UNIVERSIDAD DE VALLADOLID

E.T.S.I. TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

GRADO EN INGENIERÍA DE TECNOLOGÍAS ESPECÍFICAS DE TELECOMUNICACIÓN

APLICACIÓN DE TÉCNICAS DE APRENDIZAJE AUTOMÁTICO PARA LA CLASIFICACIÓN Y PREDICCIÓN EN DIFERENTES CASOS DE USO

Autor:
D. Pablo González Gómez
Tutora:
Dña. Noemí Merayo Álvarez

TÍTULO: **Aplicación de técnicas de aprendizaje automático para la clasificación y predicción en diferentes casos de uso**
AUTOR: **D. Pablo González Gómez**
TUTORA: **Dña. Noemí Merayo Álvarez**
DEPARTAMENTO: **Teoría de la Señal y Comunicaciones e Ingeniería Telemática**

TRIBUNAL
PRESIDENTE:
SECRETARIO:
VOCAL:
SUPLENTE:
SUPLENTE:

FECHA:
CALIFICACIÓN:

Resumen de TFG

Este Trabajo de Fin de Grado tiene como objetivo principal la implementación de modelos basados de aprendizaje automático aplicados en diversos escenarios, buscando un rendimiento óptimo de todas las métricas para todos los contextos de análisis.

En un primer bloque, se exploran modelos de clasificación y regresión para predecir la calidad del vino a partir de sus parámetros químicos.

En un segundo bloque, se aborda el procesamiento del lenguaje natural basado en aprendizaje automático para clasificar la respuesta emocional, esto es, emociones y polaridades en contextos de redes sociales. Se emplea en primer lugar una base de datos relacionada con los videojuegos en Twitch y se aplican técnicas de clasificación para analizar y categorizar las emociones y polaridades presentes en los canales de los creadores de contenidos. Posteriormente, se realiza otro análisis utilizando una base de datos relacionada con la salud mental en redes sociales, donde nuevamente se trabaja con texto como variable independiente. Se realiza el procesamiento adecuado del lenguaje natural para aplicar los distintos modelos predictivos y confrontar los resultados obtenidos.

Palabras clave

RF (Bosque de Árboles Aleatorios), SVM (Máquina de Soporte de Vectores), IA (Inteligencia Artificial), AU (Aprendizaje Automático), NLP (Procesamiento de Lenguaje Natural), RRSS (Redes Sociales), Análisis de Sentimiento.

Abstract

This Final Degree Project has as its main objective the implementation of models based on machine learning applied in various scenarios, seeking optimal performance of all metrics for all analysis contexts.

First, classification and regression models are explored to predict wine quality from its chemical parameters.

In a second block, natural language processing based on machine learning is addressed to classify the emotional response, that is, emotions and polarities in social network contexts. First, a database related to video games on Twitch is used and classification techniques are applied to analyze and categorize the emotions and polarities present in the channels of content creators. Subsequently, another analysis is carried out using a database related to mental health in social networks, where again text is used as an independent variable. Appropriate natural language processing is performed to apply the different predictive models and compare the results obtained.

Keywords

RF (Random Forest), SVM (Support Vector Machine), AI (Artificial Intelligence), DL (Deep Learning), ML (Machine Learning)

Agradecimientos

Agradezco enormemente a mis padres y a mis abuelos, ya que, gracias a su apoyo incondicional, he llegado a ser la persona que soy en la actualidad.

Quiero expresar mi profundo agradecimiento a Noemí, quien me ha brindado una valiosa ayuda en todo momento.

Índice

Índice de figuras

Índice de tablas

1

Introducción

1.1 Motivación

Este Trabajo de Fin de Grado tiene como objetivo la implementación de diferentes modelos predictivos basados en aprendizaje automático aplicados en diversos casos de uso, y se divide en tres partes distintas.

En la primera parte, se exploran modelos de clasificación y regresión para predecir la calidad del vino en función de sus parámetros químicos. Se lleva a cabo una confrontación de los resultados obtenidos por los diferentes modelos, con el propósito de determinar cuál de ellos ofrece el mejor rendimiento.

En la segunda parte, se aborda el procesamiento del lenguaje natural con el fin de clasificar emociones y polaridades en redes sociales, eso es, análisis de sentimiento. Se utiliza una base de datos relacionada con el mundo de los videojuegos en Twitch. Se aplican técnicas de clasificación para analizar y categorizar las emociones y polaridades presentes en el texto.

En la tercera parte, se realiza una prueba adicional utilizando una base de datos enfocada en la salud mental en redes sociales. Aquí también se busca clasificar emociones y polaridades en comentarios y post relacionados con la salud mental. Se emplea un procesamiento del lenguaje natural para analizar y clasificar las emociones y polaridades presentes en el texto relacionado con la salud mental.

En resumen, este Trabajo de Fin de Grado se compone de tres partes distintas que abordan diferentes casos de uso: la predicción de la calidad del vino, la clasificación de emociones y polaridades por un lado dentro del entorno de los videojuegos y por otro lado en el entorno de la salud mental. Cada parte implica la implementación de modelos predictivos y el procesamiento adecuado de los datos para lograr los objetivos establecidos.

1.2 Objetivos

Los objetivos de este Trabajo de Fin de Grado se centran en la aplicación de diversos modelos predictivos en casos de uso muy distintos. Por un lado, se busca predecir el tipo de vino en función de sus parámetros químicos, mientras que, por otro lado, se aborda la clasificación de la respuesta emocional en textos a partir del procesamiento de lenguaje natural. El objetivo principal es probar y evaluar los diferentes modelos para

determinar su rendimiento y determinar en qué circunstancias cada uno de ellos se comporta mejor. Además, se busca identificar para qué tipos de problemas cada modelo es óptimo, estableciendo así recomendaciones claras sobre su aplicación adecuada en diferentes contextos.

1.3 Fases y Métodos

En este apartado se hace una descripción de la metodología que se ha seguido en el desarrollo del documento.

1.3.1 Fase de Análisis

Durante la fase de análisis, se llevó a cabo una comprensión del funcionamiento de los diferentes modelos predictivos desarrollados, y se evaluó cómo se podían aplicar a cada problema específico y a las respectivas bases de datos correspondientes. Se realizaron investigaciones para entender las características de los modelos y determinar cuál de ellos sería más adecuado para abordar cada situación. Este análisis permitió seleccionar los enfoques más apropiados y tomar decisiones sobre su implementación en cada caso.

1.3.2 Fase de Implementación

Durante la fase de implementación, se desarrollaron los diversos modelos teniendo en cuenta los conocimientos adquiridos en la fase de análisis. Se utilizaron características comunes para todos los modelos, como la validación cruzada y el escalado de datos, con el fin de garantizar un enfoque coherente y robusto en el proceso de construcción de los modelos. Estas prácticas permitieron una evaluación del rendimiento de los modelos y aseguraron que los resultados fueran comparables en todos los casos.

1.3.3 Fase de Pruebas

Durante la fase de pruebas, se realizaron búsquedas de hiperparámetros con el objetivo de encontrar los valores óptimos para cada uno de los modelos. Se exploraron diferentes combinaciones de hiperparámetros y se llevaron a cabo evaluaciones sistemáticas para determinar qué configuraciones producían los mejores resultados en términos de rendimiento y precisión. Esta búsqueda de hiperparámetros permitió afinar y optimizar cada modelo, maximizando así su capacidad predictiva.

1.3.4 Fase de Realización de los Informes

Durante esta fase, se llevaron a cabo las comparaciones entre los modelos utilizando los resultados obtenidos en la fase de pruebas. Se analizaron y contrastaron los resultados de los diferentes modelos, evaluando sus métricas. Además, se redactó la memoria del Trabajo de Fin de Grado.

1.4 Estructura de la memoria del TFG

La memoria del proyecto está compuesta por distintos Capítulos, los cuales se introducen a continuación.

En el Capítulo 1, se introduce el problema a tratar en este Trabajo de Fin de Grado, las motivaciones, objetivos y las distintas fases aplicadas para la realización de la memoria

En el Capítulo 2, se expone la metodología usada y las herramientas necesarias para la realización de los distintos modelos predictivos

En el Capítulo 3, se llevó a cabo el desarrollo del proyecto usando modelos predictivos de regresión y clasificación, también la búsqueda de hiperparámetros y la comparación de modelos utilizando los resultados obtenidos para la clasificación de vinos en función de sus parámetros químicos.

En el Capítulo 4, se abordó el desarrollo del proyecto usando únicamente modelos de clasificación, en concreto Support Vector Machine (SVM) y RandomForest (RF). También se realizó la búsqueda de hiperparámetros y la comparación de modelos en relación con las emociones y polaridad mediante técnicas de procesamiento de lenguaje natural. Se trabajó con una base de datos recopilada en Twitch, centrándose específicamente en el ámbito de los streamers (creadores de contenido) y en concreto en el entorno de los videojuegos.

En el Capítulo 5, se probaron los modelos de clasificación anteriormente descritos en una nueva base de datos basada centrada en la salud mental en redes sociales, con el fin de predecir las emociones y los niveles de polaridad encontrados. Para ello se optimizaron los hiperparámetros en ambos modelos y finalmente se compararon los resultados de los dos modelos de aprendizaje automático.

En el Capítulo 6, se incluyen las conclusiones principales y las posibles líneas futuras de este Trabajo de Fin de Grado.

Al final de la memoria se incluye la bibliografía usada en este trabajo.

2

Metodología y herramientas de trabajo

2.1 Introducción

En este capítulo, se presentarán las herramientas utilizadas en el desarrollo de este Trabajo de Fin de Grado, así como la metodología que se siguió proporcionando una visión general de cómo se llevó a cabo el proyecto y cómo se utilizaron las diferentes herramientas y metodologías para lograr los objetivos propuestos.

2.2 Herramientas de trabajo

El trabajo se llevará a cabo en el entorno en línea de Google Colaboratory, utilizando el lenguaje de programación Python para el desarrollo de todos los modelos predictivos.

2.2.1 Google Colaboratory

Google Colaboratory, o Colab, es una plataforma en línea desarrollada por Google que permite a los usuarios colaborar de forma remota en la creación y ejecución de notebooks interactivos basados en Jupyter. Estos notebooks se ejecutan en la nube y ofrecen recursos computacionales gratuitos, como CPU y GPU, para ejecutar código y realizar análisis de datos. Colab proporciona una interfaz fácil de usar con bibliotecas y marcos preinstalados para realizar tareas de aprendizaje automático y análisis de datos de manera eficiente.

La principal ventaja de Google Colaboratory es que no requiere una configuración complicada, ya que todo el entorno de programación y análisis se encuentra en la nube. Esto permite a los usuarios colaborar en tiempo real, compartiendo y editando notebooks de forma simultánea. Además, ofrece acceso a recursos potentes, como GPUs, que aceleran el procesamiento de datos y el entrenamiento de modelos de aprendizaje automático [1].

2.2.2 Python

Python es un lenguaje de programación de alto nivel y versátil. Se utiliza ampliamente en diferentes áreas, como desarrollo web, ciencia de datos, inteligencia

artificial y automatización. Se destaca por su sintaxis clara y fácil de entender, lo que facilita la escritura y comprensión del código. Python cuenta con una amplia biblioteca estándar y una comunidad activa que desarrolla una variedad de paquetes adicionales. Es conocido por su enfoque en la legibilidad y la productividad del programador, lo que lo convierte en una opción popular tanto para principiantes como para programadores experimentados. En resumen, Python es un lenguaje de programación poderoso y accesible que permite desarrollar una amplia gama de aplicaciones de manera eficiente [2].

2.2.3 Librerías utilizadas

Se utilizarán diversas librerías ya que se trabajarán con diferentes modelos, como la regresión de capas neuronales, Support Vector Machine (SVM) y Random Forest. Algunas de las librerías que se utilizarán son:

- Pandas: Es una librería que nos ayuda a manejar y trabajar con datos de manera eficiente. Proporciona estructuras de datos especiales, como DataFrames, que nos permiten organizar y manipular nuestros datos de forma sencilla.
- Seaborn: Es una librería de visualización de datos que nos ayuda a crear gráficos estadísticos atractivos y comprensibles. Es especialmente útil para explorar patrones y relaciones en nuestros datos.
- Numpy: Es una librería esencial para el procesamiento numérico en Python. Nos proporciona herramientas para trabajar eficientemente con matrices y arreglos de datos multidimensionales.
- Matplotlib.pyplot: Es una librería de visualización en Python que nos permite crear gráficos y visualizaciones de datos de manera fácil y flexible.
- Scikit-learn (o sklearn) es una de las bibliotecas más populares y ampliamente utilizadas en Python para el aprendizaje automático (machine learning). Proporciona una amplia gama de algoritmos de aprendizaje supervisado y no supervisado, así como herramientas para la preprocesamiento de datos, evaluación de modelos, selección de características y más[3].

En resumen, estas librerías y funciones son herramientas que nos facilitan el trabajo con datos, la evaluación del rendimiento de los modelos de aprendizaje automático y la visualización de resultados. Son muy útiles para desarrollar y mejorar nuestros modelos.

2.2.4 Métricas utilizadas

En este trabajo se usarán las siguientes métricas para analizar los distintos modelos:

- Precisión (Precision): La precisión es la proporción de casos positivos que el modelo clasifica correctamente con respecto al total de casos que clasifica como positivos, es decir, es la medida de cuántos de los casos etiquetados como positivos por el modelo realmente pertenecen a la clase positiva.
- Recall (Recall): El recall, también conocido como sensibilidad o tasa de verdaderos positivos, es la proporción de casos positivos que el modelo clasifica

correctamente con respecto al total de casos positivos reales presentes en los datos, es decir, mide cuántos de los casos positivos se han identificado correctamente.

- F1-score: Es una medida que combina la precisión y el recall en un solo valor, proporcionando un equilibrio entre ambas métricas. Es útil cuando hay un desequilibrio entre las clases y se busca una métrica que considere tanto los falsos positivos como los falsos negativos.
- Support: Representa el número de muestras en el conjunto de prueba que pertenecen a cada clase. Proporciona información sobre la distribución de las clases en los datos de prueba.
- Accuracy: Es la proporción de muestras clasificadas correctamente por el modelo en el conjunto de prueba. Es una medida general del rendimiento del modelo.

2.3 Metodología del trabajo

En esta sección se describe la metodología seguida en este Trabajo de Fin de Grado.

2.3.1 Creación de los modelos predictivos

En este proyecto, se implementan diferentes modelos de aprendizaje automático adaptados a distintas bases de datos y contextos de aplicación. En el Capítulo 3, se utilizan once variables independientes de naturaleza numérica, mientras que en los Capítulos 4 y 5, se cuenta con una variable independiente que consiste en texto sin procesar. En estos dos últimos casos, se requiere aplicar técnicas de procesamiento del lenguaje natural para poder trabajar con dicha variable y utilizarla en los modelos.

2.3.2 Búsqueda de hiperparámetros

En el desarrollo de los modelos de aprendizaje automático se llevan a cabo varias iteraciones, donde se prueban diferentes valores para los hiperparámetros. El objetivo es encontrar la combinación óptima de hiperparámetros que maximice las métricas de exactitud, precisión y recall. Durante este proceso, se ajustan los hiperparámetros y se evalúa el rendimiento del modelo en función de estas métricas, buscando obtener los mejores resultados posibles.

2.3.3 Obtención de resultados finales y comparación

Los resultados obtenidos utilizando los mejores parámetros para cada modelo son evaluados mediante la generación de un informe de clasificación y una matriz de confusión. Estos resultados son luego comparados con los otros modelos estudiados en el mismo capítulo, con el objetivo de determinar qué modelo tiene un mejor rendimiento y en qué circunstancias específicas. Esta comparación nos permite entender cómo se comportan los diferentes modelos y qué modelos son más adecuados para ciertos escenarios o conjuntos de datos.

3

Aplicación de técnicas de machine learning para la clasificación del vino

3.1 Introducción

En este capítulo, se explorarán diferentes modelos para clasificar el vino en función de sus parámetros químicos. Se comenzará con un breve repaso del estado del arte actual en este campo y luego se enfocará en encontrar un modelo capaz de distinguir entre vinos blancos y tintos. Una vez logrado esto, se buscarán modelos tanto de regresión como de clasificación para cumplir con el objetivo establecido.

3.2 Estado del arte

En esta sección, se hará un análisis descriptivo del estado del arte con relación a la predicción de la calidad de los vinos utilizando diferentes modelos de aprendizaje automático y se describirán sus respectivos resultados. En dicho análisis se buscarán modelos tanto de regresión como de clasificación.

El primer trabajo que se revisará [4] y fechado en 2016, aborda el uso de validación cruzada "K-fold" y PCA (Análisis de Componentes Principales) para la reducción de variables. Los modelos considerados son SVM, KNN y Random Forest (RF). Se evaluaron múltiples medidas, como precisión, recall, F1, ROC y exactitud. Los resultados revelaron que el mejor modelo para el vino tinto fue Random Forest, con una precisión del 69.606%. Al aplicar validación cruzada, esta cifra aumentó a un 71.875%. Al emplear PCA para reducir las variables, se obtuvo una precisión del 71.232% y se alcanzó un 73.4375% con validación cruzada. Por otro lado, para el vino blanco, se logró una precisión del 70.3757% y un 68.6735% con validación cruzada. Al utilizar PCA, la precisión alcanzada fue del 69.9061%, mientras que con validación cruzada se obtuvo un 67.449%. En general, se puede apreciar que Random Forest supera significativamente al resto de los modelos evaluados. Además, se observa que tanto PCA como la validación cruzada pueden contribuir a mejorar la precisión del modelo, especialmente en el caso del vino tinto, donde se logró aumentar del 69% al 73%.

El trabajo que se presenta en [5] utiliza únicamente la base de datos de vino tinto y emplea los modelos SVM, Random Forest y Naive Bayes. El mejor modelo resultó ser SVM, con una precisión del 68.64%, seguido de Random Forest, que obtuvo un 65.46% de accuracy. Naive Bayes alcanzó un 55% de precisión.

El trabajo desarrollado por los autores en [6] con fecha de 2021, utiliza tanto la base de datos de vino blanco como la de vino tinto, y se implementa un modelo híbrido de Random Forest y SVM, logrando una precisión del 66% para el vino tinto y del 67% para el vino blanco.

El último artículo revisado [7], fechado en 2018, realiza un análisis exclusivamente sobre la base de datos de vino tinto. En este análisis, se utiliza una proporción de 80:20 para dividir los datos en conjuntos de entrenamiento y prueba. Los vinos se clasifican en dos categorías: "buenos" y "malos", siendo los vinos por encima de cinco considerados "buenos" y los vinos por debajo de cinco categorizados como "malos". Además, se lleva a cabo la estandarización de los datos y se eliminan los valores atípicos, lo que resulta en la eliminación de 438 muestras. En este estudio, se implementan dos algoritmos de aprendizaje automático: Random Forest (RF) y Regresión Logística (LR). Los resultados muestran que con Random Forest se logra una precisión (accuracy) del 84%, mientras que con Regresión Logística se obtiene un 76%. Es importante tener en cuenta que en este caso solo hay dos clases para clasificar lo que hace que se obtenga una precisión mayor que en los anteriores trabajos de investigación.

Tras analizar estos documentos, se comprueba que todavía existe margen de mejora. Tal y como se observa, los modelos más utilizados son modelos de clasificación, especialmente Random Forest y SVM, ya que son los que ofrecen mejores resultados. Muchos documentos se centran únicamente en la clasificación en dos clases, vinos buenos y malos, pero aquí estamos buscando una clasificación para todas las clases, esto es, notas (entre 0-10) que clasifiquen la calidad del vino, lo que hace más complejo el problema de optimización. Finalmente, se puede observar que la precisión se acerca al 70% en la mayoría de los documentos para las dos clases de vinos.

3.3 Bases de datos

En este trabajo de investigación, se utilizan dos bases de datos: una para el vino blanco y otra para el vino tinto. Ambas bases de datos contienen 12 parámetros, siendo uno de ellos la calidad, que representa un valor entre 0 y 10 y se refiere a la calidad del vino. Nuestro objetivo es predecir este valor utilizando un modelo entrenado con 11 parámetros de carácter químico, las cuales son: "fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates y alcohol".

Entender estas bases de datos nos permitirá explorar la relación entre los diferentes parámetros del vino y su calidad, así como desarrollar un modelo que pueda predecir la calidad del vino basándose en estos parámetros.

3.3.1 Base de Datos Vino Blanco

Esta base de datos hace referencia al vino blanco con un total de 4898 muestras, las cuales toman un valor en el parámetro calidad entre 3 y 9. La distribución de las notas de calidad se puede observar en la Figura 1 y en la Tabla 1.

Nota (Vino blanco)	Muestras (4898)	Representación en porcentaje
3	20	0.4%
4	163	3.3%
5	1457	29.747%
6	2198	44.875%
7	880	17.97%
8	175	3.57%
9	5	0.1%

Tabla 1: Muestras vino blanco

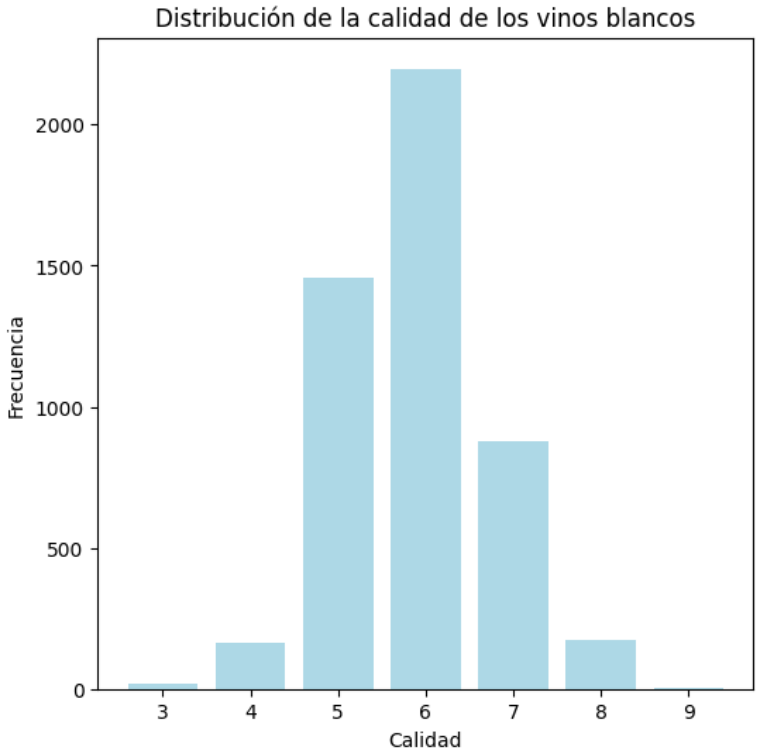

Figura 1: Gráfico muestras vino blanco

En esta base de datos, se observa que la mayoría de las muestras se concentran en las notas de calidad 5, 6 y 7. Es importante destacar que esta base de datos está significativamente desbalanceada. Este desequilibrio es comprensible, ya que los vinos con características excepcionales son menos comunes, al igual que los vinos de calidad muy baja, los cuales no serían comercializados.

