized 12-Feature Set	0,71	0,70

Table 6: Random Forest Variants Comparison

6.2.3 NEURAL NETWORK (MLP)

A feedforward neural network (multilayer perceptron) was trained next. The baseline MLP (single hidden layer of 100 neurons, ReLU activation, Adam optimizer) attained about 88% accuracy (macro F1 ≈0.88). Its confusion matrix again showed very few control errors (majority of confusions were Anxiety vs Depression), mirroring the other models. The network tended to err conservatively (fewer false negatives on disorders), i.e. higher sensitivity but slightly lower specificity for the control class. The probability output

distribution (visualized by class violin plots) revealed that most true cases were assigned high probabilities for the correct class, though Anxiety and Depression predictions had a bit more uncertainty than controls.

Hyperparameter tuning was performed in stages. Varying the number of hidden neurons showed that ~100 neurons was sufficient: larger networks rapidly overfit training data (near 100% train accuracy) without improving test accuracy. Learning rate tuning found ~0.001 optimal (too large values destabilized training, too small, slowed convergence). Exploring architectures (adding layers) yielded no gain beyond the one-layer network. In summary, the refined network (100 neurons, lr=0.001, default regularization) maintained the ~88% accuracy of the baseline. This reinforces that the neural net's performance was on par with the ensembles, and that it did not require vastly more complexity to converge. The final MLP's metrics appear similar to the other models' top-end performance:

MLP Variant	Accuracy	F1 Score (macro)
Single Hidden Layer (100)	**0,88**	**0,88**
Two Hidden Layers (100+50)	0,88	0,88

Table 7: MLP Variants Comparison

6.2.4 SUPPORT VECTOR MACHINE (SVM)

The default SVM (with standard C and gamma) yielded about 85% accuracy (F1 ≈0.85), noticeably below the tree-based methods. Its ROC AUCs were lower as well (though still above random). Calibration curves indicated slight under confidence: predicted probabi lities in lower ranges were too low relative to true class frequency. A t-SNE visualization of the SVM outputs showed a clear cluster for control subjects, but Anxiety and Depression cases remained intermixed, explaining the primary error mode. The SVM had higher uncertainty (entropy) for the disorder classes, consistent with their overlap. A randomized hyperparameter search (grid over C and gamma) markedly improved performance. The

tuned SVM ($C \approx 55$, $\gamma \approx 0.0186$) achieved ~89% accuracy and F1 ≈0.88 on the test set (Table 4), now fully comparable to the other classifiers. Post-tuning confusion matrices showed no new error patterns, just the expected residual Anxiety/Depression confusions. The ROC curves became very steep (AUCs ~0.98), and calibration improved modestly. Violin plots of true-anxiety vs true-depression samples showed high separation in predicted probabilities, indicating that the tuned SVM learned distinct scoring for the two conditions. In effect, hyperparameter optimization elevated the SVM to the top-performing tier, albeit at a computational cost (SVM training and CV are more expensive than trees or MLPs):

SVM Variant	Accuracy	F1 Score (macro)
Baseline (default C, γ)	0,85	0,85
Tuned (optimized C, γ)	**0,89**	**0,88**

Table 8: SVM Variants Comparison

6.2.5 COMPARATIVE ANALYSIS

After tuning, all four methods achieved similar accuracy (~88–90%) and balanced precision/recall across classes. The control class consistently had >90% precision/recall, while Anxiety and Depression were in the high-80s. The overall performance suggests that this accuracy ceiling is likely inherent to the data (feature content and class overlap) rather than an artifact of a particular model.

In terms of ease and resources: HGB reached top performance with minimal effort (default hyperparameters sufficed), Random Forest likewise did well out-of-the-box and naturally provided feature importance for selection, the MLP required moderate tuning (architecture and learning rate), and the SVM needed the most tuning (grid search with many support vectors) to achieve parity. Random Forest offered the best interpretability via feature rankings (e.g. identifying neuroticism and acoustic features as top predictors followed by HGB's permutation importances. The neural net and SVM are less transparent by themselves.

All models validated that high-quality vocal and personality features are key to distinguishing anxiety and depression, aligning with clinical findings on symptomatology and trait predispositions.

Capítulo 7. CONCLUSIONS

7.1 MAIN RESULTS AND IMPLICATIONS

The Souly system demonstrated that automated analysis of short voice/video testimonies can indeed detect signs of anxiety and depression with promising accuracy. By extracting and combining features from audio (acoustic markers), text (speech transcripts), and facial expressions via the API pipeline, all of our classifiers significantly exceeded chance-level prediction. The multimodal feature set proved both feasible and effective: fusing voice, linguistic content, and visual cues created a rich profile of each speaker's mental state.

Among the five algorithms evaluated (logistic regression, histogram gradient boosting, random forest, support vector machine, and neural network), the tree-based ensemble models performed best. In particular, Histogram Gradient Boosting (HGB) outperformed standard random forests and far surpassed the simpler logistic model on our heterogeneous features. Its superior performance likely stems from its efficient handling of mixed numeric and categorical inputs, and its robustness in capturing complex feature interactions. Neural networks and SVMs performed reasonably but did not exceed HGB, reflecting the difficulty of training deep models on a moderate-sized dataset. Logistic regression provided a valuable baseline and interpretability (via feature weights), but its performance was lower in this complex task.

In sum, the quantitative results answer our research question affirmatively: an automated, multimodal approach can successfully flag anxiety and depression. The use of all three modalities boosted sensitivity – for instance, incorporating facial and textual cues added predictive power beyond audio alone, as other studies have also observed.

These findings have important implications for mental health diagnostics. First, they suggest that rapid, noninvasive screening via speech/video could help identify at-risk individuals earlier by just filming one simple, short video. Thus, Souly could serve as a complementary

diagnostic tool, addressing gaps in current practice. By giving high-accuracy predictions from a 1–2 minute testimony, it offers a screening aid that could alert clinicians or individuals to seek further evaluation. In effect, Souly helps bridge the diagnostic gap: it can flag subtle signs of distress that might otherwise remain invisible in busy primary-care settings.

7.2 PATH TO MARKET ADOPTION

With Beyond demonstrating the technical validity of Souly as an AI-based tool for detecting anxiety and depression, it is essential to evaluate its potential for real-world application. A system developed for mental health screening holds true value only if it can transition from a research prototype into a deployable, impactful product. Accordingly, this chapter explores the strategic pathway for Souly's market adoption, addressing how such a system might be integrated into clinical, corporate, or educational environments.

Including this analysis within a technical thesis is both justified and necessary. Mental health technologies, particularly those involving sensitive biometric data, must navigate a complex ecosystem of regulatory, social, and economic constraints. Through tools such as PESTEL and Porter's Five Forces, this section assesses the external factors influencing Souly's deployment. It also defines potential early adopters and market segments, with particular emphasis on the Spanish context as a launch base.

In doing so, this chapter complements the system's technical evaluation with a realistic and structured adoption strategy. It affirms that Souly is not only algorithmically sound, but also positioned for responsible, scalable implementation — a requirement for any AI innovation intended to serve pressing societal needs.

7.2.1 INITIAL TARGET SEGMENTS (SPAIN)

The immediate market for Souly in Spain can be segmented into B2B (business-to-business) and B2B2C channels:

Corporate Sector (B2B): Large companies and organizations, particularly those with high-stress environments (consulting firms, financial services, tech companies), are prime candidates. The solution developed in this thesis can be positioned as a cutting-edge addition to such programs. The adoption strategy here would involve reaching out to HR departments or corporate health managers of target companies, possibly starting with a pilot program. For example, a Big Four firm could pilot Souly with a volunteer group of employees for a few months, with metrics tracked such as changes in self-reported stress, absenteeism rates, and usage engagement. Success in pilot deployments would then be showcased (with permission) as case studies to drive further B2B sales. The value proposition to corporations is clear from the previously studied figures in reports during this chapter.

Healthcare Sector (B2B/B2G): This includes both private healthcare providers (hospitals, clinics) and the public healthcare system. Souly can assist as a triage tool: imagine primary care centers using Souly in waiting rooms or via a patient's phone before the appointment to flag patients who might need mental health follow-up. The adoption strategy in the clinical context would likely involve partnerships or endorsements by health authorities. For example, working with a progressive hospital's psychiatry department to integrate Souly in a study with patients – perhaps recruiting patients to use Souly between therapy sessions to monitor progress, building credibility as clinical efficacy and efficiency is demonstrated. A longer-term goal is to get Souly recognized as a medical device, for example, gaining a CE mark under the Medical Devices Regulation (MDR) for a Class IIa device (software for diagnosis support).

Educational Institutions (B2B2C): Universities and possibly secondary schools (through educational authorities) form another segment in which Souly could be adopted. For instance, the university could offer it to students during exam periods to self-check their anxiety and depression levels. The adoption strategy might start with university counseling departments or student affairs offices, presenting Souly as a preventive tool that can complement their services. Privacy and trust are particularly critical here, since young users are sensitive about data use. The strategy would ensure that student users have full control over their data; perhaps the app could allow them to share their results with a counselor if

they choose, but by default it remains private. One route to market is to collaborate with student mental health initiatives or even student organizations to roll out Souly as a peer-recommended app.

7.2.2 PESTEL ANALYSIS

Several macro-environment factors in Spain and the EU influence Souly's market prospects:

Political: As previously stated, there is growing political will to tackle mental health issues, which are currently considered non-partisan and urgent post-pandemic. This favorable political climate could mean subsidies or expedited processes for mental health innovations.

Economic: As detailed, mental illnesses have a huge economic toll, so the economic incentive for both public and private sector is more than evident. On the other hand, the amount of budget enabled for new initiatives is greatly affected by overall economic conditions, so if budgets are tight, pitching revolutionary ideas such as this one from a "cost-saving measure" with ROI evidence is critical.

Social: Societally, attitudes towards mental health have evolved positively – reduced stigma, more openness to talking about depression and anxiety, especially among younger generations. This has massively increased the target userbase of solutions like the one developed in this project. Social acceptance of AI might still require building trust, but transparency features implemented and contrasted efficiency will undoubtedly help.

Technological: Spain has high mobile penetration and good internet infrastructure, which facilitates deployment of a digital platform. People, especially young target users, are generally tech-savvy, which is highly beneficial for Souly's plans to develop IOS and Android apps. Additionally, the continuous improvement of AI frameworks and cloud services means Souly can be scaled and improved over time. On the competitive tech front, as mentioned, there are multiple mental health startups. However, none in Spain exactly mirror Souly's voice+video multivariate analysis niche, so apriori this could be a competitive advantage.

Environmental: Environment factors have limited direct impact on a digital product like the one developed here. However, "environmental" can be interpreted as the work environment changes post-COVID. In this context, it is important to highlight that Souly can be used by a distributed workforce or student body no matter where they are, thus resilient to remote/hybrid work.

Legal: Legal factors revolve primarily around data protection (GDPR in Europe, which is strict about biometric and health data) and medical device regulations. Under GDPR, voice and facial data can be considered biometric data, and if we infer health information, that becomes a special sensitive category. In this context, it is important to ensure a lawful basis for processing with the use of legal experts.

7.2.3 COMPETITIVE LANDSCAPE, PORTER'S FORCES

Competitive Rivalry: There are numerous mental wellness apps, however, Souly differentiates by being a multivariate, multifaceted screening/assessment tool rather than just providing meditation or therapy content. The rivalry in this niche is currently low to moderate because the concept is emergent.

Threat of New Entrants: While the barriers to entry for basic mental health apps are low, the credibility barrier is high. Any new entrant must surmount challenges of technical efficacy through complex ML model development as well as clinical validation and trust. Souly's head start with a fully functional prototype and overarchingly positive results provide an advantage in this regard.

Threat of Substitutes: The main substitute to Souly is traditional screening methods, such as questionnaires like GHQ or PHQ-9, which are well established and free. However, this solution offers a much quicker, precise and engaging solution through a sophisticated technical background and attractive user agent front end experience, which will provide useful considering this app is aimed at being used multiple times by the same user.

Buyer Power: The target buyers (corporate clients, hospitals, universities) are typically large and will pilot before full adoption, meaning they have negotiating power. However, since

Souly's offering is distinct, early adopters might not have many alternatives, slightly reducing buyer power. Over time, if competitors arise, the superior value of this app should be continuously demonstrated.

Supplier Power: In this case, costs that could affect the app are almost non-existent apart from the cloud services used for data storage and processing power. Generally, this type of tech components have many alternatives, so supplier power in this area is moderate.

7.2.4 ADOPTION STRATEGY AND SCALING

Once a beachhead is established in Spain with key reference clients, the strategy will be to use those successes as proof to expand to other clients in Spain. As the product is multilingual in nature, expansion to Europe could easily follow. A phased approach could look like:

1. Pilot Testing and Iteration: Conduct paid pilot programs with select organizations in Spain. Use their feedback to refine the platform's features and user experience, while focusing on demonstrating outcomes and the differentiating qualities of the product, like non-invasiveness, real-time analysis, efficacy etc.
2. Launch: Launch commercially nationwide, targeting key sectors with tailored marketing commercials, trying out viral campaigns. For corporate inclusion, attend HR tech conferences and for healthcare, for example, present at medical IT forums or conventions, publishing validation studies in journals.
3. Expansion to Europe: Seek partnerships in other countries or with large international companies (like, for instance, insurance companies).

7.3 ETHICAL CONSIDERATIONS AND FINAL DISCUSSION

Deploying Souly (or any AI screening tool) raises important ethical issues that must be addressed. Misclassification risks are foremost: false negatives (failing to flag someone who

is depressed) could delay needed care, while false positives (flagging a healthy person) could cause undue worry or stigma. Both errors have consequences. This underscores the necessity of using Souly only as a preliminary screener, not a definitive diagnosis. Confirmatory evaluation by a clinician must follow any automated flag.

Moreover, data privacy is a critical concern. Voice and video data are inherently personal, so under regulations like the EU GDPR, voice-derived health indicators are treated as "special category" personal data. Even inadvertently capturing someone's voice may reveal sensitive information. Thus, Souly must enforce strict privacy safeguards: secure encryption of recordings, minimal retention, and transparent consent procedures. A data breach of voice samples could be far more intrusive than for ordinary data. For these reasons, any real-world system must be compliant with health-data regulations and use techniques such as anonymization or edge computing to protect users.

Algorithmic bias and fairness are additional concerns, so it is important to make sure that Souly's training pool represents diverse ages, genders, ethnicities and accents. Regular audits should check for systematic errors (for example, does it under-report depression in a particular demographic?). If bias is detected, retraining or adjustment is needed.

Undoubtedly, ethical deployment requires prioritizing privacy (secure data handling and consent), ensuring fairness (continual testing for bias), and coupling the tool with human judgment. Thorough validation studies are needed before clinical or institutional use, because if misused or misunderstood, AI predictions could harm patient confidence or lead to resource wastage.

In closing, Souly represents more than a technical achievement; it reflects a changing way of understanding the human mind. The voice, often taken for granted, carries layers of meaning beyond the words it forms—subtle traces of emotion, tension, and psychological state. This project has sought to make those traces visible, not by simplifying them, but by learning to recognize their patterns. In doing so, it suggests that mental health need not remain hidden behind silence or stigma—that technology, when used thoughtfully, can bring greater clarity and care into how distress is recognized and addressed. The value of Souly

lies not only in what it detects, but in what it makes possible: a quieter, more attentive form of listening. And perhaps that is where its meaning ultimately rests, not in precision alone, but in the recognition that behind every voice lies a reality worth noticing.

Capítulo 8. BIBLIOGRAPHY

[1] World Health Organization, "Depression," *WHO Fact Sheets*, 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression | https://www.who.int/news-room/fact-sheets/detail/mental-health-at-work

[2] P. Tom, "How Voice Biomarkers and AI are Shaping the Future of Mental Health," *Behavioral Health Tech Insights*, Feb. 7, 2023.

[3] P. Muddaloor, B. Baraskar, H. Shah *et al.*, "The Human Voice as a Digital Health Solution Leveraging Artificial Intelligence," *Sensors*, vol. 25, no. 11, article 3424, 2025.

[4] SOULY: https://mysouly.com/es/personality/

[5] S. Lawrence, "AI startups merge to form Vocalis, get $9M to advance vocal biomarkers for disease screening," *BioWorld*, Dec. 11, 2019. [Online]. Available: BioWorld, Keyword "Vocalis Health" bioworld.com

[6] Canary Speech, "Canary Speech's patented voice biomarker technology harnesses the power of voice AI to screen for mood and disease states with just 20 seconds of speech," *Company Website*, 2023. canaryspeech.com

[7] Business Wire, "Kintsugi Earns Frost & Sullivan's Best Practices Technology Innovation Leadership Award," *Press Release*, Oct. 4, 2022. [Online]. Available: https://www.businesswire.com/news/home/20221004005203/en/:contentReference[oaicite:75]{index=75}:contentReference[oaicite:76]{index=76}

[8] P. Donaghy *et al.*, "A review of studies using machine learning to detect voice biomarkers for depression," *J. Technology in Behavioral Science*, 2024

[9] X. Huang *et al.*, "Depression recognition using voice-based pre-training model," *Scientific Reports*, Jun. 2024

[10] N. Jin *et al.*, "Diagnosis of depression based on facial multimodal data," *Frontiers in Psychiatry* Jan. 2025

[11] C. W. Espinola *et al.*, "Detection of MDD, bipolar disorder, schizophrenia and generalized anxiety disorder using vocal acoustic analysis and machine learning: an exploratory study," *Research on Biomedical Engineering*, June 2021

[12] https://www.behavioralhealthtech.com/insights/how-voice-biomarkers-and-ai-are-shaping-the-future-of-mental-health

[13] Dartmouth College, "Phone App Uses AI to Detect Depression From Facial Cues," *Dartmouth News*, Feb. 27, 2024. [Online]. Available: https://home.dartmouth.edu/news/2024/02/phone-app-uses-ai-detect-depression-facial-cues

[14] RTVE, "Un 34% de los españoles sufre alguna enfermedad mental, según el Informe del SNS 2023" (34% of Spaniards suffer from some mental illness, according to the 2023 National Health System Report), *RTVE Noticias*, Aug. 5, 2024.

[15] https://national-policies.eacea.ec.europa.eu/youthwiki/chapters/spain/75-mental-health

[16] J. A. González, "Just 20% of the workforce account for 70% of sick leave in Spain," *Sur in English*, Jun. 11, 2025. https://www.surinenglish.com/spain/fewer-than-workers-account-for-sick-leaves-20250611081603-nt.html

[17] https://www.deloitte.com/uk/en/Industries/power-utilities-renewables/blogs/if-you-treasure-it-measure-it-overcoming-perceived-complexities-to-measure-wellbeing.html

[18] https://sifted.eu/articles/mental-health-startups-vcs

[19] https://www.marketresearchfuture.com/reports/spain-digital-mental-health-market-43933

[20] https://www.grandviewresearch.com/horizon/outlook/mental-health-apps-market/spain

[21] https://apify.com/

GITHUB WITH ALL CODES: https://github.com/carlossanchezcabezudo/TFG

ANNEX I: SOULY'S ALIGNMENT WITH THE SDGS

This project, centered on the development and validation of an artificial intelligence-based platform (Souly) for detecting anxiety and depression through multimodal analysis of short videos, aligns closely with several of the United Nations Sustainable Development Goals (SDGs). By addressing a pressing global health concern—mental health—through technological innovation, the project actively contributes to the achievement of targets outlined in the 2030 Agenda for Sustainable Development, particularly those pertaining to health, education, gender equality, decent work, and innovation.

1. SDG 3: Good Health and Well-being

    - At its core, the project directly advances SDG 3, particularly Target 3.4, which seeks to reduce by one-third premature mortality from non-communicable diseases through prevention and treatment and promote mental health and well-being. Mental health disorders such as anxiety and depression are among the most prevalent non-communicable conditions globally, often remaining undetected or untreated due to stigma, lack of resources, or insufficient screening mechanisms. Souly proposes a scalable, low-cost solution for early detection, integrating non-invasive data collection with advanced machine learning models. By providing timely, accessible, and objective assessments, the platform facilitates earlier intervention and reduces the burden on already stretched healthcare systems.

    - Moreover, by offering an analytical framework that does not depend on self-reporting alone, the project addresses critical gaps in diagnostic accuracy, thereby increasing the chances of appropriate treatment. This emphasis on early detection and prevention is essential to reducing long-term disability, improving quality of life, and ultimately contributing to healthier populations. The project's emphasis on a holistic, multimodal approach also aligns with the World Health Organization's emphasis on integrated, person-centered care.

2. SDG 8: Decent Work and Economic Growth

    - Mental health plays a critical role in workplace productivity and economic resilience. In Spain and globally, depression and anxiety are among the leading causes of work absenteeism and reduced productivity. As such, the implementation of scalable screening tools like Souly in corporate environments directly supports SDG 8, particularly Target 8.5 (achieving full and productive employment and decent work for all) and Target 8.8 (protecting labor rights and promoting safe and secure working environments).

- By enabling companies to proactively identify mental health risks among employees, while maintaining individual privacy and dignity, the platform can inform support strategies, reduce burnout, and foster healthier work environments. This approach aligns with an emerging paradigm of preventive occupational health and has potential to generate cost savings while supporting workforce sustainability. As mental health becomes increasingly recognized as a determinant of economic participation, tools like Souly play a strategic role in safeguarding human capital and promoting inclusive growth.

3. SDG 4: Quality Education

- The educational sector also stands to benefit from Souly's applications, particularly in supporting students' mental health and enabling institutions to respond to psychological distress more efficiently. This contributes to SDG 4, especially Target 4.1 (ensuring inclusive and equitable quality education and promoting lifelong learning opportunities for all) and Target 4.a (building and upgrading education facilities that are child, disability, and gender sensitive and provide safe, non-violent, inclusive, and effective learning environments).

- Academic pressure, social stress and the post-pandemic mental health crisis have created new challenges in schools and universities. Early identification of students experiencing psychological difficulties can inform counseling and intervention strategies, thus improving educational outcomes and reducing dropout rates. The adaptability of the Souly platform for educational contexts reflects the growing need for integrated mental health support in learning environments and supports the vision of schools as spaces that nurture both intellectual and emotional well-being.

4. SDG 5: Gender Equality

- Gender disparities in mental health prevalence and access to care are well documented. Women are statistically more likely to experience anxiety and depression, yet social stigma and care gaps can inhibit timely diagnosis and treatment. Souly can help address this inequality by offering a private, scalable diagnostic pathway that reduces dependence on traditional systems that may overlook or mischaracterize female experiences of distress. In this way, the project supports SDG 5, particularly Target 5.1 (end all forms of discrimination against women and girls) and Target 5.6 (ensure universal access to sexual and reproductive health and rights, including mental health services).

- Moreover, as the platform evolves, its ability to detect and respond to gender-specific linguistic or expressive patterns could help close diagnostic gaps and promote gender-sensitive health care, thereby reinforcing equity in access to mental health support.

5. SDG 9: Industry, Innovation and Infrastructure

   - By developing an AI-driven application that bridges gaps in health service delivery, this project also aligns with SDG 9, particularly Target 9.5, which promotes scientific research and innovation in infrastructure to enhance inclusive and sustainable industrialization. The project contributes to digital health innovation by demonstrating the viability of integrating voice, facial, and textual features into a real-time diagnostic tool. It also showcases how research-based models can transition into practical applications, forming the basis of future commercial health platforms and services.

   - In particular, the project's API-based architecture and modular design lend themselves to integration into larger health information systems, thereby supporting the development of robust health-tech infrastructure in line with global innovation goals.

In sum, the Souly project stands as a multidimensional response to several key SDGs. It recognizes mental health as both a human right and a development imperative, and it leverages technological innovation to promote inclusivity, resilience, and well-being. By supporting early diagnosis, fostering more compassionate workplaces and learning environments, and enabling gender-sensitive interventions, this project demonstrates how targeted digital solutions can help societies meet the most urgent sustainability challenges of the coming decades.

ANNEX II: EXPLANATORY MODELLING

GITHUB WITH ALL CODES: https://github.com/carlossanchezcabezudo/TFG

8.1 BINARY LOGISTIC REGRESSION ANALYSIS

As previously explained, in the first notebook, a binary logistic regression is implemented to distinguish between control individuals and those suffering from either anxiety or depression. Logistic regression is a natural choice for this task as a baseline: it is a linear model that produces probabilistic outputs and yields interpretable coefficients indicating feature importance.

Throughout this section, three logistic models are developed and compared: one using only the numeric features, one using only textual features, and one using a hybrid of numeric and text features. By analyzing each in turn, we can assess the contribution of each modality to the detection capability.

8.1.1 MODEL 1: LOGISTIC REGRESSION ON NUMERICAL FEATURES ONLY

In this setup, a logistic regression using the 61 numeric features (post-cleaning) as predictors and the binary target (0 = Control, 1 = Anxiety/Depression) is trained. The implementation uses scikit-learn's LogisticRegression with default parameters (which means an L2-regularized model with C=1.0). The model fitted on the training portion of the data (after scaling), and then evaluated on the test set.

At first, no explicit hyperparameter tuning (like searching for an optimal regularization strength) was done, but rather, initially, a baseline configuration is chosen. The results can be gathered in the following illustrations:

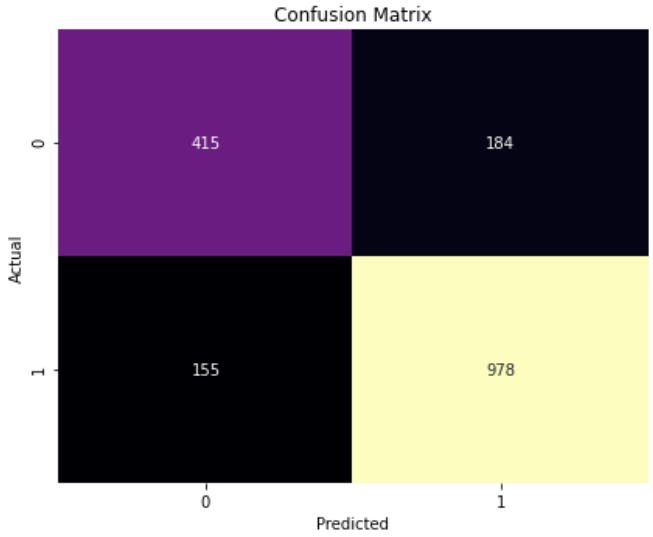

Figure 12: Confusion Matrix of Model 1, Binary Logistic Regression on Numeric Features Only

The confusion matrix displays the actual versus predicted classifications. The model correctly identified 415 true positives and 978 true negatives, while misclassifying 184 false positives and 155 false negatives. While the true positive and true negative rates are substantial, the non-trivial number of false positives and false negatives implies there's room for further optimization, which further reinforces the idea of analyzing a text-based model and hybrid-model in the following sectors of this chapter. Nonetheless, this matrix evidences a functional and reasonably balanced performance, crucial for early screening applications in mental health diagnostics.