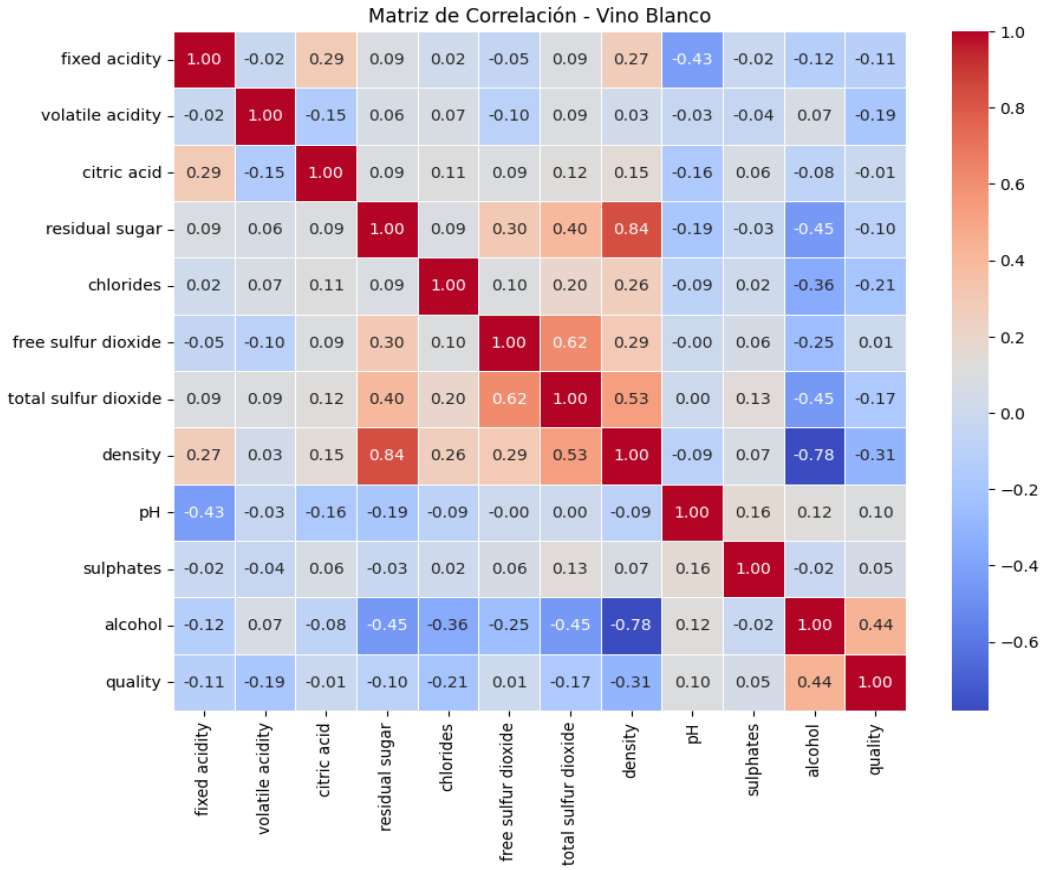

Figura 2: Matriz de correlación vino blanco

Por otro lado, también se ha enfocado en el análisis en la relación entre los parámetros químicos y la calidad del vino. Para visualizar esta relación, se ultiliza una matriz de correlación que incluye todas las variables, como se muestra en la Figura 2. La matriz de correlación nos permitirá identificar las posibles asociaciones entre las diferentes variables y la calidad del vino.

En la Figura 2, se pueden observar varias relaciones de variables con una correlación significativa. Algunos ejemplos destacados son:

1. Densidad y azúcar residual: Existe una correlación intensa de 0.84 entre la densidad del vino y el contenido de azúcar residual. Esto sugiere que a medida que aumenta el contenido de azúcar residual, la densidad del vino tiende a incrementar.
2. Densidad y alcohol: Hay una correlación fuerte de -0.78 entre la densidad y el contenido de alcohol. Esto indica que a medida que aumenta el porcentaje de alcohol, la densidad del vino tiende a disminuir.

3. Dióxido de azufre libre y dióxido de azufre total: Estas dos variables presentan una correlación significativa de 0.62. Esto sugiere que existe una relación entre la cantidad de dióxido de azufre libre y la cantidad total de dióxido de azufre presente en el vino.
4. Dióxido de azufre total y densidad: Se observa una correlación de 0.53 entre el dióxido de azufre total y la densidad del vino. Esto implica que a medida que aumenta la concentración de dióxido de azufre total, la densidad también tiende a incrementar.

En cuanto a la calidad del vino, la variable que muestra la mayor correlación es el alcohol, con un valor de 0.44. Esto indica que existe una relación moderada pero positiva entre el contenido de alcohol y la calidad del vino, lo que sugiere que un mayor porcentaje de alcohol puede estar asociado con una mayor calidad percibida.

3.3.2 Base de Datos Vino Tinto

La base de datos utilizada contiene muestras de vino tinto, con un total de 1599 muestras. Estas muestras han sido evaluadas en términos de calidad, que varía en un rango de 3 a 8. La distribución de las notas de calidad se puede observar en la Tabla 2 y en la Figura 3.

Nota (Vino Tinto)	Muestras (1599)	Representación en porcentaje
3	10	0.625%
4	53	3.31%
5	681	42.58%
6	638	39.9%
7	199	12.45%
8	18	1.126%

Tabla 2: Muestras vino tinto

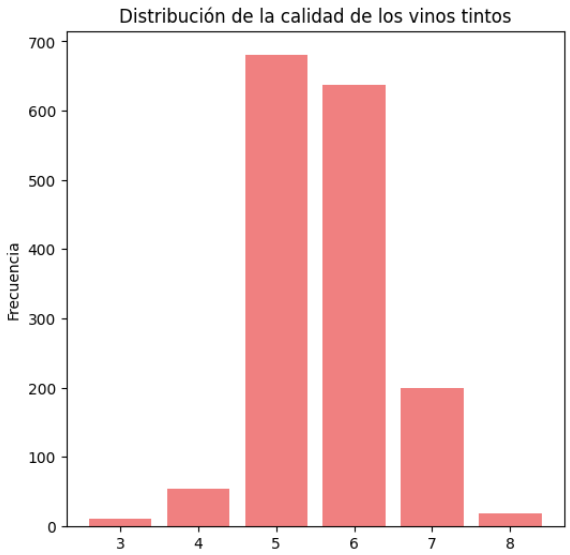

Figura 3:Gráfico muestras vino tinto

Es interesante observar que, al igual que en el caso del vino blanco, la mayoría de las muestras de vino tinto se concentran en los valores de calidad 5, 6 y 7. Esto indica que, al igual que el vino blanco, la base de datos de vino tinto también está desbalanceada.

Además, es importante destacar que la base de datos de vino tinto contiene significativamente menos muestras en comparación con la base de datos de vino blanco. Esta diferencia en la cantidad de muestras puede tener implicaciones en el análisis y en la generalización de los resultados obtenidos.

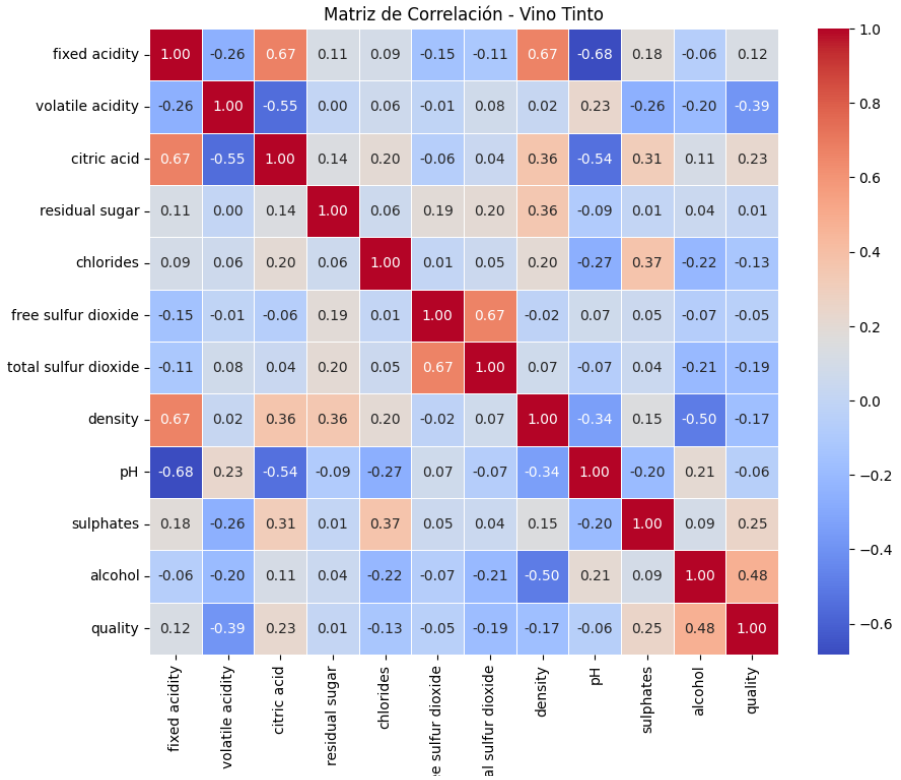

Figura 4: Matriz de correlación vino tinto

En la Figura 4, se pueden observar las correlaciones más intensas entre las variables en la base de datos de vino tinto. Algunas correlaciones destacadas son:

1. pH y acidez fija: Existe una correlación intensa de -0.68 entre el pH del vino y la acidez fija. Esto sugiere que a medida que aumenta la acidez fija, el pH tiende a disminuir.
2. Acidez fija y densidad: Hay una correlación fuerte de 0.67 entre la acidez fija y la densidad del vino. Esto indica que a medida que aumenta la acidez fija, la densidad también tiende a aumentar.
3. Acidez fija y ácido cítrico: Se observa una correlación significativa de 0.67 entre la acidez fija y el contenido de ácido cítrico. Esto sugiere una relación positiva entre ambas variables.
4. Acidez volátil y ácido cítrico: Estas dos variables presentan una correlación de -0.55. Esto indica una asociación inversa entre la acidez volátil y el contenido de ácido cítrico en el vino.
5. Ácido cítrico y pH: Se puede observar una correlación de -0.54 entre el ácido cítrico y el pH del vino. Esto implica una relación inversa entre estas dos variables.

6. Densidad y alcohol: Existe una correlación de -0.5 entre la densidad y el contenido de alcohol. Esto sugiere que a medida que aumenta el porcentaje de alcohol, la densidad del vino tiende a disminuir.

En cuanto a la calidad del vino, se encuentran correlaciones intensas con el alcohol (0.48) y la acidez volátil (-0.39). Esto indica que el contenido de alcohol y la acidez volátil pueden tener una influencia significativa en la calidad percibida del vino tinto.

Estas correlaciones destacadas proporcionan información valiosa sobre las relaciones entre las variables químicas y la calidad del vino tinto en la base de datos analizada.

3.4 Clasificación en función del tipo de vino

Tras analizar las bases de datos de vino tinto y vino blanco por separado, ahora se confrontarán los parámetros químicos para identificar posibles diferencias significativas entre ambas variedades. En este análisis, se han identificado tres variables que presentan

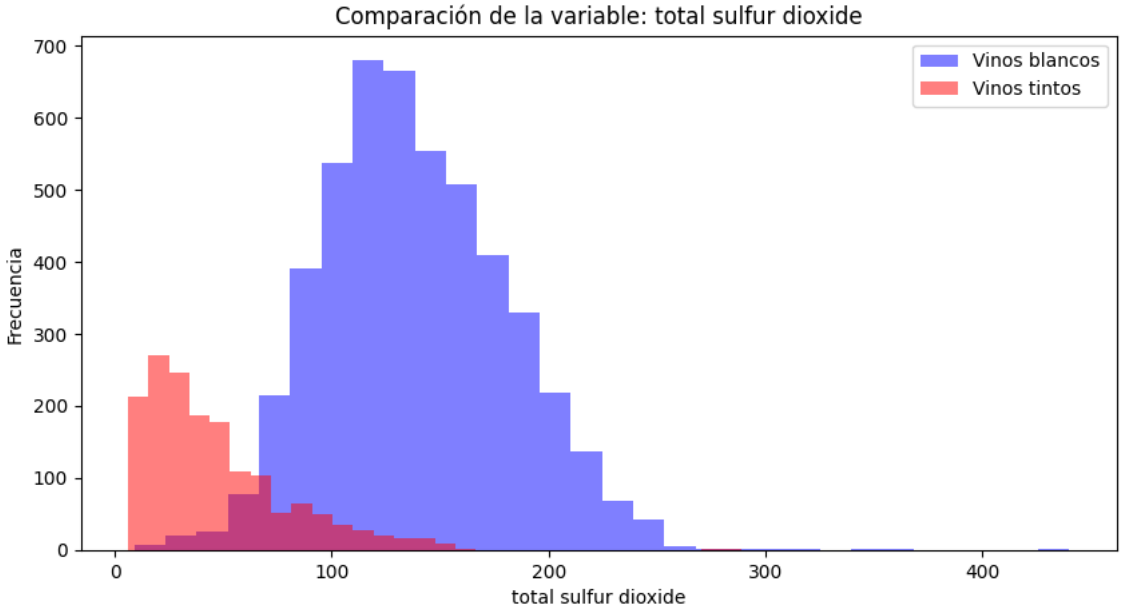

Figura 5: Variable total sulfur dioxide

las mayores diferencias: total sulfur dioxide (dióxido de azufre total), free sulfur dioxide (dióxido de azufre libre) y volatile acidity (acidez volátil). Estas diferencias se ilustran en las Figuras 5, 6 y 7, respectivamente.

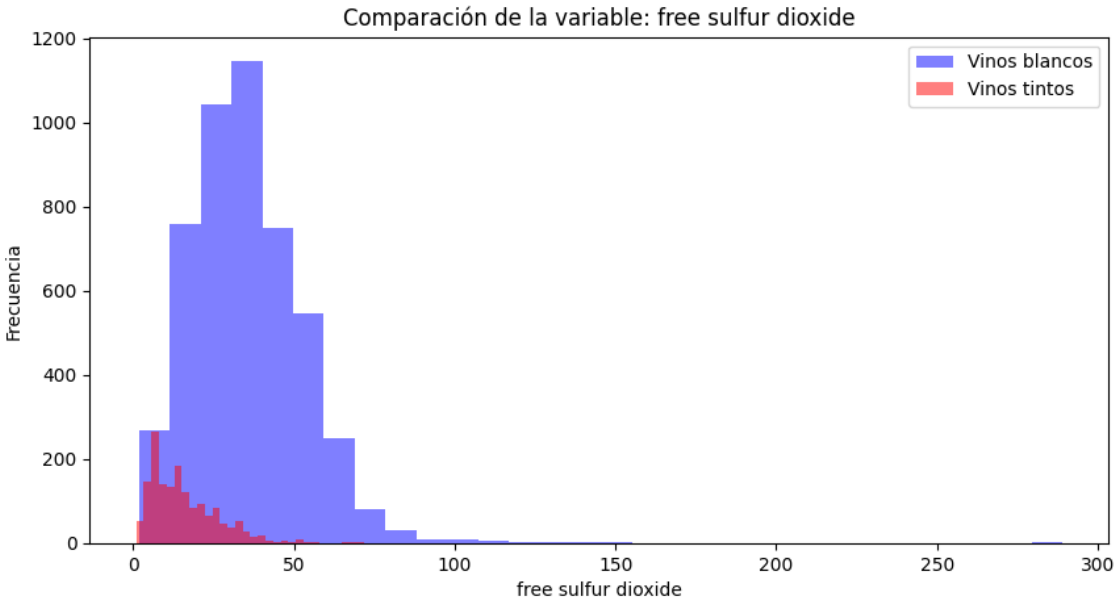

Figura 7:Variable free sulfur dioxide

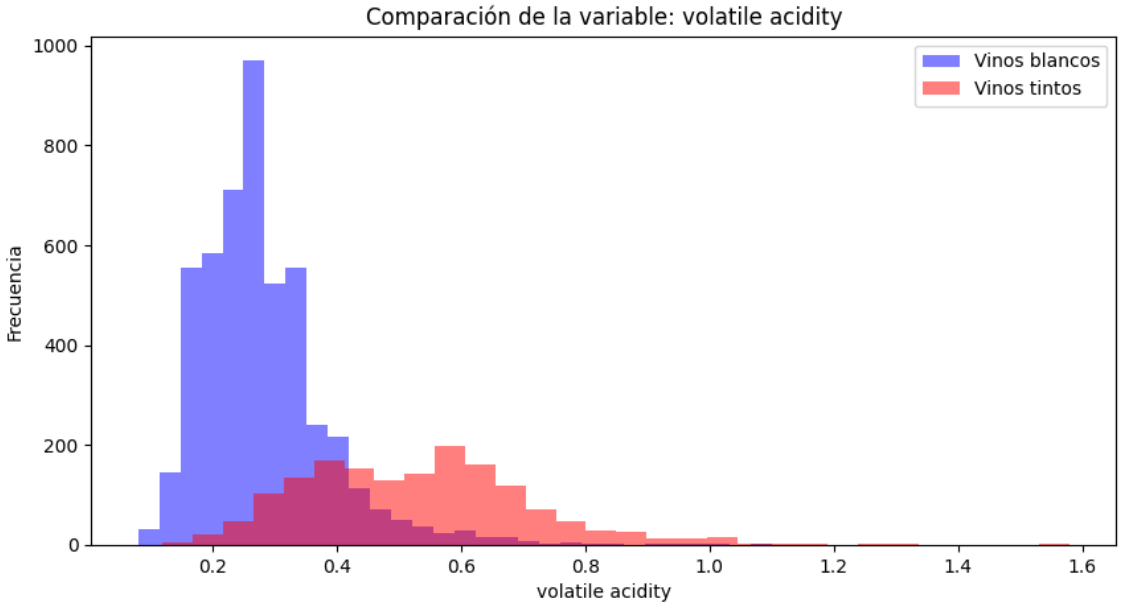

Figura 6:Variable volatile acidity

	Precision	Recall	F1-score	Support	Accuracy
Tinto	93%	86%	89%	477	-
Blanco	95%	98%	97%	1473	-
Global	95%	-	95%	1950	95.12%

Tabla 3: Clasificación PCA 3 variables

A continuación, se realizará una prueba inicial utilizando un modelo de Random Forest con PCA, utilizando las tres variables mencionadas anteriormente: total sulfur dioxide, free sulfur dioxide y volatile acidity. Se establecerá un tamaño de prueba del 30% y se emplearán 200 árboles de decisión. Los resultados se presentan en la Tabla 3.

Se han obtenido resultados muy prometedores con una precisión del 95% utilizando únicamente tres variables. Ahora, se procede a realizar unas pruebas adicionales utilizando PCA con las seis variables más importantes y posteriormente con las nueve variables más relevantes. Los resultados de estas pruebas se mostrarán en las Tablas 4 y 5, respectivamente.

	Precision	Recall	F1-score	Support	Accuracy
Tinto	95%	91%	93%	477	-
Blanco	97%	99%	98%	1473	-
Global	97%	-	97%	1950	97%

Tabla 4: Clasificación PCA 6 variables

	Precision	Recall	F1-score	Support	Accuracy
Tinto	98%	95%	96%	477	-
Blanco	98%	100%	99%	1473	-
Global	98%	-	98%	1950	98%

Tabla 5: Clasificación PCA 9 variables

Se ha observado una progresión en el rendimiento a medida que aumentamos el número de variables utilizadas en nuestro modelo. Por lo tanto, se repetirá el ejemplo sin utilizar PCA, considerando las once variables proporcionadas en la base de datos. Los resultados de esta prueba se muestran en la Tabla 6.

	Precision	Recall	F1-score	Support	Accuracy
Tinto	100%	98%	99%	477	-
Blanco	99%	100%	100%	1473	-
Global	100%	-	100%	1950	99.5%

Tabla 6: Clasificación 11 variables

Con las once variables se obtiene el mejor resultado, con casi un 100% de precisión y exactitud para ambos tipos de vino, pero cabe destacar que usando únicamente tres variables los resultados también eran bastante buenos.

Se puede concluir que las once variables presentan independencia entre sí, lo que significa que todas son útiles para clasificar si un vino es tinto o blanco. Sin embargo, el dióxido de azufre total, el dióxido de azufre libre y la acidez volátil tienen la capacidad de clasificar el vino con una precisión del 95% por sí solas, lo que permite obtener un modelo muy sencillo. Por otro lado, al utilizar las once variables, se aumenta ligeramente la complejidad del modelo y se logra casi el 100% de precisión, lo cual es notable. Por lo tanto, es recomendable utilizar todas las variables, ya que cada una aporta información valiosa (algunas en menor cantidad) para mejorar la precisión de la clasificación del vino.

3.5 Clasificación de la calidad del vino mediante técnicas de regresión

En el punto anterior se ha conseguido encontrar un modelo que nos permite diferenciar con casi el 100% de exactitud si un vino es tinto o blanco. Gracias a esto se pueden crear nuevos modelos más concretos y diferenciados para vino tinto o blanco y clasificarlos en función de su posible nota, que corresponde a la variable calidad.

La regresión se emplea en un enfoque estadístico para estudiar y comprender la asociación entre diferentes variables [8]. Su objetivo es proporcionar predicciones o explicaciones acerca de una variable dependiente la cual en este trabajo será la calidad y para ello, se usarán sus once parámetros químicos, esperando encontrar una relación lineal entre ellas. Por lo tanto, se han aplicado técnicas de regresión lineal para predecir la calidad del vino basándonos en las bases de datos previamente descritas.

3.5.1 Red de capas neuronales

Una red de capas neuronales es un modelo de aprendizaje automático que se basa en la estructura del cerebro humano. En esta red, los nodos, que se llaman neuronas, se agrupan en capas. Los datos de entrada se procesan a través de capas intermedias conocidas como capas ocultas, donde se extraen características más complejas y abstractas. Finalmente, en la capa de salida se generan las predicciones o resultados deseados. Durante el entrenamiento de la red, los pesos de las conexiones entre las neuronas se ajustan para mejorar la precisión de las predicciones, permitiendo que la red aprenda a partir de los datos y realice predicciones más precisas a medida que avanza el proceso de entrenamiento[9]. Los resultados que se obtendrán se redondearán a la clase más próxima, si se obtiene una salida de 7.2, será aproximada a la clase de calidad 7. Por lo que se usaran la red de capas neuronales en forma de un modelo de regresión.

En la búsqueda de hiperparámetros de dicho modelo, se enfocará en dos aspectos clave: el número de capas y el número de neuronas en la red neuronal. Se utilizará la técnica de búsqueda en cuadrícula (*grid search*) para explorar diferentes combinaciones de estos valores.

Además, para evaluar el rendimiento de cada configuración de hiperparámetros, se implantará la validación cruzada con 5 splits. Esto permitirá obtener una estimación más robusta y precisa del rendimiento del modelo.

Para asegurarse de que el modelo no se sobreajuste durante el entrenamiento, se aplicará la técnica de *early stopping*. Para ello, se configura un límite de paciencia (*patience*) de 30 épocas, lo que significa que, si el modelo no mejora después de 30 épocas consecutivas, se detendrá automáticamente el entrenamiento y pasará al siguiente split.

En cuanto a la división de los conjuntos de datos, se utilizará un tamaño de prueba (*test size*) del 20% para ambos conjuntos de datos, y se estandarizarán las variables mediante la técnica *Standarscaler* el cual transforma las características de modo que tengan media cero y desviación estándar de uno.

El número total de épocas para el entrenamiento será de 10000. Sin embargo, gracias al *early stopping*, el modelo se detendrá antes si no se observan mejoras adicionales, se ha elegido un número tan grande de épocas para forzar siempre el *early stopping.*

En resumen, se utilizará *grid search* con validación cruzada de 5 splits para explorar diferentes combinaciones de número de capas y neuronas, implementando *early stopping* con una paciencia de 30 épocas para evitar el sobreajuste. Se dividirán los conjuntos de datos en una proporción de 80:20 para entrenamiento y prueba respectivamente. Y finalmente, se realizará un entrenamiento de hasta 10000 épocas, aunque este nunca se dará, ya que se detendrá antes cuando el *early stopping* se active y el modelo haya convergido. La capa final será siempre 1. La validación cruzada con n=5, el tamaño de *test size*, el uso de *standarscaler*, y la semilla se repetirá en todos los modelos, para que sean comparables en las mismas condiciones.

3.5.1.1 Resultados de redes neuronales para el vino blanco

En primer lugar, se inicia el proceso de búsqueda de hiperparámetros para el vino blanco. Para esta tarea, se establecerá un máximo de cuatro capas y se examinará en la métrica de exactitud (*accuracy*) para comparar todos los modelos generados. Los resultados obtenidos se presentan en la Tabla 7.

	Capa Inicial	Capa 2	Capa 3	Capa 4	Capa final	accuracy
1- Neuronas	2	-	-	-	1	48%
2- Neuronas	4	-	-	-	1	50%
3- Neuronas	4	2	-	-	1	49%
4- Neuronas	8	-	-	-	1	54%
5- Neuronas	8	4	-	-	1	53%
6- Neuronas	8	4	2	-	1	52%
7- Neuronas	16	-	-	-	1	50%
8- Neuronas	16	8	-	-	1	53%
9- Neuronas	16	8	4	-	1	52%
10-Neuronas	16	8	4	2	1	53%
11- Neuronas	32	-	-	-	1	54%
12- Neuronas	32	16	-	-	1	54%
13- Neuronas	32	16	8	-	1	54%
14- Neuronas	32	16	8	4	1	54%
15- Neuronas	64	-	-	-	1	55%
16- Neuronas	64	32	-	-	1	53%
17- Neuronas	64	32	16	-	1	53%
18- Neuronas	64	32	16	8	1	54%
19- Neuronas	128	-	-	-	1	56%
20-Neuronas	128	64	-	-	1	56%
21- Neuronas	128	64	32	-	1	55%
22- Neuronas	128	64	32	16	1	54%
23- Neuronas	256	-	-	-	1	56%
24- Neuronas	256	128	-	-	1	53%
25- Neuronas	256	128	64	-	1	54%
26- Neuronas	256	128	64	32	1	52%
27- Neuronas	512	-	-	-	1	58%
28- Neuronas	512	256	-	-	1	56%
29- Neuronas	512	256	128	-	1	57%
30-Neuronas	512	256	128	64	1	56%
31- Neuronas	1024	-	-	-	1	58%

32- Neuronas	1024	512	-	-	1	58%
33- Neuronas	**1024**	**512**	**256**	**-**	**1**	**59%**
34- Neuronas	1024	512	256	128	1	57%

Tabla 7: Modelo de red neuronal para el dataset del vino blanco

Durante la búsqueda de hiperparámetros, identificamos el valor óptimo en la iteración 33 utilizando una estructura de red neuronal compuesta por 4 capas, con tamaños respectivos de 1024, 512, 256 y 1. Esta configuración de capas mostró un rendimiento ligeramente superior en comparación con los demás modelos, tal y como se observa en la Tabla 7, con un nivel de predicción del 57%.

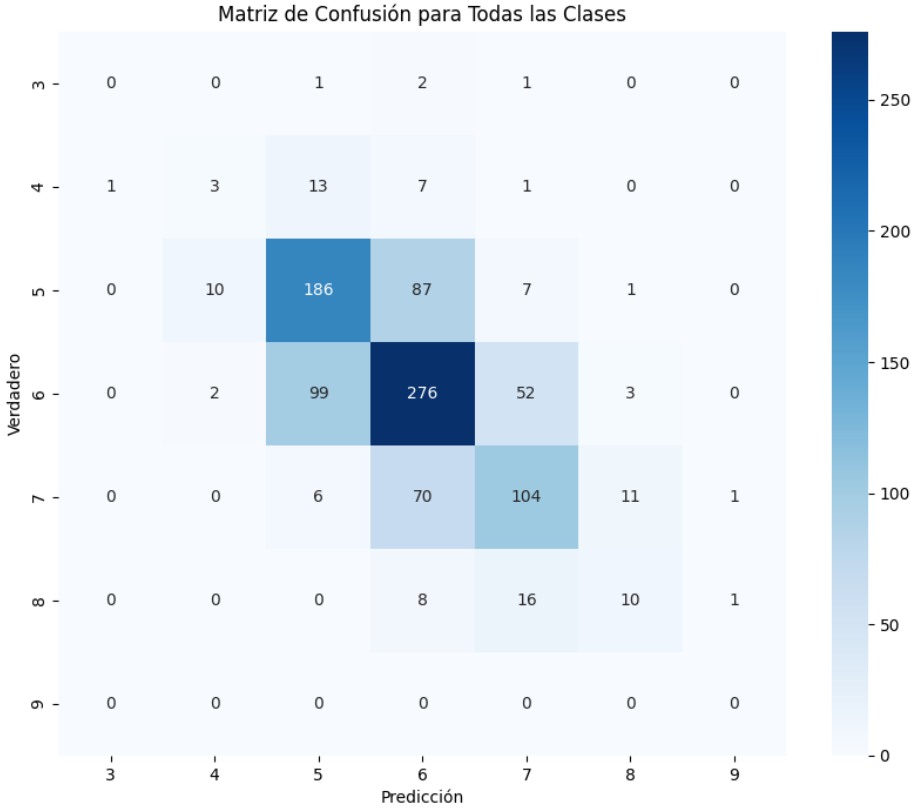

Figura 8: Matriz de confusión vino blanco red neuronal

	Precision	Recall	F1-score	Support	Accuracy
3	0%	0%	0%	5	0%
4	29%	12%	15%	25	12%
5	61%	64%	62%	291	64%
6	61%	64%	63%	432	64%
7	57%	54%	56%	192	54%
8	40%	29%	33%	35	29%
Global	58%	59%	59%	980	59%

Tabla 8: Informe de clasificación para el vino blanco mediante red neuronal

La matriz de confusión, representada en la Figura 8, proporciona una visualización detallada de cómo el modelo clasificó correcta e incorrectamente las muestras de datos.

La Tabla 8, por otro lado, muestra el informe de clasificación correspondiente, que incluye las medidas de precisión, recall, f1-score, numero de muestras y exactitud para cada clase. Todo ello referido a la iteración 33 de la Tabla 7. Se observa, tanto en la Tabla 8 como en la Figura 8, que este modelo muestra una precisión inferior al 60% al predecir la calidad del vino blanco. Las métricas de precisión y recall son bastante similares para las clases 5, 6 y 7, lo cual indica que el modelo puede clasificar adecuadamente esas categorías. Sin embargo, las clases 4 y 8 muestran una disminución drástica en el rendimiento. Solo se predicen 15 muestras en la clase 4, de las cuales solo 3 son clasificadas correctamente, como se puede apreciar en la Figura 8. Esto resulta en valores muy bajos tanto de precisión como de recall para esas categorías. Se observa una situación similar para la clase 8. En cuanto a la clase 3, el modelo no logra clasificar ninguna instancia, lo cual puede deberse a la escasez de muestras disponibles para esa clase.

3.5.1.2 Resultados de redes neuronales para el vino tinto

Siguiendo la misma metodología, en primer lugar, se realiza la búsqueda de hiperparámetros igual que en el modelo anterior, pero en este caso para el vino tinto mostrando los resultados en la Tabla 9.

	Capa Inicial	Capa 2	Capa 3	Capa 4	Capa final	accuracy
1- Neuronas	2	-	-	-	1	44%
2- Neuronas	4	-	-	-	1	48%
3- Neuronas	4	2	-	-	1	51%
4- Neuronas	8	-	-	-	1	54%
5- Neuronas	8	4	-	-	1	54%
6- Neuronas	8	4	2	-	1	55%
7- Neuronas	16	-	-	-	1	55%
8- Neuronas	16	8	-	-	1	51%
9- Neuronas	16	8	4	-	1	52%
10-Neuronas	16	8	4	2	1	54%
11- Neuronas	32	-	-	-	1	55%
12- Neuronas	32	16	-	-	1	53%
13- Neuronas	32	16	8	-	1	52%
14- Neuronas	32	16	8	4	1	57%
15- Neuronas	64	-	-	-	1	56%
16- Neuronas	64	32	-	-	1	59%
17- Neuronas	64	32	16	-	1	57%
18- Neuronas	64	32	16	8	1	60%
19- Neuronas	128	-	-	-	1	60%
20-Neuronas	128	64	-	-	1	61%
21- Neuronas	128	64	32	-	1	61%
22- Neuronas	**128**	**64**	**32**	**16**	**1**	**62%**
23- Neuronas	256	-	-	-	1	59%
24- Neuronas	256	128	-	-	1	59%
25- Neuronas	256	128	64	-	1	59%
26- Neuronas	256	128	64	32	1	58%
27- Neuronas	512	-	-	-	1	57%
28- Neuronas	512	256	-	-	1	58%
29- Neuronas	512	256	128	-	1	57%
30-Neuronas	512	256	128	64	1	56%
31- Neuronas	1024	-	-	-	1	57%
32- Neuronas	1024	512	-	-	1	57%
33- Neuronas	1024	512	256	-	1	58%
34- Neuronas	1024	512	256	128	1	58%

Tabla 9: Modelo de red neuronal para el dataset del vino tinto

En el presente caso, se ha encontrado el valor óptimo para este modelo en la iteración 22, utilizando una estructura de red neuronal compuesta por 5 capas con tamaños respectivos de 128, 64, 32, 16 y 1. La Figura 9 y la Tabla 10 muestran la matriz de confusión y el informe de clasificación correspondiente, respectivamente. Estos resultados nos brindan una visión detallada del rendimiento del modelo en términos de clasificación y nos permiten evaluar su precisión en cada clase.

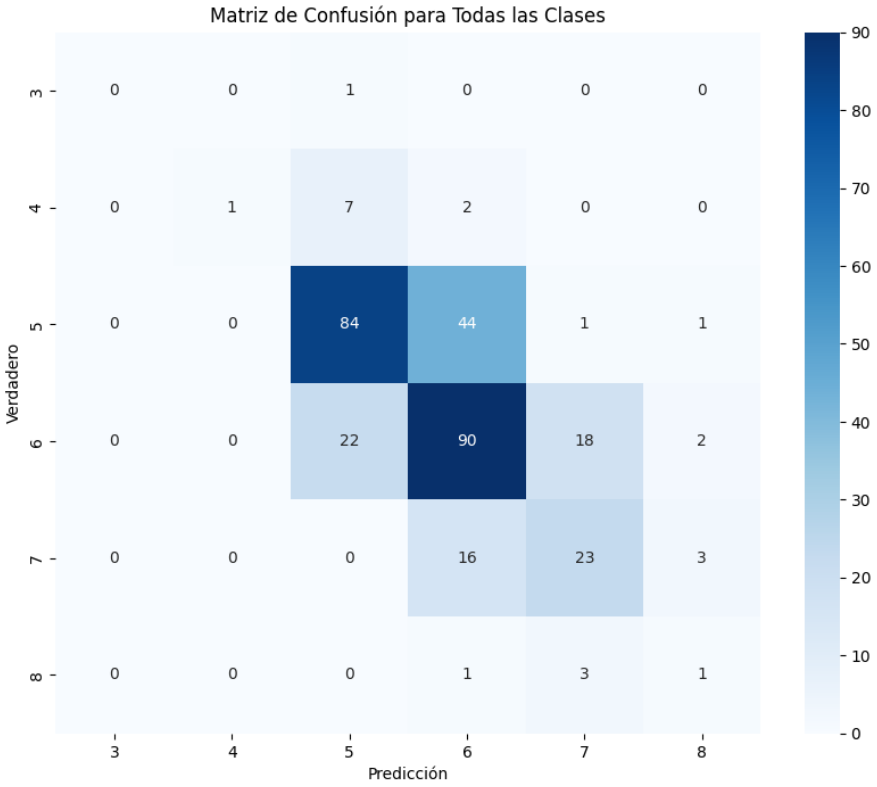

Figura 9: Matriz de confusión vino tinto red neuronal

	Precision	Recall	F1-score	Support	Accuracy
3	0%	0%	0%	1	-
4	100%	10%	18%	10	-
5	74%	65%	69%	130	-
6	59%	68%	63%	132	-
7	51%	55%	53%	42	-
8	14%	20%	17%	5	-
Global	64%	-	62%	320	62%

Tabla 10: Informe de clasificación de la red neuronal para el vino tinto

Se comprueba que este modelo utiliza significativamente menos neuronas en comparación con el modelo para el vino blanco. A pesar de eso, logra una exactitud superior del 62%. Sin embargo, observamos que las clases 3, 4 y 8 no son clasificadas correctamente, posiblemente debido a la escasez de muestras disponibles en la base de datos. También cabe destacar que en la clase 4 se consigue una precision del 100%, pero

no es tan reseñable ya que viendo la Figura 9 solo se ha predicho una instancia en esta clase (hay diez), pero esta ha sido acertada, por lo que se obtiene también un recall de 0.1.

3.6 Clasificación de la calidad del vino mediante técnicas de clasificación

Los modelos de clasificación son técnicas que se utilizan en el aprendizaje automático con el fin de asignar las muestras a diferentes categorías. En este trabajo las clases serán las diferentes notas de la variable calidad, números entero del 0 (muy malo) al 10 (excelente). Estos modelos se construyen entrenando diferentes algoritmos con datos etiquetados para que aprendan patrones y relaciones entre características y etiquetas. Se usarán los modelos de SVM y Random Forest.

3.6.1 SVM (Support Vector Machine)

SVM (*Support Vector Machine*) es un algoritmo de aprendizaje supervisado utilizado para clasificar datos. Busca un plano óptimo que pueda separar las diferentes clases de manera efectiva, en nuestro caso las clases son los valores de calidad. Puede manejar conjuntos de datos de alta dimensionalidad y utilizar funciones de kernel para encontrar límites de separación más complejos, este será uno de los hiperparámetros que se buscarán. En resumen, como se mostró en el estado del arte, SVM es un algoritmo muy interesante para este problema de clasificación[9], [10].

Los hiperparámetros que se buscaran optimizar dentro de este modelo de aprendizaje automático son:

1. Parámetro de regularización (C): Controla el equilibrio entre el ajuste de los datos de entrenamiento y la generalización del modelo. Un valor más alto de C dará como resultado un modelo que se ajusta más a los datos de entrenamiento, pero puede ser propenso al sobreajuste. Un valor más bajo de C da lugar a un modelo más regularizado.
2. Tipo de kernel: SVM utiliza diferentes tipos de kernels para mapear los datos en espacios de mayor dimensionalidad. Los tipos de kernel comunes incluyen lineal, polinomial y radial (RBF). La elección del kernel depende de la naturaleza de los datos y la complejidad de la relación entre las clases.
3. Parámetros del kernel: Los kernels como el polinomial y el RBF tienen parámetros adicionales que afectan la forma de la función de decisión. Por ejemplo, en el kernel polinomial, se debe especificar el grado del polinomio, y en el kernel RBF, se debe especificar la anchura de banda [9].

El orden que se seguirá para realizar la búsqueda de los hiperparámetros óptimos será: tipo de kernel, parámetros de kernel y finalmente el parámetro de regularización.

También se usará validación cruzada y se estandarizarán las variables con el método *StandarScaler*, esto se repetirá en todos los modelos.

3.6.1.1 Resultados de SVM para vino blanco

En primer lugar, se inicia la búsqueda de hiperparámetros, comenzando con la variación del kernel y reflejando los resultados en la Tabla 11.

kernel	Accuracy
Linear	51%
Poly	52%
Rbf	**56%**
sigmoid	41%

Tabla 11: Optimización del parámetro kernel para el modelo SVM en el vino blanco

Se han obtenido los mejores resultados utilizando el kernel RBF. A continuación, se procederá a buscar el mejor valor para el parámetro C, cuyos resultados se muestran en la Tabla 12.

C	Accuracy RBF
1	56%
5	58%
10	59%
100	61%
1000	63%
10000	62%
0.1	53%
0.01	44%
0.001	44%
500	**63%**
250	62%
1250	62%
1500	62%
600	63%
400	63%
300	62%

Tabla 12: Optimización del parámetro C para el modelo SVM en el vino blanco

Para el parámetro de regularización, se encuentra que el valor óptimo es C=500, lo que proporciona una exactitud del 63%. A continuación, se procederá a buscar el hiperparámetro gamma utilizando el kernel RBF y C=500. Los resultados de esta búsqueda se presentan en la Tabla 13.

gamma	Accuracy, rbf,C=500
1	67%
10	64%
100	62%
0.1	63%
0.01	56%
2	66%
1.5	68%
1.25	68%
1.1	67%
0.9	67%
0.8	67%
0.5	66%

1.6	68%
1.7	67%
1.3	68%
1.4	**68%**

Tabla 13: Optimización del parámetro Gamma para el modelo SVM en el vino blanco

Tal y como se observa, el mejor valor para el hiperparámetro gamma resultó ser 1.4, lo cual brinda una exactitud del 68%. Utilizando estos tres hiperparámetros optimizados, presentamos la Tabla 14 y la Figura 10. La Tabla 14 muestra el informe de clasificación para todas las clases. Por otro lado, la Figura 10 representa la matriz de confusión, brindando una visualización detallada de cómo el modelo ha clasificado correcta e incorrectamente las muestras de datos en cada clase.

	Precision	Recall	F1-score	Support	Accuracy
3	0%	0%	0%	5	-
4	44%	16%	24%	25	-
5	77%	58%	66%	291	-
6	61%	85%	71%	432	-
7	78%	55%	64%	192	-
8	94%	46%	62%	35	-
Global	70%	-	66%	980	68%

Tabla 14: Informe de clasificación para el modelo SVM para el vino blanco

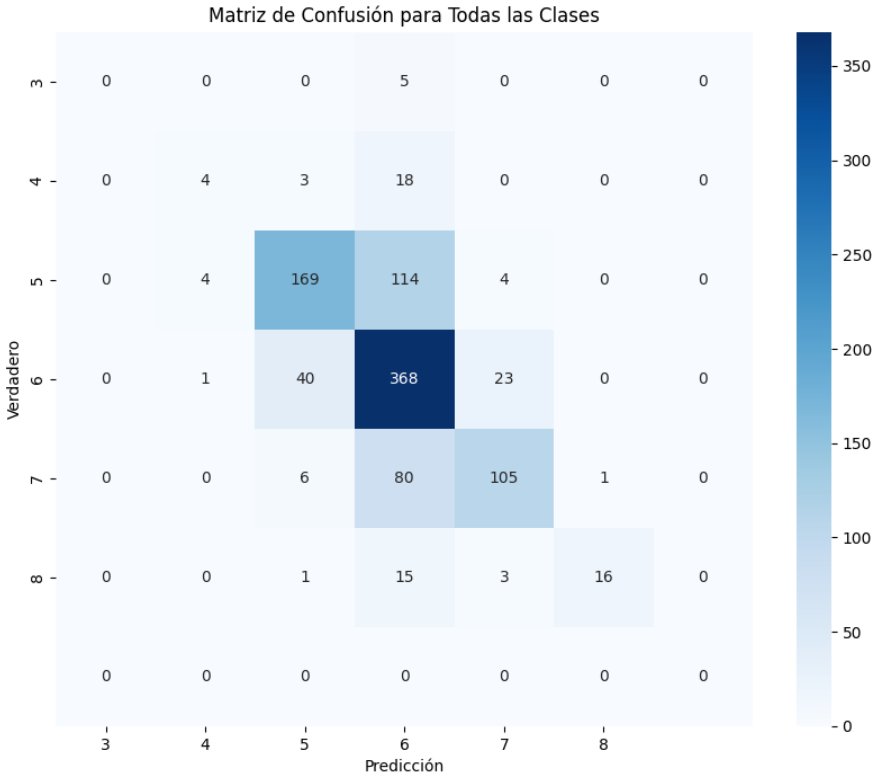

Figura 10: Matriz de confusión del modelo SVM para el vino blanco

Con todos estos valores optimizados, se consigue un sistema que clasifica con un 68% de exactitud para el resultado exacto, y con un 94.9% de exactitud con margen de error de 1 punto, es decir confundir una clase X con otra clase que sea X + 1 o X - 1.

Son resultados significativamente mejores que la red neuronal analizada en el modelo anterior de regresión, pasamos de un 58% a un 68% de exactitud, mejorando la precisión para todas las clases. Centrándose en este modelo presenta bastante variación entre la precision y el recall en general, en la clase 8 obtiene un 94% de precision pero tan solo un 46% de recall, debido a que el modelo predice 17 muestras en esta clase (16 correctas) pero hay 35 en total, esto se repite en todas las clases a excepción de la clase 6 (la cual también es la que más muestras tiene) viendo la disparidad es más interesante quedarse con el parámetro de f1-score que con la precision o recall por separado en este caso.

3.6.1.2 Resultados de SVM para el vino tinto

Como en el modelo de vino blanco anterior, se inicia la búsqueda de hiperparámetros, comenzando con la variación del kernel y reflejando los resultados en la Tabla 15.

kernel	Accuracy
Linear	56%
Poly	57%
Rbf	**60%**
sigmoid	48%

Tabla 15: Optimización del parámetro Kernel para el modelo SVM para el vino tinto

Al emplear el kernel RBF, se alcanzan los resultados más destacados con una exactitud de 0.6. A continuación, se procede a determinar el valor óptimo para C, cuyo análisis se muestra en la Tabla 16.