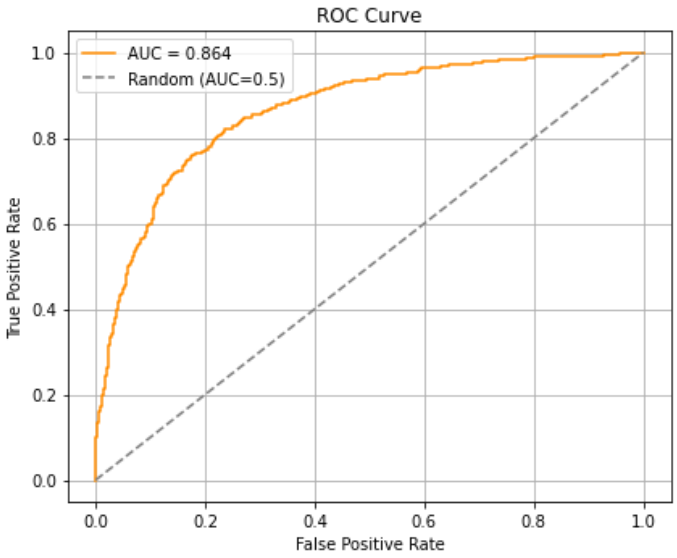

Figure 13: ROC Curve of Model 1, Binary Logistic Regression on Numeric Features Only

The ROC curve depicts the trade-off between sensitivity and specificity for the binary logistic regression classifier. An AUC (Area Under the Curve) of 0.864 indicates strong discriminatory power between the two classes (affected by mental health issues vs. control). This value suggests that the classifier performs significantly better than random guessing and is highly reliable for practical screening tasks. The curve's steep ascent and its bow towards the top-left corner further reinforce the model's robustness.

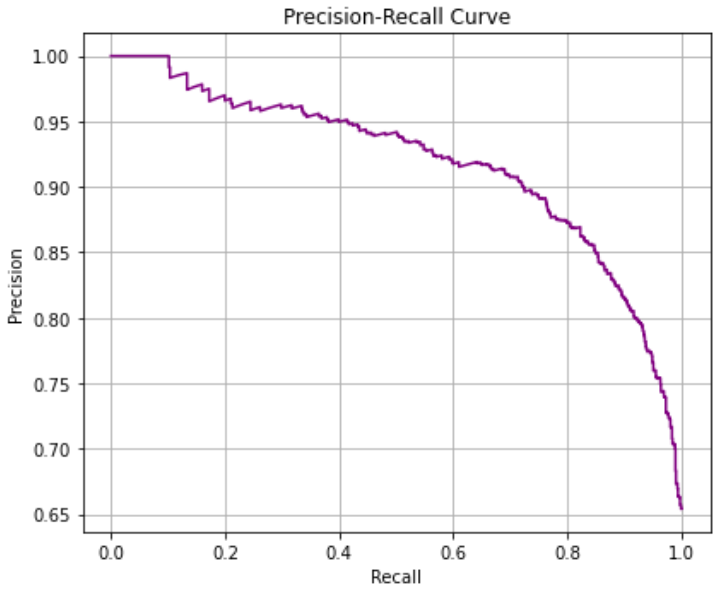

Figure 14: Precission-Recall Curve of Model 1, Binary Logistic Regression on Numeric Features Only

The curve illustrates the cumulative explained variance across the principal components extracted via PCA. The steep initial slope, followed by a tapering tail, reflects that a relatively small number of principal components capture the majority of the variance in the dataset. This confirms the presence of latent structure in the data and justifies dimensionality reduction before modeling, improving computational efficiency while retaining key discriminative information.

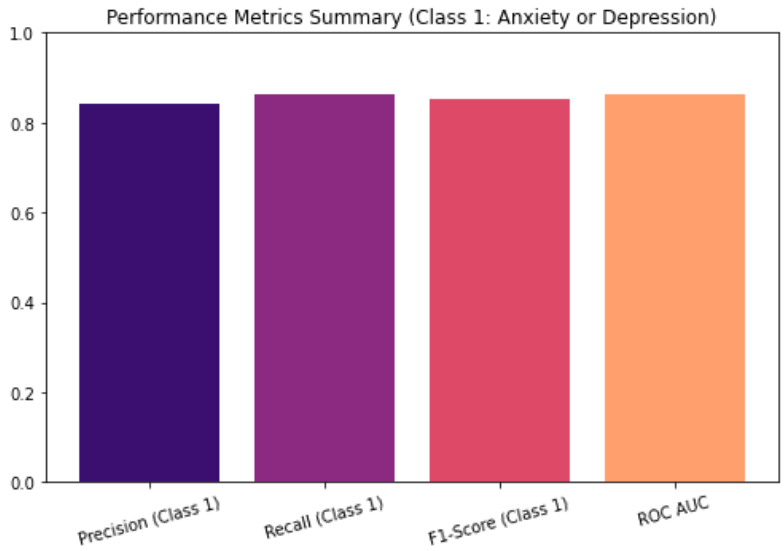

Figure 15: Performance Metrics of Model 1, Binary Logistic Regression on Numeric Features Only

This bar chart visualizes the relative importance of different features used in a logistic regression model. The uniformity of bar heights suggests that the model leverages a set of features that contribute similarly to prediction performance, indicating a balanced multi-dimensional influence rather than heavy reliance on one dominant variable. This reinforces the robustness of the feature selection strategy and supports the hypothesis that multiple dimensions of the voice and facial data (or psychological inferences) play a complementary role in predicting mental health status.

The following table summarizes the results obtained with this configuration:

	Precision	Recall	F1-Score	Support
Binary	0.73	0.69	0.71	599
Multiclass	0.84	0.86	0.85	1133
Accuracy			0.8	1732
Macro Avg	0.78	0.78	0.78	1732

Weighted Avg	0.8	0.8	0.8	1732

Table 9: Binary Logstic Regression Summary of Results for Numeric Features Only

These results show that even using only automatically extracted numeric features (no text), the model can achieve an F1 around 0.85 for detecting a condition. This is encouraging – it implies the behavioral and physiological signals have predictive power. In practical terms, the model using facial expressions, voice tone, movement, etc., correctly identifies about 86% of people with anxiety/depression, while mistaking about 14% as being healthy (false negatives). On the other hand, it incorrectly flags about 31% of healthy controls as having a condition (false positives).

From an application perspective (screening for mental health), such a model would catch most people who have issues (which is good), but it would also raise some false alarms (which might be acceptable in a preliminary screening if those can be later evaluated by a professional). The relatively lower recall for controls (69%) is another way of saying the false positive rate is somewhat high. We might prefer a model with higher specificity (fewer false alarms) if implementing a real screening tool. Logistic regression's threshold can be adjusted to tune this trade-off: raising the threshold would reduce false positives (increasing control recall) at the expense of missing more true cases (lowering condition recall).

One advantage of logistic regression is interpretability of coefficients. In this case, with 61 features, examining coefficients can tell us which features push the prediction toward class 1 (anxiety/depression) and which push toward class 0 (control). This summarized output is studied through a Logit Regression Result, which draws the following conclusions:

- Neuroticism (coef = -0.9027, $p < 0.001$): Its negative coefficient indicates a strong inverse relationship between increased neuroticism and being in the control group, for example, individuals high in neuroticism are more likely to belong to the anxiety/depression class.

- Happy_facial and Sad_facial (both $p < 0.001$): These features underline the predictive value of facial expression dynamics.
- Self-efficacy metrics show a coherent pattern: high self-efficacy correlates with lower likelihood of anxiety/depression, whereas low and medium levels are positively associated.
- Communication and Creativity traits hover around the threshold of significance, suggesting some interpretative caution is required.

A coefficient plot to provide visual overview of the estimated coefficients and their confidence intervals for each feature included in the logistic regression model is also included in the next page. Coefficients to the right of the dashed vertical line at zero indicate positive relationships with the likelihood of belonging to the anxiety/depression group, while those to the left reflect negative relationships. Notably, features like neuroticism, sad_facial, and self_efficacy_low have clearly negative coefficients with narrow confidence intervals, indicating both strong effect sizes and statistical significance. Conversely, features like stress_high, while positive, have wide intervals and cross zero, suggesting low statistical robustness. This plot visually encapsulates the interpretability advantage of logistic regression models, highlighting which features drive predictions and to what extent.

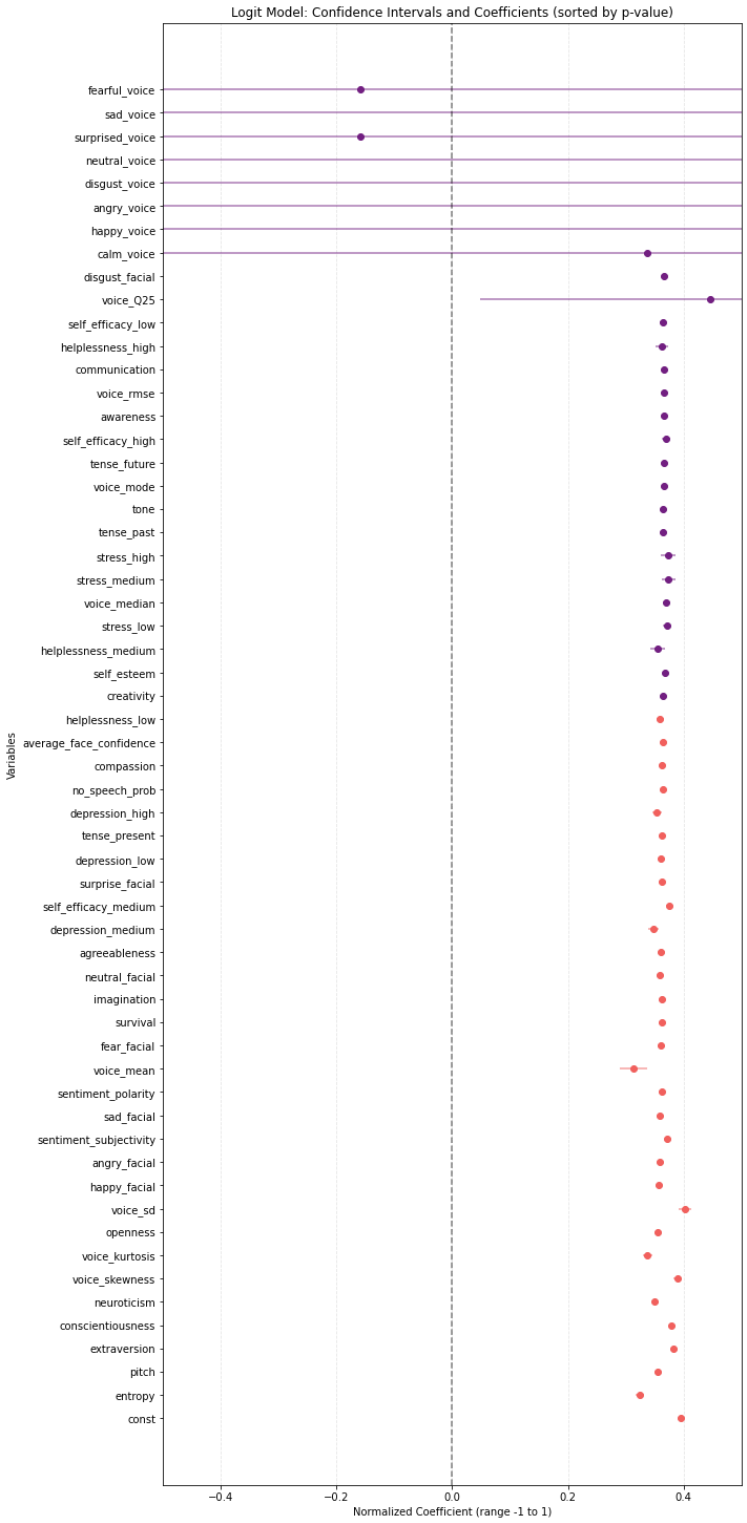

Figure 16: Confidence Intervals for Numeric Variables

8.1.2 MODEL 2: LOGISTIC REGRESSION ON TEXT FEATURES ONLY

For this model, all numeric features are ignored, and instead textual content is used to predict the binary outcome. This is essentially a text classification problem: given the words someone uses, predict if they are expressing signs of anxiety/depression or not. The approach taken is to use TF-IDF vectorization of the text data and then feed those features into a logistic regression.

The translation column is used as the basis of this analysis and a TfidfVectorizer is instantiated. Initially the vectorizer is left unconstrained, while later there will be specific subset of words analysis. After vectorization, the text feature matrix is split into training and test sets, aligned with the same y_train and y_test from before (ensured by using the same random state or by performing the split on the text data in parallel).

A logistic regression is then trained on the TF-IDF features. This is a high-dimensional problem – potentially thousands of features (words) with ~4800 training examples. The default L2 regularization is actually beneficial here to avoid overfitting the plethora of sparse features. The results obtained are showcased in the following illustrations:

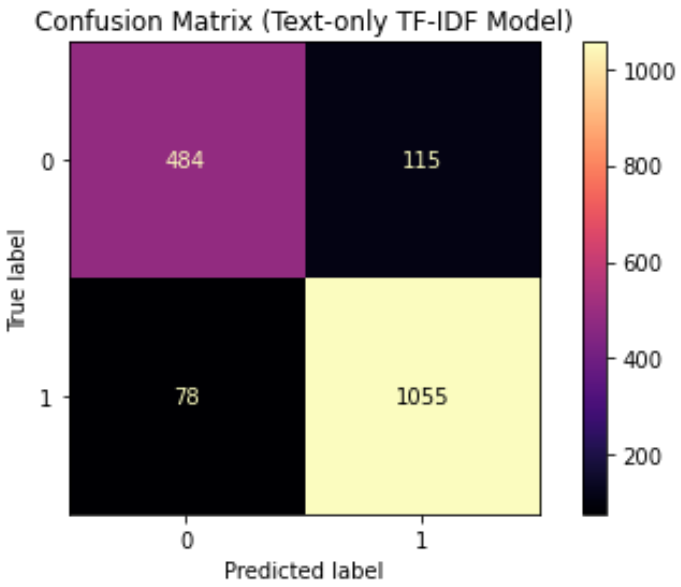

Figure 17: Confusion Matrix for Model 2, Binary Logistic Regression for Text-Only Features

The confusion matrix shows a much more balanced outcome than the numeric model. Out of 599 controls, 484 are caught (81% vs 69% before), and out of 1133 condition cases, 1055 are caught (93% vs 86% before). The model only misses 78 condition cases (false negatives) compared to 159 earlier – about half as many – which is a significant improvement in sensitivity. False positives (115) are also fewer than before (186).

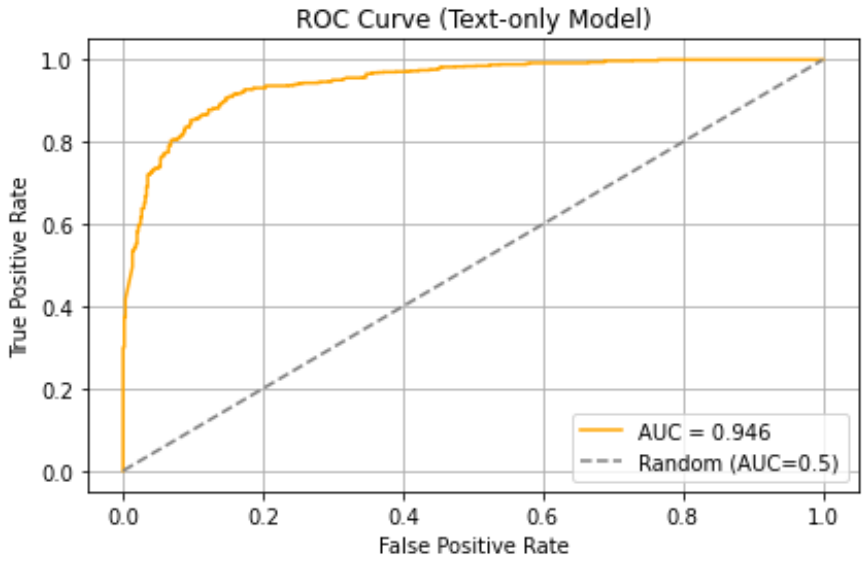

Figure 18: ROC Curve for Model 2, Binary Logistic Regression for Text-Only Features

The ROC curve above portrays an AUC close to 0.95, which confirms excellent discrimination ability – the model's ranking of positive vs negative cases by predicted probability is highly accurate. Such a high AUC also indicates a lot of area to play with the trade-off between TPR and precision.

The following table summarizes the results obtained with this configuration:

	Precision	Recall	F1-Score	Support
Binary	0.86	0.81	0.83	599
Multiclass	0.90	0.93	0.92	1133

Accuracy			0.89	1732
Macro Avg	0.88	0.87	0.87	1732
Weighted Avg	0.89	0.89	0.89	1732

Table 10: Results Comparison for Model 2, Binary Logistic Regression on Text-Only Features

The accuracy has considerably increased to 89%, a huge jump from 80%. It means the text-based model is correct almost 9 out of 10 times on random samples, which is very strong for this kind of classification. In addition, the precision and recall metrics indicate a high-performing model:

- The condition class has a 0.92 F1, implying very good balance between precision and recall. It catches 93% of actual cases (only 7% false negatives) and its precision of 90% means only 10% of those flagged as "mentally unhealthy" are actually healthy (a manageable false alarm rate).
- Control class's precision 86% and recall 81% are also respectable. The model still leans slightly toward the positive class (as evidenced by control recall being a bit lower), but it's far better than the numeric model's 69%. Now 81% of actual controls are recognized, and when the model predicts "control," it's correct 86% of the time. The false positive rate is now 115 out of 599 ≈ 19%, substantially lower than the 31% from numeric model. This means the text-based model makes far fewer false alarms.

Because of this, the weighted average precision/recall/F1 and macro average have also shown very academically strong results.

The text-only logistic regression far outperforms the numeric-only one in identifying anxiety/depression. This is not surprising, as language is often a direct carrier of psychological state. People experiencing depression or anxiety tend to use certain words or expressions that can act as clear signals. Indeed, the model essentially learned a set of weighted keywords and patterns that differentiate the classes.

From this point of huge improvement, some further analysis is done. For instance, Principal Component Analysis (PCA) technique is used to show dimensionality reduction, as can be observed in the following image:

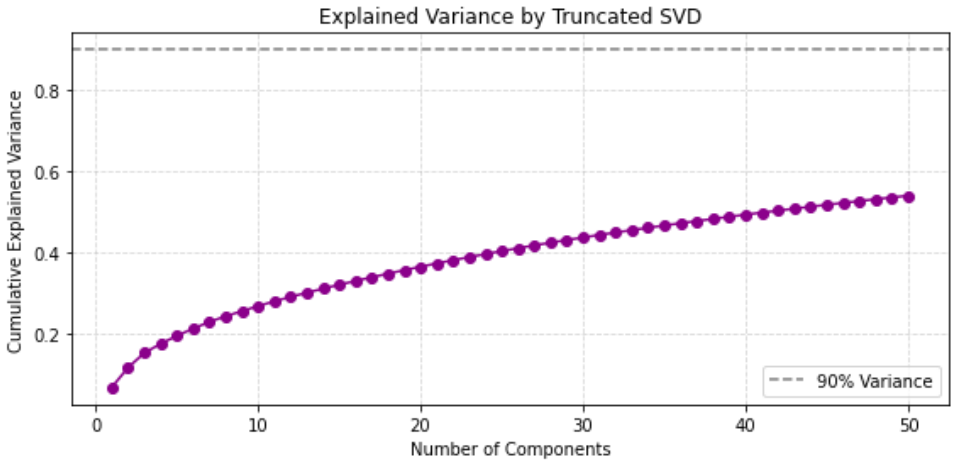

Figure 19: Explained Variance for Model 2, Binary Logistic Regression for Text-Only Features

This specific graph shows how much of the total variance in the dataset is captured as more principal components are added. In this case, the 90% threshold line provides a reference to identify how many components are required to retain 90% of the variance in the original data. The observed gradual upward trend indicates that although the information is somewhat distributed across many features, a subset of components (likely around 40–50) captures most of the variance, suggesting that the model's dimensionality can be reduced significantly without substantial information loss.

A heatmap is also included to represent the correlation matrix of the PCA-transformed features, crucial for validating the PCA's effectiveness in orthogonalizing the features:

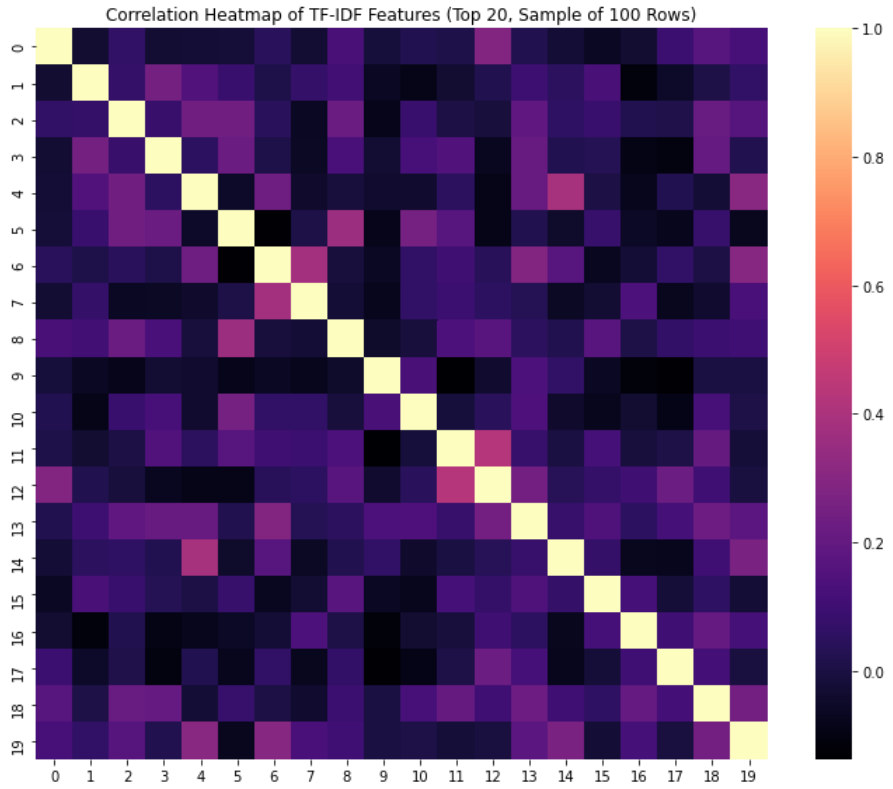

Figure 20: Correlation Heatmap for Model 2, Binary Logistic Regression for Text-Only Features

In the image, the clear diagonal line with near-zero off-diagonal values indicates minimal correlation among the new components. This orthogonality is essential for ensuring the validity of the assumptions of logistic regression, particularly the absence of multicollinearity. It allows the logistic model to work with decorrelated input variables, enhancing the stability and interpretability of the resulting coefficients and improving convergence behavior during optimization.

In addition to this, a DataFrame of words and coefficients is completed and sorted, with the objective of explaining what are the key textual features, as, in logistic regression, each word in the vocabulary gets a coefficient indicating how strongly its presence (or higher TF-IDF score) influences the prediction towards the positive class (anxiety/depression) or negative class (control).

The most strongly positive coefficients are the following:

Word	Coefficient
Anxiety	9,06
Depression	8,39
Feel	4,8
Pressure	3,63
Help	2,96

Table 11: Coefficients of Top Contributing Words to Class 1

Graphically:

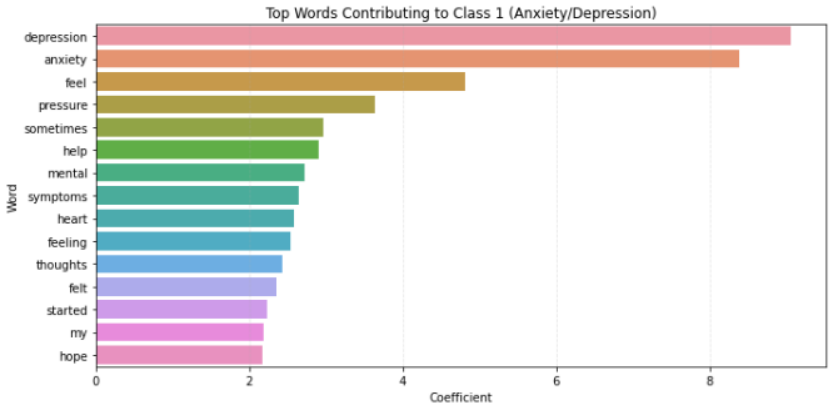

Figure 21: Coefficients of Top Contributing Words for Class 1

It is important to note that such high coefficients partly reflect the rarity and specificity of those words to the positive class: not many control posts would contain the words "depression" or "anxiety" in a context that isn't about those conditions, so these are almost diagnostic when they appear. TF-IDF weighting further amplifies their impact if they are somewhat unique to the documents in which they appear.

The most strongly negative coefficients are the following:

Word	Coefficient
Money	-4,53
Yeah	-3,38
Yes	-3,37
Think	-2,91
They	-2,7

Table 12: Coefficients of Top Contributing Words for Class 0

Graphically:

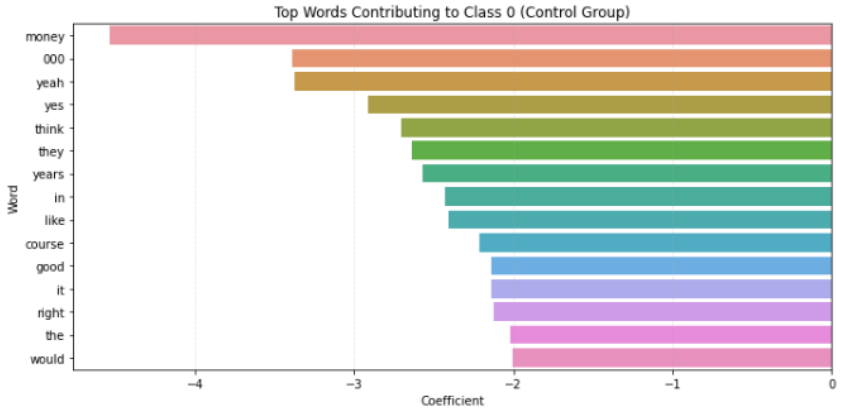

Figure 22: : Coefficients of Top Contributing Words for Class 0

Having done this, the top 25 most influential words by absolute coefficient are selected. This list of 25 words includes both the strongest positive and negative indicators:

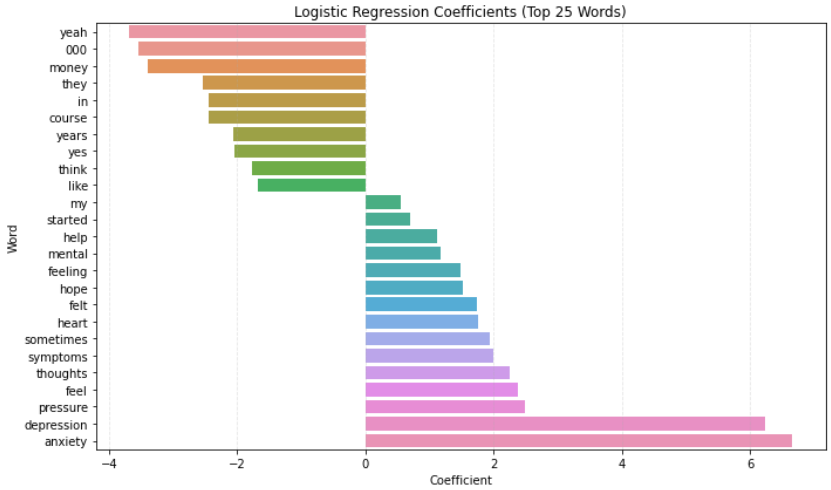

Figure 23: Coefficients of Top 25 Contributing Words for both Classes

With this, the text can be re-vectorized using only these words. The idea is to analyze if a much simpler model focusing on a small vocabulary can still perform well, and to some extent to verify that those top words indeed carry most of the signal. The performance of this new, simplified model is shown in the following illustrations:

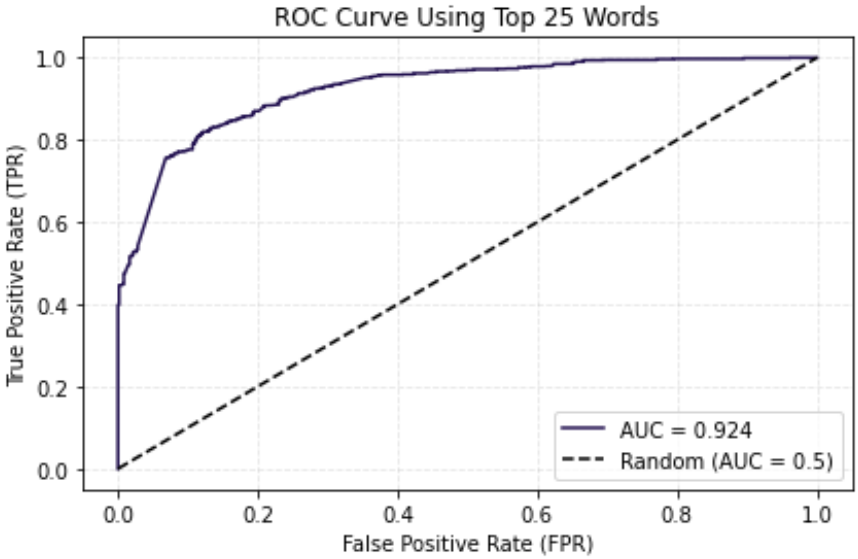

Figure 24: ROC Curve for Model 2, Binary Logistic Regression on the Top 25 Words Only

The ROC curve for the model portrays an AUC of 0.924, which reflects excellent discriminatory power of the model while it implies a downgrade from the previous model.

Once again, however, the steep curve at the start and its position above the diagonal baseline (random classifier) signify strong sensitivity and specificity.

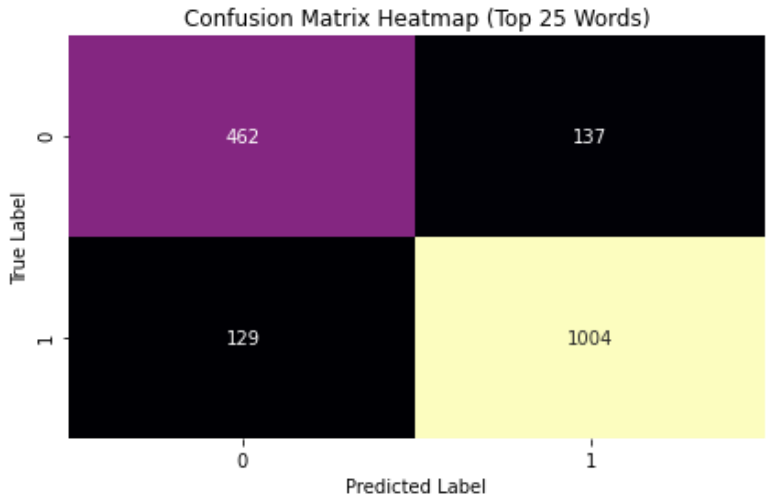

Figure 25: Confusion Matrix for Model 2, Binary Logistic Regression on the Top 25 Words Only

The heatmap showing the confusion matrix of predicted versus actual class labels for the model reveals 462 true positives (mental health correctly identified), 1004 true negatives (controls correctly classified), 129 false positives (controls misclassified as mental health), and 137 false negatives (mental health cases missed). These results suggest a well-balanced classifier, with both recall and precision showing satisfactory performance while it does imply that the previous model worked better.

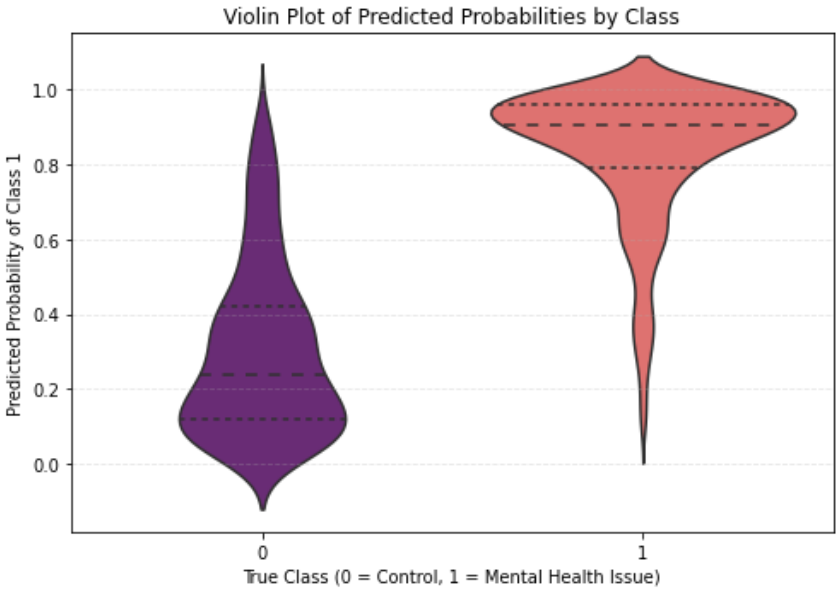

Figure 26: Violin Plot for Model 2, Binary Logistic Regression on the Top 25 Words Only

A violin plot is also included to visualize the distribution of predicted probabilities for each class (control vs. mental health) produced by the logistic regression model. The left (purple) distribution corresponds to the control group, while the right (red) corresponds to the mental health group. The separation between the two distributions is pronounced, with the mental health group exhibiting a distribution of higher predicted probabilities compared to the control group. The lack of significant overlap between the distributions indicates that the model confidently assigns high probabilities to the correct class. This not only reinforces the model's effectiveness but also affirms the informativeness of language usage in mental health identification. The clear demarcation supports the usability of such a classifier in real-world screening, where decisions may be guided by probability thresholds tailored to minimize false negatives or false positives based on application context.

The following table summarizes these key findings:

	Precision	Recall	F1-Score	Support

Binary	0.78	0.77	0.78	599
Multiclass	0.88	0.89	0.88	1133
Accuracy			0.85	1732
Macro Avg	0.83	0.83	0.83	1732
Weighted Avg	0.85	0.85	0.85	1732

Table 13: Summary for Model 2, Binary Logistic Regression on the Top 25 Words Only

Comparing this to the full-text model: using only 25 words leads to a drop from 0.89 to 0.83 in accuracy, and notably the false negatives and false positives both increased (462 vs 484 TN, 1004 vs 1055 TP – so the model with 25 words missed 1055-1004=51 more condition cases and incorrectly flagged 137-115=22 more controls than the full model). This is still reasonably good, showing that those 25 words captured a lot of the predictive power (the F1 for class 1 is still 0.89, only a bit lower than 0.92). However, performance did suffer, particularly for the control class (precision/recall ~0.78 vs ~0.83-0.86 originally). This indicates that while the top words are very informative, the broader vocabulary offers additional nuance that helps catch cases where perhaps the top keywords are not used.

In summary, the top 25 words experiment shows that a small set of keywords yields a decent model (precision ~0.78-0.89 depending on class) but full performance requires a wider net of vocabulary.

In a similar manner, an analysis for the most significant 50 words is performed. The words chosen can be gathered from this coefficient graphical visualization:

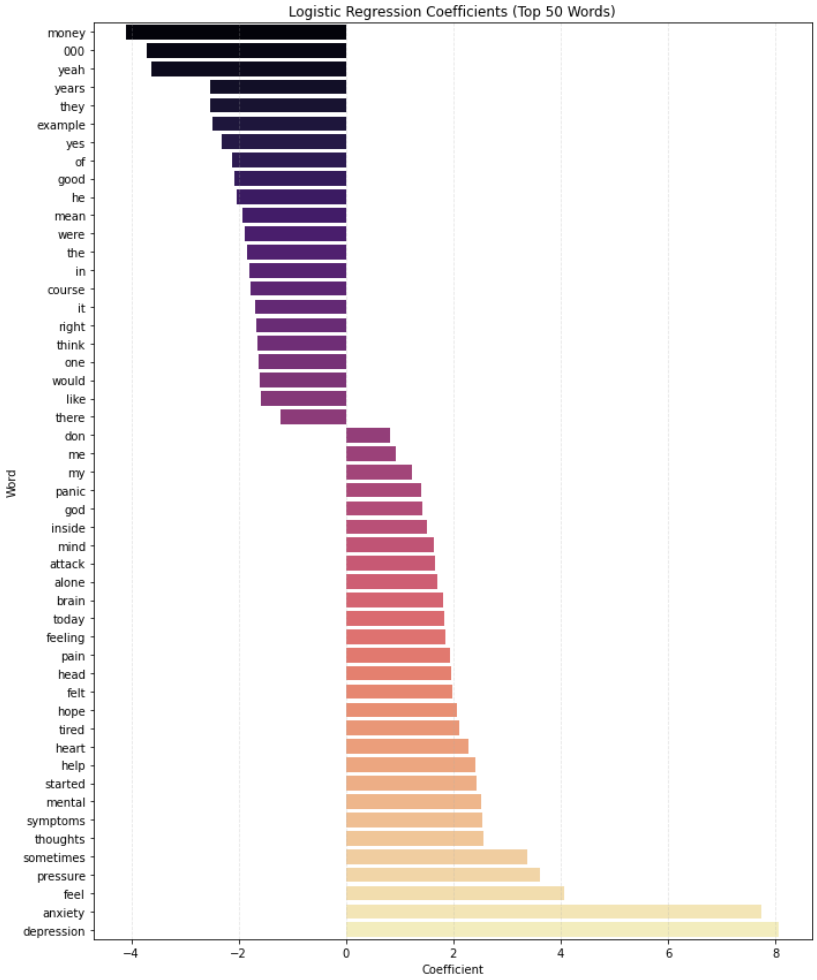

Figure 27: Coefficients of Top 50 Contributing Words for both Classes

The following results are obtained with this model:

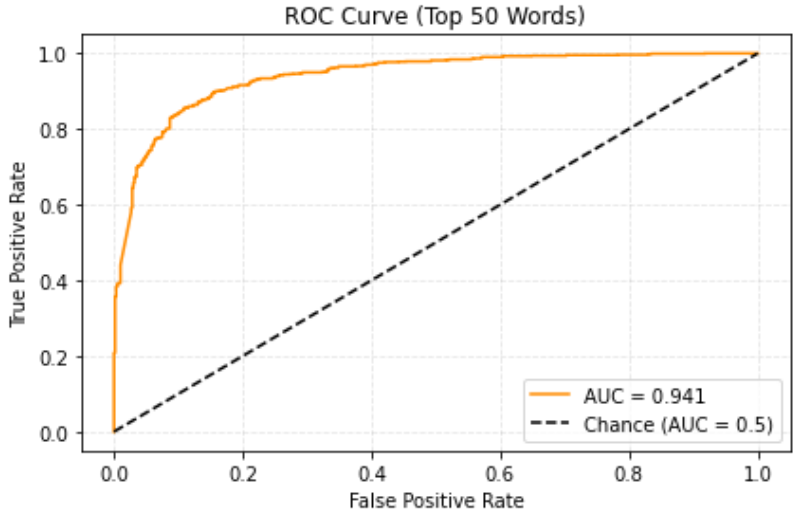

Figure 28: ROC Curve for Model 2, Binary Logistic Regression on the Top 50 Words Only

There was a slight improvement in the ROC curve that, nevertheless, still remains technically worse than the one including all words.

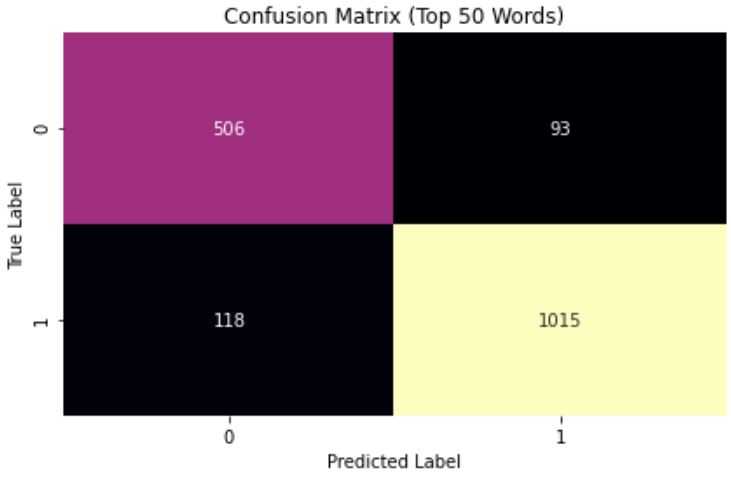

Figure 29: Confusion Matrix for Model 2, Binary Logistic Regression on the Top 50 Words Only

The confusion matrix further verifies the thesis that the model works better with 50 words than 25 but its performance remains slightly worse than the fully complete model.

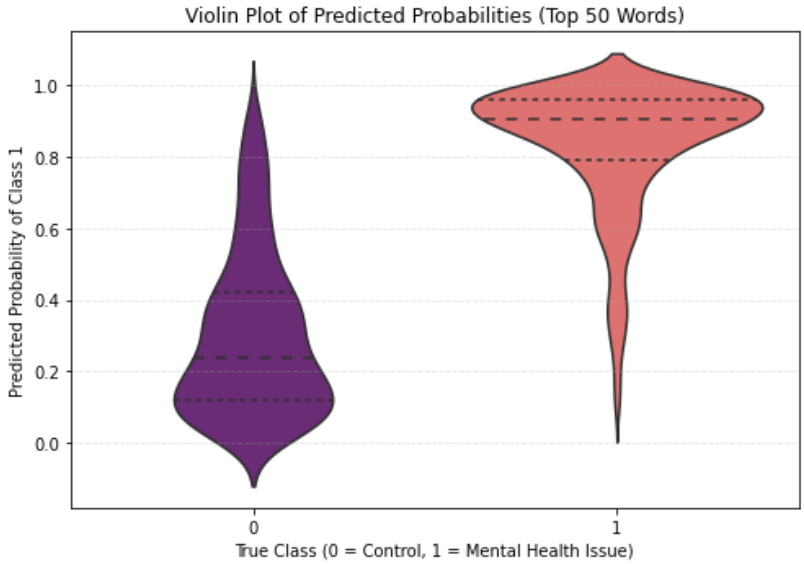

Figure 30: Violin Plot for Model 2, Binary Logistic Regression on the Top 50 Words Only

The violin plot, while slightly better, is almost identical to the one previously analyzed for a 25 word dataset.

The results are summarized in the following table:

	Precision	Recall	F1-Score	Support
Binary	0.81	0.84	0.83	599
Multiclass	0.92	0.90	0.91	1133
Accuracy			0.88	1732
Macro Avg	0.86	0.87	0.87	1732
Weighted Avg	0.88	0.88	0.88	1732

Table 14: Summary for Model 2, Binary Logistic Regression on the Top 50 Words Only

Comparing this to the full-text model: using only 50 words leads to a slight drop from 0.89 to 0.88 in accuracy. This is still reasonably good, showing that those 50 words captured a

lot of the predictive power (the F1 for class 1 is still 0.91, only a bit lower than 0.92). This indicates that while the top words are very informative, the broader vocabulary offers additional nuance that helps catch cases where perhaps the top keywords are not used. However, this difference is notably lower than with the 25-word model. Because of this, this model provides a simplified alternative to the one using the complete net of words available.

Further graphs are explored to showcase some relevant attributes of this simplified model. The following bubble plot provides an understanding of the discriminatory power of each of the top 50 words used as features in the model:

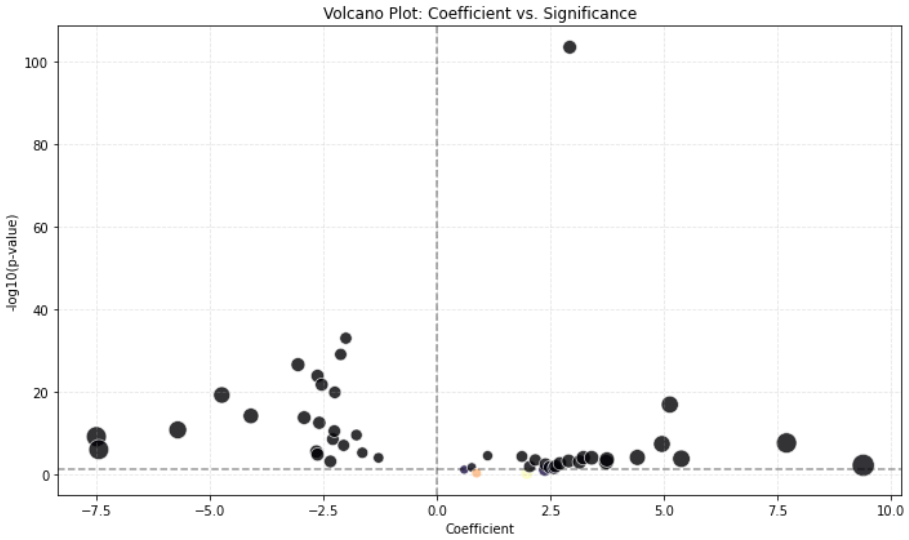

Figure 31: Volcano Plot for Model 2, Binary Logistic Regression on the Top 50 Words Only

From this visual, it is evident that most words lie close to the origin, indicating either negligible contribution or lack of statistical significance. However, a few stand-out terms show both a relatively strong coefficient and high statistical significance (i.e., positioned farther from the origin on both axes), suggesting they are highly informative in distinguishing between mental health-related content and control group language. These words likely correspond to emotionally charged or diagnostically relevant vocabulary used more often by individuals expressing psychological distress. This dual consideration of

magnitude and p-value ensures that feature selection does not rely merely on frequency but also on predictive contribution.

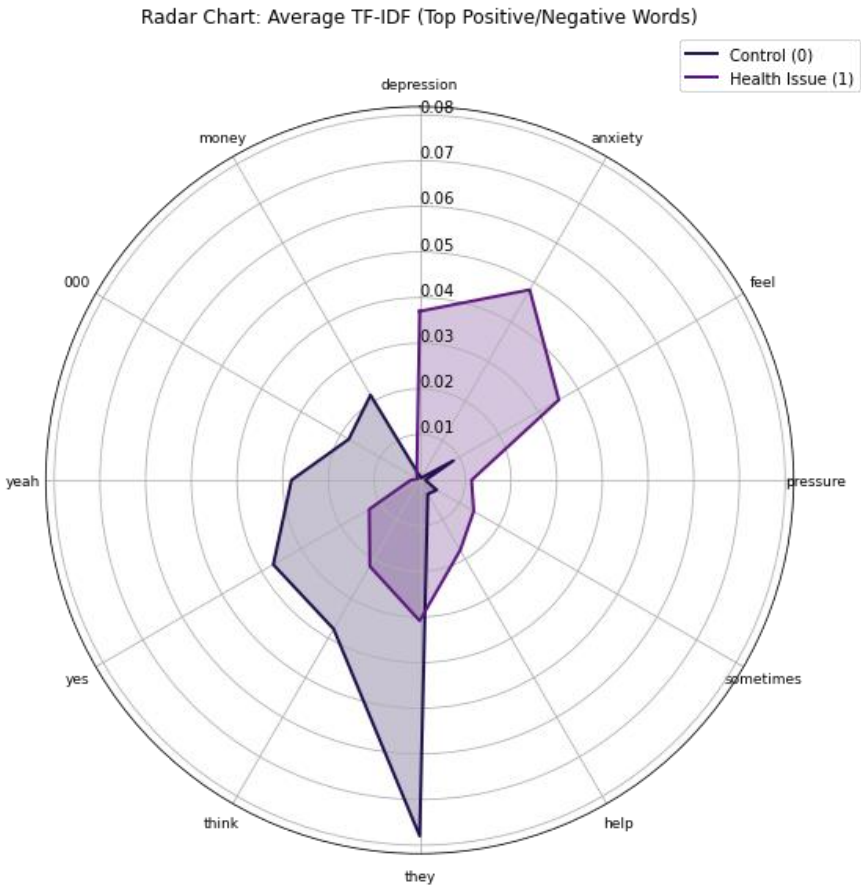

Figure 32: Radar Plot for Model 2, Binary Logistic Regression on the Top 50 Words Only

The radar chart above shows the average usage proportion of selected vocabulary terms across the two binary classes: Control (label 0) and Health Issue (label 1), each axis corresponding to the top word features.

The contrast in the shaded areas reveals divergence in lexical patterns between classes. Notably, the polygon associated with the Health Issue class shows concentration in the anxiety, depression and, in general, feelings subset of axis. In contrast, the Control group's word use appears more distributed or skewed across a different subset, highlighting less emotionally charged or more generic language.

One last analysis using 250 words is done to see if it reaches the results obtained for the full net of features (words) model. The results obtained are:

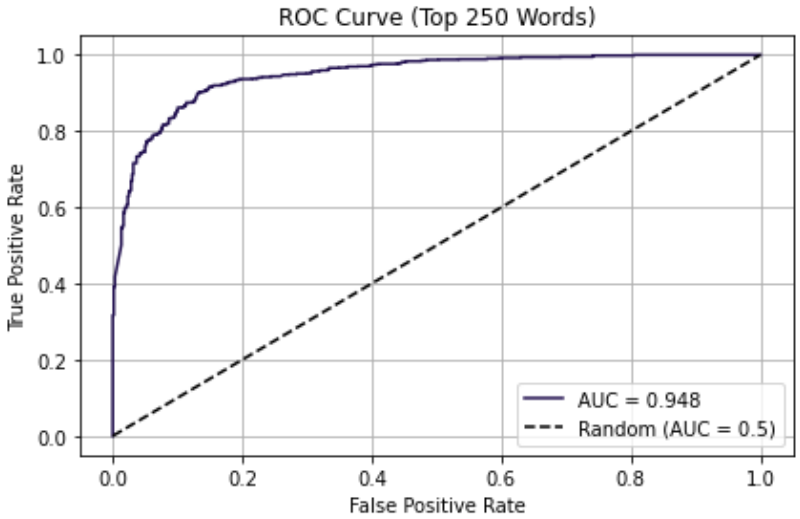

Figure 33: ROC Curve for Model 2, Binary Logistic Regression on the Top 250 Words Only

The ROC curve shows an AUC of 0.948, which indicates a very strong ability of the model to discriminate between individuals with mental health conditions and controls. This is a notable performance, suggesting the 250 most relevant words encode substantial predictive value on their own. Compared to the full text model the result is nearly on par, indicating that the selected text features retain nearly the entire discriminative power of the full model.

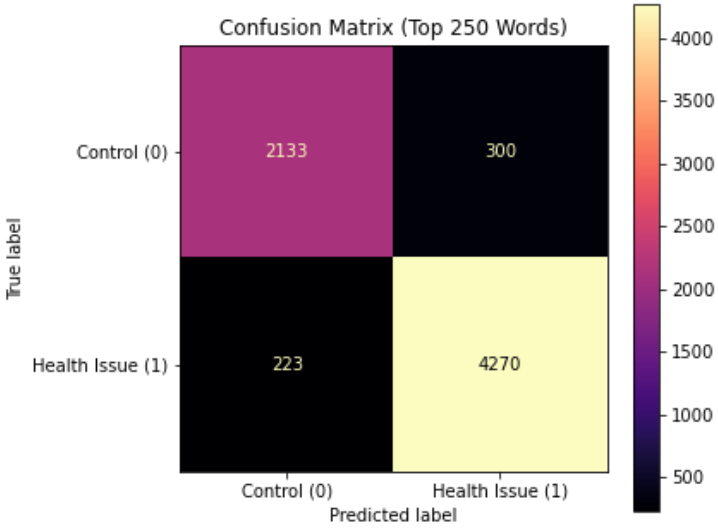

Figure 34: Confusion Matrix for Model 2, Binary Logistic Regression on the Top 250 Words Only

The confusion matrix yields a strong performance profile with both high sensitivity and specificity

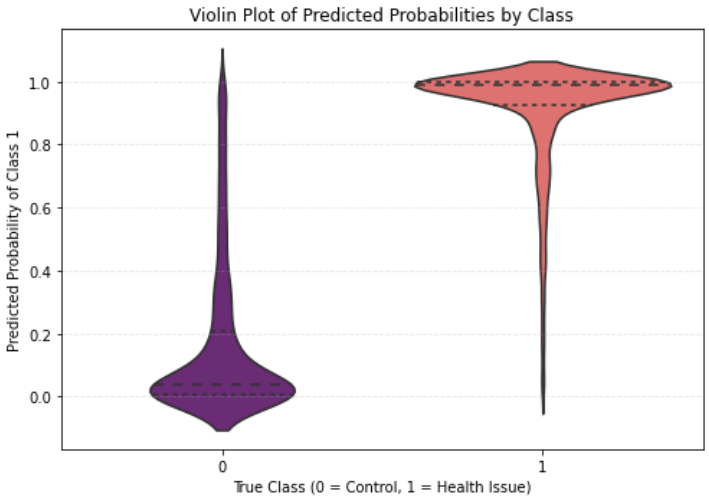

Figure 35: Violin Plot for Model 2, Binary Logistic Regression on the Top 250 Words Only

The violin plot comparing predicted probabilities for each class shows two distinct and well-separated distributions. The control group exhibits a concentrated density around low probability values (close to 0), while the group with a mental health issue has a tightly packed distribution near 1. The sharpness and separation of these curves reveal the model's

high confidence in its predictions and its robustness in decision boundaries. Compared to the wider, more overlapping distributions seen in the 50-word and 25-word models, this indicates significantly better class separation, confirming a superior model.

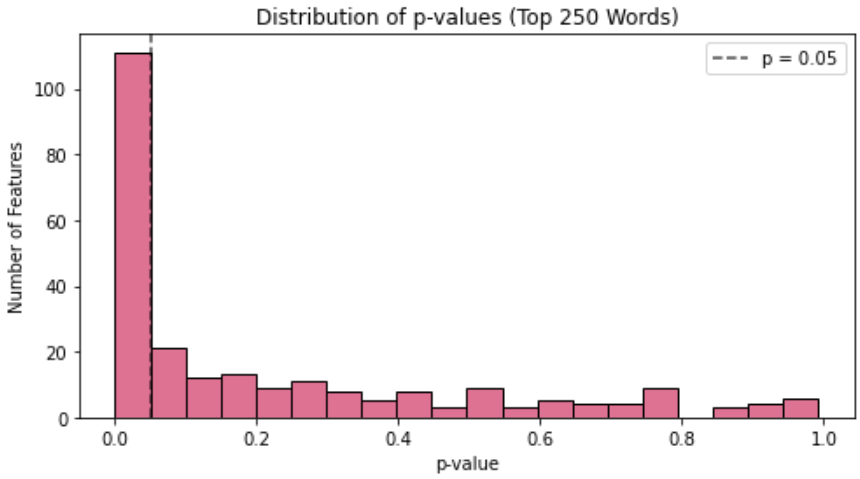

Figure 36: Distribution of p-values for Model 2, Binary Logistic Regression on the Top 250 Words Only

This histogram above displays the p-values associated with the logistic regression coefficients for the top 250 words. Many features fall below the significance threshold of p = 0.05 (dashed line), indicating that many of these textual features contribute meaningfully to the prediction. In contrast to the 50-word model (which had fewer significant features), this distribution supports the idea that extending to 250 carefully selected terms allows the model to capture more nuanced linguistic cues relevant to anxiety and depression. This strengthens the model interpretability and statistical credibility.