C	Accuracy
1	60%
5	60%
10	60%
100	**63%**
1000	60%
10000	59%
0.1	55%
0.01	41%
0.001	41%
500	61%
250	60%
200	61%
150	61%
80	62%
120	62%
50	61%

Tabla 16: Optimización del parámetro C con el modelo SVM para el vino tinto

Después de determinar que el valor óptimo para el parámetro de regularización C es 100, logrando una exactitud del 63%, se continúa buscando los hiperparámetros de

gamma utilizando el kernel RBF con C=100. A continuación, se presentan los resultados en la Tabla 17.

gamma	Accuracy, C=100
1	68%
10	53%
100	52%
0.1	62%
0.01	60%
2	63%
1.1	67%
1.2	67%
1.05	67%
0.9	67%
0.8	67%
1.15	68%
1.9	64%
1.8	64%
1.7	65%

Tabla 17: Optimización del parámetro Gamma para el modelo SVM en el vino tinto

El mejor valor de gamma aparece en 1.15, obteniendo un 68% de exactitud. Usando estos tres hiperparámetros, a continuación, se muestra la Tabla 18 y la Figura 11 que referencian el informe de clasificación y la matriz de confusión para todas las clases usando estos hiperparámetros.

	Precision	Recall	F1-score	Support	Accuracy
3	0%	0%	0%	1	-
4	0%	0%	0%	10	-
5	70%	77%	73%	130	-
6	64%	71%	67%	132	-
7	82%	55%	66%	42	-
8	0%	0%	0%	5	-
Global	65%	-	66%	320	68%

Tabla 18: Informe de clasificación para el modelo SVM en el vino tinto

Utilizando los valores mencionados, se ha logrado desarrollar un sistema de clasificación con una exactitud del 68% para la clasificación precisa. Además, se obtiene una exactitud del 96.25% con un margen de error de 1 punto, lo que significa que es capaz de confundir una clase X con otra clase que sea X + 1 o X - 1.

Es importante destacar que este rendimiento es similar al obtenido para el modelo de vino blanco y representa una mejora significativa en comparación con el modelo de regresión utilizando una red neuronal para el vino tinto, el cual solo alcanzó un valor de 62% para la exactitud. En este modelo tanto precision como recall tienen valores más próximos, a excepción de la clase 7, la cual se presenta en la Tabla 7 que consigue un 82% de precision y en la Figura 11 se observa que predice 28 muestras (23 correctas), pero el test cuenta con 42, prácticamente predice la mitad, pero en las predicciones acierta en gran porcentaje.

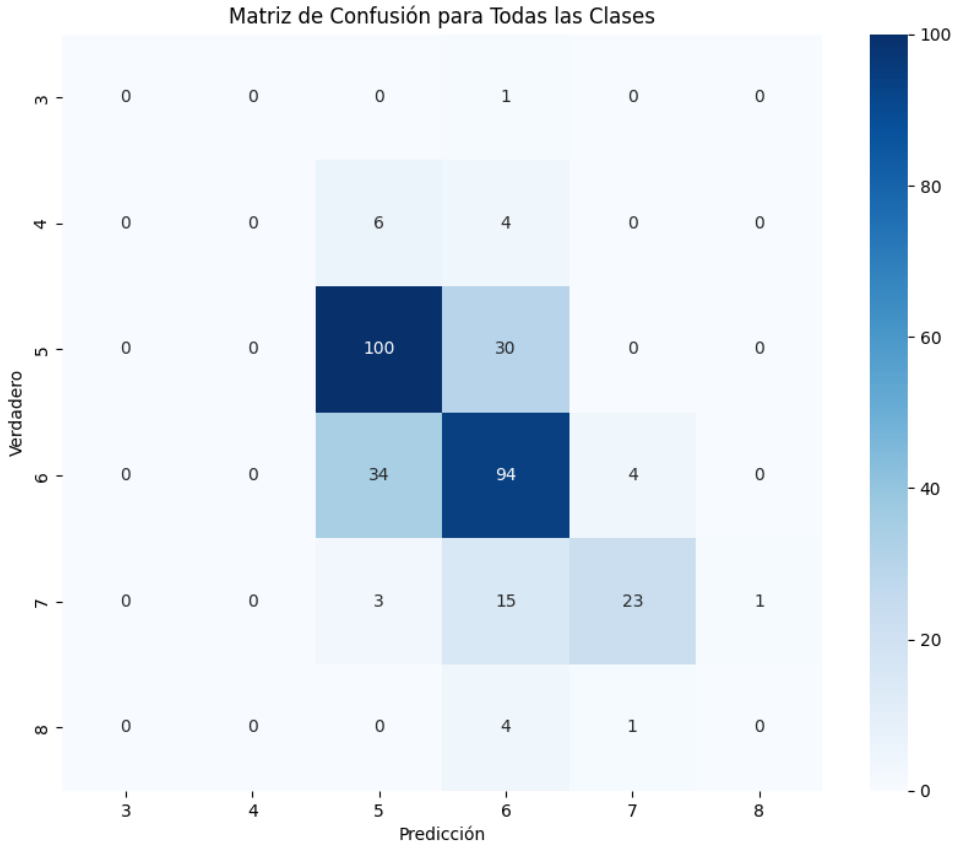

Figura 11: Matriz de confusión para el modelo SVM en el vino tinto

3.6.2 RandomForest

Random Forest es un algoritmo de aprendizaje automático que se puede utilizar para predecir características o propiedades de los vinos. En lugar de depender de un solo árbol de decisión, este algoritmo combina múltiples árboles para obtener una predicción más precisa.

Cada árbol de decisión en Random Forest se construye utilizando muestras aleatorias de los datos de entrenamiento, lo que significa que cada árbol ve una porción diferente de los datos. Esto ayuda a reducir la influencia de características irrelevantes o ruidosas y mejora la calidad de las predicciones.

Una vez que se han construido todos los árboles, para predecir características de un vino en particular, cada árbol emite su propia predicción. La predicción final se determina mediante votación, donde se elige la opción que obtiene más votos de los árboles [10], [11]

Para buscar un modelo óptimo se necesitará en primer lugar realizar una búsqueda de hiperparámetros. El modelo base usará validación cruzada con kfold=5, y un grid donde poder ir modificando los hiperparámetros, y en Randomforest se buscará el número de estimadores “n_stimators” el cual se refiere al número de árboles que se utilizan en el modelo.

3.6.2.1 Resultados de RF para el vino blanco

Como en el resto de las modelos se comienza con la búsqueda de hiperparámetros, cuyos resultados se reflejan en la Tabla 19.

Árboles	Accuracy
10	61.6%
100	64.4%
1000	69.25%
10000	70.5%
500	68.25%
750	68.8%
900	68.94%
1100	70.54%
1200	**71.2%**
1300	71.12%
1500	70.9%
1800	70.8%
2000	70.8%
1150	71.08%
1250	71.1%

Tabla 19: Resultados del modelo de RF para el vino blanco

Esta base de datos al tener más muestras y una clase más es significativamente más compleja, tiene sentido observar que para este caso se necesitan más árboles de decisión para conseguir su valor óptimo, el cual lo consigue en 1200 árboles con un 71.2% de exactitud, teniendo un margen de error de un punto; es decir, confundir un vino de una nota X con otro de X+1 o X-1, consigue un 96.73% de exactitud. En la Tabla 20 se muestran las métricas para cada clase y el global usando los 1200 árboles y en la Figura 12 su matriz de confusión para todas las clases.

	Precision	Recall	F1-score	Support	Accuracy
3	0%	0%	0%	5	-
4	64%	28%	39%	25	-
5	73%	72%	72%	291	-
6	68%	81%	74%	432	-
7	77%	60%	67%	192	-
8	83%	43%	57%	35	-
Global	71%	-	70%	980	71%

Tabla 20: Informe de clasificación del modelo RF para el vino blanco

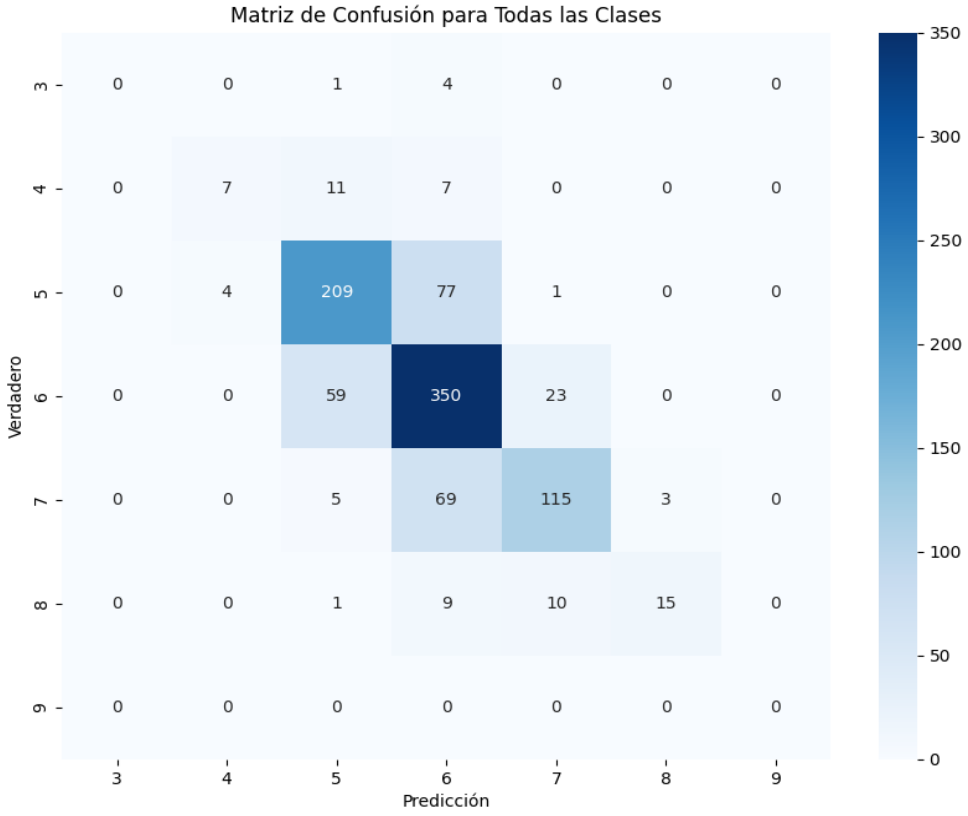

Figura 12: Matriz de confusión del modelo de RF para el vino blanco

Se observa que las clases con menos muestras (4 y 8) presentan mayor desviación en sus valores de precision y recall, esto es debido a que el modelo predice pocas muestras para la clase, como se ve en la matriz de confusión Figura 12, pero las pocas que predice las clasifica correctamente en su mayoría.

3.6.2.2 Resultados de RF para el vino tinto

Esta base de datos presenta vinos con la variable calidad con valores de tres a ocho, como se aprecia en la Figura 2, estos valores son nuestras clases. Se comenzará con la búsqueda de hiperparámetros, mostrados en la Tabla 21.

Árboles	Accuracy
10	64.1%
100	74.375%
1000	72.43%
200	72.81%
125	74.06%
110	74.43%
105	73.75%
50	73.42%
75	73.43%
90	74.0625%
93	74.375%
94	**74.6875%**
95	74.52%
96	74.375%

Tabla 21: Resultados del modelo de RF para el vino tinto

Después de buscar diferentes hiperparámetros, se ha descubierto que el modelo presenta un mejor rendimiento con 94 árboles de decisión. Para este valor, se obtienen los siguientes resultados de las diferentes métricas, que se detallan en la Tabla 22, y la matriz de confusión en la Figura 13.

	Precision	Recall	F1-score	Support	Accuracy
3	0%	0%	0%	1	-
4	0%	0%	0%	10	-
5	73%	84%	78%	130	-
6	78%	78%	78%	132	-
7	60%	56%	57%	42	-
8	100%	60%	75%	5	-
Global	73%	-	74%	320	75%

Tabla 22: Informe de clasifcación del modelo RF para el vino tinto

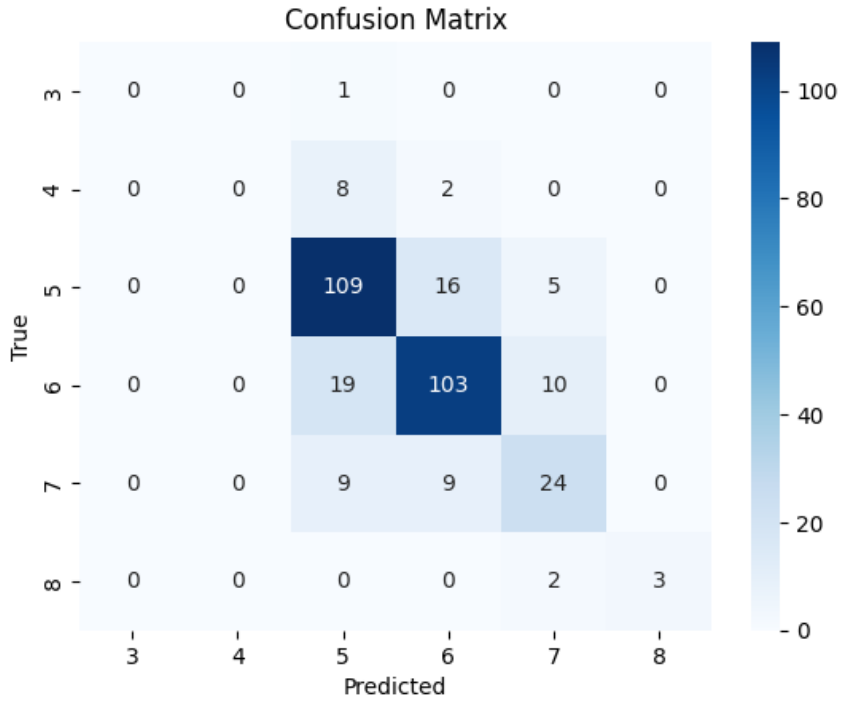

Figura 13: Matriz de confusión del modelo RF para el vino tinto

Como se puede observar, el modelo está poco balanceado y hay clases de las cuales se predicen muy pocas muestras. Pese a ello el modelo es capaz de conseguir un 74.68% de exactitud, pero teniendo un margen de error de un punto; es decir, confundir un vino de una nota X con otro de X+1 o X-1, consigue un 98.4375% de exactitud.

3.7 Comparación de modelos y resumen de resultados

Después de analizar los tres modelos, se ha identificado un claro ganador: el modelo Random Forest. Este modelo ha demostrado obtener los mejores resultados tanto para el vino blanco como para el vino tinto, con una exactitud del 71% y 75%, respectivamente. Estos valores se pueden verificar en la Tabla 23 en la cual se muestra la exactitud global.

ACCURACY	Red neuronal	SVM	RANDOM FOREST
Vino blanco	59%	68%	71%
Vino tinto	62%	68%	75%

Tabla 23: Resumen del parámetro accuracy al comparar los tres modelos

Si se evalúan los modelos basándonos únicamente en su precisión, nos enfocamos en la proporción de instancias positivas clasificadas correctamente. En otras palabras, nos estamos centrando en la capacidad del sistema para identificar correctamente las instancias positivas y evitar clasificar erróneamente las instancias negativas como positivas. Al analizar los resultados específicos para cada nota de calidad para el vino blanco, se observa que las diferencias en SVM son muy pequeñas, pero ligeramente mejores para Random Forest, como se muestra en la Tabla 24. En cuanto al vino tinto, Tabla 25, se encuentra una diferencia significativa entre los tres métodos, siendo Random Forest el mejor en términos de precisión y de exactitud como vimos antes.

BLANCO PRECISION	Red neuronal	SVM	RANDOM FOREST
3	0%	0%	0%
4	20%	44%	64%
5	61%	77%	73%
6	61%	61%	68%
7	57%	78%	77%
8	40%	94%	83%
Global	58%	70%	71%

Tabla 24: Resumen métrica precision para vino blanco

TINTO PRECISION	Red neuronal	SVM	RANDOM FOREST
3	0%	0%	0%
4	100%	0%	0%
5	74%	70%	73%
6	59%	64%	78%
7	51%	82%	60%
8	14%	0%	100%
Global	64%	65%	73%

Tabla 25: Resumen métrica precision para el vino tinto

En la Tabla 26 para el vino blanco y en la Tabla 27 para el vino tinto, en las cuales se evalúa el recall para todas clases, se observa que hay desbalanceo en las clases más minoritarias en ambos casos. Sin embargo, a nivel global, se evidencia que los valores de recall y precisión son similares. Esto indica que se consigue un sistema bastante deseable en todos los modelos, ya que existe un equilibrio al identificar correctamente los casos positivos y evitar clasificar incorrectamente los casos negativos.

BLANCO RECALL	Red neuronal	SVM	RANDOM FOREST
3	0%	0%	0%
4	12%	16%	28%
5	64%	58%	72%
6	64%	85%	81%
7	54%	55%	60%
8	29%	46%	43%
Global	59%	68%	71%

Tabla 26: Resumen métrica recall vino blanco

TINTO RECALL	Red neuronal	SVM	RANDOM FOREST
3	0%	0%	0%
4	10%	0%	0%
5	65%	77%	84%
6	68%	71%	78%
7	55%	55%	56%
8	20%	0%	60%
Global	62%	68%	75%

Tabla 27: Resumen métrica recall vino tinto

Finalmente, en las Tablas 28 y 29 se muestra la distribución del F1 score para todas las clases. Esta métrica puede ser interesante si se analiza de forma individual, pero al haber examinado la precisión y el recall en detalle, no proporciona tanta información adicional. Sin embargo, se observa lo mencionado anteriormente: las clases menos dominantes pueden tener una precisión cercana al 100%, como en el caso de la red neuronal para el vino tinto en la clase 4. No obstante, su F1 score es únicamente del 18%.

BLANCO F1	Red neuronal	SVM	RANDOM FOREST
3	0%	0%	0%
4	15%	24%	39%
5	62%	66%	72%
6	63%	71%	74%
7	56%	64%	67%
8	33%	62%	57%
Global	59%	66%	70%

Tabla 28: Resumen métrica F1 vino blanco

TINTO F1	Red neuronal	SVM	RANDOM FOREST
3	0%	0%	0%
4	18%	0%	0%
5	69%	73%	78%
6	63%	67%	78%
7	53%	66%	57%
8	17%	0%	75%
Global	62%	66%	74%

Tabla 29: Resumen métrica F1 vino tinto

4

Técnicas de machine learning para el análisis de la respuesta emocional: caso de uso Twitch y videojuegos

4.1 Introducción

En este capítulo, se probarán dos modelos de clasificación (SVM y Random Forest) desarrollados previamente en un contexto completamente diferente: el procesamiento de lenguaje natural.

Se llevará a cabo un breve análisis de la base de datos que se utilizará y luego se probarán los dos modelos en tres categorías diferentes, con el objetivo de clasificar la respuesta emocional en función de la polaridad de los mensajes, en concreto dos y tres polaridades y en emociones, en concreto en siete emociones. En este caso concreto se buscará analizar la respuesta emocional en el entorno de videojuegos dentro de la plataforma Twitch.

4.2 Base de datos de Twitch y videojuegos

Esta base de datos contiene 2006 muestras y se compone de tres variables: Texto, Polaridad y Emociones.

1. En la columna "Texto", se encuentra una variable independiente, que representa la característica de entrada. Estos son comentarios en bruto de usuarios de Twitch reaccionando al streaming. Esta variable requerirá un procesamiento adicional para poder analizarla, y los detalles sobre este procesamiento se proporcionarán más adelante.
2. En la columna "Polaridad", se encuentra una variable objetivo que puede tener tres valores:
    a. Positivo
    b. Negativo
    c. Indeterminado

3. En la columna "Emociones", se muestra la segunda variable dependiente, que puede tener siete valores de emoción:
    a. Aprobación/Empatía/Confianza
    b. Desinterés/Tedio
    c. Decepción/Tristeza
    d. Desaprobación
    e. Enfado/Ira
    f. Interés/Aceptación/Hype
    g. Indeterminado

Hay que tener en cuenta que la variable independiente (Texto) representa los comentarios en crudo, mientras que las variables dependientes (Polaridad y Emociones) se utilizan para clasificar y etiquetar los comentarios según su polaridad y las emociones asociadas.

Se comenzará explicando las variables dependientes, iniciando con la variable "Polaridad". En esta variable se aprecia que las 2006 muestras se reparten principalmente entre Negativo y Positivo y en la Figura 14 y Tabla 30 se muestra la distribución detalladamente.

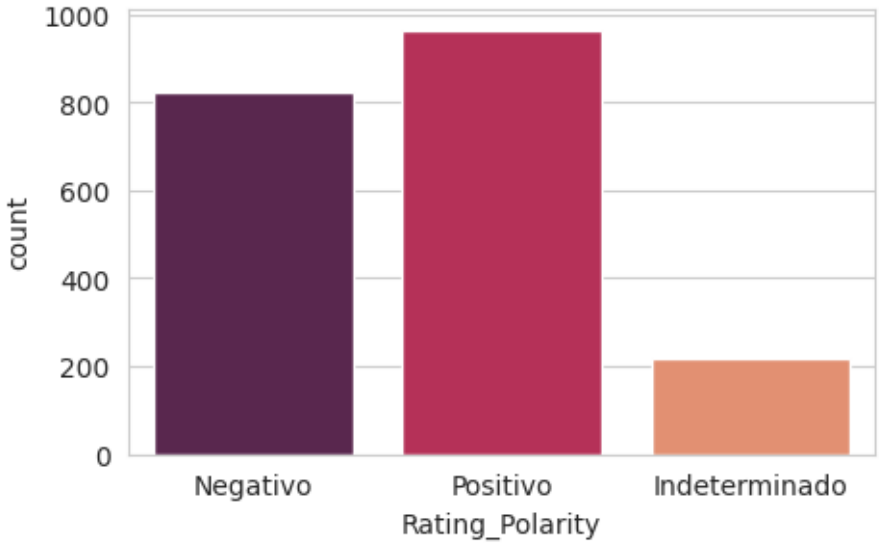

Figura 14:Distribución polaridad Twitch

Polaridad	Muestras	Representación en porcentaje
Positivo	966	48.16%
Negativo	822	41%
Indeterminado	218	10.87%

Tabla 30: Distribución polaridad Twitch

En cuanto a la variable dependiente de “Emoción”, asigna la mayoría de las muestras a la clase “Aprobación” concentrando un total del 35.4% de todas las muestras. El resto de las emociones varían desde el 8% al 13% como se observa en la Figura 15 y Tabla 31.

Polaridad	Muestras	Representación en porcentaje
Aprobación/Empatía/Confianza	710	35.4%
Desinterés/Tedio	171	8.52%
Decepción/Tristeza	182	9.07%
Desaprobación	246	12.26%
Enfado/Ira	168	8.37%
Interés/Aceptación/Hype	268	13.36%
Indeterminado	261	13.01%

Tabla 31: Distribución Twitch emociones

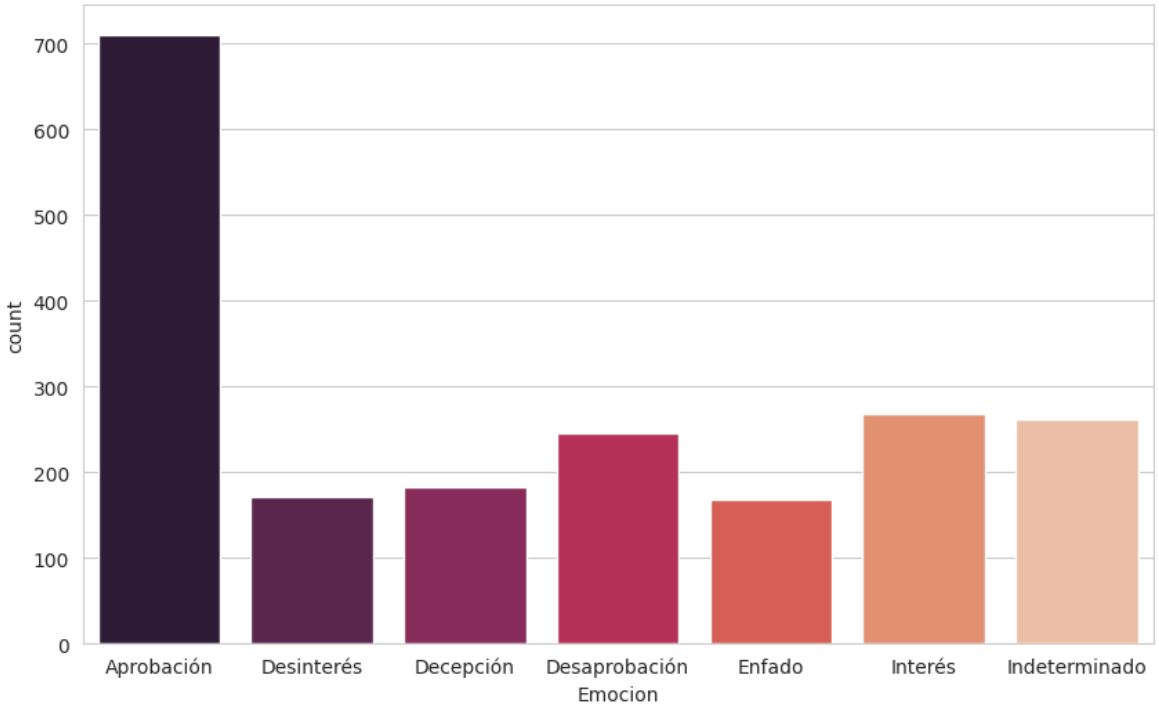

Figura 15: Distribución Twitch emociones

4.2.1 Procesado de texto

El primer paso crucial al abordar este tipo de problemas es el preprocesamiento de los textos, en este caso comentarios en las redes sociales, para asegurarnos de que los modelos de aprendizaje automático los comprenda de manera uniforme. En el proceso de preprocesamiento, se han tenido en cuenta diversos aspectos, como:

1. Normalización de mayúsculas y minúsculas: Se convierten todas las letras a un formato consistente, ya sea mayúsculas o minúsculas, para evitar discrepancias.
2. Eliminación de tildes y vocales diacríticas: Se remueven los acentos y caracteres diacríticos de las palabras, de modo que se reduzca la variabilidad en la representación de las mismas.
3. Reducción de la repetición de caracteres: Se busca reducir la repetición excesiva de caracteres en las palabras, como "holaaaaaa" por "hola", para simplificar el vocabulario.
4. Normalización de la risa: Se trata de convertir las expresiones de risa o emoticonos en una forma más estandarizada, de modo que la red pueda interpretarlas de manera consistente.
5. Normalización de las jergas: Se realizan adaptaciones específicas para normalizar las jergas o expresiones coloquiales utilizadas en el contexto de las redes sociales, de manera que se unifiquen y se eviten ambigüedades.
6. Eliminación de menciones, hashtags y enlaces: Se eliminan las menciones a usuarios, los hashtags y los enlaces, ya que no aportan información relevante para el análisis de sentimientos.
7. Eliminación de signos de puntuación: Se remueven los signos de puntuación, como puntos, comas o signos de interrogación, para evitar que interfieran en la interpretación de las palabras.

Una vez completado el preprocesamiento, se procede a la tokenización, que consiste en considerar cada palabra de cada mensaje como un "token".

4.3 Clasificación de la respuesta emocional con SVM

4.3.1 Resultados de SVM para dos polaridades

En este modelo, se eliminan las muestras con polaridad indeterminada. Esto reduce el conjunto a 1788 muestras, distribuidas como se muestra en la Tabla 32 y la Figura 16. Es importante destacar que esta modificación es una versión simplificada de la Figura 14 y la Tabla 30.

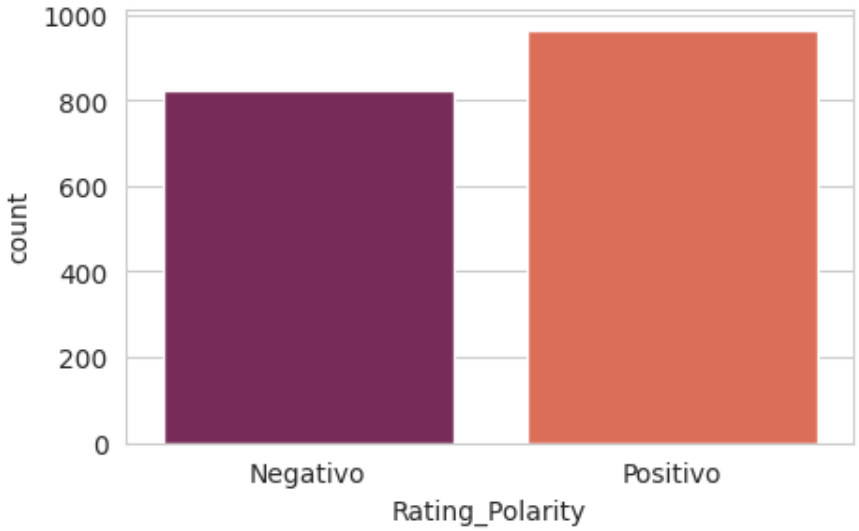

Figura 16: Polaridad P/N Twitch

Polaridad	Muestras	Representación en porcentaje
Positivo	966	54.02 %
Negativo	822	45.98%

Tabla 32: Polaridad P/N Twitch

En la búsqueda de hiperparámetros, se seguirá el mismo procedimiento utilizado en la clasificación de vinos con SVM. Primero, buscar el kernel adecuado y luego determinar el valor óptimo del parámetro de regularización C. Al analizar la Tabla 33, se observa que el kernel RBF vuelve a destacarse por encima de los demás, al igual que sucedía en los modelos de clasificación de vino.

kernel	Accuracy
Poly	65%
Rbf	**71%**
sigmoid	68%

Tabla 33: Parámetro Kernel SVM 2 polaridades

Al determinar el valor del parámetro de regularización (C), se descubre que los mejores resultados se obtienen con valores por debajo de 1. Específicamente, para este modelo, se logra el valor óptimo de C en 0.02, tal como se refleja en la Tabla 34. Cabe destacar que al usar valores de C muy pequeños (0.01) la exactitud bajaba de forma notable, mientras que para valores grandes (100-1000), también descendía, pero no de una forma tan significativa.

C	Accuracy RBF
1	71.1%
5	68.9%
10	71.1%
100	69%
1000	67%
10000	66%
0.1	72.8%
0.01	70%
0.001	54%
0.15	72.4%
0.14	72.4%
0.13	72.4%
0.12	72.6%
0.11	72.6%
0.09	72.8%
0.08	72.8%
0.07	72.8%
0.06	72.8%
0.05	72.8%
0.04	72.9%
0.03	72.9%
0.02	**73.1%**

Tabla 34: Parámetro C SVM 2 polaridades

Por último, se presenta el informe de clasificación y la matriz de confusión en la Tabla 35 y la Figura 17, respectivamente. Se logró una exactitud del 73.1% en la clasificación. Es importante destacar que tanto la polaridad positiva como la negativa se predicen en términos de exactitud muy similares (Tabla 35), pero para la clase positivo apreciamos como la precisión es notablemente inferior al recall, y en el caso de la clase negativo pasa lo contrario, la precision es un 10% mayor al recall, lo que implica que el modelo tiene tendencia a realizar menos falsos positivos en comparación con los falsos negativos.

En el primer caso, la precisión superior al recall indica que el modelo tiene mayor capacidad para identificar correctamente los casos positivos, pero podría perder algunos casos positivos en el proceso. El caso de la clase negativo podría significar que el modelo

tiene una mayor capacidad para detectar la totalidad de los casos positivos, aunque pueda haber más casos negativos clasificados como positivos. Se aprecia en la Figura 17 como la diagonal concentra la mayoría de muestras, las cuales han sido correctamente predichas y como se obtiene 89 falsos positivos y 55 falsos negativos.

	Precision	Recall	F1-score	Support	Accuracy
Positivo	73%	81%	77%	293	-
Negativo	74%	64%	68%	244	-
Global	73%	-	73%	573	73.1%

Tabla 35: Informe de clasificación SVM 2 polaridades

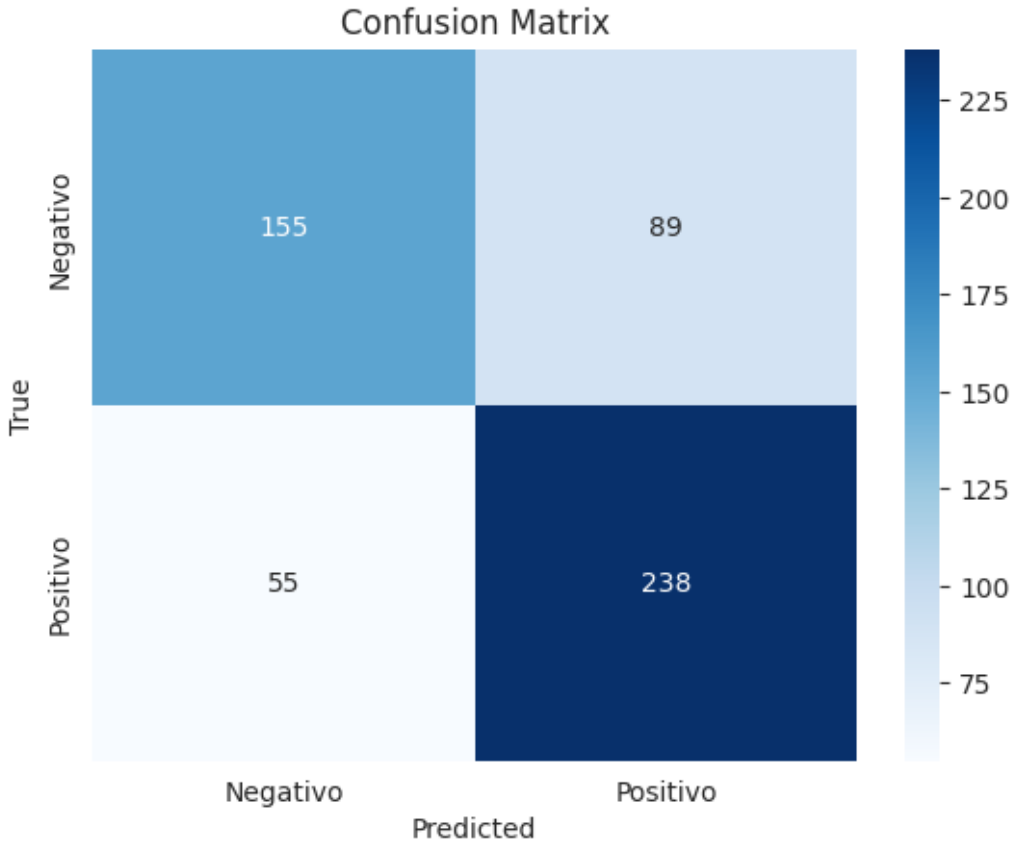

Figura 17: Matriz de confusión SVM 2 polaridades

4.3.2 Resultados de SVM para tres polaridades

Se procede ahora a repetir la búsqueda de hiperparámetros del clasificador SVM, pero esta vez incluyendo la categoría "indeterminado". En la Tabla 36, se observa que el kernel RBF sigue mostrando una exactitud superior. Sin embargo, se encuentra una disminución drástica en la exactitud en comparación con tener solo dos polaridades.

kernel	Accuracy
Poly	48%
Rbf	**53%**
sigmoid	34%

Tabla 36: Parámetro Kernel SVM 3 polaridades

A continuación, se procede a buscar el parámetro de regularización en la Tabla 37 y se descubre que los valores de exactitud apenas aumentan, encontrando el valor optimo en 0.55. Como en el caso anterior para dos polaridades, el modelo se comporta ligeramente mejor con valores de C por debajo de 1.

C	Accuracy RBF
1	53%
5	54%
10	54%
100	54%
1000	54%
10000	50%
0.1	50%
0.01	46%
0.001	46%
25	54%
50	54%
75	53%
125	54%
150	**55%**
175	54%
200	54%
300	54%
400	53%
500	53%

Tabla 37: Parámetro C SVM 3 polaridades

Utilizando el kernel RBF y el parámetro de regularización C=150, se obtiene el informe de clasificación y la matriz de confusión para las tres clases mostradas en la Tabla 38 y la Figura 18. Sin embargo, el sistema empeora significativamente, alcanzando solo un 55% de exactitud.

Es importante destacar que para la nueva clase "Indeterminado", solo logra obtener un 10% de exactitud y 32% de precisión, acertando solamente 8 de las 75 instancias, aunque también se aprecian diferencias significativas en las métricas de recall y precisión. Se observa como en la matriz de confusión (Figura 18) solo se consiguen predecir 25 muestras dentro de la clase indeterminado y de las cuales solo 8 correctamente, cuando se tiene un total de 75, lo cual indica que el modelo clasifica notablemente mal esta clase, ya que confunde mayoritariamente con la clase positivo (11 muestras) y en menor medida con la clase negativo (6 muestras).

	Precision	Recall	F1-score	Support	Accuracy
Positivo	54%	80%	64%	279	-
Negativo	60%	40%	48%	248	-
Indeterminado	32%	11%	16%	75	-
Global	54%	-	52%	602	55%

Tabla 38:Informe de clasificación SVM 3 polaridades

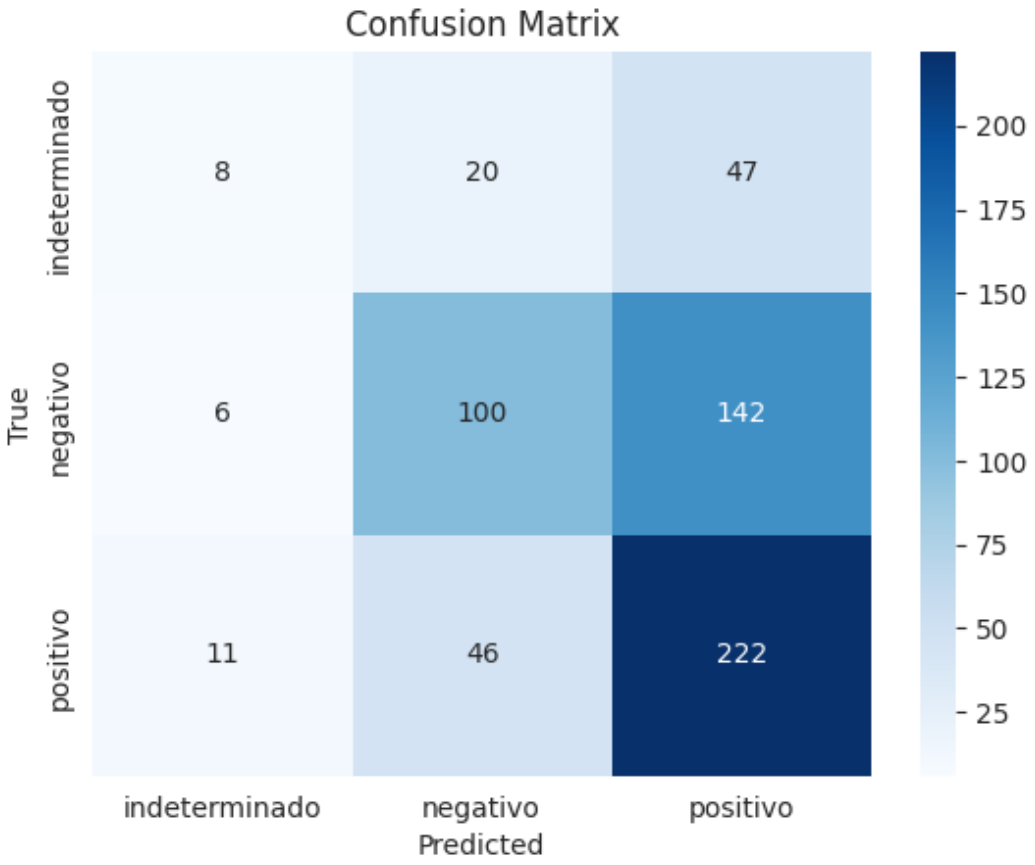

Figura 18: Matriz de confusión SVM 3 polaridades

4.3.3 Resultados de SVM para clasificación de Emociones

En este apartado se procede a predecir las siete emociones mencionadas, cuyos resultados se muestran en la Tabla 31 y la Figura 15. Siguiendo el procedimiento anterior, se busca el kernel a utilizar. Sin embargo, en la Tabla 39 se observa que la exactitud disminuye significativamente, obteniendo resultados muy similares en los tres kernels. Además, ninguno de ellos clasifica correctamente las emociones.

kernel	Accuracy
Poly	335
Rbf	**35%**
sigmoid	31%

Tabla 39: Parámetro kernel SVM emociones

Al probar con el kernel RBF, se aprecia algo interesante en la Tabla 40. Para valores de parámetro de regularización C inferiores a 10, el sistema prácticamente no clasifica nada y todas las instancias se asignan a la variable "aprobación", lo que resulta en un 35% de precisión, ya que el 35% de las instancias pertenecen a esta categoría. Por otro lado, para valores extremadamente altos de C, la exactitud disminuye, pero el sistema comienza a clasificar las emociones. Se tomará este valor como referencia para la evaluación.

C	Accuracy RBF
1	35%
5	35%
10	33%
100	31%
1000	29%
10000	29%
100000	**31%**
0.1	35%
0.01	35%
500000	29%
500	30%
1500	28%
5000	28%
50000	29%
75000	29%

Tabla 40: Parámetro C SVM emociones

Utilizando los parámetros mencionados anteriormente (kernel RBF y parámetro de regularización en 100.000), se obtiene la Tabla 41 y la Figura 19. En esta configuración, el modelo muestra una precisión que varía entre el 30% y el 10% para cada clase, ya que asigna la mayoría de las instancias a la categoría "aprobación". Aunque se alcanza una exactitud del 73% para esta categoría, la precisión es solo del 37%, lo cual indica que el modelo clasifica de manera deficiente al arrojar la mayoría de las muestras a la clase de "aprobación". Además, como se ha mencionado durante la búsqueda de hiperparámetros, los valores de recall, precisión y F1 score para las demás clases son bastante bajos, lo que refleja un rendimiento poco satisfactorio.

	Precision	Recall	F1-score	Support	Accuracy
Aprobación	37%	735	49%	208	-
Decepción	17%	11%	13%	55	-
Desaprobación	22%	16%	18%	77	-
Desinterés	15%	10%	12%	42	-
Enfado	12%	6%	8%	47	-
Indeterminado	11%	3%	5%	91	-
Interés	29%	9%	13%	82	-
Global	25%	-	24%	602	31%

Tabla 41: Informe de clasificación SVM emociones

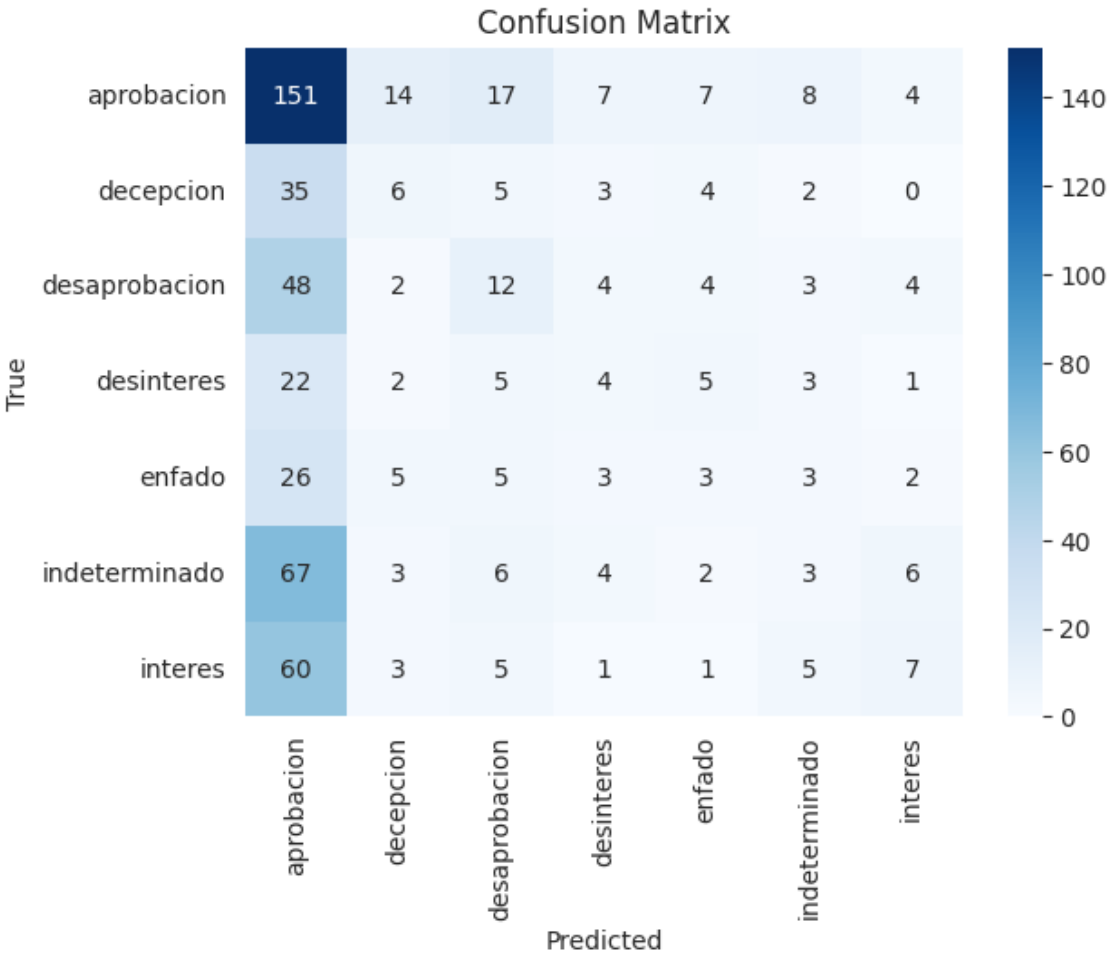

Figura 19: Matriz de confusión SVM emociones

4.4 Clasificación de la respuesta emocional con RandomForest

4.4.1 Resultados de RF para dos polaridades

En este apartado, se procede a utilizar el modelo de Random Forest, el cual obtuvo los mejores resultados en la clasificación de vinos basado en sus parámetros químicos. Para el caso de dos polaridades, se utilizarán los valores presentados en la Tabla 32 y la Figura 16. Comenzando con la búsqueda de parámetros, se encuentra que el mejor valor para el parámetro "max_features" es nuevamente "sqrt", como se muestra en la Tabla 42.