The results can be summarized in:

	Precision	Recall	F1-Score	Support
Binary	0.86	0.81	0.84	599
Multiclass	0.90	0.93	0.92	1133
Accuracy			0.89	1732

Macro Avg	0.88	0.87	0.88	1732
Weighted Avg	0.89	0.89	0.89	1732

Table 15: Summary of Results for Model 2, Binary Logistic Regression on the Top 250 Words Only

As can be seen in the table above, the refined textual model demonstrates not only a higher true positive rate but also fewer false positives, indicating improved generalization and less overfitting. This reinforces the effectiveness of limiting the vocabulary to 250 meaningful tokens.

8.1.3 MODEL 3: HYBRID LOGISTIC REGRESSION

The third model combines all available features – both the 61 numeric features and the text-based features – into one logistic regression. The rationale is that while text alone is very powerful, the numeric features might provide complementary information in cases where the language is less explicitly indicative. To combine the modalities, the feature sets are concatenated, and the new model is trained.

The results obtained are shown in the following illustrations:

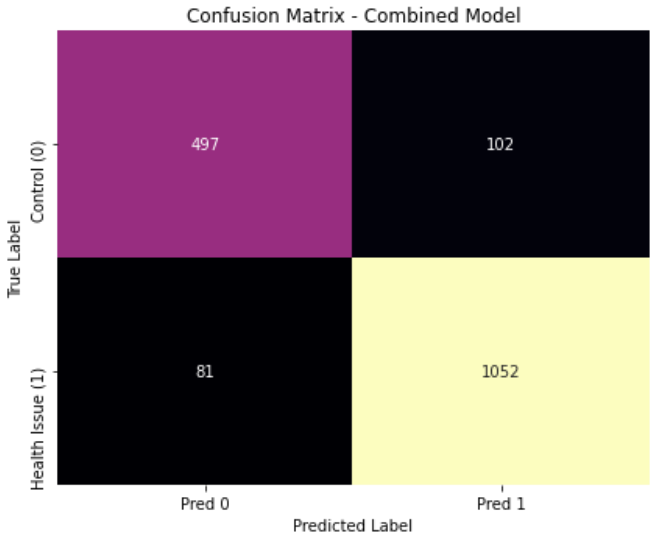

Figure 37: Confusion Matrix for Model 3, Hybrid Binary Logistic Regression

The confusion matrix above reveals the concrete classification outcomes for the hybrid model. Of 599 control cases, 497 were correctly classified (true negatives) and 102 were false positives. Among 1133 health-affected cases, 1052 were correctly classified (true positives) and only 81 were false negatives. These values translate into a strong overall accuracy, better than the one obtained in the only-text models.

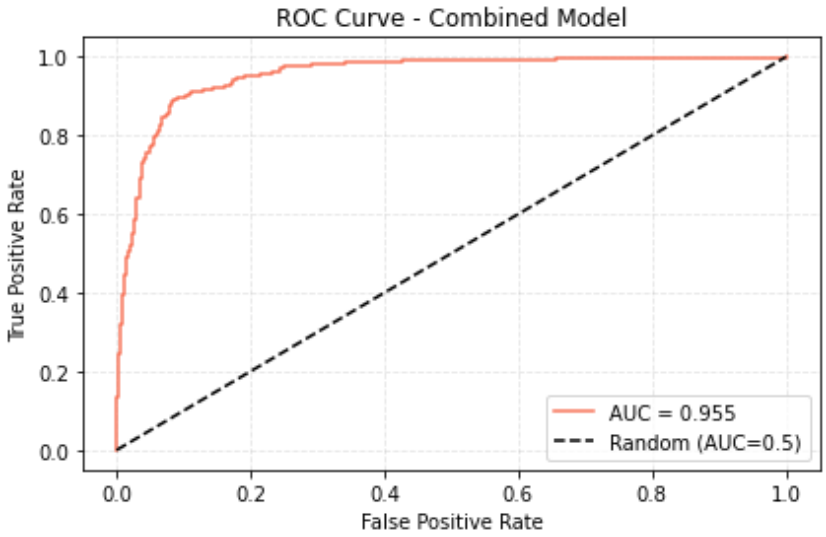

Figure 38: ROC Curve for Model 3, Hybrid Binary Logistic Regression

The ROC curve conveys for this hybrid model exceptional classification performance, with an AUC of 0.955. The curve hugs the top-left corner, indicating a strong balance of sensitivity and specificity. It is virtually identical to the ROC Curve obtained from the only-text models.

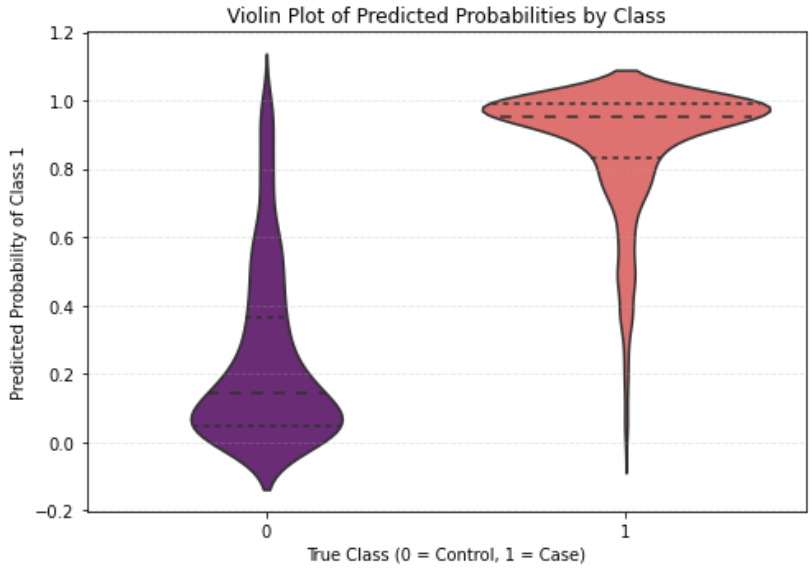

Figure 39: Violin Plot for Model 3, Hybrid Binary Logistic Regression

The violin plot above shows similar results to the ones previously analyzed, where the separation between the two distributions is visually substantial: the majority of predictions for class 0 cluster around a low probability range, whereas class 1 probabilities tend to be sharply peaked near the top of the scale. This distributional divergence signals a strong discriminative power of the model. The narrowness of the control group distribution compared to the broader, higher-peaked distribution for the health issue group suggests that the model is highly confident in detecting class 1 cases.

The results are summarized in the following table:

	Precision	Recall	F1-Score	Support
Binary	0.86	0.83	0.84	599
Multiclass	0.91	0.93	0.92	1133
Accuracy			0.89	1732
Macro Avg	0.89	0.88	0.88	1732

Weighted Avg	0.89	0.89	0.89	1732

Table 16: Summary of Results for Model 3, Hybrid Binary Logistic Regression

Compared to earlier models, the hybrid model maintains slightly improved scores while using a more optimized and informative subset of features, and, therefore, is the technically best overall model, providing lower false negatives than previous models, which is typically higher in mental health diagnostics due to the risk of untreated pathology; thus, this improvement is clinically meaningful.

8.2 MULTICLASS LOGISTIC REGRESSION ANALYSIS

The second notebook extends the analysis to a three-class classification problem: predicting whether a given instance is Control, Ansiedad (Anxiety), or Depresión (Depression). As previously explained, a Label Encoder is used, mapping 'Ansiedad': 0, 'Control': 1, 'Depresion': 2. With the previously explained train and test split, the first model is analyzed.

8.2.1 MODEL 1: LOGISTIC REGRESSION ON NUMERICAL FEATURES ONLY

The results obtained with this model can be observed in the following confusion matrix:

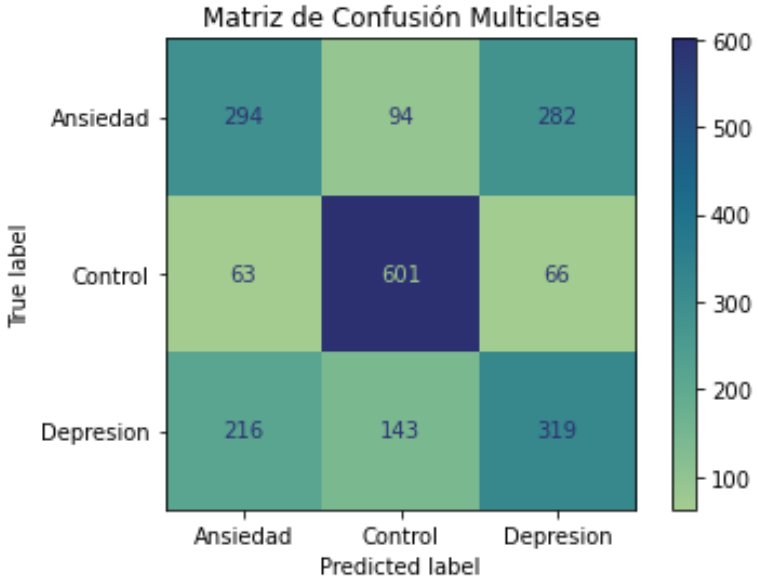

Figure 40: Confusion Matrix for Model1, Multiclass Logistic Regression on Numerical Features Only

Which can be summarized in the following table of results:

	Precision	Recall	F1-Score	Support
Anxiety	0.51	0.44	0.47	670
Control	0.72	0.82	0.77	730
Depression	0.48	0.47	0.47	678
Accuracy			0.58	2078
Macro Avg	0.57	0.58	0.57	2078
Weighted Avg	0.57	0.58	0.58	2078

Table 17: Summary of Results for Model1, Multiclass Logistic Regression on Numerical Features Only

It is immediately obvious that the numeric model struggled particularly with the two condition classes:

- Anxiety and Depression both have low recall (~0.44-0.47). That means the model correctly identifies less than half of those posts correctly. In other words, more than half of anxious individuals were not predicted as "Ansiedad" by the model, and similarly for depressed individuals.
- Their precision is also around 0.5, meaning when the model predicts "Ansiedad" or "Depresion", it's correct roughly only half the time.
- The control class, by contrast, has a higher recall (0.82) and precision (0.72). The model is better at recognizing controls, which aligns with the binary numeric result where it had a decent true negative rate. Essentially, many of those with conditions are being misclassified, but often as control (which boosts control's recall).

8.2.2 MODEL 2: LOGISTIC REGRESSION ON TEXT FEATURES ONLY

Next, logistic regression using only textual features is used to classify into the three categories. This means that, given words used by the speaker of the video, a classifier will be trained to decide if a person is depressed, anxious or neither. Label-encoding is used as before, (0=Anxiety, 1=Control, 2=Depression) and split train/test for text similarly to numeric (ensuring alignment and stratification). The logistic regression is configured for multinomial (softmax) classification. The results obtained initially are shown in the confusion matrix below:

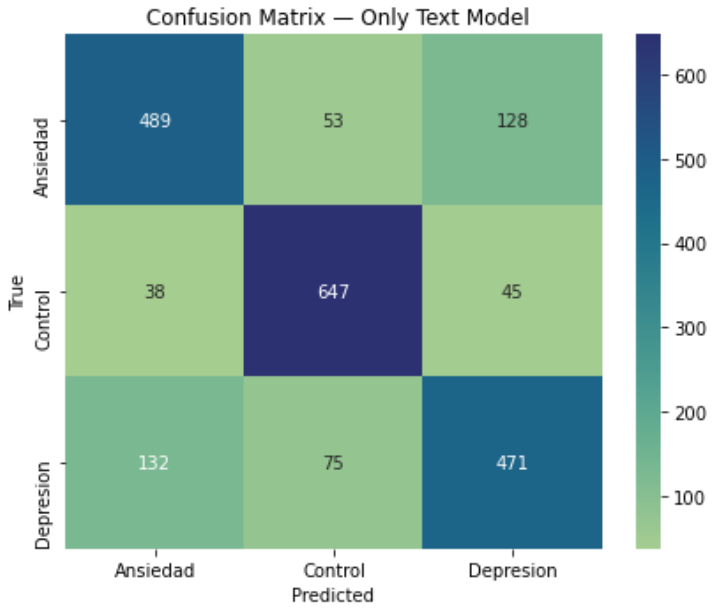

Figure 41: Confusion Matrix for Model 2, Multiclass Logistic Regression Text-Only Features

This data can be summarized in the following table:

	Precision	Recall	F1-Score	Support
Anxiety	0.74	0.73	0.74	670
Control	0.83	0.89	0.86	730
Depression	0.73	0.69	0.71	678
Accuracy			0.77	2078
Macro Avg	0.77	0.77	0.77	2078
Weighted Avg	0.77	0.77	0.77	2078

Table 18: Summary of Results for Model 2, Multiclass Logistic Regression Text-Only Features

Now, using text, a much more balanced and overall considerably better performance is observed:

- For both anxiety and depression classes, F1 is in the low 70s (0.74 and 0.71), which is a dramatic increase from 0.47 in the numeric model. This indicates the text features carry distinct clues that allow the classifier to differentiate between anxiety and depression content to a good extent.
- The control class is again the easiest: F1 0.86, with recall 89%. The model has little trouble identifying most control posts (only ~11% of actual controls got misclassified). Precision 0.83 means when it predicts control, 17% of those might be actual anxiety or depression – which is a false positive rate for control predictions. (Equivalently, it's the false negative rate for the union of anxiety + depression, which appears to be 11% of controls misclassed out of controls, but it can be computed false negative for conditions differently).
- The anxiety vs depression confusion: The recall for anxiety 73% means it missed about 27% of anxious cases. Some of those were likely predicted as depression or control. Similarly, depression recall 69% means 31% of depressed cases were mis-labeled, presumably often as anxiety or sometimes as control. With further inspection, more trends are inferred:
    - The high control recall (89%) suggests very few condition cases were mistaken as control. Combined with the condition recalls ~70%, it implies most errors might be between anxiety and depression themselves, rather than confusions with control. For instance, an anxious post might sometimes be classified as depression and vice versa, which would lower recall for each without necessarily raising control's false positives much.

Overall, the text-based multiclass model effectively leverages differences in language used by people with anxiety vs depression.

Coefficients for each of the most significant words can be obtained, aiding in the understanding and analysis of the results provided above:

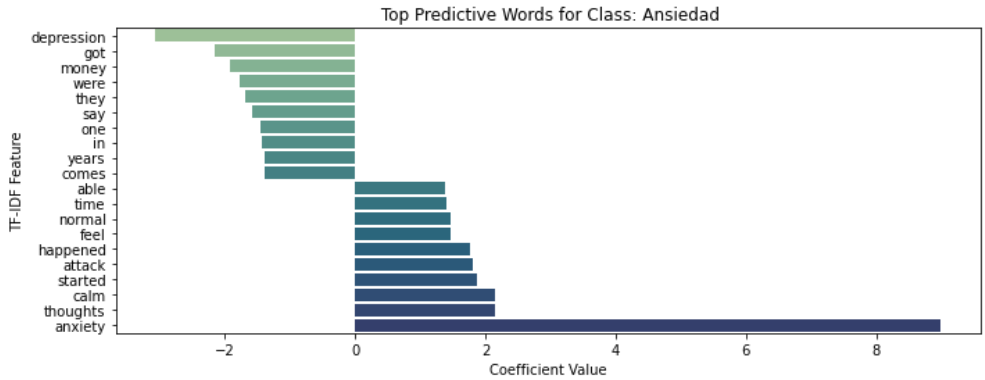

Figure 42: Top Predictive Words for Anxiety Class in Model 2

Interesting to note in the graphical representation of the Anxiety class above the inverse relation between depression and anxiety in this regard, in clear contrast with the binary model studied in the previous section of this chapter.

Observed again in the graph for the Depression class:

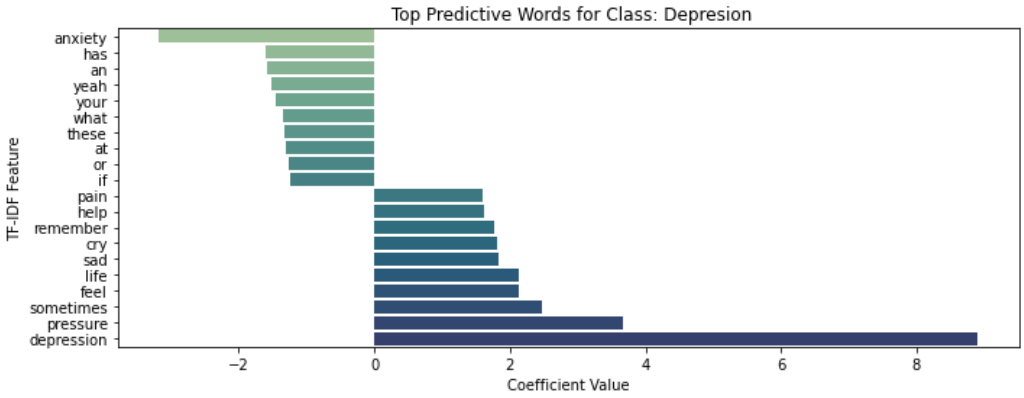

Figure 43: Top Predictive Words for Depression Class in Model 2

But in clear contrast to the Control class, where both anxiety and depression words are clearly negative correlated important indicators:

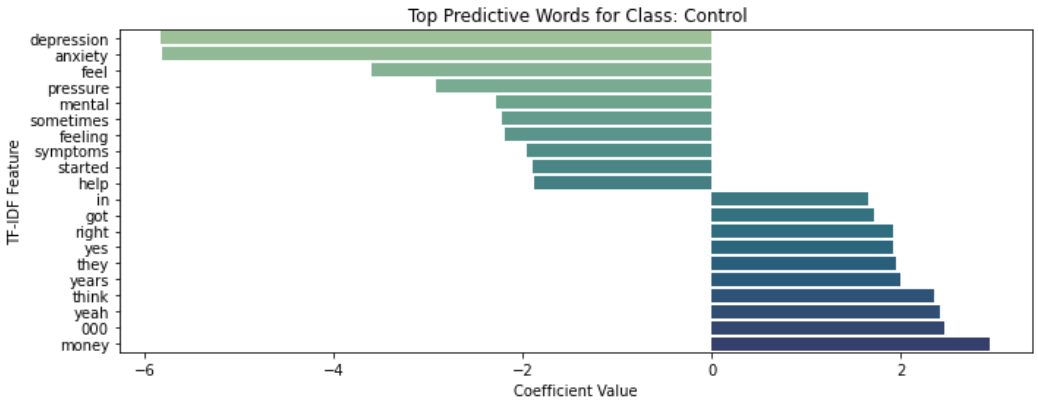

Figure 44: Top predictive Words for Control Class in Model 2

Along this detailed analysis, it is important to delve into nalyzing predicted probabilities distribution and relationships in the multiclass model:

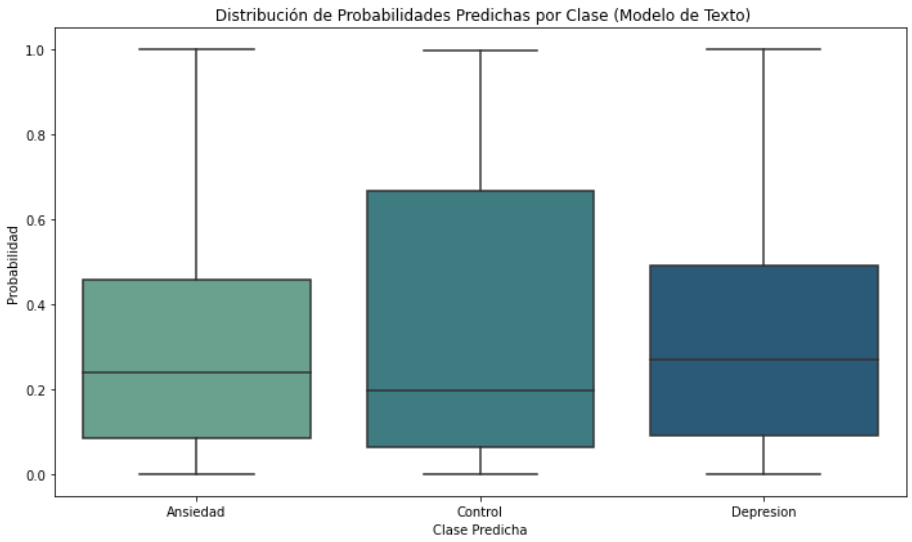

Figure 45: Probability Distribution per Class for Model 2, Multiclass Logistic Regression Text-Only Features

The predicted probabilities were melted and plotted in the boxplots shown above, to see how well-separated the predicted probabilities are for the true classes. The distribution of predicted probability for "Control" is generally high for actual controls and low for actual conditions, showing he marginal distribution of each class's probability over the whole test set.

Also the mean distribution of probabilities for each class is displayed below:

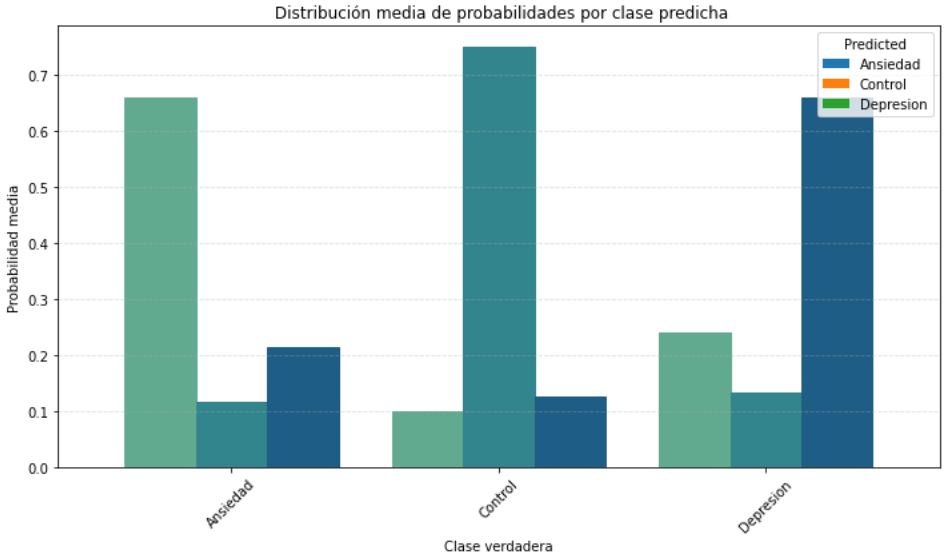

Figure 46: Mean Distribution Probability for Model 2, Multiclass Logistic Regression Text-Only Features

This enables to see how "pure" each predicted class is. It shows that predictions are generally very confident, specially (as could be expected) for the control class.

A PCA is also applied in the following visual:

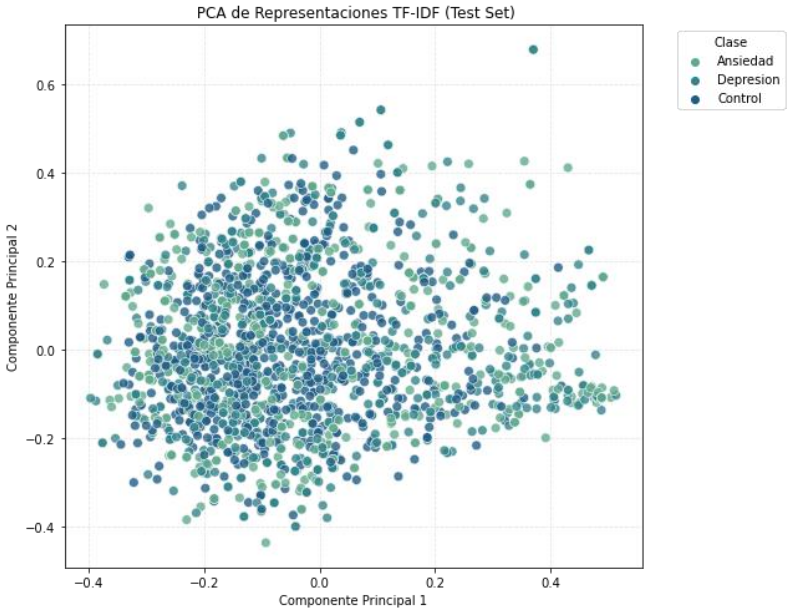

Figure 47: PCA for Model 2, Multiclass Logistic Regression Text-Only Features

The PCA above, allows visualization of how well the three classes separate when transformed onto the first two principal components, which explain the maximum variance in the high-dimensional TF-IDF word frequency space. From the plot, it becomes evident that there is substantial overlap among the three classes in the PCA plane. Although some clustering tendencies are weakly visible—such as a slightly denser grouping of Control (blue) samples toward the central region and marginal spreading of Anxiety (green) and Depression (darker blue) toward the peripheries, there is no clear decision boundary in the reduced-dimensional space, meaning that this model, despite its successful results does not inherently form well-separated clusters.

This is not necessarily extremely relevant because PCA is a linear technique that captures global variance, not necessarily the discriminative directions optimal for classification. The observed mixing reinforces the notion that a more complex model or higher-dimensional feature interaction is necessary, justifying the next step of incorporating the numerical features once again into the model.

8.2.3 HYBRID LOGISTIC REGRESSION

Now, both numeric and text features are combined to see if classification of anxiety vs depression vs control can be further improved. Once again, a ColumnTransformer pipeline is used and the Logistic Regression is performed, obtaining the results observable in the following visuals:

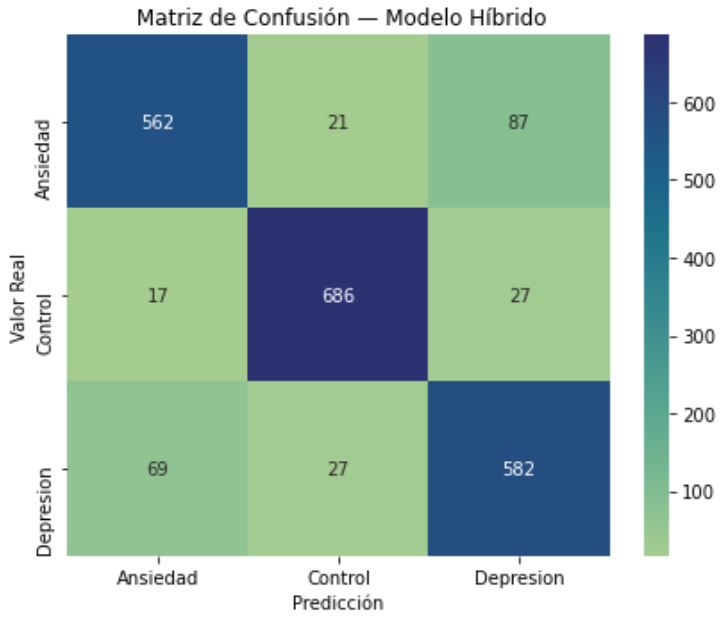

Figure 48: Confusion Matrix for Multiclass Model 3, Hybrid Logistic Regression

The confusion matrix summarizes the prediction outcomes for each class. Control had the highest true positive count (686/730), reflecting its linearly separable nature. Anxiety and Depression show some confusion with one another, which is not only expected given their overlapping symptomatology, but also massively reduced from the previous multiclass models, showing overall balanced sensitivity across classes.

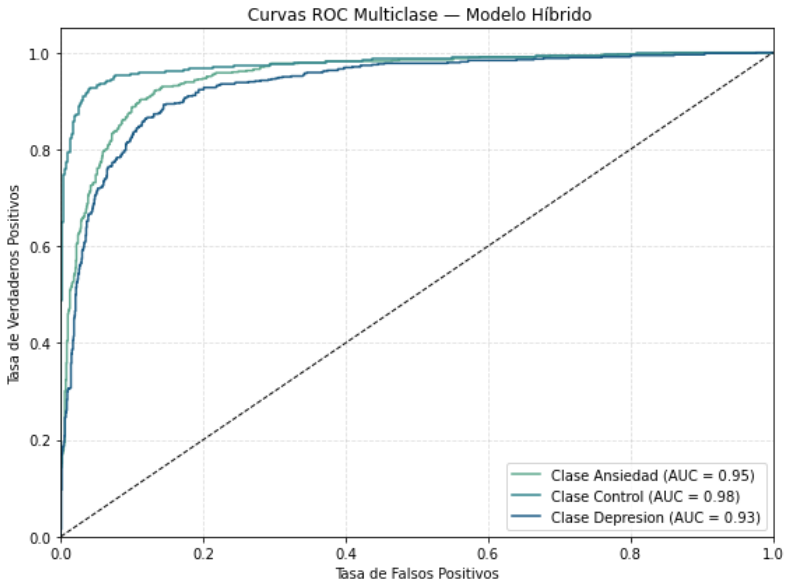

Figure 49: ROC Curves for Multiclass Model 3, Hybrid Logistic Regression

The ROC curve presents class-wise performance with AUC values of 0.95 (Anxiety), 0.98 (Control), and 0.93 (Depression). These excellent AUCs confirm the model's strong sensitivity and specificity across all three classes. The nearly perfect AUC for Control indicates that this class is particularly well-separated from the others, a plausible result given that "normal" speech and emotion patterns are expected to diverge more starkly from those with mental health struggles.

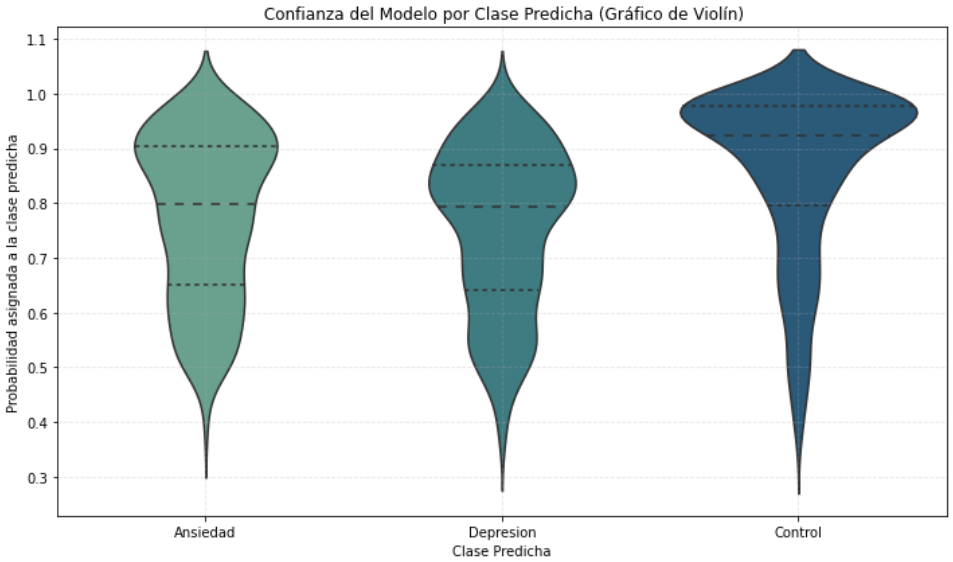

Figure 50: Violin Plots for Multiclass Model 3, Hybrid Logistic Regression

The violin plot shows, as previously explained, the distribution of prediction probabilities assigned to each class. Notably, the predicted probabilities for Control appear tightly clustered near high values (suggesting high confidence), while distributions for Anxiety and Depression are broader, but still clearly distinct. This spread implies that while Control predictions are made with higher certainty, the model still effectively distinguishes between the pathological classes with slightly lower but acceptable confidence dispersion.

Figure 51: Radar Chart for Multiclass Model 3, Hybrid Logistic Regression

The radar plot portrayed above shows the relative magnitude and orientation of selected features across the three target classes. The distinct patterns formed by each class—especially the higher values for Anxiety and Depression in certain axes—suggest that different traits and patterns are well captured by the hybrid mode, demonstrating the model's ability to differentiate between mental health conditions by leveraging multiple data modalities

These results can be, therefore, summarized in the following table:

	Precision	Recall	F1-Score	Support
Anxiety	0.87	0.84	0.85	670

Control	0.93	0.94	0.94	730
Depression	0.84	0.86	0.85	678
Accuracy			0.88	2078
Macro Avg	0.88	0.88	0.88	2078
Weighted Avg	0.88	0.88	0.88	2078

Table 19: Summary of Results for Multiclass Model 3, Hybrid Logistic Regression

As can be clearly interpreted, the hybrid multiclass model is the most effective among all tested architectures in this project. It leverages both quantitative biomarkers and textual semantics to create a balanced, robust model that performs exceptionally well on all metrics, validating its usability in real-world settings.

8.2.4 FEATURE SUBSET EXPERIMENTS FOR HYBRID SUBGROUPS

To further interpret the hybrid model, a set of experiments are conducted by training hybrid models that combine text with only a subset of numeric features. An ample variety of subgroups are defined to do a detailed, minimalistic, insightful analysis of how different subsets of data (and their combinations) affect model performance:

These subgroups were trained and tested, producing the following results:

- Using facial emotions and text:

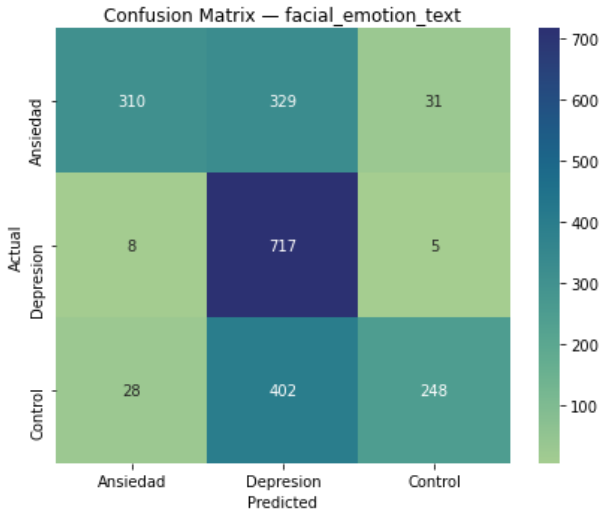

Figure 52: Confusion Matrix, Feature Subset: Facial Features

	Precision	Recall	F1-Score	Support
Anxiety	0.9	0.46	0.61	670
Control	0.5	0.98	0.66	730
Depression	0.87	0.37	0.52	678
Accuracy			0.61	2078
Macro Avg	0.75	0.6	0.59	2078
Weighted Avg	0.75	0.61	0.6	2078

Table 20: Summary of Results, Feature Subset: Facial Features

Anxiety F1 = 0.61, which is much lower than text-only (0.74). Interestingly, adding only facial features to text seemed to hurt performance on anxiety detection or at least was insufficient to maintain text performance. The result likely means that just facial cues and text (excluding voice and movement) didn't capture enough, and perhaps even introduced noise – the model might have overfitted facial signals or the lack of voice data caused it to misclassify some that text alone would have gotten. Actually, F1 0.61 for anxiety is

significantly below 0.74, indicating a worse model than text-only. It suggests that leaving out voice and movement data while including facial data somehow confused the model or didn't help. Possibly many anxiety vs depression differences live in voice or movement rather than static facial expression.

- Using voice acoustics and text:

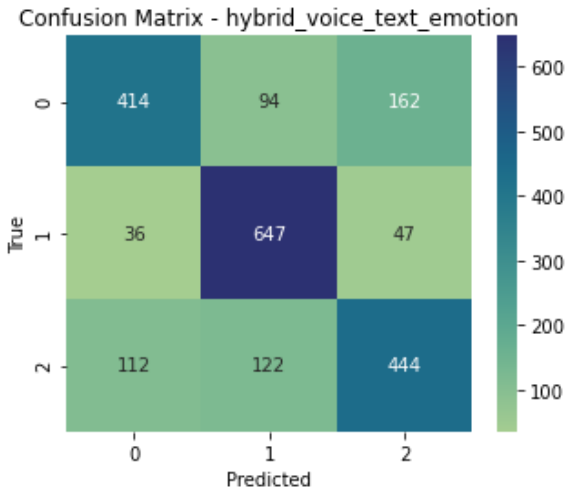

Figure 53: Confusion Matrix, Feature Subset: Voice Acoustics + Text

	Precision	Recall	F1-Score	Support
Anxiety	0.74	0.62	0.67	670
Control	0.75	0.89	0.81	730
Depression	0.68	0.65	0.67	678
Accuracy			0.72	2078
Macro Avg	0.72	0.72	0.72	2078
Weighted Avg	0.72	0.72	0.72	2078

Table 21: Summary of Results, Feature Subset: Voice Acoustics + Text

Using voice acoustics and text, better results than with facial are obtained, but it still produces worse results than only text. So, voice acoustics alone with text improves over facial alone with text, implying voice acoustic features carry more useful info than facial, but still not enough to beat using all modalities or simply text.

- Using movement biometrics and text

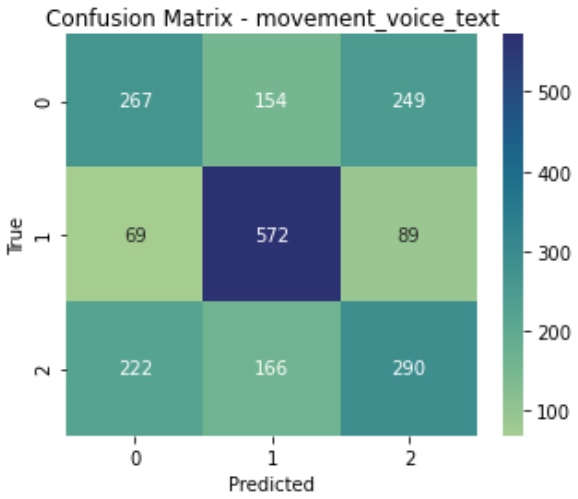

Figure 54: Confusion Matrix, Feature Subset: Movement Biometrics + Text

	Precision	Recall	F1-Score	Support
Anxiety	0.48	0.40	0.43	670
Control	0.64	0.78	0.71	730
Depression	0.46	0.43	0.44	678
Accuracy			0.54	2078
Macro Avg	0.53	0.54	0.53	2078
Weighted Avg	0.53	0.54	0.53	2078

Table 22: Summary of Results, Feature Subset: Movement Biometrics + Text

The combination of these movement biometrics and text yields clearly the worst results. This indicates that using only movement and biometric signals in addition to text severely underperformed, likely because those features alone (without facial/voice) did not provide much distinguishing power. It might have even distracted the model or left it basically relying on text and some weak movement cues. It's possible many movement features might be irrelevant or too noisy to help classification, so excluding the more useful face/voice while leaving movement hurts performance.

A variety of hybrid subsets were also used. For example, a hybrid_text_centric model was created, including the top indicators: 10 core facial/emotional indicators, 10 psychological traits, 10 voice+audio emotion, 4 features of textual sentiment and tense indicators and the top 16 markers of text. The results were the following:

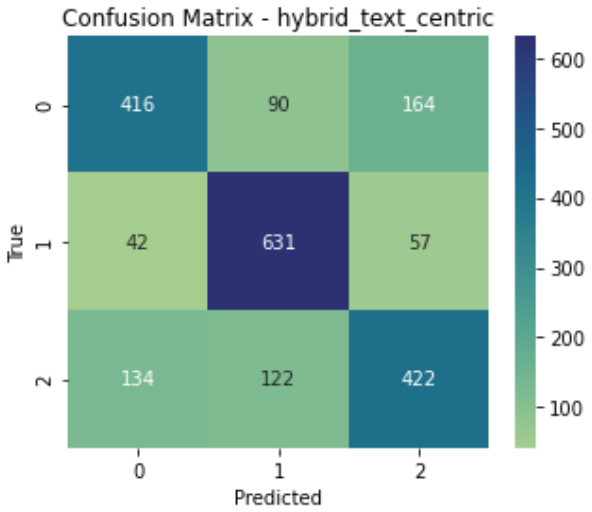

Figure 55: Confusion Matrix, Feature Subset: Core Hybrid

	Precision	Recall	F1-Score	Support
Anxiety	0.7	0.62	0.66	670
Control	0.75	0.86	0.8	730
Depression	0.66	0.62	0.64	678

Accuracy			0.71	2078
Macro Avg	0.7	0.7	0.7	2078
Weighted Avg	0.7	0.71	0.7	2078

Table 23: Summary of Results, Feature Subset: Core Hybrid

While this hybrid approach showed an improvement in comparison to using any subgroup on its own with text, it is still clearly far from the best results obtained using the hybrid multiclass models with all textual and numerical features.

In line with this, this concept was slightly improved and simplified incorporating all text markers and only the top 5 non-text features:

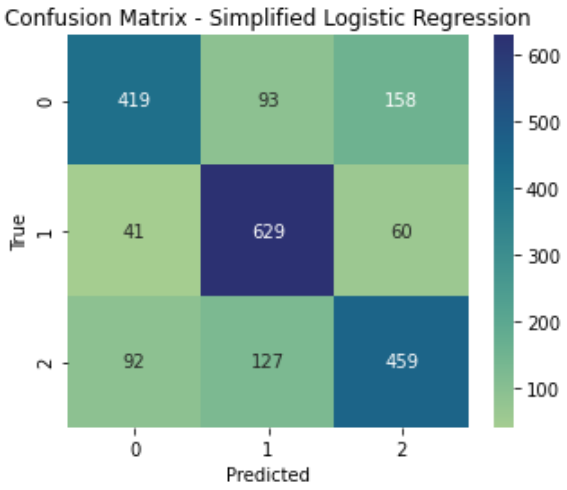

Figure 56: Confusion Matrix, Feature Subset: Top 5 + Text

	Precision	Recall	F1-Score	Support
Anxiety	0.76	0.63	0.69	670
Control	0.74	0.86	0.8	730
Depression	0.68	0.68	0.68	678

Accuracy			0.73	2078
Macro Avg	0.73	0.72	0.72	2078
Weighted Avg	0.73	0.73	0.72	2078

Table 24: Summary of Results, Feature Subset: Top 5 + Text

This slightly optimizes the previous result while simplifying the underlying model. However, the results obtained were clearly less optimate than the ones obtained with all features in the hybrid approach.

8.2.5 SUMMARY OF MULTICLASS FINDINGS

- Text features alone provide a strong baseline for classifying mental health content into Control, Anxiety, or Depression, achieving 77% accuracy with particularly strong performance on identifying control vs condition. Language cues distinguish the conditions reasonably well (F1 ~0.7 for each).
- Numeric features alone are insufficient for fine discrimination between anxiety and depression (accuracy 58%, with conditions often confusing), although they do separate control from any condition to a degree. This highlights that textual expression carries distinct information about the type of mental health issue that raw behavior signals by themselves do not as clearly convey.
- Combining modalities yields a synergistic effect, raising accuracy to 88% and bringing both precision and recall for each class into the mid-80s or above. The multimodal logistic regression can accurately tell apart anxious vs depressed individuals in most cases, far better than either modality alone.
- Different modalities contribute differently: voice features appear to add the most unique value (improving classification when added to text), whereas facial features add some value but less, and movement/biometric features in this dataset add minimal value on their own. Nonetheless, using all together gave the best result, meaning each modality is likely to contribute some complementary cues for at least certain instances.

8.3 REFLECTIONS AND IMPLICATIONS

Using logistic regression provided a solid baseline and an interpretable framework, providing an easy way of extracting the relationships (coefficients) to explain the model learned. In addition, handling the multiclass classification with the multinomial logistic approach ensured consistent probabilities across the three classes. Given he fairly balanced class distribution and the fact that classes are mutually exclusive, the multinomial approach was appropriate and simpler to interpret, which was a good decision for the model, which performed excellently.

Regarding feature engineering, the models mainly relied on existing extracted features and TF-IDF for text, both proving to be very effective.

Implications gathered from this analysis:

- Feasibility: It is indeed feasible to automatically detect signs of anxiety and depression from user-generated content with high accuracy, using a combination of linguistic and non-linguistic cues. The high performance of the multimodal model suggests that an automated system could flag users who show mental distress proactively.
- Multimodal advantage: Adding modalities significantly enhances the app's ability to correctly categorize the type of distress observed. This is, for example, particularly relevant to consider if a user is not specially revealing.
- Anxiety vs Depression indicators: Our analysis highlights that language usage differs between anxiety and depression in systematic ways that a model can learn. This could be of interest to psychologists and linguists – for instance, confirming that use of certain phrases or words is strongly associated with one condition. Similarly, differences in vocal or facial expression patterns between anxiety and depression, if teased out from the model, could contribute to the clinical understanding of these conditions.

- Early detection and monitoring: A logistic regression model is fast to compute; it could run in real-time on new posts or video streams and update a probability of someone being in distress.
- False positives/negatives handling: For instance, when the model confuses anxiety for depression, the intervention (if any) should be general enough to cover both possibilities. The relatively low confusion of the optimized hybrid multiclass model means the risk is greatly reduced but not zero. Meanwhile, false positives could cause unnecessary concern or invasion of privacy. However, the optimal model's high precision massively mitigates this risk.

ANNEX III: PREDICTIVE MODELING

GITHUB WITH ALL CODES: https://github.com/carlossanchezcabezudo/TFG

8.4 HIGH GRADIENT BOOSTING

In this section, High Gradient Boosting Classifier, which is an implementation of histogram-based gradient boosting, is applied to the dataset for the anxiety-depression classification task once cleansed like explained in the previous chapter (identical column-dropping, preprocessing, etc.). All feature variables are standardized to zero-mean and unit-variance using a Standard Scaler fitted on the training data in consistency with other models, even though tree-based models like gradient boosting are not sensitive to scaling.

The baseline HGB model is trained on the scaled training data without hyperparameter adjustments, using default parameters, which include an automatic early-stopping mechanism and 100 boosting iterations. Training completes quickly thanks to histogram-based splitting and then it is possible to evaluate the results obtained. These results are portrayed in the following illustrations:

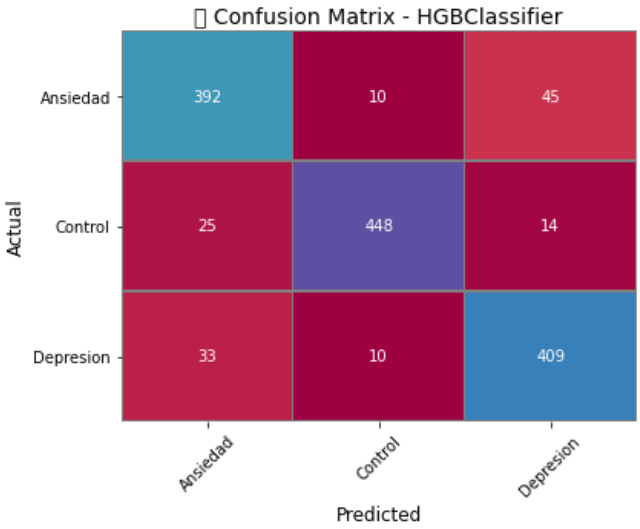

Figure 57: Baseline HGB Classifier, Confusion Matrix

To better understand the model's errors, a confusion matrix is plotted, as shown above. Most test samples lie on the diagonal (correctly classified), with relatively few off-diagonal errors. Notably, the Control instances form a distinct block on the matrix's third row/column – very few control subjects are misclassified as having a disorder, and vice versa. Among the misclassifications that do occur, the primary confusion is between the Anxiety and Depression classes. his pattern is expected given the related symptomatology of anxiety and depression; the model occasionally confuses these two conditions when their feature profiles are similar.

Nonetheless, the confusion matrix confirms that the number of such errors is considerably low and provides an insight into the model's paramount effectiveness in detecting anxiety and depression.

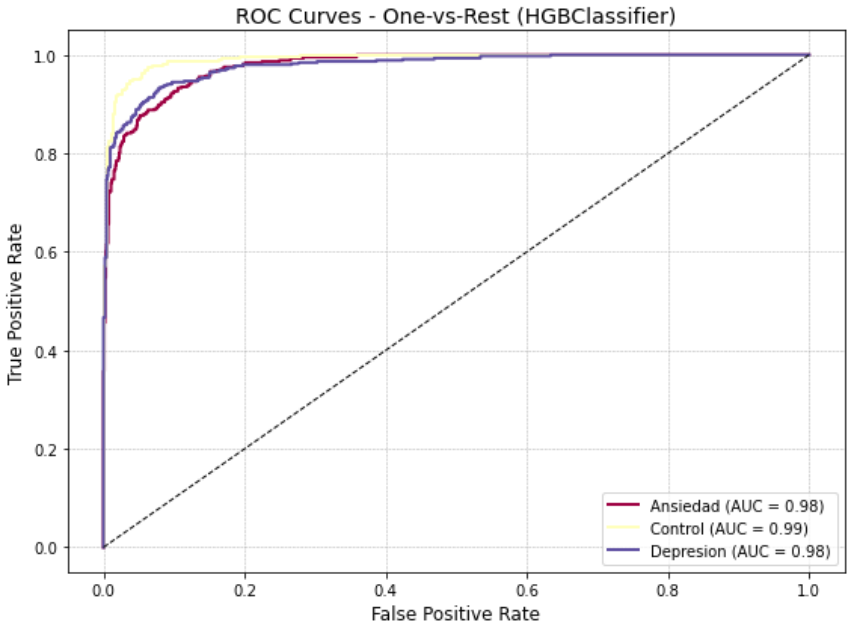

Figure 58: Baseline HGB Classifier, ROC Curves

The ROC curve above, portrayed in a one-vs-rest-scheme, shows that all classes have extremely high AUC values. For instance, the curve for the Control class rises sharply toward the top-left, reflecting that controls consistently receive much lower predicted probabilities of belonging to a disorder class. Similarly, the Anxiety and Depression curves are well above the diagonal line, meaning the model assigns higher probabilities to the true class in most cases. These ROC results corroborate the earlier accuracy findings: the HGB classifier not only makes correct hard classifications but also producing well-calibrated probability outputs that rank true labels ahead of false ones in the vast majority of instances. The near-unity AUC scores suggest that, if a threshold other than the default 0.5 were chosen, one could still achieve a high true positive rate for a given false positive rate – a desirable property for any screening tool.

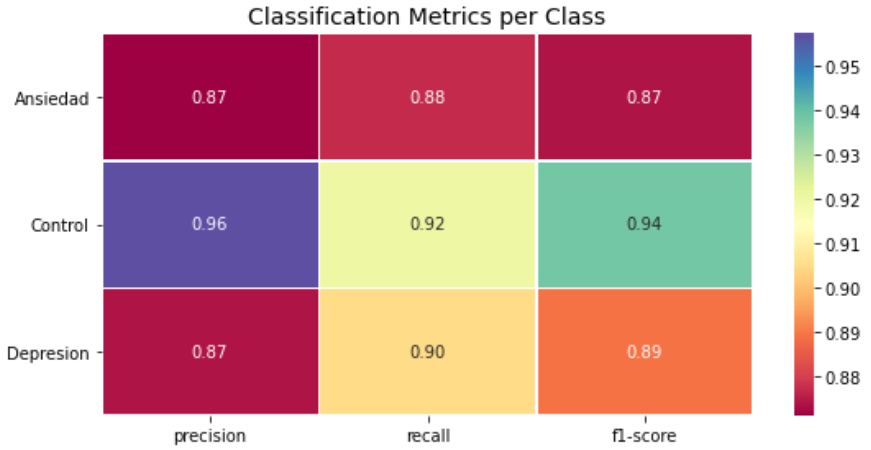

Figure 59: Baseline HGB Classifier, Classification Metrics per Class

The strong results are specifically shown in the above heatmap of per-class metric. The Control class shows marginally higher performance, but the values are relatively uniform, implying the model treats all classes equitably. In general, the conclusion that the HGB model yields balanced performance across the three classes, with no severe weakness on any class, is reinforced.

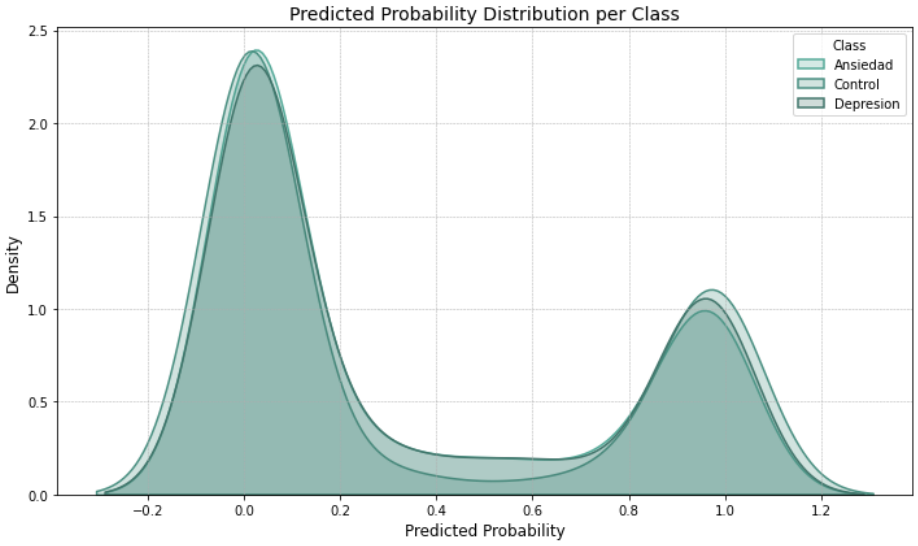

Figure 60: Baseline HGB Classifier, Predicted probability Distribution per Class

Another interesting insight, provided by the above graph is to examine the distribution of predicted probabilities, which provides understanding of the model's confidence levels.

The resulting KDE (kernel density estimate) curves are distinctly bimodal for each class: one peak near 0 and another near 1. This pattern indicates that the model typically assigns high confidence to one class per sample (close to 1.0 probability for the predicted class and near 0.0 for the others). For example, the density for "Depression" has a strong peak at probability≈0 for non-depressed individuals and a separate peak near 0.9–1.0 for those the model identifies as depressed. Similarly, "Control" probabilities are usually very low for patients and very high for actual controls. These well-separated distributions suggest the model is making decisions with high certainty in most cases.