Max_features	Accuracy
Log2	76%
Sqrt	**79%**

Tabla 42: Parámetro kernel RF 2 polaridades

Para finalizar, se lleva a cabo la búsqueda del mejor número de árboles de decisión para este modelo (Tabla 43), descubriendo que el mejor valor es de 500 árboles de decisión, con una exactitud del 80%. También se observa que, en este modelo, el número de árboles no tiene un impacto significativo en la precisión del modelo.

Árboles	Accuracy
10	76%
100	78%
1000	79%
10000	79%
500	**80%**
750	79%
400	80%
300	79%
200	79%
600	79%
700	79%
450	79%
550	80%
475	79%
525	79%

Tabla 43:Parámetro Árboles RF 2 polaridades

Con los valores del kernel "sqrt" y "n_estimators" en 500, se procede a generar el informe de clasificación y visualizar la matriz de confusión para las dos clases de polaridad (negativo y positivo). Estos resultados se muestran en la Tabla 44 y la Figura 20. En la Figura 20 se observa que la mayoría de las muestras se encuentran en la diagonal de la matriz, por lo que han sido correctamente predichas, a excepción de 55 falsos negativos y 52 falsos positivos. Una vez más, se observa una precisión similar para ambas clases, pero esta vez ligeramente superior para la clase "positivo". El modelo mejora significativamente los resultados del SVM, que obtenía un 70% de exactitud, al lograr un

80% de exactitud con Random Forest. Además, los valores de recall y precision obtenidos son similares, por lo que se presenta un equilibrio entre identificar correctamente los casos positivos y evitar clasificar incorrectamente los negativos.

	Precision	Recall	F1-score	Support	Accuracy
Positivo	82%	81%	82%	293	-
Negativo	78%	79%	78%	244	-
Global	80%	-	80%	573	80%

Tabla 44: Informe de clasificación RF2 polaridades

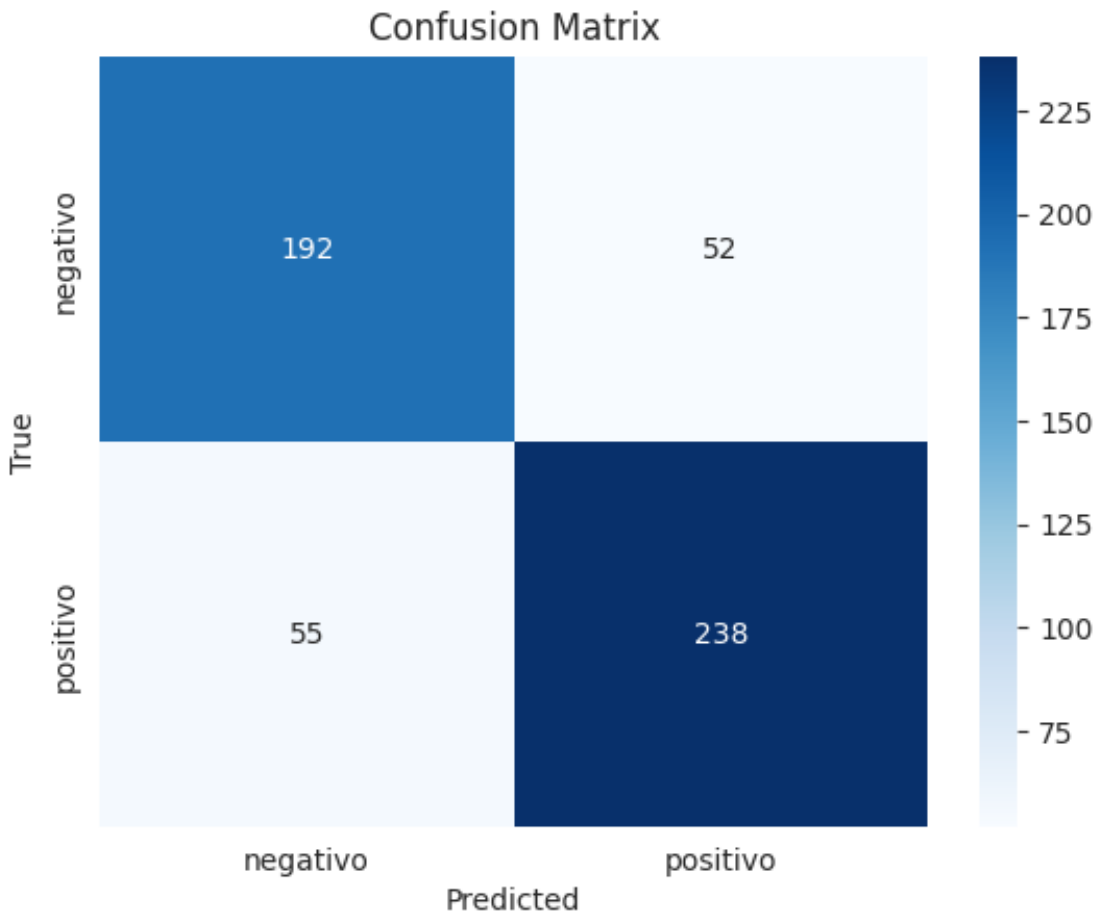

Figura 20: Matriz de confusión RF 2 polaridades

4.4.2 Resultados de RF para tres polaridades

A continuación, se procede a estudiar la base de datos completa, con las tres polaridades (positivo, negativo e indeterminado) representadas en la Tabla 30 y Figura 14. De nuevo se repite el mismo procedimiento para buscar hiperparámetros, donde el mejor valor para max_features en este caso es sqrt con un 69% de exactitud en el global como se ve en la Tabla 45.

Max_features	Accuracy
Log2	66%
sqrt	**69%**

Tabla 45: Parámetro kernel RF 3 polaridades

A continuación, se procede a buscar el mejor valor para el número de árboles de decisión (Tabla 46). Al igual que en el modelo de 2 polaridades, el número de árboles de decisión no tiene un impacto significativo en la precisión. Sin embargo, el valor óptimo se alcanza con 500 árboles utilizando el kernel RBF.

Árboles	Accuracy
10	66%
100	70%
1000	70%
10000	66%
500	68%
750	70%
400	68%
300	69%
200	69%
600	69%
700	69%
450	69%
550	**70%**
475	70%
525	70%
560	70%
540	70%

Tabla 46: Parámetro Árboles RF 3 polaridades

Con los resultados óptimos obtenidos en las Tablas 45 y 46, se procede a realizar el informe de clasificación y la matriz de confusión en la Tabla 47 y la Figura 21. Se observa que nuestro modelo logra una precisión cercana al 70% para las clases "positivo" y "negativo", y un 78% para la clase "Indeterminado". Sin embargo, en el otro extremo, la exactitud para la clase "Indeterminado" es del 33%. Esto indica que, si bien el modelo acierta en la mayoría de las muestras que clasifica como "Indeterminado", introduce muy pocas instancias en esta clase. Como se puede apreciar en la Figura 21, solo se etiquetan 29 muestras (de las cuales 25 son correctas) cuando en realidad hay un total de 75 muestras en esta clase. Respecto a la clase negativo, se aprecia que el modelo obtiene 180 muestras bien catalogadas, pero se confunde mayoritariamente con la clase positivo, a la que asigna 57 muestras, y de un total de 248 muestras consigue predecir 180, lo que implica un 72% de exactitud. Por último, la clase con mejores resultados es “Positivo” ya que como se ve en la Figura 21, del total de muestras, 217 son predichas en su clase y 89 de forma errónea, siendo la clase con la que más se confunde negativo.

	Precision	Recall	F1-score	Support	Accuracy
Positivo	71%	78%	74%	279	-
Negativo	68%	73%	70%	248	-
Indeterminado	78%	33%	47%	75	-
Global	71%	-	69%	602	70%

Tabla 47: Informe de clasificación RF 3 polaridades

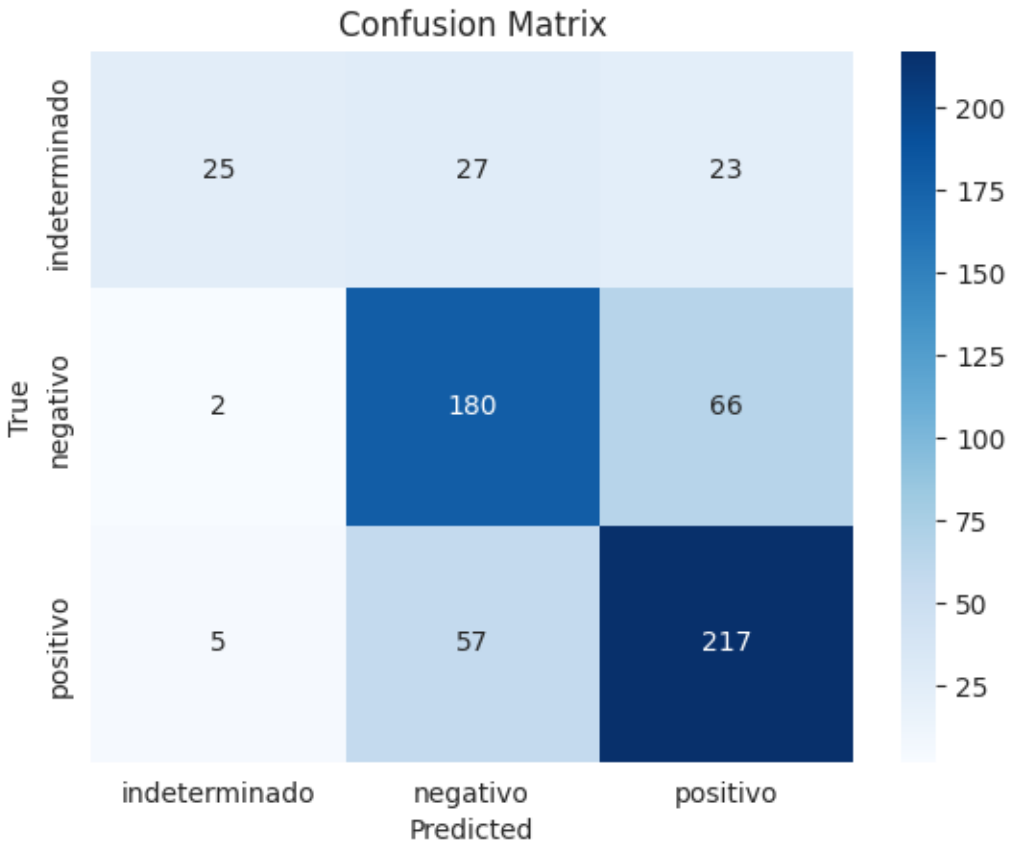

Figura 21: Matriz de confusión RF 3 polaridades

4.4.3 Resultados de RF para la clasificación de emociones

Concluye el estudio de Random Forest con el modelo para predecir las siete emociones presentadas en la Tabla 31 y la Figura 15. Para ello, se procede a buscar los hiperparámetros comenzando por "max_features", como se muestra en la Tabla 48. Una vez más, encontramos que el mejor resultado se obtiene con "sqrt", logrando una exactitud del 42%.

Max_features	Accuracy
Log2	40%
sqrt	**42%**

Tabla 48: Parámetro kernel RF emociones

Se finaliza la búsqueda de hiperparámetros con el número de árboles de decisión, como se muestra en la Tabla 49. Se observa que los valores de exactitud se modifican ligeramente, encontrando el mejor resultado con 400 árboles y una exactitud del 44%.

Árboles	Accuracy
10	39%
100	42%
1000	42%
10000	41%
5000	42%
500	43%
750	42%
250	42%
300	43%
350	43%
400	**44%**
450	43%
550	43%
600	42%
650	42%
700	42%

Tabla 49: Parámetro kernel RF emociones

Utilizando los parámetros obtenidos en las Tablas 48 y 49, se genera el informe de clasificación y la matriz de confusión para las siete clases de la base de datos, como se muestra en la Tabla 50 y la Figura 22, respectivamente. En la tabla 50, se aprecia que todas las clases tienen una precisión aproximada del 40% al 50%. Sin embargo, al examinar la matriz de confusión, notamos que la gran mayoría de las muestras se predicen dentro de la clase "aprobación", obteniendo un 75% de exactitud en esta categoría, mientras que en el resto de las clases oscila en torno al 30%.

En cuanto al recall se aprecia que es inferior en todas las clases salvo la de aprobación, esto es debido a que la mayoría de las muestras caen en esta clase. Es importante tener en cuenta que, al clasificar las emociones, algunas clases no están claramente diferenciadas, lo que dificulta, por ejemplo, distinguir entre un mensaje de enfado y uno de desinterés o desaprobación, esto puede apreciarse en la Figura 22 ya que las predicciones se confunden principalmente entre estas clases, ya que estas emociones tienen una estrecha relación y a veces un mensaje puede contener más de una emoción de este tipo.

Finalmente se observa que disponemos de muy pocas muestras de la clase "indeterminado", la cual, como se muestra en la Figura 22, se confunde con casi todas las demás clases, excepto con la clase de “enfado”.

	Precision	Recall	F1-score	Support	Accuracy
Aprobación	43%	75%	55%	208	-
Decepción	47%	27%	34%	55	-
Desaprobación	45%	23%	31%	77	-
Desinterés	39%	31%	35%	42	-
Enfado	46%	36%	40%	47	-
Indeterminado	38%	16%	23%	91	-
Interés	53%	38%	44%	82	-
Global	44%	-	41%	602	44.2%

Tabla 50: Informe de clasificación RF emociones

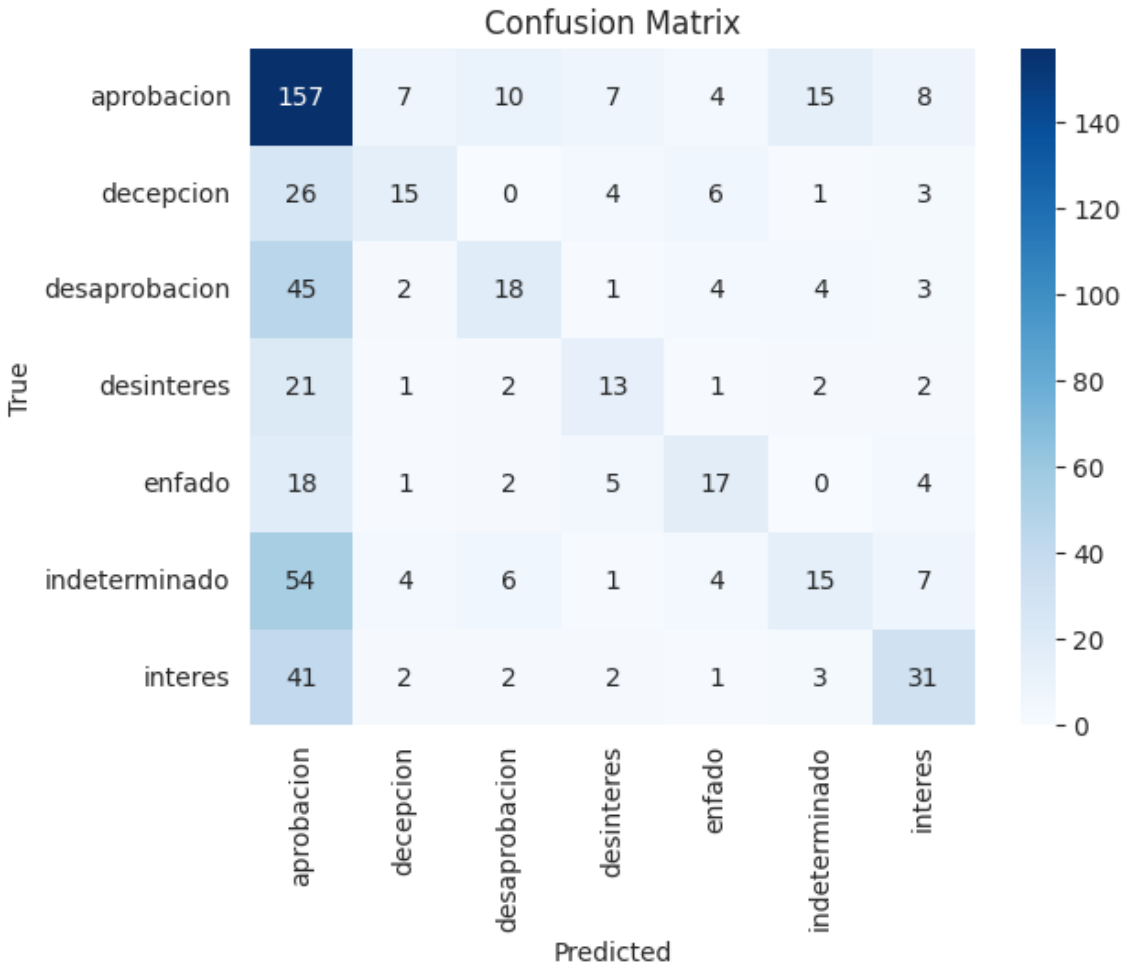

Figura 22: Matriz de confusión RF emociones

4.5 Comparación de modelos y resumen de resultados

En este apartado, se llevará a cabo una comparación detallada de los dos modelos estudiados, analizando punto por punto y métrica por métrica. Para comenzar, se presenta la Tabla 51, que muestra la métrica de exactitud global. Se observa que Random Forest obtiene consistentemente un mayor valor de exactitud en comparación con SVM.

En el caso de la clasificación en dos polaridades, Random Forest logra alcanzar un 80% de exactitud, mientras que SVM obtiene un 73.15%. Ambos modelos experimentan un ligero descenso en la exactitud al considerar la clasificación en tres polaridades, pero Random Forest logra mantenerse en un 70%, mientras que SVM disminuye al 55%. Se destaca que SVM sufre una disminución de aproximadamente un 20% en exactitud en comparación con la clasificación de dos polaridades.

Cuando se analiza la clasificación en siete emociones, se observa un decrecimiento significativo en la exactitud para ambos modelos. Random Forest alcanza un 44.2% y SVM obtiene un 31%. Estos resultados brindan una idea más profunda del rendimiento de los modelos y establecen una base para el estudio de las métricas restantes.

Accuracy Global	SVM	RF
Dos Polaridades	73.15%	80%
Tres polaridades	55%	70%
Emociones	31%	44.2%

Tabla 51: Resumen exactitud Twitch

A continuación, se procederá a analizar en detalle las demás métricas para obtener una comprensión más completa de la efectividad y desempeño de los modelos. Se procede al análisis de las métricas de precisión, recall y F1-score para el modelo de dos polaridades, presentadas en las Tablas 52, 53 y 54, respectivamente. En este contexto, se observa consistentemente un mejor desempeño del modelo Random Forest, y se evidencia un equilibrio entre los valores de recall y precisión para RF. Por otro lado, en el caso de SVM, se observa una ligera inferioridad en la precisión en comparación con el recall, y se presenta una situación opuesta en la clase negativa.

PRECISION	SVM	RF
Positivo	73%	82%
Negativo	74%	78%
Global	73%	80%

Tabla 52: Resumen precision 2 polaridades Twitch

RECALL	SVM	RF
Positivo	81%	81%
Negativo	64%	79%
Global	73%	80%

Tabla 53:: Resumen recall 2 polaridades Twitch

F1 SCORE	SVM	RF
Positivo	77%	82%
Negativo	68%	78%
Global	73%	80%

Tabla 54: Resumen F1-score 2 polaridades Twitch

Pasamos a comparar la clasificación de tres polaridades, que incluye la clase "indeterminado". Se presentan las Tablas 55, 56 y 57, que contienen las métricas de precisión, recall y F1-score, respectivamente. Al analizar la precisión, se observa que el modelo Random Forest mantiene un 70% para las clases "positivo" y "negativo", y alcanza un 80% para la clase "indeterminado". Por otro lado, SVM disminuye a un 54% y 60% en las clases "positivo" y "negativo", respectivamente, y solo logra un 32% de precisión para la nueva clase "indeterminado". Estos resultados indican que Random Forest tiene una mayor capacidad para identificar correctamente los casos positivos, aunque puede perder algunos casos que no se clasifican correctamente. Por otro lado, SVM muestra un desempeño significativamente deficiente en la clasificación de la clase "indeterminado". Al analizar los valores de recall, se observa que Random Forest presenta valores similares para las clases "positivo" y "negativo", mientras que para la clase "indeterminado" obtiene solo un 33%. Esto indica que el modelo tiene una mayor capacidad para identificar correctamente los casos positivos, pero pierde algunos casos en la clasificación de la clase "indeterminado". En contraste, SVM muestra un bajo desempeño en la clasificación de la clase "indeterminado".

PRECISION	SVM	RF
Positivo	54%	71%
Negativo	60%	68%
Indeterminado	32%	78%
Global	54%	71%

Tabla 55:: Resumen precision 3 polaridades Twitch

RECALL	SVM	RF
Positivo	80%	78%
Negativo	40%	73%
Indeterminado	11%	33%
Global	55%	70%

Tabla 56: Resumen recall 3 polaridades Twitch

F1-SCORE	SVM	RF
Positivo	64%	74%
Negativo	48%	70%
Indeterminado	16%	47%
Global	52%	69%

Tabla 57: Resumen F1-score 3 polaridades Twitch

En el problema de clasificar las siete emociones, se evalúan las métricas para los modelos SVM y Random Forest. Como se observó anteriormente durante la búsqueda de hiperparámetros, SVM presenta un desempeño deficiente en este contexto, ya que no logra clasificar adecuadamente la mayoría de las instancias, asignándolas todas a la clase "Aprobación". Por otro lado, Random Forest muestra una mejor capacidad de clasificación, aunque también tiende a asignar la mayoría de sus predicciones a la clase

"Aprobación". Se observan ligeras variaciones entre las métricas de recall y precisión para las demás clases, lo que hace interesante considerar el F1-score. En general, se puede apreciar que el modelo presenta un promedio de entre el 30% y el 40% para las otras clases, siendo la clase "indeterminado" la más perjudicada.