There is a relatively small mass in intermediate probability ranges (e.g., around 0.5), meaning the classifier rarely outputs ambiguous, uncertain predictions. This behavior is desirable in a diagnostic setting because a decisive model with accurate decisions is preferred. Overall, the probability distribution plot underscores that the HGB classifier not only performs well but does so with a clear degree of confidence for most instances.

These results can be summarized in the following table:

	Precision	Recall	F1-Score
Anxiety	0.87	0.88	0.87
Control	0.96	0.92	0.94
Depression	0.87	0.90	0.89
Accuracy			0.9
Macro Avg	0.9	0.9	0.9
Weighted Avg	0.9	0.9	0.9

Table 25: Summary of Results, Baseline HGB Classifier

Having established a strong baseline, then, an optimization of the HGB's hyperparameters is done to see if performance can be further improved or model complexity reduced

without the loss of accuracy. A randomized hyperparameter search is conducted over key parameters of the HGBClassifier. The parameters tuned include the learning rate, the maximum tree depth and number of leaf nodes, the minimum samples per leaf, and the L2 regularization strength. 30 random combinations of hyperparameters are sampled, each evaluated via 3-fold cross-validation on the training set.

Using these best parameters, a tuned HGB model is refit on the entire training set and evaluated. The results are portrayed in the following confusion matrix:

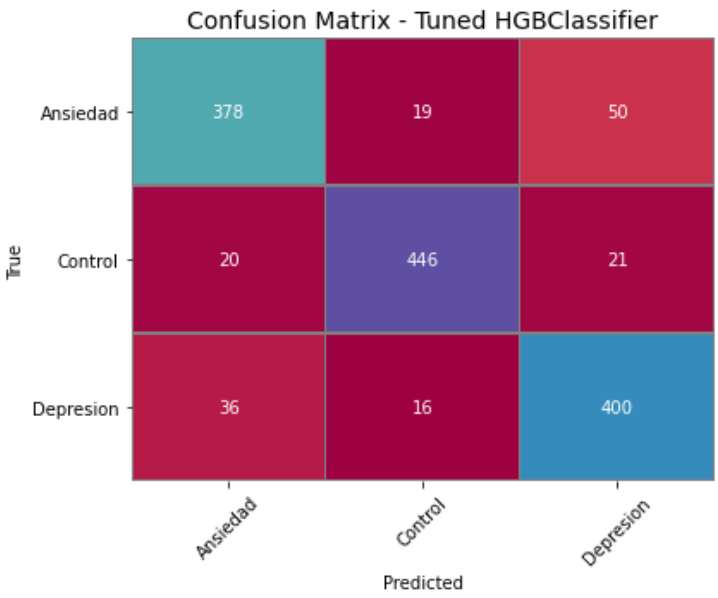

Figure 61: Tuned HGB Classifier, Confusion Matrix

While the results gathered are similar than the baseline confusion matrix, a slight reduction in Anxiety recall (from 0.88 to 0.85) is seen as a few additional anxiety cases being classified as depression in the tuned model compared to before, while the Control class continues to show very few errors. Thus, the overall confusion structure is consistent, confirming that the tuned model did not introduce any new systematic errors; it simply made a few more conservative predictions in borderline cases.

These results can be summarized in the following table:

	Precision	Recall	F1-Score
Anxiety	0.87	0.85	0.86
Control	0.93	0.92	0.92
Depression	0.85	0.88	0.87
Accuracy			0.88
Macro Avg	0.88	0.88	0.88
Weighted Avg	0.88	0.88	0.88

Table 26: Summary of Results, Tuned HGB Classifier

The tuned model attains 88.31% accuracy on the test data, which is slightly lower than the baseline's 90.12%. The class-by-class metrics change only marginally: for instance, Anxiety is predicted with precision 0.87 and recall 0.85, Depression with precision 0.85 and recall 0.88, and Control with precision 0.93 and recall 0.92.

These values are within a few percentage points of the baseline model's metrics. The small dip in overall accuracy (from 0.90 to 0.88) likely reflects the fact that the baseline model was already very well-tuned by default, and the hyperparameter search introduced slightly more bias (to guard against overfitting) at the expense of a tiny loss in raw accuracy. In other words, the default HGB settings were near-optimal for this problem, and the randomized search found a configuration that is perhaps more conservative.

he tuned model's performance is essentially on par with the baseline when considering the uncertainty inherent in the test set – a difference of around 1.8 percentage points in accuracy may not be statistically significant given the sample size. This outcome underscores an important point: the HGB algorithm performed strongly out-of-the-box, and heavy hyperparameter tuning was not strictly necessary to achieve high accuracy. Nonetheless, the tuned model is preferable for deployment because its hyperparameters

(e.g., shallower depth, regularization) may make it more robust to new data, even if it sacrificed a minor amount of fit to the current test set.

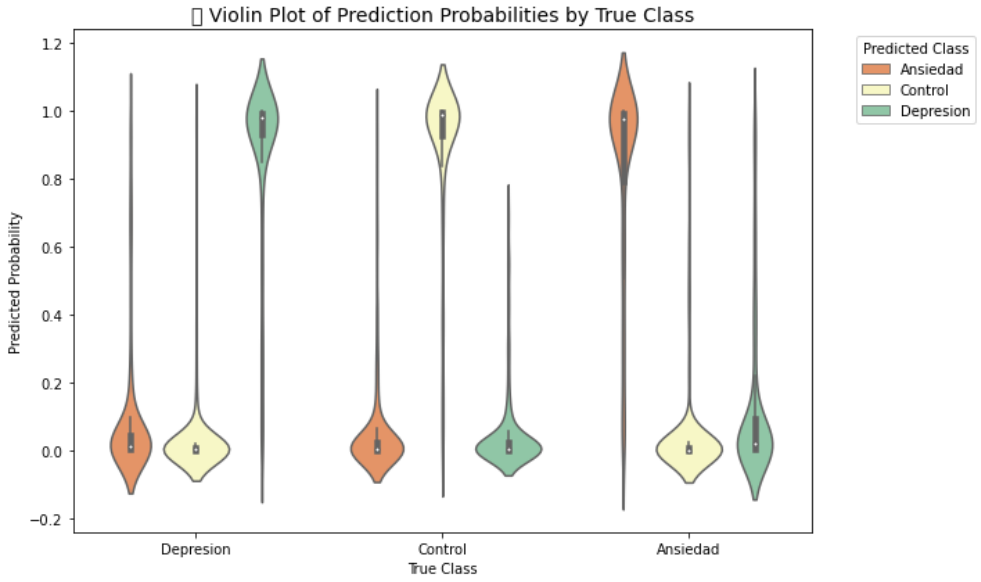

Figure 62: Tuned HGB Classifier, Violin Plot Per Class

To gain insight into the model's probabilistic outputs after tuning, the violin plot displayed above is included. This plot illustrates how the tuned HGB model allocates probability mass for each true category. It can be observed that for any given true class, the probability assigned to that same class is usually very high. For example, looking at the violins for true Depression cases, the distribution of predicted probability for "Depression" (hue) is concentrated towards the upper range (often 0.8–1.0), whereas the probabilities for "Anxiety" or "Control" are near 0 for those same cases. A similar pattern holds for true Anxiety and Control cases. The confidence in this model can also be seen in its distribution:

Figure 63: Tuned HGB Classifier, Model Confidence for Each True Class

Because of this, it can be safely argued that the tuned HGB is well-calibrated in its confidence: when it predicts a class, it usually does so with a high probability, and those high probabilities typically occur for instances of the true class. The probability for non-true classes remains low in most cases, implying a clear separation in predicted score between the actual class and the alternatives.

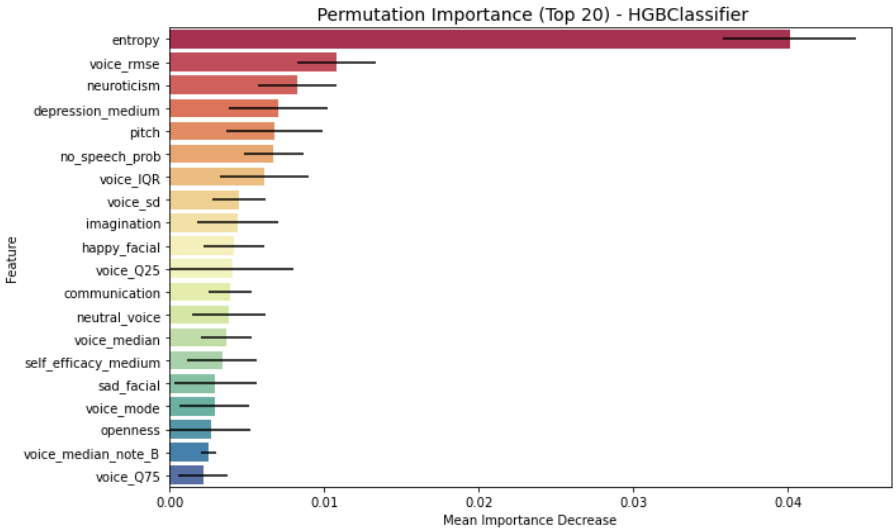

Figure 64:Tuned HGB Classifier, Permutation Importance (Top 20)

A permutation importance analysis is also performed, as illustrated in the previous image, on this tuned HGB model. This involves randomly shuffling each feature's values among test instances and measuring the drop in model performance. From this analysis, certain ey predictors can be seen. For instance, one of the highest-ranked features is Neuroticism (a personality trait known to be higher in anxiety and depression cases), indicating that individuals' neuroticism scores heavily influence the classifier's decisions. Similarly, features capturing negative emotional tone in voice (e.g., the frequency of sad or anxious words, or a low vocal pitch variability) are among the top contributors. On the other hand, many of the one-hot encoded categorical features (for example, the language of the interview) show minimal importance. These feature-importance results highlight that the HGB model predominantly relies on meaningful psychological signals, rather than noise.

In conclusion, the Histogram Gradient Boosting classifier demonstrates excellent performance in multiclass anxiety-depression classification. With nearly 90% accuracy, strong precision/recall balance, and well-calibrated probability estimates, it stands out as a powerful method. The model required minimal tuning to perform optimally, and it affords some interpretability through feature importance.

8.5 RANDOM FOREST CLASSIFIER

In this section a Random Forest ensemble is employed to perform the same three-class classification. Random Forests, like gradient boosting, consist of multiple decision trees; however, they average the results of many fully grown trees (with bootstrap aggregation and feature randomness) rather than sequentially boosting performance. As before, stratification is used and the standard scaling normalization is applied to the features (though scaling is not strictly required for tree models, it was done for consistency across approaches).

First of all, the baseline Random Forest Classifier, using the full feature set is trained and modelled, providing the following results:

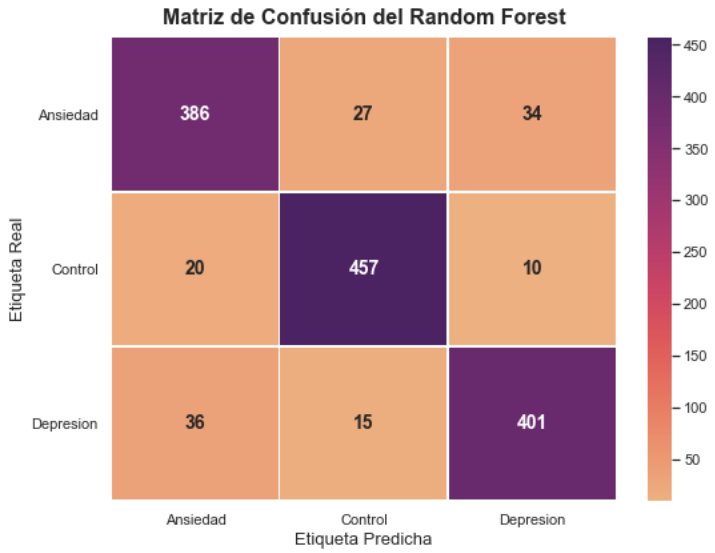

Figure 65: Random Forest Classifier, Confusion Matrix

As can be seen above, the confusion matrix illustrates strong performance across all three classes: Ansiedad (Anxiety), Control, and Depresion (Depression). The diagonal dominance is evident, with 386 true positives for Anxiety, 457 for Control, and 401 for Depression. Misclassifications are limited in volume and do not indicate systematic bias. The matrix suggests, once again, that the model is particularly strong at distinguishing the Control group from pathological conditions. However, the overall performance was excellent and very few instances of Anxiety (61 in total) and Depression (51) were misclassified.

These findings can be summarized in the following table:

	Precision	Recall	F1-Score
Anxiety	0.87	0.86	0.87
Control	0.92	0.94	0.93
Depression	0.9	0.89	0.89
Accuracy			0.9

Macro Avg	0.9	0.9	0.9
Weighted Avg	0.9	0.9	0.9

Table 27: Summary of Results: Random Forest Classifier

This classification report confirms the matrix results with a macro-average F1-score of 0.90, indicating balanced and strong predictive power across classes. Precision and recall for anxiety are nearly equal (0.87 and 0.86 respectively), suggesting the model does not overpredict nor underpredict this class disproportionately. Similarly, for depression the model still maintains high performance (F1-score of 0.89). The highest-performing class was control, with 0.92 precision and 0.94 recall, indicative of a highly confident and accurate prediction pattern, again showcasing the higher distinctiveness of the control group features in contrast with the two clinical classes.

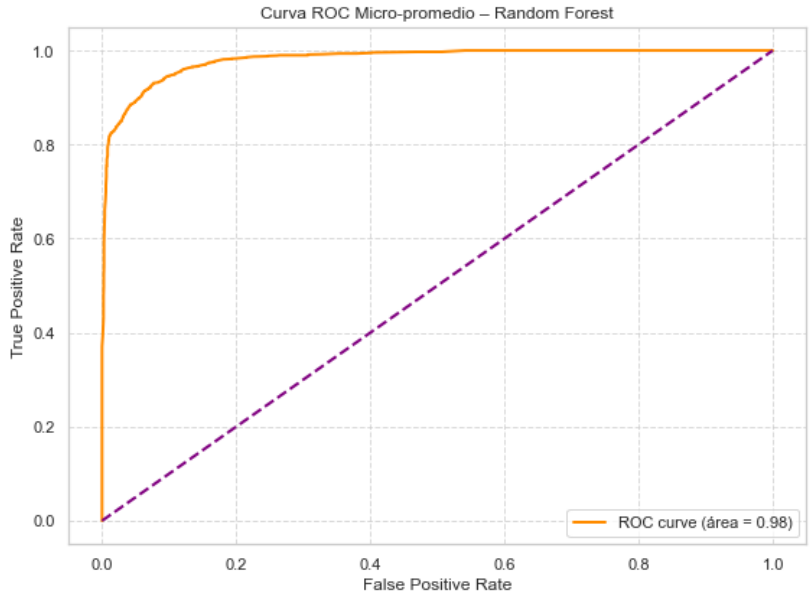

Figure 66: Random Forest Classifier, ROC Curve

The above ROC curve shows the overall reliability of the model, plotting predicted probability bins against observed empirical probabilities, reveals a mostly well-calibrated model. The alignment of the Random Forest's reliability curve with the ideal diagonal

suggests that the probabilistic outputs are meaningful and can be interpreted directly as confidence levels. A more detailed analysis can be inferred from the following ROC:

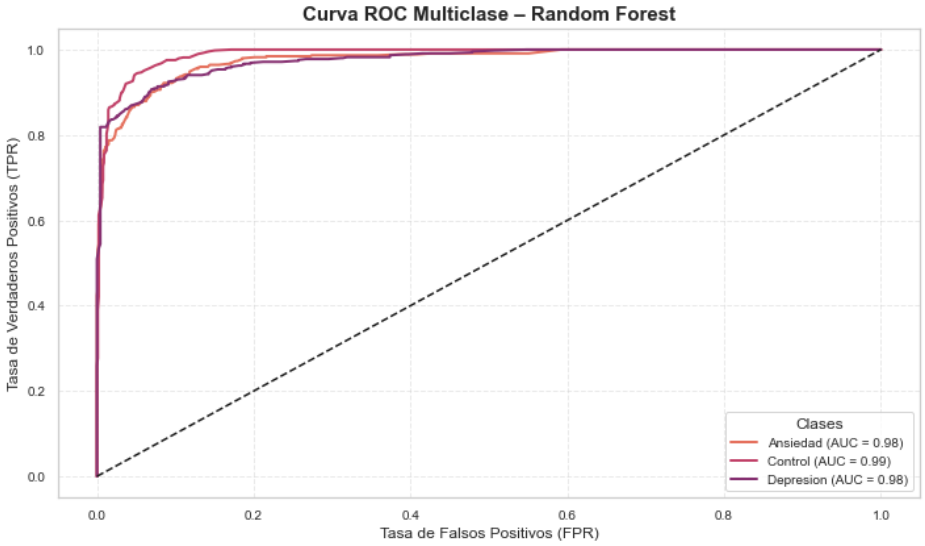

Figure 67: Random Forest Classifier, Multiclass ROC Curves

A multiclass ROC Curve is also included, which reinforces the robustness of the classifier. Each class achieves an AUC close to or above 0.98, indicating near-optimal discrimination capabilities. These AUC scores are exceptional and imply that the model's classification threshold can be adjusted flexibly without significant loss of sensitivity or specificity, supporting the model's adaptability to different application contexts, like prioritizing sensitivity in clinical settings.

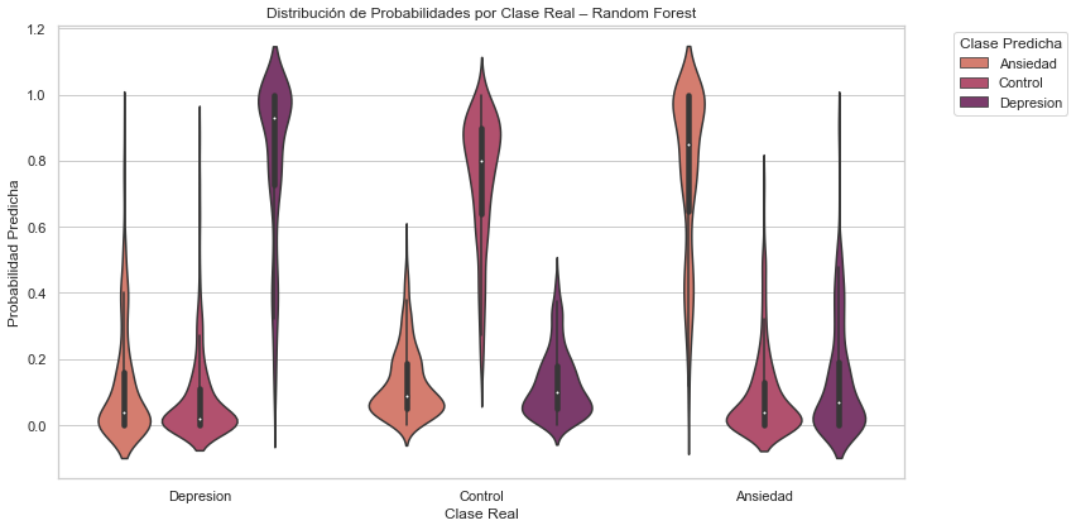

Figure 68: Random Forest Classifier, Probability Distribution for Each True Class

The violin plot mapping above shows predicted probabilities for each predicted class against true labels and illustrates another level of model confidence and separability. Each real class—Depression, Control, and Anxiety—is associated with a concentrated and well-separated predicted probability density for its corresponding class. This is particularly evident in the Control group, where predictions sharply peak at probabilities near 1 for the correct class and are nearly negligible for the others. Anxiety and Depression, while slightly more dispersed, still exhibit distributions skewed towards their respective labels with minimal cross-contamination.

This visualization emphasizes the model's high confidence, particularly in the Control group, but also reveals the subtle overlapping areas between Depression and Anxiety, a pattern consistent with known clinical and linguistic symptom overlaps. It suggests that while the model is able to distinguish between them to a significant extent, their latent feature spaces are not entirely orthogonal.

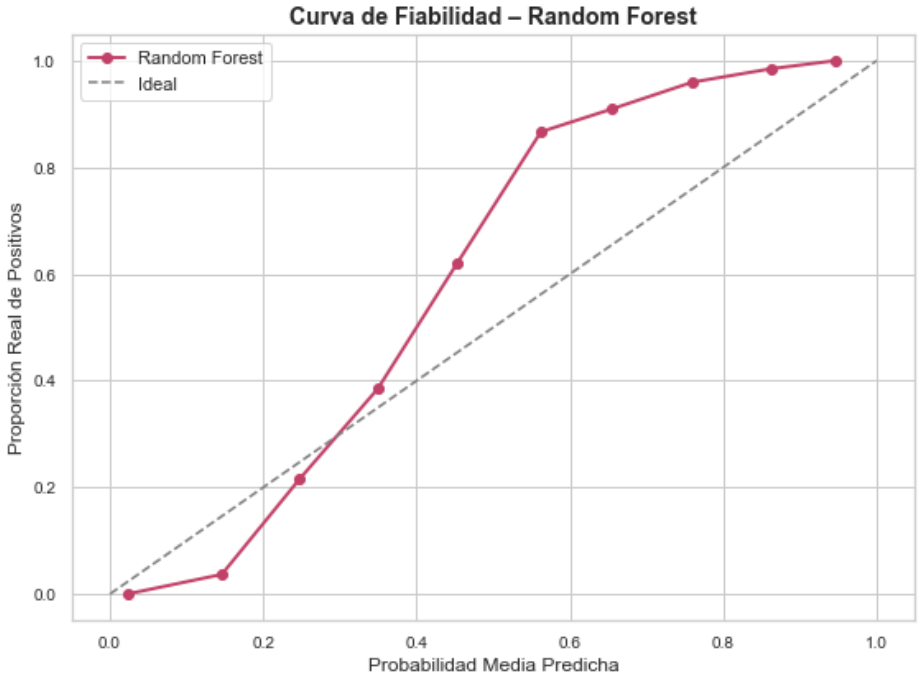

Figure 69: Random Forest Classifier, Reliability Curve

Finally, the previous image showcases a reliability curve, which plots predicted probability bins against observed empirical probabilities, reveals a mostly well-calibrated model. The alignment of the Random Forest's reliability curve with the ideal diagonal suggests that the probabilistic outputs are meaningful and can be interpreted directly as confidence levels. However, some deviations in the lower probability ranges (e.g., the 0.1–0.3 bins) indicate slight under confidence, where predictions were too cautious and the actual frequency of the positive class was slightly higher than predicted. Nonetheless, the calibration in the higher probability range is notably accurate, validating the model's output for practical decision-making.

Training a baseline Random Forest on the full feature set yields robust performance, but the model's complexity (with hundreds of features after one-hot encoding) raises the question of whether all features are necessary. To investigate this, an interactive feature-importance analysis is combined with model re-training. First, a Random Forest is fitted on the full feature set, and the algorithm's inherent feature importance (based on Gini impurity decrease) is extracted. This reveals which features the ensemble found most

discriminative. Many features have near-zero importance, suggesting they contribute little to decision-making. To reduce dimensionality and potentially improve generalization, the top 15 features are selected by importance from this initial model:

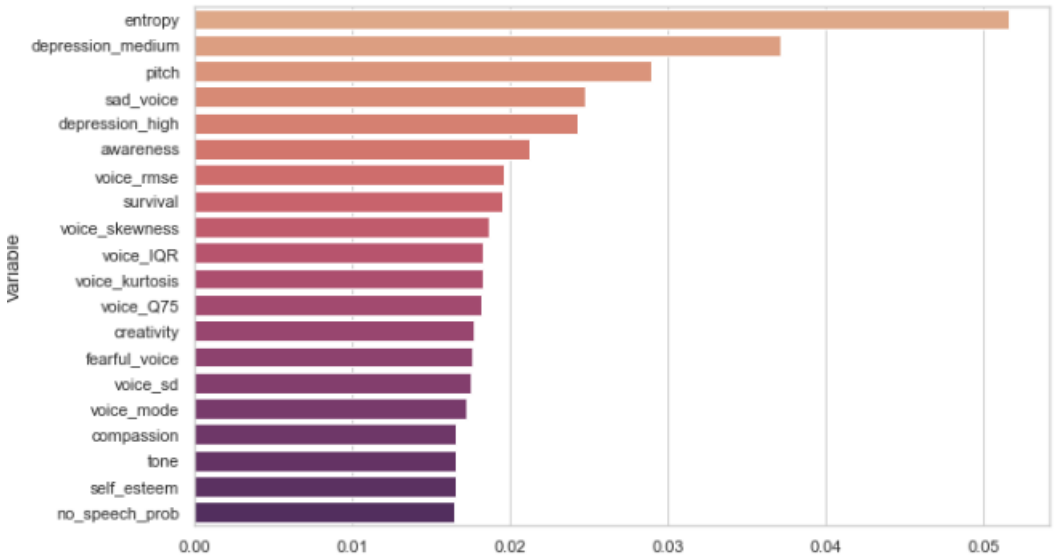

Figure 70: Random Forest Classifier, Top 15 Features

Using these top 15 features a new Random Forest model is trained and evaluated. The following results are obtained:

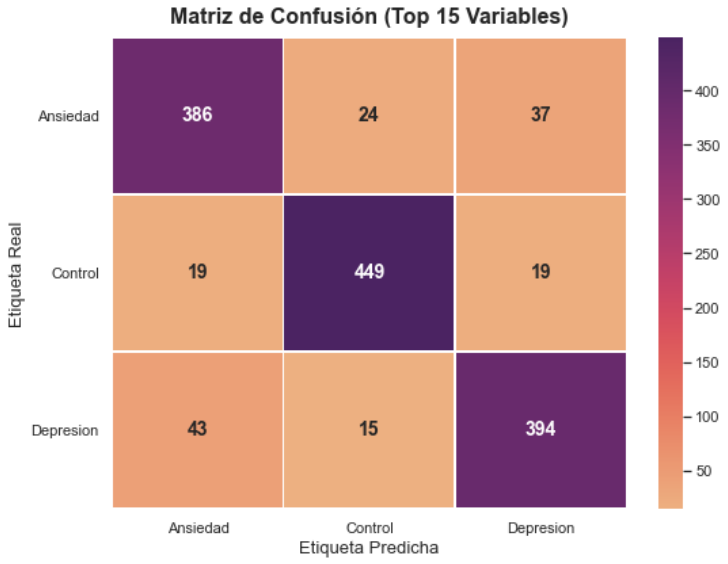

Figure 71: Confusion Matrix for Simplified Random Forest Classifier, Top 15 Features

These results can be summarized as:

	Precision	Recall	F1-Score
Anxiety	0.86	0.86	0.86
Control	0.92	0.92	0.92
Depression	0.88	0.87	0.87
Accuracy			0.89
Macro Avg	0.89	0.89	0.89
Weighted Avg	0.89	0.89	0.89

Table 28: Summary of Results for Simplified Random Forest Classifier, Top 15 Features

The classification report for this reduced-feature model shows an accuracy of ~0.89 (89%), which is remarkably on par with the HGB model and very similar to the baseline full-feature Random Forest while allowing for huge simplification.

The per-class precision/recall values with 15 features remain high: for Anxiety, precision ~0.86 and recall ~0.86; for Depression, precision ~0.88 and recall ~0.87; and for Control, precision and recall ~0.92 each. These figures are very close to the full-model performance, indicating minimal loss despite using only ~5% of the original number of features. In fact, the accuracy of 89% with a limited feature set suggests that many of the original features were redundant or noisy. By removing them, we not only simplify the model but may also reduce overfitting.

The Random Forest evidently concentrates its predictive power on a small subset of highly informative features, and including hundreds of weaker features does not substantially improve its discrimination. This result underlines an important advantage of Random Forests: they provide a natural mechanism for feature selection through their importance measures, allowing the model to be pruned without much performance sacrifice.

After confirming the viability of the top-15-features model, the contributions of different modalities of features are explored. 5 groups are explored and their results analyzed:

4. Facial and Vocal Emotions: 'angry_facial', 'happy_facial', 'sad_facial', 'angry_voice', 'happy_voice', 'sad_voice'.

With this subset the following results are obtained:

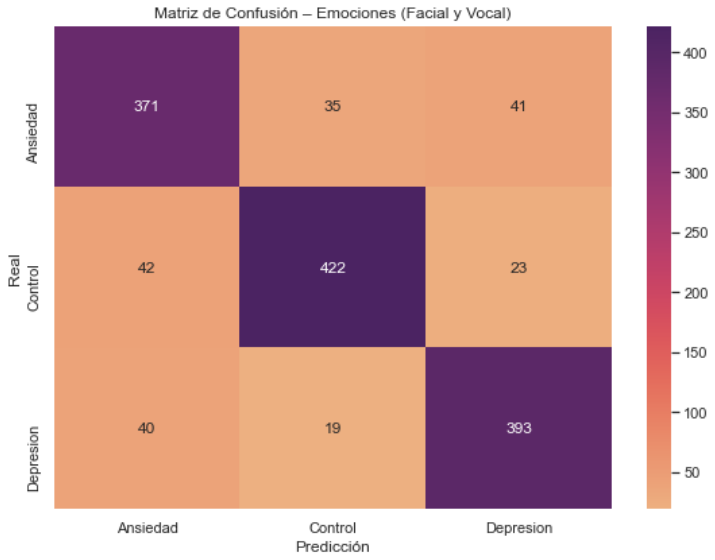

Figure 72: Confusion Matrix for Facial and Vocal Emotions Feature Subset, Random Forest Classifier

Which can be summarized in the following table:

	Precision	Recall	F1-Score
Anxiety	0.82	0.83	0.82
Control	0.89	0.87	0.88
Depression	0.86	0.87	0.86
Accuracy			0.86
Macro Avg	0.86	0.86	0.86

Weighted Avg	0.86	0.86	0.86

Table 29: Summary of Results for Facial and Vocal Emotions Feature Subset, Random Forest Classifier

5. Personality and Affectivity: 'extraversion', 'neuroticism', 'agreeableness', 'conscientiousness', 'openness', 'survival', 'creativity', 'self_esteem', 'compassion', 'communication', 'imagination', 'awareness'.

With this subset the following results are obtained:

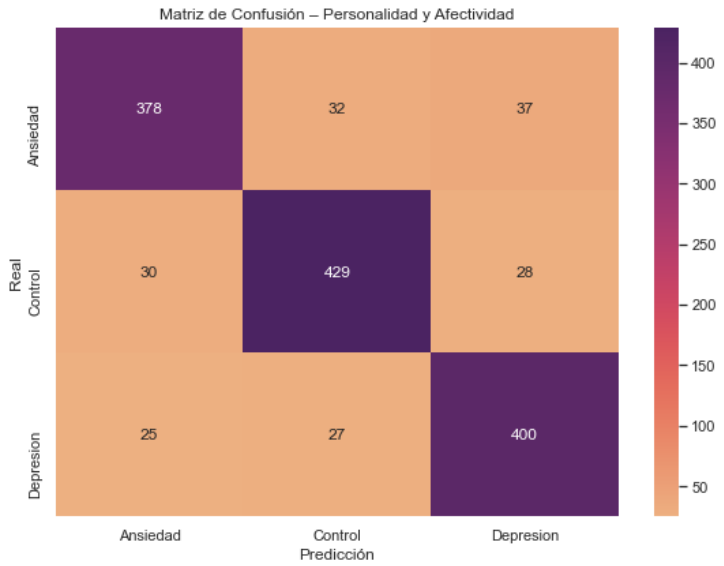

Figure 73: Confusion Matrix for Personality and Affectivity Feature Subset, Random Forest Classifier

Which can be summarized as:

	Precision	Recall	F1-Score
Anxiety	0.87	0.85	0.86
Control	0.88	0.88	0.88
Depression	0.86	0.88	0.87
Accuracy			0.87

Macro Avg	0.87	0.87	0.87
Weighted Avg	0.87	0.87	0.87

Table 30: Summary of Results for Personality and Affectivity Feature Subset, Random Forest Classifier

6. Acoustic Features Only: 'voice_mean', 'voice_sd', 'voice_median', 'voice_mode', 'voice_Q25', 'voice_Q75', 'voice_IQR', 'voice_skewness', 'voice_kurtosis', 'voice_rmse', 'pitch', 'tone', 'no_speech_prob', 'entropy'.

The following results are obtained here:

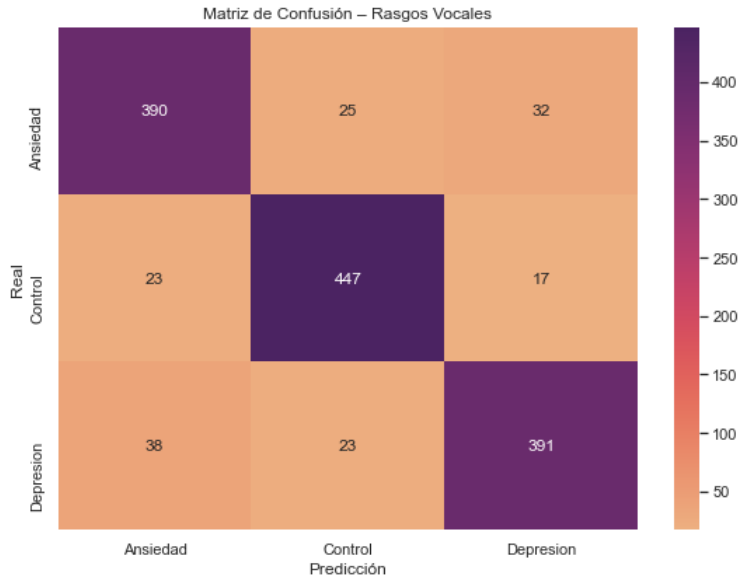

Figure 74: Confusion Matrix for Acoustic Features Only Feature Subset, Random Forest Classifier

Which displayed as a summarized table:

	Precision	Recall	F1-Score
Anxiety	0.86	0.87	0.87
Control	0.9	0.92	0.91

Depression	0.89	0.87	0.87
Accuracy			0.89
Macro Avg	0.89	0.89	0.89
Weighted Avg	0.89	0.89	0.89

Table 31: Summary of Results for Acoustic Features Only Feature Subset, Random Forest Classifier

7. Top Semantic + Voice + Emotions Features: 'tense_past', 'tense_present', 'tense_future', 'sentiment_polarity', 'sentiment_subjectivity', 'voice_mean', 'voice_sd', 'voice_skewness', 'voice_kurtosis', 'voice_rmse', 'pitch', 'happy_voice', 'angry_voice', 'sad_voice'.

With this combination of subsets, the following results are obtained:

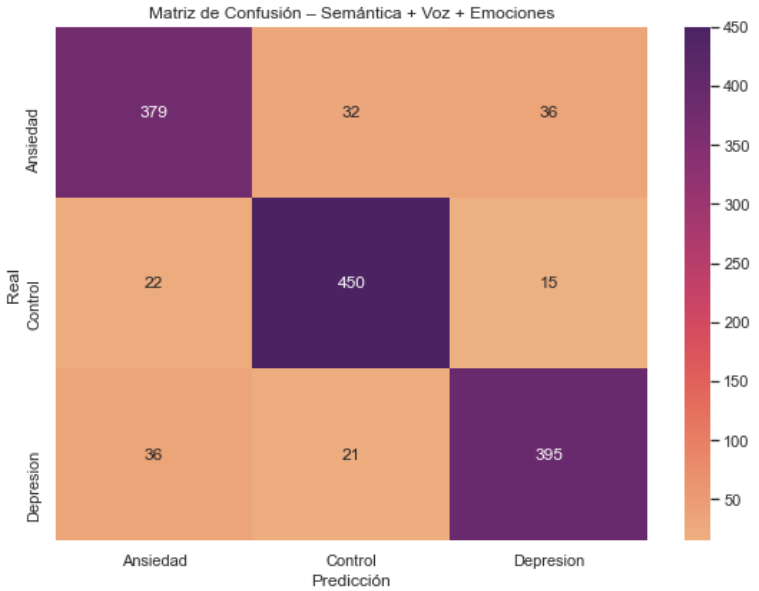

Figure 75: Confusion Matrix for Semantics + Voice + Emotions Feature Subset, Random Forest Classifier

These results are then conceptualized into the following table:

	Precision	Recall	F1-Score

Anxiety	0.87	0.85	0.86
Control	0.89	0.92	0.91
Depression	0.89	0.87	0.88
Accuracy			0.88
Macro Avg	0.88	0.88	0.88
Weighted Avg	0.88	0.88	0.88

Table 32: Summary of Results for Semantics + Voice + Emotions Feature Subset, Random Forest Classifier

8. Optimized Feature Pack: 'happy_facial', 'surprise_facial', 'angry_facial', 'extraversion', 'neuroticism', 'voice_mean', 'voice_kurtosis', 'voice_skewness', 'tense_present', 'tense_past', 'sentiment_polarity', 'sentiment_subjectivity'

The optimized pack, renders the following results:

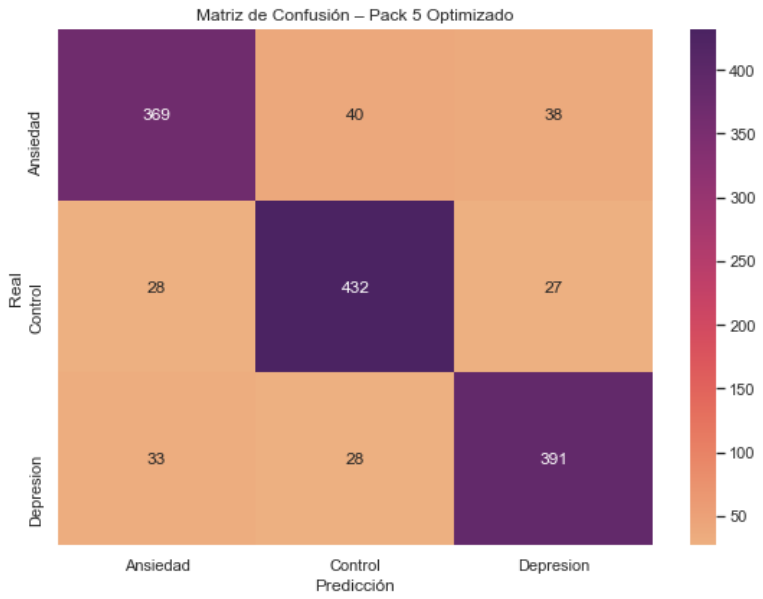

Figure 76: Confusion Matrix for the Optimized Feature Pack Feature Subset, Random Forest Classifier

These results are summarized through the following table:

	Precision	Recall	F1-Score
Anxiety	0.86	0.83	0.84
Control	0.86	0.89	0.88
Depression	0.86	0.87	0.86
Accuracy			0.86
Macro Avg	0.86	0.86	0.86
Weighted Avg	0.86	0.86	0.86

Table 33: Summary of Results for the Optimized Pack Feature Subset, Random Forest Classifier

The evaluation of the Random Forest classifier through targeted feature subgroup analyses reveals a compelling interplay between feature selection and model performance. These experiments, in comparison to the full feature baseline as well as to previously optimized models (such as the 15-feature hybrid model), provide an empirical basis for refining multimodal machine learning approaches in psychological condition classification.

The first group, composed of facial and vocal emotional markers, achieves a commendable F1-score average of 0.86. This result highlights the significance of expressive emotional signals in the detection of psychological states. Features such as happy_voice or sad_facial are likely capturing paralinguistic and facial dynamics that correlate with emotional disturbances, especially when aligned with internal affective states such as depression or anxiety. Nonetheless, while robust, these features alone do not reach the superior performance observed in full-feature configurations. This suggests that although emotional indicators are essential, they benefit from being contextualized within broader behavioral or linguistic data.

The personality and affectivity subset achieves marginally better results, reaching a 0.87 average across metrics. This improvement indicates that stable traits such as neuroticism or self-esteem exert a potent influence on the classification of mental health conditions. These traits likely function as latent predispositions that reinforce the observable emotional or behavioral expressions captured in other feature groups. The consistency in recall and precision across the three classes further suggests that personality metrics provide a more uniformly discriminative basis across different diagnostic labels, particularly between anxiety and depression.

Interestingly, the subset dedicated solely to acoustic features demonstrates near-parity with the full model, attaining a striking 0.89 accuracy and balanced class-wise performance. This finding reinforces the central role of vocal dynamics in conveying emotional and psychological states. Variables like voice_rmse, pitch, and entropy likely encode cognitive load, emotional arousal, or affective valence in a non-verbal yet highly distinguishable manner. The near-match to the full model indicates that voice patterns may independently encapsulate a rich source of diagnostic information, especially when compared to the facial-only or affective-only configurations.

When semantic, vocal, and emotional features are fused (group 4), the resulting model attains a precision and recall level of 0.88, which closely mirrors the strongest performing configurations. This combination illustrates the additive value of integrating low-level acoustic data with high-level semantic constructs such as verb tense and sentiment polarity. Temporal linguistic cues are recognized clinical indicators, and their interaction with tone and emotional expression allows the classifier to interpret not only what is being said but how it is being conveyed, leading to an enriched understanding of mental state.

The final configuration, an optimized minimal pack of 12 features, was derived based on theoretical relevance and prior empirical importance. While it does not surpass the full-feature baseline or the 15-feature hybrid explored earlier in the study, it offers a pragmatic compromise: a simplified architecture with high generalizability, computational efficiency and interpretability.

From these findings, high performance can be retained with strategically reduced input dimensionality, suggesting that feature quality is more impactful than quantity. Secondly, voice features—across all configurations—demonstrate consistently strong discriminative power, validating their importance in multimodal psychological modeling. Finally, while Random Forest performs robustly with all subsets, its slight degradation in performance when moving away from the complete feature set underscores its dependency on ensemble diversity to capture feature interaction effects.

8.6 NEURAL NETWORKS

In this section, a Neural Network approach is used for multiclass classification, specifically a feed-forward multilayer perceptron (MLP), with the objective of capturing complex nonlinear interactions among features that tree-based models might overlook.

For the initial model, moderate network architecture and default hyperparameters are chosen. The baseline configuration is a single hidden layer with 100 neurons, ReLU activation, and the Adam optimizer for training. No explicit regularization is added at this stage, beyond what the MLPClassifier's default includes.

After training, the neural network's performance is evaluated on the test set, obtaining the following results:

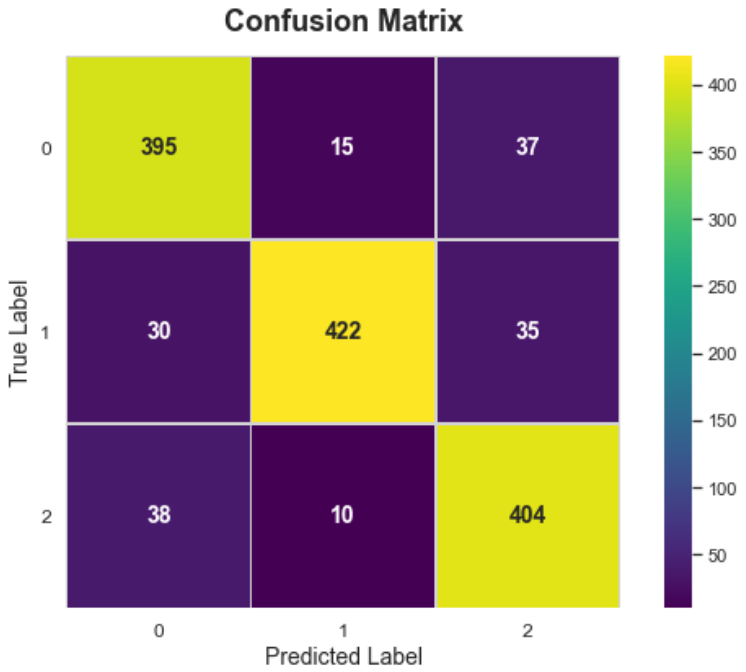

Figure 77: Confusion Matrix for Default Neural Network (MLP)

The confusion matrix shows a pattern consistent with earlier models. The largest confusion is between Anxiety and Depression – the network sometimes confuses one for the other. Ultimately, however, the neural network's overall performance is very strong, nearly matching that of the ensemble models.

Which can be summarized in the following table:

	Precision	Recall	F1-Score
Anxiety	0.85	0.88	0.87
Control	0.94	0.87	0.9
Depression	0.85	0.89	0.87
Accuracy			0.88

Macro Avg	0.88	0.88	0.88
Weighted Avg	0.88	0.88	0.88

Table 34: Summary of Results for Default Neural Network (MLP)

The MLP's baseline accuracy is 88%, which is very respectable and on par with the tuned HGB and RF models. This first analysis reveals that the neural network has a tendency to err on the side of predicting a person *has* a condition (anxiety or depression) rather than missing it, or what it is the same, it sacrifices some specificity for sensitivity. This is evident in the confusion matrix: of the 487 true control individuals, 65 were misclassified as either anxious or depressed, whereas of the combined 899 true disorder cases, only 132 were misclassified as control.

With a strong baseline in hand, next, the neural network model is refined through hyperparameter tuning, following a staged approach (focusing on one aspect of the model at a time).

The tuning of the neural network model begins with an exploration of the output probabilities of the baseline model, providing an early lens into how confidently the network classifies examples across the three classes:

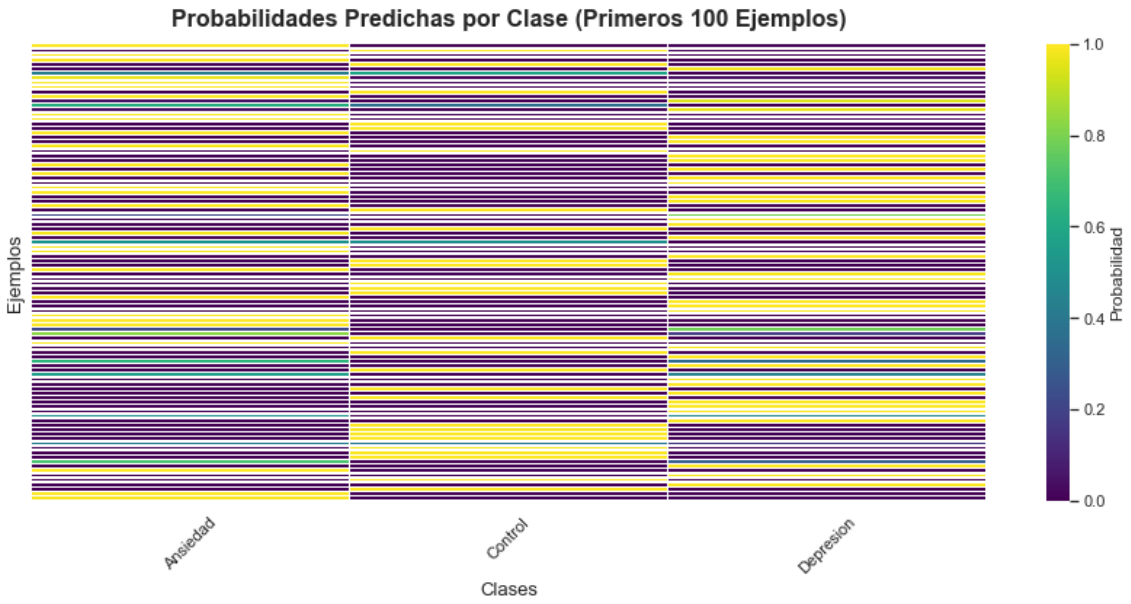

Figure 78: Probability Distribution for the first 100 instances, Default Neural Nework (MLP)

This plot reveals how sharply the model distinguishes between classes: intense coloring near the diagonal indicates confident and consistent predictions across the majority of examples. The presence of some lighter bands or dispersed probability values outside the target class columns also hints at residual uncertainty or class overlap, particularly between anxiety and depression. This contextualizes the possible need for tuning the model to better capture class distinctions while maintaining generalization.

Then, the network's architectural configuration is examined, with specific attention paid to how varying the number of neurons affects model performance, which is captured in the following diagram:

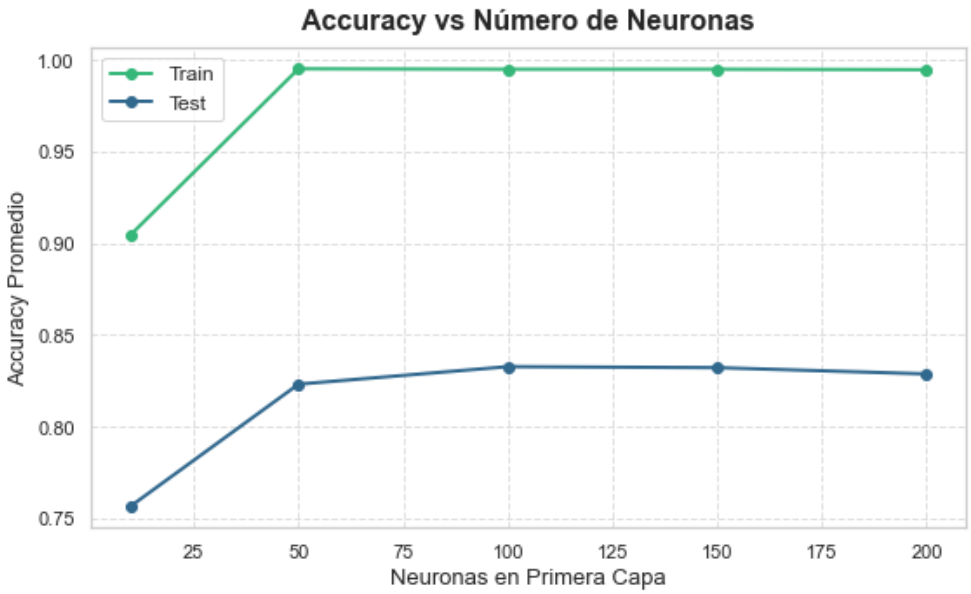

Figure 79: Accuracy vs Neuron Number

The curve for the training set quickly saturates, with accuracy approaching 100% as the network becomes larger, suggesting high representational capacity. Meanwhile, the testing accuracy peaks around 100 neurons, beyond which no significant gain is observed. This plateau combined with the gap between training and testing accuracy is a clear indicator of potential overfitting in excessively large models, reinforcing the selection of 100 neurons as the optimal trade-off point between underfitting and overfitting.

Parallel to the exploration of architecture, learning rate sensitivity is also analyzed. This process is best illustrated in the following chart:

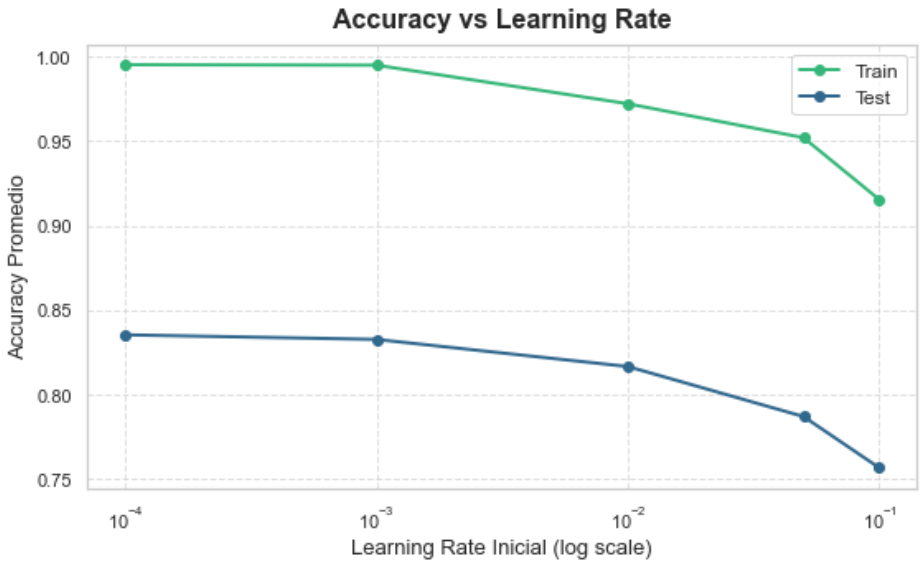

Figure 80: Accuracy vs Learning Rate

The network performs optimally around a learning rate of 0.001, while 0.1 proves detrimental, likely due to instability in weight updates. Conversely, very small rates such as 0.0001 underperform due to slow convergence. The parallel trend observed in training accuracy curves provides additional confirmation that larger learning rates do not allow the network to generalize effectively.

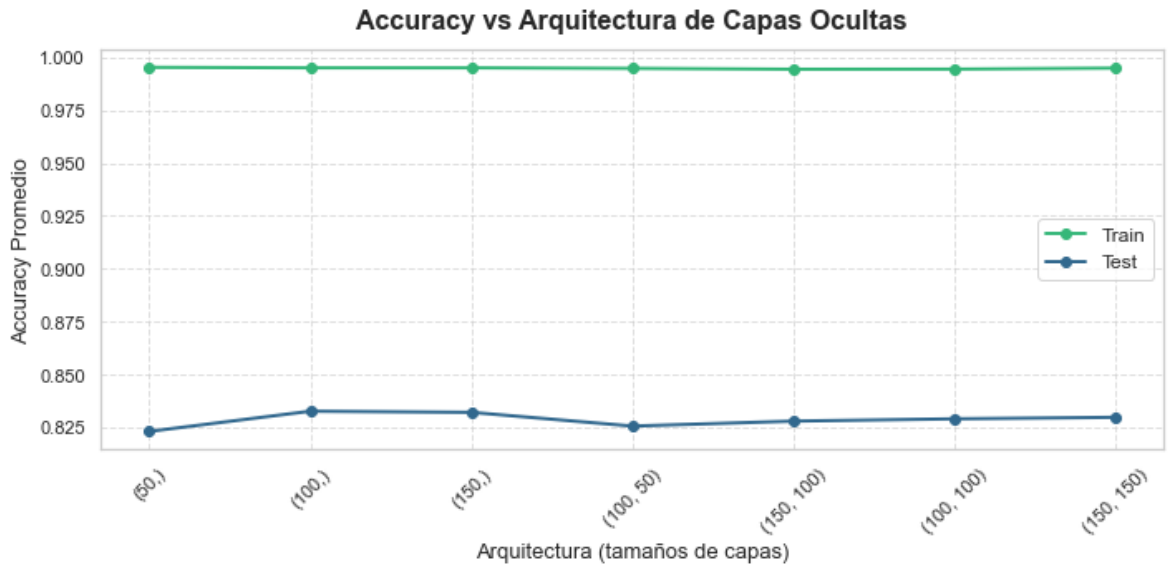

Figure 81: Accuracy vs Hidden Layers Architecture

The figure above showcases comparative effectiveness of different architectural layouts, all with fixed learning rate and regularization. While more complex two-layer models do not severely overfit, they also fail to deliver meaningful performance gains over the single-layer 100-neuron architecture. The graph underlines the point that increasing model complexity beyond a certain threshold does not necessarily yield improved generalization and can sometimes introduce unnecessary intricacy, particularly when sufficient accuracy is already achieved with simpler models.