PRECISION	SVM	RF
Aprobación	37%	43%
Decepción	17%	47%
Desaprobación	22%	45%
Desinterés	15%	39%
Enfado	12%	46%
Indeterminado	11%	38%
Interés	29%	53%
Global	25%	44%

Tabla 58: Resumen precision emociones Twitch

RECALL	SVM	RF
Aprobación	73%	75%
Decepción	11%	27%
Desaprobación	16%	23%
Desinterés	10%	31%
Enfado	6%	36%
Indeterminado	3%	16%
Interés	9%	38%
Global	31%	44%

Tabla 59: Resumen recall emociones Twitch

F1-SCORE	SVM	RF
Aprobación	49%	55%
Decepción	13%	34%
Desaprobación	18%	31%
Desinterés	12%	35%
Enfado	8%	40%
Indeterminado	5%	23%
Interés	13%	44%
Global	24%	41%

Tabla 60: Resumen F1-score emociones Twitch

En conclusión, el modelo Random Forest muestra un mejor rendimiento en todas las métricas evaluadas. Tanto en la clasificación de vinos como en la clasificación de emociones, Random Forest supera a los otros modelos considerados, incluyendo SVM. Proporciona resultados más satisfactorios en términos de precisión, recall, F1-score y exactitud general.

Por otro lado, sería recomendable establecer una categorización más clara de para las distintas clases, dado que algunas emociones muestran una relación cercana y pueden confundirse en las predicciones debido a esta dualidad, como es el caso de las clases "Desinterés" y "Decepción" o “Enfado” y “Desaprobación”. También resultaría importante aumentar y equilibrar el número de muestras, ya que la clase "indeterminado" se confunde significativamente con las demás clases, mientras que la clase "aprobación" presenta una sobrerrepresentación que afecta tanto en la fase de entrenamiento como en la de prueba.

5

Técnicas de machine learning para el análisis de la respuesta emocional: caso de uso salud mental

5.1 Introducción

En este capítulo se realizarán pruebas de los modelos de clasificación SVM y RandomForest en una base de datos diferente, en concreto en el contexto de la salud metal en redes sociales. El objetivo es predecir la polaridad en las categorías de positivo/negativo y positivo/negativo/indeterminado. Asimismo, se buscará predecir la emoción que se presenta catalogada en dicha variable dependiente.

5.2 Base de Datos

Esta base de datos contiene 2286 muestras y se compone de tres variables principales: Texto, Polaridad y Emociones. Cabe decir que contiene más variables, pero no interesan en el estudio que se realizará. En concreto, las variables del corpus de salud mental son:

1. En la columna "Texto", se nuestra variable independiente, que representa la característica de entrada. Esta variable requerirá un procesamiento adicional para poder analizarla, tal y como se hizo en el capítulo anterior.
2. En la columna "Polaridad", se encuentra una variable objetivo que puede tener tres valores:
    a. Positivo
    b. Negativo
    c. Indeterminado
3. En la columna "Emociones", se muestra la segunda variable dependiente, que puede tener seis valores:
    a. Comprensión/Empatía/Identificación
    b. Amor/Admiración
    c. Enfado/Desprecio/Burla
    d. Gratitud

e. Tristeza/Pena
f. Indeterminado

Hay que tener en cuenta que la variable independiente (Texto) representa los comentarios en crudo, mientras que las variables dependientes (Polaridad y Emociones) se utilizan para clasificar y etiquetar los comentarios según su polaridad y las emociones asociadas. El análisis se comenzará explicando las variables dependientes, comenzando con la variable de polaridad, mostrando su distribución de muestras en la Figura 23 y Tabla 61.

Polaridad	Muestras	Representación en porcentaje
Positivo	1526	66.75%
Negativo	587	25.67%
Indeterminado	173	7.57%

Tabla 61: Distribución muestras 3 polaridades Salud Mental

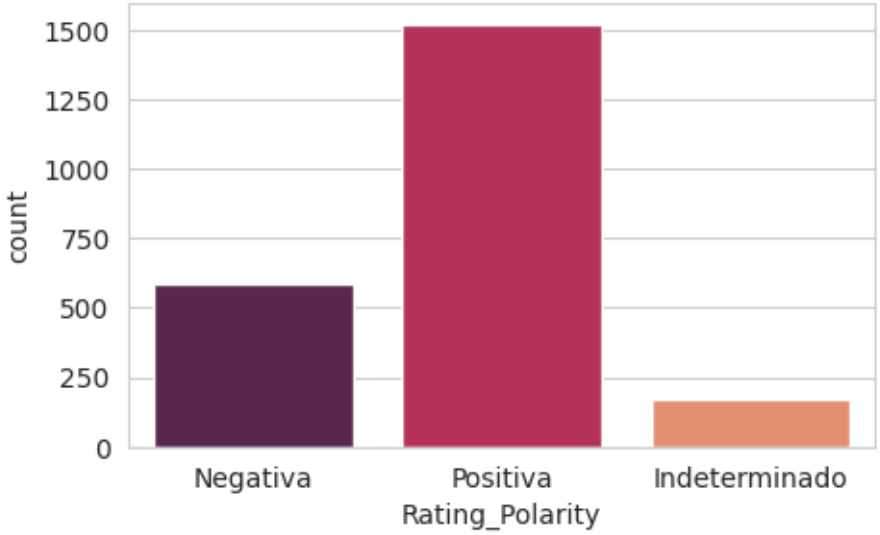

Figura 23: Gráfico 3 polaridades Salud Mental

En esta variable, se ha observado que la mayoría de las muestras pertenecen a la clase positiva, que representa el 66.75% del total, mientras que la clase negativa abarca el 25.67% de las muestras y la clase Indeterminado comprende el 7.57%. En cuanto a la variable dependiente de Emoción la distribución se muestra en la Tabla 62 y Figura 24. Esta variable está notablemente más equilibrada, presenta tres clases que superan el 20% de las muestras, las cuales son: Amor/Admiración, Comprensión/Empatía/Identificación y Enfado/Desprecio/Burla. Por otro lado, las otras tres clases se encuentran por debajo del 10%, siendo la más pequeña Tristeza/Pena con un 5.34% de las muestras.

Polaridad	Muestras	Representación en porcentaje
Amor/Admiración	640	28%
Gratitud	227	9.93%
Comprensión/Empatía/Identificación	659	28.82%
Tristeza/Pena	122	5.34%
Enfado/Desprecio/Burla	466	20.38%
Indeterminado	173	7.57%

Tabla 62: Distribución muestras emociones Salud Mental

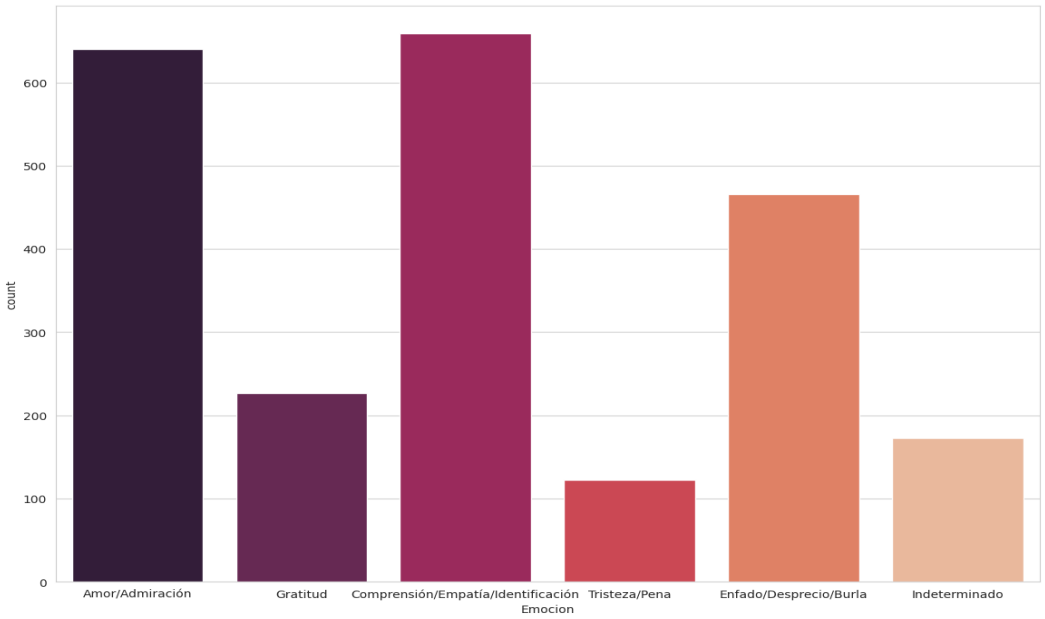

Figura 24:Gráfico Emociones Salud Mental

5.3 Clasificación de la respuesta emocional con SVM

5.3.1 Resultados de SVM para dos polaridades

Con el fin de desarrollar este modelo, se realizarán modificaciones en la Tabla 61 y en la Figura 23. En estas modificaciones, se eliminará la clase "Indeterminado" ya que nuestro objetivo es estudiar exclusivamente la polaridad positiva y negativa. Estos cambios se reflejarán tanto en la Figura 25 como en la Tabla 63.

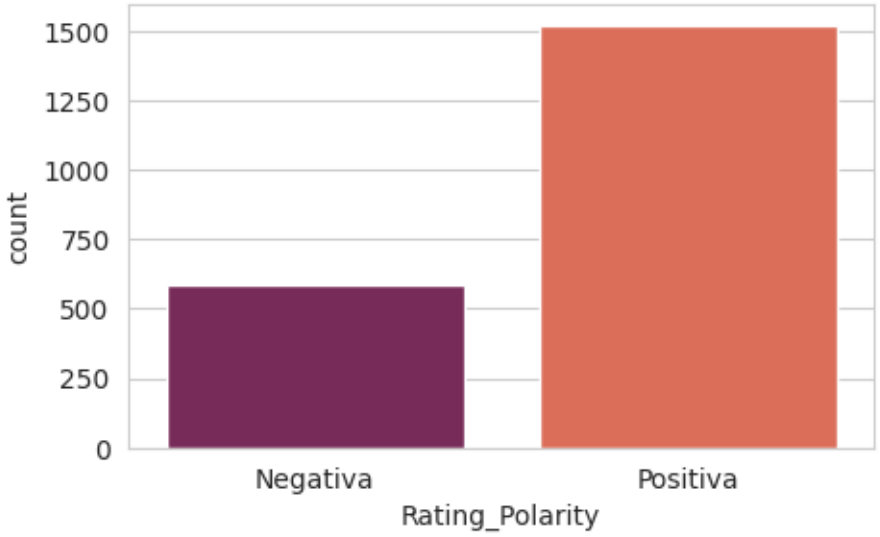

Figura 25: Gráfico P/N Salud mental

Polaridad	Muestras	Representación en porcentaje
Positivo	1526	72.22%
Negativo	587	27.78%

Tabla 63: Distribución muestras P/N Salud Mental

Al eliminar la clase "Indeterminado", el desbalanceo se acerca a una proporción de 3 a 1 a favor de la clase positiva. Se inicia la búsqueda de hiperparámetros con el objetivo de encontrar el kernel adecuado para este modelo, reflejando los resultados en la Tabla 64. Se destaca ligeramente el valor del kernel RBF, ya que tanto Poly como Sigmoid no logran detectar correctamente ningún caso negativo.

Kernel	Accuracy
Poly	71%
Rbf	**72%**
sigmoid	71%

Tabla 64: Parámetro kernel SVM 2 polaridades Salud Mental

A continuación, se procede a la búsqueda del parámetro de regularización y se registra en la Tabla 65. Se observa que el valor óptimo es 2, con una exactitud de 73%. Se destaca que para valores inferiores a 1 no se logra clasificar correctamente ninguna instancia de la clase Negativa, mientras que, con valores muy altos, todas las métricas indican una predicción casi aleatoria.

C	Accuracy Rbf
1	72%
5	71%
10	70%
100	67%
1000	66%
0.1	72%
0.01	72%
0.001	72%
3	71%
2.8	71%
2.6	72%
2.4	72%
2.2	72%
2	**73%**
1.8	73%
1.6	72%
1.4	72%
1.2	72%
0.8	71%
0.6	71%
0.4	71%
0.2	71%

Tabla 65: Parámetro C SVM 2 polaridades Salud Mental

Con la configuración de parámetros mencionada, se obtiene la Tabla 66 y la Figura 26. En la Figura 26 se observa que la mayoría de las muestras han sido clasificadas en la clase “Positiva” 585, de las cuales 432 han sido clasificadas correctamente, por lo cual conseguimos un recall muy bueno de 96% pero la precision cae significativamente hasta el 74%. En cuanto a la clase “Negativa”, solo se han clasificado 49 muestras en esta clase, de las cuales solo 25 son verdaderos negativos. Esto resulta insuficiente considerando que hay un total de 182 muestras en esta clase, luego este modelo realiza predicciones muy conservadoras para esta clase, dejando fuera la mayoría de las instancias, que se refleja en la Tabla 66 con un 16% de recall para esta clase.

	Precision	Recall	F1-score	Support	Accuracy
Positivo	74%	96%	83%	452	-
Negativo	59%	16%	25%	182	-
Global	70%	-	67%	634	73%

Tabla 66: Informe de clasificación SVM 2 polaridades Salud Mental

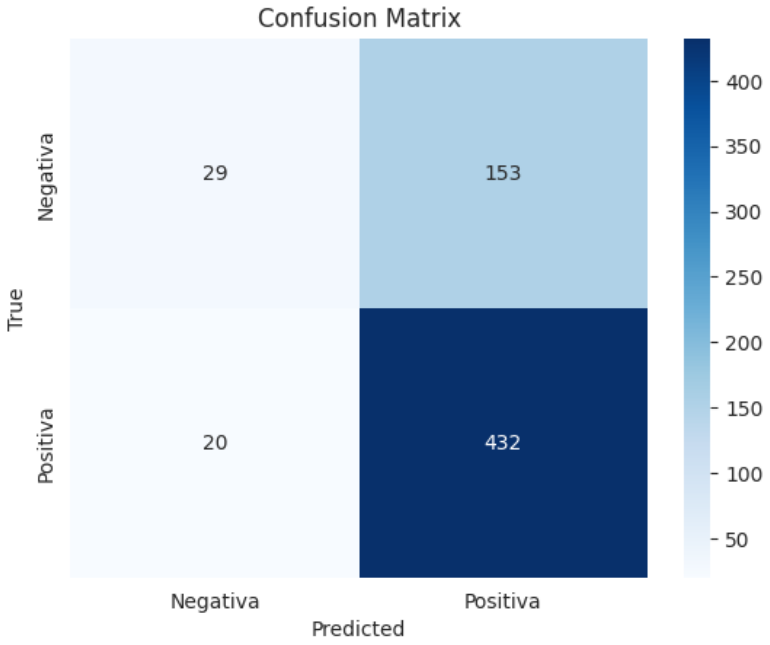

Figura 26: Matriz de confusión SVM 2 polaridades Salud Mental

5.3.2 Resultados de SVM para tres polaridades

Se inicia la búsqueda de hiperparámetros para el caso de 3 polaridades utilizando las muestras de la Tabla 61 y la Figura 23. El proceso comienza con la exploración del kernel y los resultados se registran en la Tabla 67. Se confirma una vez más que el kernel RBF presenta los mejores resultados en términos de rendimiento.

kernel	Accuracy
Poly	62%
Rbf	**63%**
Sigmoid	62%

Tabla 67:Parámetro Kernel SVM 3 polaridades Salud Mental

Después de seleccionar el kernel RBF para el caso de 3 polaridades, se realiza la búsqueda del parámetro de regularización, cuyos resultados se reflejan en la Tabla 68. En este caso, se obtiene el valor óptimo de 1.4.

C	Accuracy Rbf
1	63%
5	61%
10	61%
100	59%
1000	54%
0.1	63%
0.01	63%
0.001	63%
3	61%
2.8	61%
2.6	61%
2.4	62%
2.2	62%
2	63%
1.8	63%
1.6	64%
1.4	**65%**
1.2	64%
0.8	63%
0.6	63%
0.4	63%
0.2	63%

Tabla 68: Parámetro C SVM 3 polaridades Salud Mental

Con los valores de hiperparámetros mencionados, se obtiene la Figura 23 y la Tabla 69. Al considerar el cambio de 2 a 3 polaridades y tener en cuenta el desbalance significativo entre las tres clases, se observa que el modelo presenta un rendimiento deficiente en la predicción.

En la Figura 27 se puede apreciar que la mayoría de las muestras son clasificadas incorrectamente en la clase “Positiva”. Aunque se logra un 90% de Recall en esa clase, esto se logra a expensas de una precisión del 68%. En cuanto a la clase “Negativa” consigue acertar 35 muestras de las 84 que predice logrando un 42% de precisión, pero su recall cae hasta el 19% ya que como se ve en la Tabla 69 que tienen 186 muestras. Para la clase “Indeterminado”, se obtiene una precisión del 50%, pero solo un 2% de recall, ya que el modelo predice solo 2 muestras en esta clase y solo una de ellas es correcta.

	Precision	Recall	F1-score	Support	Accuracy
Positivo	68%	90%	77%	451	-
Negativo	42%	19%	26%	186	-
Indeterminado	50%	2%	4%	49	-
Global	59%	-	58%	686	65%

Tabla 69: Informe de clasificación SVM 3 polaridades Salud Mental

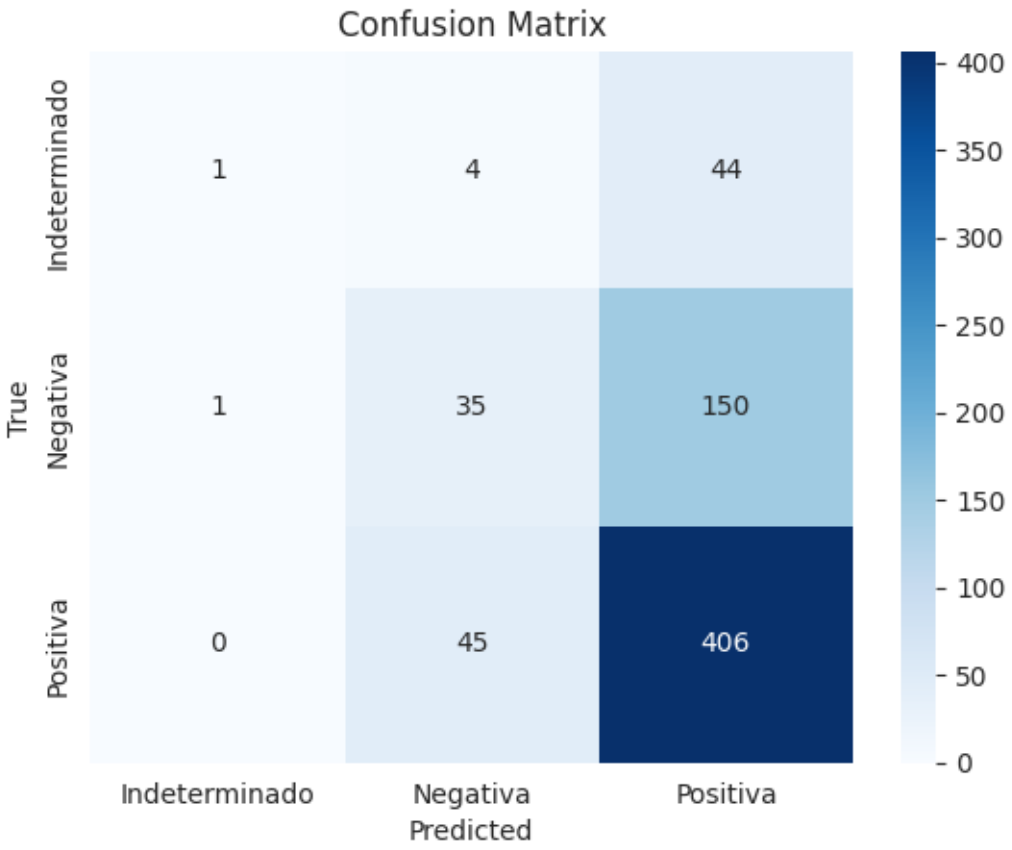

Figura 27: Matriz de confusión SVM 3 polaridades Salud Mental

5.3.3 Resultados de SVM para la clasificación de emociones

Para concluir el estudio utilizando SVM, se analizan las seis emociones presentes en la base de datos, como se muestra en la Figura 24 y la Tabla 62. El primer paso consiste en buscar el kernel adecuado, cuyos resultados se encuentran en la Tabla 70, donde se destaca la elección del kernel RBF.

Kernel	Accuracy
Poly	33%
Rbf	**38%**
Sigmoid	32%

Tabla 70: Parámetro Kernel SVM emociones Salud Mental

A continuación, se realiza la búsqueda del parámetro de regularización para el modelo. Durante este proceso, se observa que valores demasiado altos conducen a predicciones aleatorias, mientras que valores demasiado bajos solo logran clasificar las clases mayoritarias, como Amor y Comprensión. Finalmente, se encuentra el valor óptimo de 0.8 para este modelo, el cual está reflejado en la Tabla 71.

C	Accuracy Rbf
1	38%
5	35%
10	34%
100	32%
1000	31%
0.1	37%
0.01	37%
0.001	37%
3	35%
2.8	35%
2.6	35%
2.4	36%
2.2	36%
2	36%
1.8	37%
1.6	38%
1.4	38%
1.2	38%
0.8	**38%**
0.6	38%
0.4	37%
0.2	37%

Tabla 71: Parámetro C SVM emociones Salud Mental

En la Tabla 72 y la Figura 28 se presenta el informe de clasificación y la matriz de confusión para el caso estudiado, que corresponde a la predicción de las 6 emociones.

De las tres clases menos frecuentes (“Gratitud”, “Indeterminado” y “Tristeza”), solo la clase “Indeterminado” logra ser predicha, con una precisión notable del 63% pero un recall del 12%. Al observar la Figura 28, se aprecia que solo se asignaron 8 muestras

a esta clase, de las cuales 5 fueron correctamente clasificadas. Por otro lado, la clase "Amor" obtiene una precisión del 70%, pero solo un 38% de precisión, ya que la mayoría de las muestras se agrupan en esta clase. La clase "Comprensión" presenta un mejor equilibrio, con un 40% de precisión y un 56% de recall. En general, el modelo presenta un rendimiento deficiente en la predicción de las clases, ya que se observan resultados insatisfactorios en la mayoría de ellas.