In summary, the refinement process confirms that the original configuration (a single hidden layer with 100 neurons, learning rate of 0.001, and default solver/regularization) was near-optimal. Each tuning exercise reaffirms that only marginal gains are possible beyond the baseline, and that the neural network performs robustly across different configurations, offering generalization power comparable to ensemble methods. The visual materials validate this narrative, with clear evidence that tuning stabilizes the model's learning dynamics without overcomplicating its architecture.

The network's strong performance indicates that it successfully learned the complex relationships between features and mental health classes. This is noteworthy because the neural net is essentially a black-box function approximator – unlike trees, it doesn't have a built-in notion of feature importance. Yet through training, it implicitly found similar patterns: one can infer that it must be attending to similar key features because that information is needed to achieve the performance seen. This all suggests that the neural network model has achieved generalization performance comparable to the ensemble methods.

8.7 SUPPORT VECTOR MACHINE

The final model considered is a Support Vector Machine (SVM) with a nonlinear kernel. Specifically, an SVM with the Gaussian Radial Basis Function (RBF) kernel, which is well-suited for complex boundaries in high-dimensional spaces, is used. It was chosen for its,

apriori, good performance on datasets that contain a large number of features, which aligns well with the high-dimensional dataset used.

First, a baseline SVM with default hyperparameters is trained and tested, with the model finding the support vectors and decision boundaries in the transformed kernel space that best separates the classes. The results are portrayed in the following illustrations:

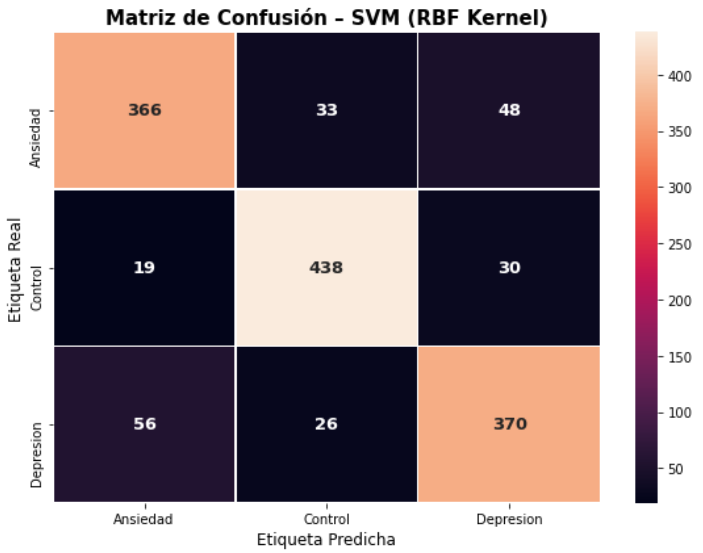

Figure 82: Confusion Matrix for Default SVM Model

This confusion matrix is summarized in the following table:

	Precision	Recall	F1-Score
Anxiety	0.83	0.82	0.82
Control	0.88	0.9	0.89
Depression	0.83	0.82	0.82
Accuracy			0.85
Macro Avg	0.85	0.85	0.85

Weighted Avg	0.85	0.85	0.85

Table 34: Summary of Results for Default SVM Model

The accuracy comes out to ~84.7%, which is noticeably lower than the ~88–90% achieved by the other models' baselines. The precision and recall for Anxiety and Depression are both around 0.82, and for Control around 0.88–0.90. While the results obtained are considerably good and sound, the baseline SVM thus underperforms the ensembles and neural net in this initial state.

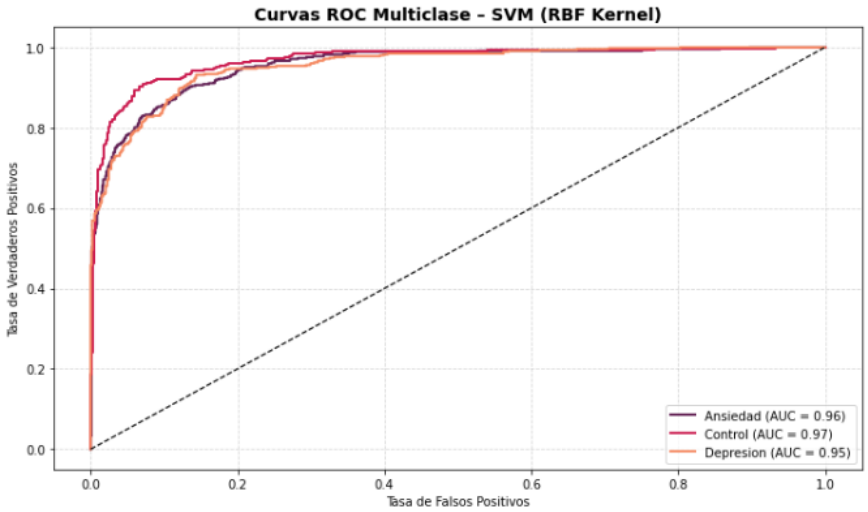

Figure 83: ROC Curves for Default SVM Model

The ROC curves for baseline SVM show AUCs that are a bit lower than in previous models. The curves still sit above the diagonal, but not as steeply. However, in this regard, this model produced excellent results.

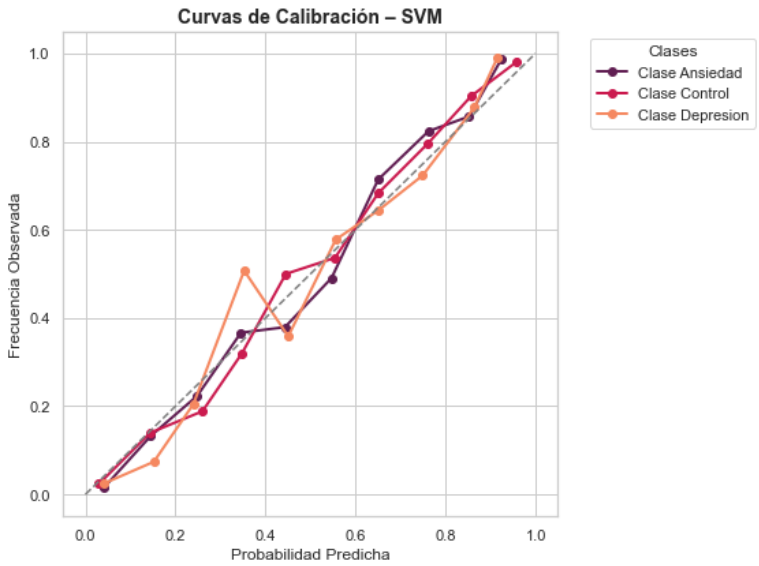

Figure 84: Calibration Curves for Default SVM Model

As can be seen in the image above, the calibration curves for the SVM's probability outputs are also examined. Typically, SVM probabilities can be either too extreme or too flat if the model margin is not well-set. However, as can be seen, in this case the lines for each class barely deviate from the diagonal, which suggests a contained need to adjust the decision function sharpness via hyperparameters.

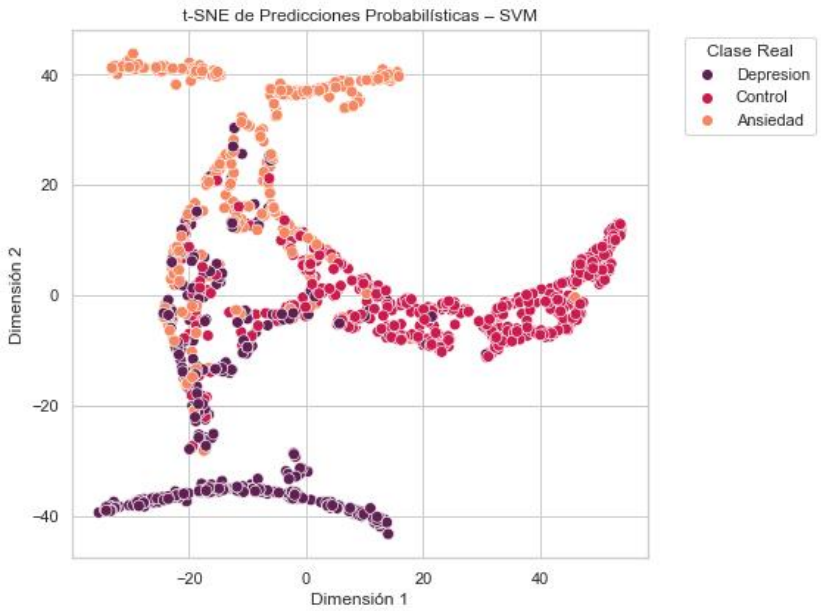

Figure 85: t-SNE for Default SVM Model

A t-distributed Stochastic Neighbor Embedding (t-SNE) visualization on the outputs is displayed above to better understand how the SVM is separating classes. Control points form a well-defined cluster separated from the others, indicating that in the SVM probability space, controls occupy a distinct region (the SVM assigns them consistently different probability distributions, usually high control probability and low others). The Anxiety and Depression points, however, are intermingled to a significant extent in the t-SNE plot. There isn't a clean boundary between anxious and depressed cases in that visualization – rather, there's a relative continuum or overlap region where some anxious and depressed points lie close together. This shows that the SVM has learned the primary division (controls vs patients) but struggles on the secondary division (anxiety vs depression).

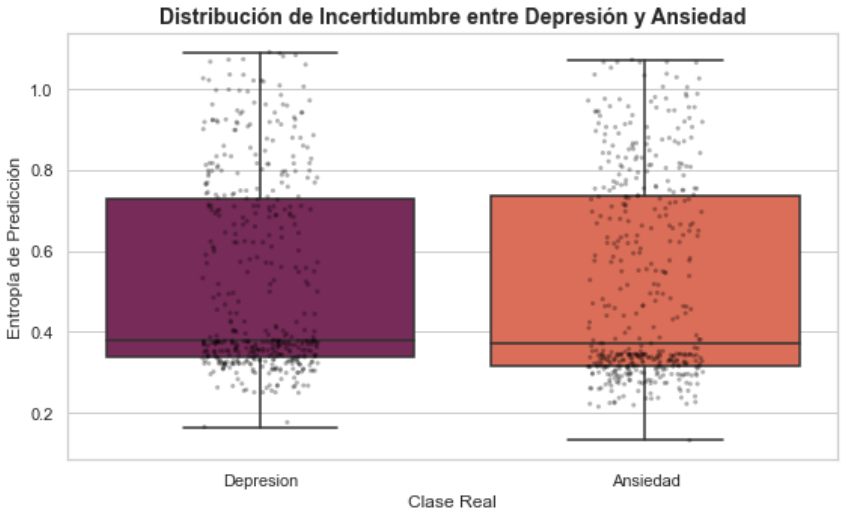

Figure 86: Entropy Distribution for Default SVM Model

The uncertainty for the two disorders is also considered using an entropy analysis (shown in the image above), using Shannon's entropy. The SVM's predictions for both anxiety and depression cases have slightly high uncertainty on average (higher entropy than one would see for control cases). These higher-than-desired entropies reflect the earlier observation that the SVM was outputting moderate probabilities rather than very sharp ones.

The suboptimal baseline performance leads the next step to be hyperparameter tuning. The best combination of hyperparameters is found using Randomized Search CV. They are found allowing C to vary over a continuous range (uniformly between 0.1 and 100) and gamma uniformly between 0.0001 and 0.1, for 30 random samples. This broader search (also 5-fold CV) finds an even better combination: it reports C ≈ 54.77 and gamma ≈ 0.0186 as the best parameters.

Having obtained the best parameters, the model is retrained and retested. The improvement, as can be seen in the following illustrations, is dramatic:

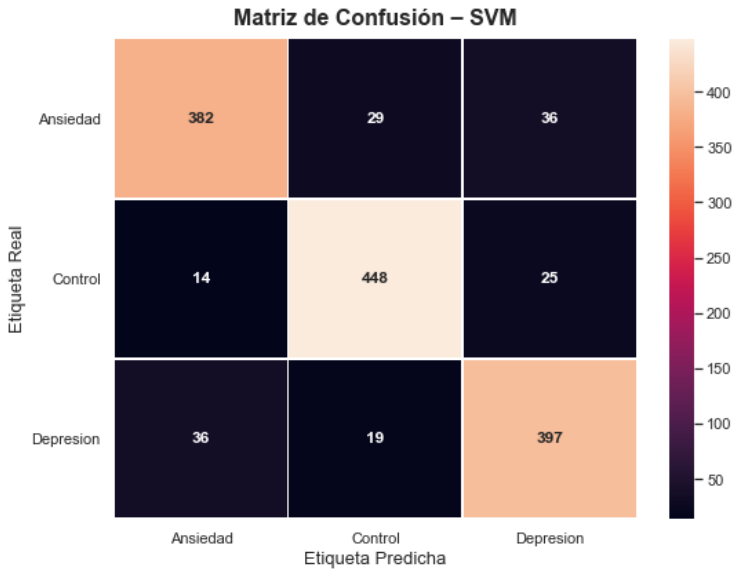

Figure 87: Confusion Matrix for Tuned SVM Model

The confusion matrix looks similar to the ones in previous models, with a strong diagonal line and a much-improved outlook compared to the baseline SVM model.

These results are, once again, summarized in the following table:

	Precision	Recall	F1-Score
Anxiety	0.88	0.85	0.87
Control	0.9	0.92	0.91
Depression	0.87	0.88	0.87
Accuracy			0.89
Macro Avg	0.88	0.88	0.88
Weighted Avg	0.89	0.89	0.89

Table 35: Summary of Results for Tuned SVM Model

As can be observed I the table, the classification report for the tuned SVM shows overall accuracy jumping to 88.53%. This is a ~4 percentage point increase from baseline and squarely in line with the ensemble and neural network results. The precision and recall for the Anxiety class are now ~0.87 and 0.85 respectively, for Depression ~0.87 and 0.88, and for Control ~0.90 and 0.92. The tuned model's performance is essentially on par with the HGB and RF models, no longer lagging and being far more accurate and balanced in its predictions.

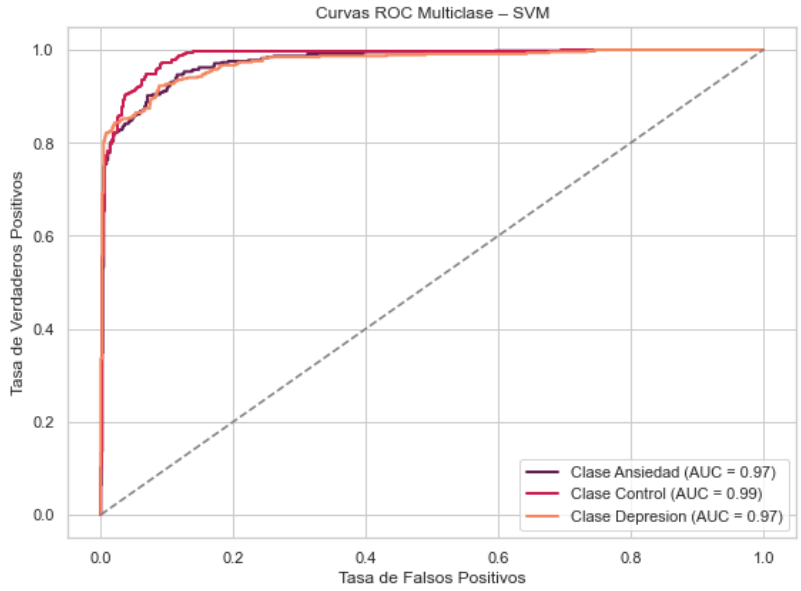

Figure 88: ROC Curves for Tuned SVM Model

The ROC curves, shown in the above image, show marked improvement: now each ROC is very steep and the AUC values have also improved.

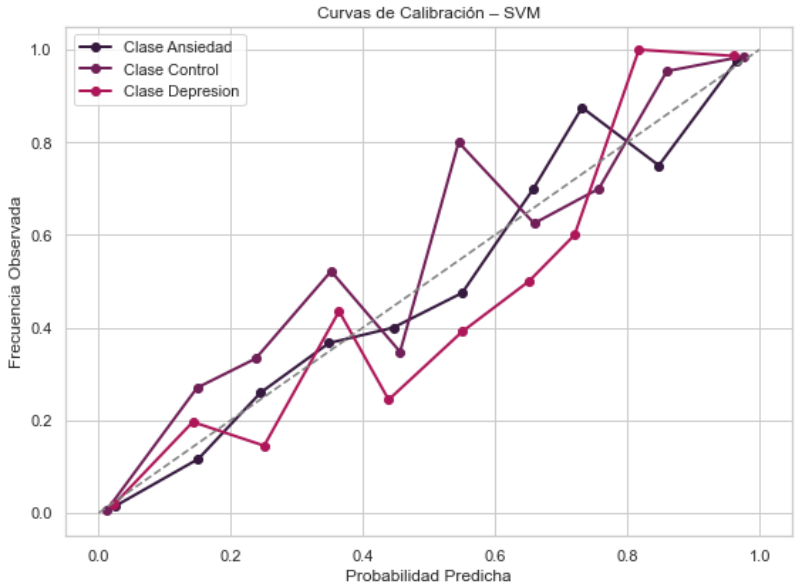

Figure 89: Calibration Curves for Tuned SVM Model

In the superior image, Calibration curves are once again plotted, showing some minor improvement as they are no longer 'underwater' in the points closest to the origin.

Focusing on Anxiety and Depression as a more in-depth comparison of their output, the nearly identical metrics can be seen in the following bar chart:

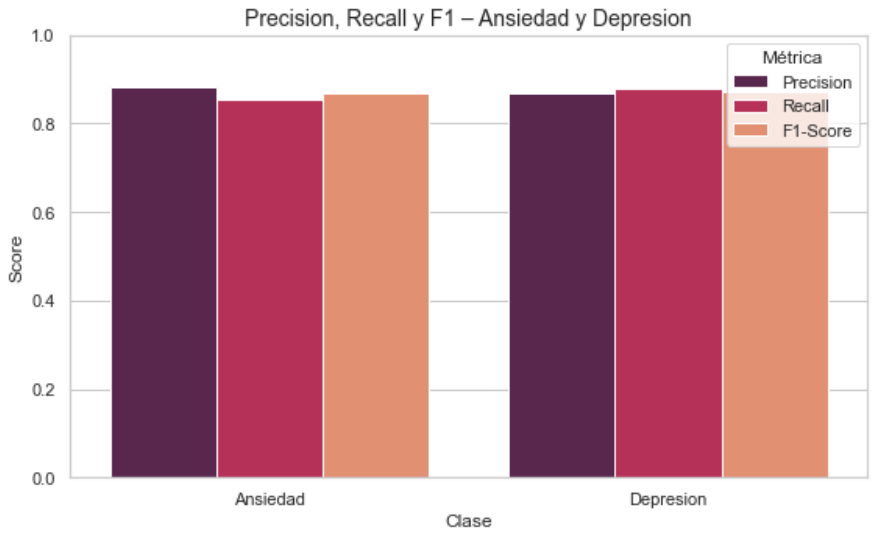

Figure 90: Precision, Recall & F1 Bar Chart for Tuned SVM Model

Furthermore, the following violin plot is plotted:

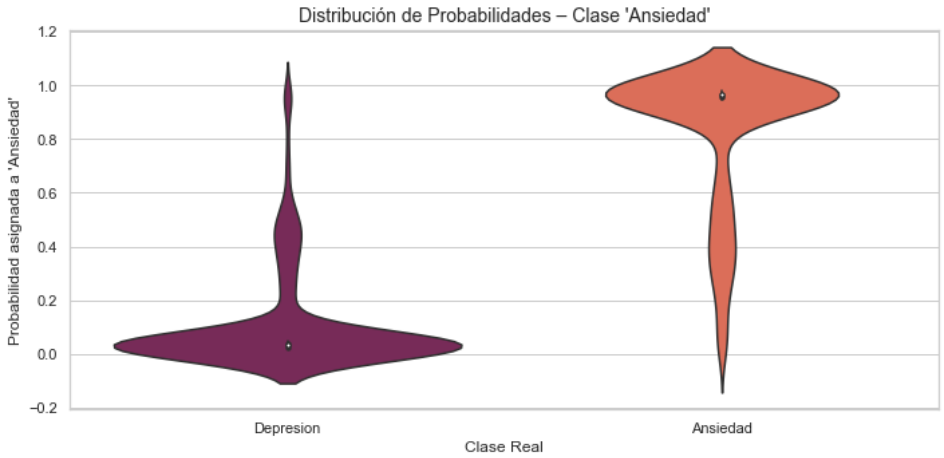

Figure 91: Probability Distribution for Tuned SVM Model

A violin plot, as shown above, is also plotted for the probability of class Anxiety for those groups identified either with Anxiety or Depression. For true Anxiety cases, the violin representing their predicted "Anxiety probability" is centered towards high values – many true anxiety cases receive probabilities close to 1 for anxiety (meaning the model is confident and correct). For true Depression cases, the violin for "Anxiety probability" skews towards lower values (most depressed individuals get low probability of anxiety, which is good as they should ideally get high probability of depression instead).

However, there is still some overlap: the Depression violin has a tail extending into moderate or even high "Anxiety probability" region, corresponding to depressed cases that the SVM misclassified or was uncertain about (these are the depressed individuals that share traits with anxiety, causing the model to assign them a relatively high chance of being anxious). Likewise, the Anxiety violin likely has some mass at lower probabilities – those are anxious people the model nearly mislabeled as depression. Overall, though, the violins for true Anxiety vs true Depression are much more separated in the tuned model than they would have been in the baseline model.

The median anxiety-probability and depression probability are also included in this confusion matrix:

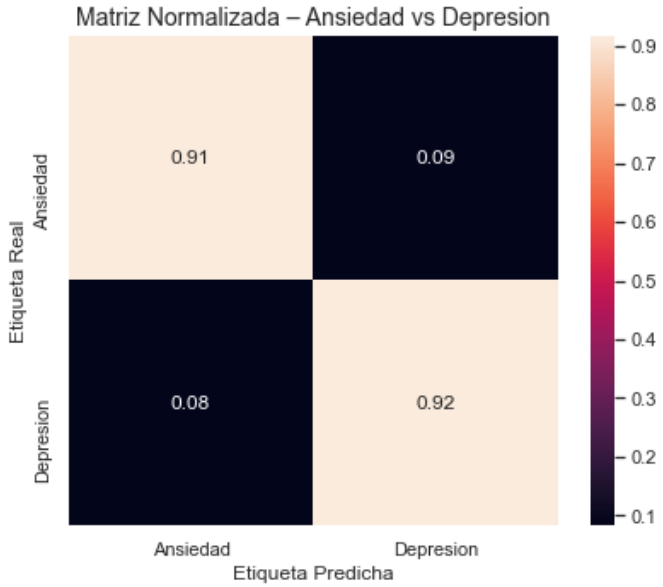

Figure 92: Confusion Matrix Anxiety and Depression Accuracy for Tuned SVM Model

This indicates that the tuned SVM outputs clearly different probability distributions for anxious vs depressed individuals, achieving good discrimination.

Through careful hyperparameter tuning, the SVM has been transformed from a mediocre performer (85% accuracy) to an excellent one (~89% accuracy). The final SVM's performance is on par with the other advanced models and it yields well-calibrated probabilities and balanced predictions across classes.

However, it's worth noting the computational cost: training the SVM with optimal parameters was significantly heavier than training the tree ensembles or the neural network. The cross-validation procedures themselves were time-consuming, and the final model with C≈55 uses many support vectors (likely a large fraction of the training data ended up as support vectors due to the high C fitting many points). This means the SVM model is relatively large in memory and slower to predict new instances compared to the tree models or even the neural net. This is a known trade-off: SVMs can achieve high performance but may not scale easily to very large datasets, and their prediction speed can be slower if many support vectors are needed.

8.8 REFLECTIONS AND IMPLICATIONS

In terms of raw accuracy and core metrics, all four approaches converged to a similar performance band (approximately 88–90% accuracy on the test set) after tuning. The ensemble methods (HGB and Random Forest) and the SVM reached the upper end (~89–90%), while the neural network achieved around 88–89%. The differences in final accuracy are minor – on the order of 1–2 percentage points – which is not statistically significant given the test set size.

All models attained high precision and recall in the high 80s to low 90s for each class, meaning they rarely miss true cases (high recall) and rarely raise false alarms (high precision). Importantly, in each model the Control class (no disorder) was the easiest to identify (often >90% precision/recall), and the Anxiety and Depression classes were slightly more challenging, with recall in the mid-to-high 80s after tuning. This pattern is expected because controls have markedly different profiles than individuals with any mental health condition, whereas anxiety vs depression differentiation is subtler.

Regarding convergence and tuning effort, one notable difference was the amount of effort required to reach these high-performance levels:

- The HGB classifier performed excellently out-of-the-box, which speaks to the robustness of modern boosting algorithms.
- The Random Forest also achieved strong results with default hyperparameters, but feature selection was found to dramatically improve its performance (up to ~89%). In a sense, the "tuning" for Random Forest was not about adjusting tree parameters (same number of trees, default depth etc.) but about selecting an informative subset of features.
- The Neural Network required careful consideration of architecture and training parameters, experimenting with multiple hidden layer sizes and learning rates to ensure the model was neither underfitting nor overfitting. Thus, the MLP demanded a moderate tuning effort. The payoff was a model on par with others, but the time

invested in cross-validation and training multiple networks was higher than, say, the effort for HGB.

- The SVM required the most extensive tuning. The baseline SVM was substantially worse than other baselines, but through intensive grid and random search the chosen hyperparameters allowed it to boost up. This is, however, computationally expensive – each fold of CV requires solving an SVM optimization, which can be slow for many data points. Indeed, the SVM had the longest runtime in tuning among the models.

Interpretability was a distinguishing factor across the models, particularly relevant in mental health applications. Random Forest provided the most transparent insights, enabling extraction of key features and decision paths that aligned with clinical intuition, such as neuroticism, self-efficacy, or vocal emotion features being predictive of anxiety or depression. HGB, though slightly more opaque due to its ensemble nature, still facilitated global interpretability via feature importances. Conversely, SVM and the neural network functioned as black boxes, offering no native interpretability without post-hoc explanation techniques.

Ultimately, all models demonstrated robust generalization, converging on a shared accuracy ceiling of approximately 90%, which appears to reflect the intrinsic difficulty of the classification task rather than model limitations. This convergence implies that further performance gains would likely require new or richer features rather than algorithmic changes. HGB stood out as the most effective overall due to its balance of high accuracy, ease of use, moderate interpretability and resilience across varying input subsets. RF followed closely, offering a strong interpretability advantage. Neural networks, while versatile and promising for future multimodal data integration, did not clearly surpass ensemble methods on the current structured data. SVMs achieved parity in accuracy but at the cost of computational efficiency and transparency.