	Precision	Recall	F1-score	Support	Accuracy
Amor/Admiración	38%	71%	49%	207	-
Comprensión/Empatía/Identificación	40%	56%	46%	190	-
Enfado/Desprecio/Burla	28%	5%	8%	147	-
Gratitud	0%	0%	0%	61	-
Indeterminado	62%	12%	20%	41	-
Tristeza/Pena	0%	0%	0%	41	-
Global	32%	-	31%	687	38.5%

Tabla 72: Informe de clasificación SVM emociones Salud Mental

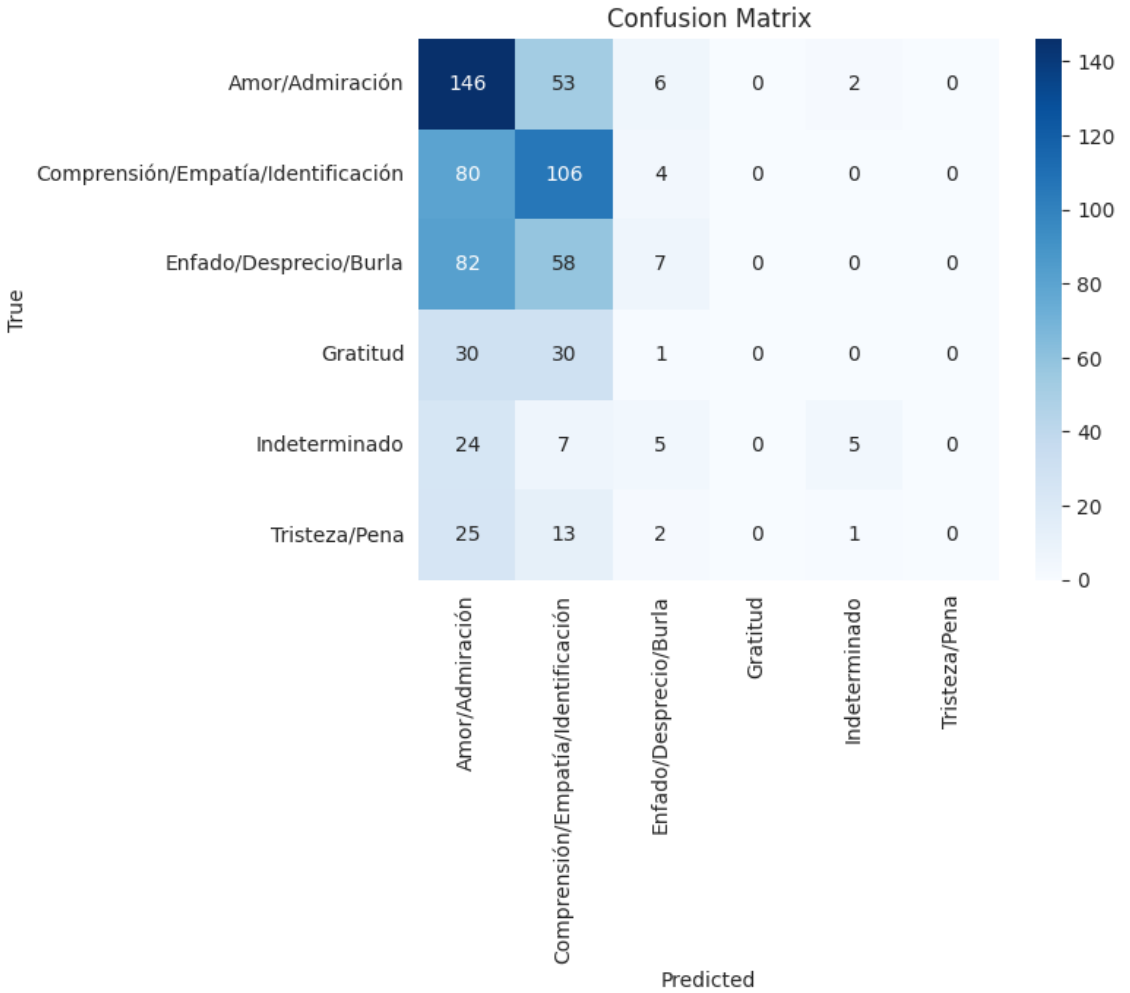

Figura 28: Matriz de confusión SVM emociones Salud Mental

5.4 Clasificación de la respuesta emocional con RandomForest

5.4.1 Resultados de RF para dos polaridades

A continuación, se procede a repetir los casos anteriores utilizando un modelo de Random Forest. Comenzando con el caso de dos polaridades, los resultados se muestran en la Tabla 63 y la Figura 25. En primer lugar, se buscará el mejor parámetro max_features, cuyos resultados se presentan en la Tabla 73. Nuevamente, se observa que la opción óptima es "Sqrt" con una precisión del 70%.

Max_features	Accuracy
Log2	68%
Sqrt	**70%**

Tabla 73: Parámetro Max_features RF 2 polaridades Salud Mental

A continuación, se procede a optimizar el parámetro de regularización en la Tabla 74, donde se encuentra el mejor valor en 475 árboles de decisión con una exactitud del 74%. Sin embargo, se observa que, para valores muy superiores como 1000 y 10000, el problema se sobredimensiona y se obtienen resultados más deficientes.

Árboles	Accuracy
10	67%
100	73%
1000	72%
10000	71%
500	74%
750	73%
400	73%
300	73%
200	73%
600	74%
700	73%
450	74%
550	73%
475	**74%**
525	73%
560	73%
540	73%

Tabla 74: Árboles RF 2 polaridades Salud Mental

Con los hiperparámetros anteriores, se obtiene la Tabla 75 y la Figura 29, que representan el informe de clasificación y la matriz de confusión para el caso de dos polaridades utilizando Random Forest. El modelo alcanza una exactitud global de 74.3%.

La mayoría de las muestras se asignan a la clase "Positivo", la cual obtiene un recall del 94%. Sin embargo, se observa una precisión del 76% en esta clase, lo que indica que hay algunas muestras mal clasificadas y que la mayoría de las muestras predichas están en esta clase como se aprecia en la Figura 29, donde encontramos 561 predichas en esta clase, frente a las 73 muestras predichas en la clase negativa.

En cuanto a la clase "Negativo", se logra una precisión del 63% y un recall del 25% lo que nos dice que el modelo tiene una buena capacidad para predecir correctamente esa clase, pero a costa de ser muy selectivo en la predicción de esta clase, dejando muchas muestras fuera. Como hemos mencionado antes en la Figura 29 se muestra que se predicen 73 instancias en esta clase, de las cuales 46 son clasificadas correctamente. Sin embargo, cabe destacar que en realidad existen 182 muestras pertenecientes a esta clase, lo que resulta en un recall del 25% para la misma ya que este modelo es muy conservador en las predicciones para esta clase.

	Precision	Recall	F1-score	Support	Accuracy
Positivo	76%	94%	84%	452	-
Negativo	63%	25%	36%	182	-
Global	72%	-	70%	634	74.3%

Tabla 75: Informe de clasificación RF 2 polaridades Salud Mental

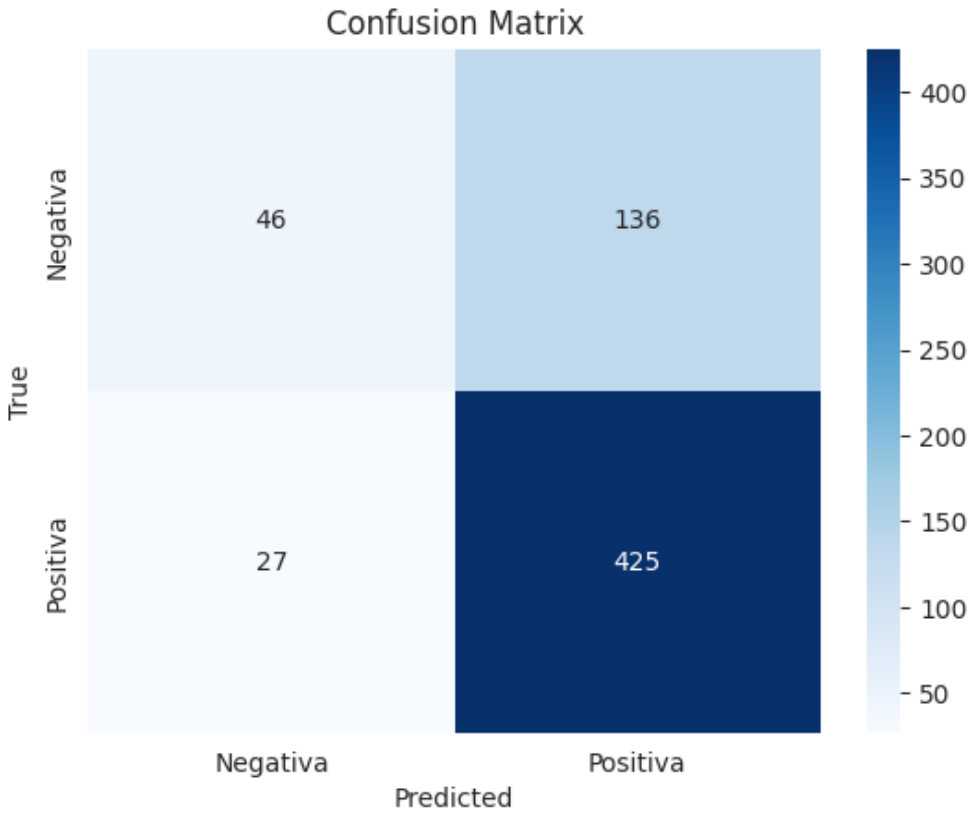

Figura 29: Matriz de confusión RF 2 polaridades Salud Mental

5.4.2 Resultados de RF para tres polaridades

En estas pruebas se añade la clase “Indeterminado”, lo que resulta en un caso de 3 polaridades, tal como se muestra en la Tabla 61 y la Figura 23. A continuación, se procede a buscar el parámetro max_features en la Tabla 76.

Max_features	Accuracy
Log2	66%
Sqrt	**69%**

Tabla 76: Parámetro Max_features RF 3 polaridades Salud Mental

Se encuentra que el mejor valor para el parámetro max_features es "sqrt", logrando una exactitud del 69%. Con este parámetro determinado, se procede a buscar el mejor número de árboles de decisión para abordar el problema. En la Tabla 77 se muestra que el valor óptimo es de 610 árboles de decisión, obteniendo una exactitud del 71%.

Árboles	Accuracy
10	67%
100	70%
1000	69%
10000	69%
500	71%
750	71%
400	70%
300	70%
200	70%
600	71%
700	71%
450	70%
550	71%
475	70%
525	71%
560	71%
540	71%
520	70%
510	70%
490	70%
480	70%
610	**71%**
620	71%

Tabla 77: Árboles RF 3 polaridades Salud Mental

Se finaliza este apartado con la presentación del informe de clasificación en la Tabla 78 y la matriz de confusión en la Figura 30, utilizando los parámetros mencionados anteriormente. Se observa un patrón similar al caso de dos polaridades, donde la clase “Positiva” es dominante en nuestras predicciones como se ve en la Figura 30, ya que la mayoría de muestras (560) caen en la columna de predicción para esta clase obteniendo una exactitud del 92.4%. Sin embargo, la precisión en esta clase es del 74%, lo que indica

que existen una considerable cantidad de muestras mal clasificadas. En cuanto a la clase "Negativo", se logra una precisión del 59%, acertando 69 de las 116 muestras predichas. No obstante, el recall en esta clase es bajo, alcanzando solo un 37%, dado que en realidad hay un total de 186 muestras pertenecientes a esta clase. Por último, la clase "Indeterminado" logra una precisión del 40%, pero solo predice correctamente 4 muestras de las 9 que ha predicho, como se muestra en la Figura 30.

	Precision	Recall	F1-score	Support	Accuracy
Positivo	74%	92%	82%	451	-
Negativo	59%	37%	46%	186	-
Indeterminado	40%	8%	14%	49	-
Global	68%	-	68%	686	71.43%

Tabla 78: Informe de clasificación RF 3 polaridades Salud Mental

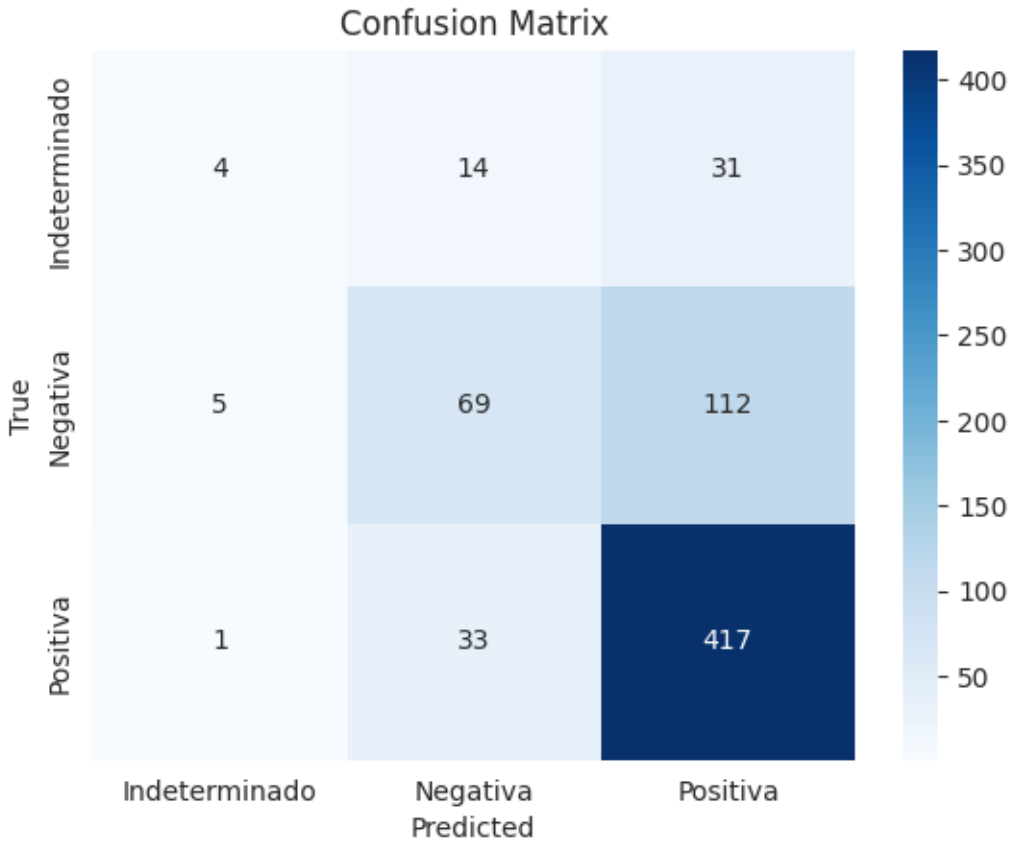

Figura 30: Matriz de confusión RF 3 polaridades Salud Mental

5.4.3 Resultados de RF para la clasificación de emociones

Para concluir el estudio utilizando RF, se analizan las seis emociones presentes en la base de datos, como se muestra en la Figura 24 y la Tabla 62. El primer paso consiste en buscar el valor del parámetro Max_Features adecuado, cuyo resultado se encuentra en la Tabla 79, donde se destaca la elección del sqrt con el cual se consigue una exactitud de 46%.

Max_features	Accuracy
Log2	44%
Sqrt	**46%**

Tabla 79: Parámetro Max_features RF emociones Salud Mental

Con este parámetro determinado, se procede a buscar el mejor número de árboles de decisión para abordar el problema. En la Tabla 80 se muestra que el valor óptimo es de 510 árboles de decisión, obteniendo una exactitud del 48%. Cabe destacar que no se encuentra una diferencia muy significativa al modificar el número de árboles de decisión en este modelo.

Árboles	Accuracy
10	45%
100	46%
1000	46%
10000	47%
500	48%
750	46%
400	48%
300	47%
200	47%
600	47%
700	47%
450	47%
550	48%
475	48%
525	48%
560	47%
540	47%
520	47%
510	**48%**
490	47%
480	47%

Tabla 80: Árboles RF emociones Salud Mental

Con los resultados obtenidos en las dos Tablas anteriores, se procede a mostrar el informe de clasificación en la Tabla 81 y la matriz de confusión en la Figura 31. En la Tabla 81, se puede apreciar que la precisión para todas las clases oscila alrededor del 40%. La clase con la precisión más alta es "Gratitud" con un 87%. Al examinar la matriz de confusión de las 38 muestras predichas en esta clase, se observa que 33 se clasificaron correctamente. Por otro lado, la clase "Tristeza" muestra una precisión del 40% pero un recall de tan solo el 10%, ya que en Figura 31 se aprecia que solo se acertaron 4 de las 41 muestras de esta clase, por lo que se tiene una precision relativamente baja y además esta precision se consigue por medio de una clasificación muy selectiva, ya que se dejan muchas muestras fuera. Las demás clases presentan valores de recall y precisión más equilibrados. En general, se logra una exactitud del 48.5% para todas las clases. También se observa en la Figura 31 que las clases que se confunden más fácilmente en las predicciones son “Comprensión” con “Amor” y “Enfado”, y “Enfado” con “Amor”.

	Precision	Recall	F1-score	Support	Accuracy
Amor/Admiración	48%	56%	51%	207	-
Comprensión/Empatía/Identificación	45%	70%	55%	190	-
Enfado/Desprecio/Burla	51%	25%	34%	147	-
Gratitud	87%	54%	67%	61	-
Indeterminado	37%	27%	31%	41	-
Tristeza/Pena	40%	10%	16%	41	-
Global	50%	-	47%	687	48.5%

Tabla 81: Informe de clasificación RF emociones Salud Mental

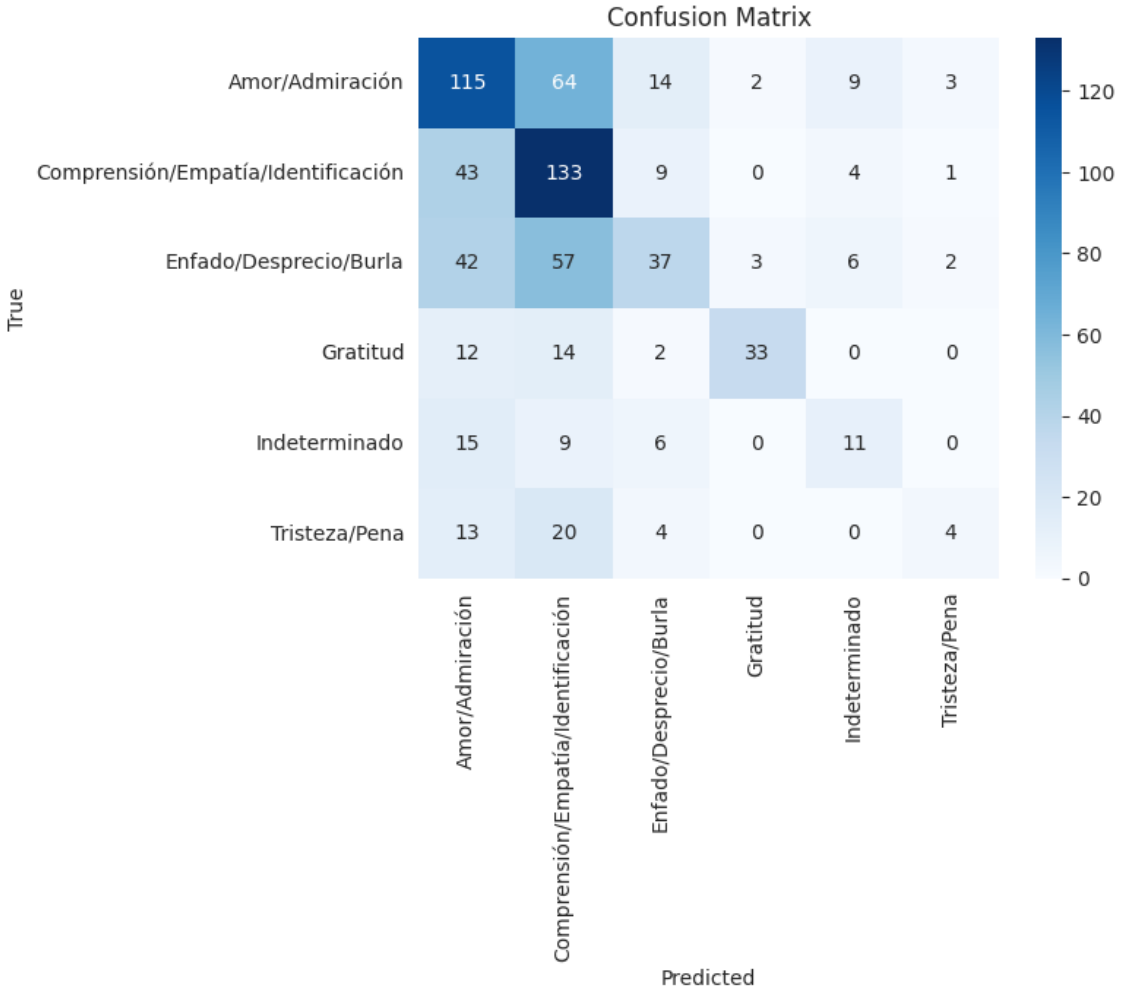

Figura 31: Matriz de confusión RF emociones Salud Mental

5.5 Comparación de modelos y resumen de resultados

Se finaliza este capítulo con una comparación entre los dos modelos utilizados, SVM y Random Forest. Es importante destacar que la base de datos presentaba un desbalance significativo al predecir las polaridades, como se evidencia en las Tablas 23 y 25. En la Tabla 82 se resume la métrica global de la exactitud (accuracy), donde se observa que Random Forest tiene un mejor rendimiento en todos los casos. Sin embargo, se profundizará en los detalles, punto por punto y métrica por métrica a continuación.

EXACTITUD Global	SVM	RF
Dos Polaridades	73%	74.3%
Tres polaridades	65%	71.4%
Emociones	38.5%	48.5%

Tabla 82: Resumen exactitud Salud Mental

Se procederá a comparar las métricas de precisión, recall y puntuación F1 (Tablas 83, 84 y 85 respectivamente) para el caso de dos polaridades (negativa y positiva), donde el 72% de las muestras corresponden a la clase positiva. En este escenario, se observan valores muy similares en ambos modelos, excepto en la métrica de recall para la clase negativa, donde hay una diferencia del 9%. Esta disparidad se debe a que ambos modelos asignan la mayoría de las muestras predichas a la clase positiva, pero en el caso de SVM, esta asignación es aún más pronunciada, dejando muy pocas para la clase negativa. Por lo tanto, a pesar de tener una precisión del 59%, su recall es solo del 16%.

PRECISION	SVM	RF
Positivo	74%	76%
Negativo	59%	63%
Global	70%	72%

Tabla 83: Resumen precision 2 polaridades Salud Mental

RECALL	SVM	RF
Positivo	96%	94%
Negativo	16%	25%
Global	73%	74%

Tabla 84: Resumen recall 2 polaridades Salud Mental

F1-SCORE	SVM	RF
Positivo	83%	84%
Negativo	25%	36%
Global	67%	70%

Tabla 85: Resumen F1-score 2 polaridades Salud Mental

En el caso del problema de tres polaridades, el desbalanceo se agrava aún más, ya que una clase tiene el 67% de las muestras (positiva), mientras que la nueva clase añadida,

"Indeterminado", solo representa un 7%. En este escenario, los valores de las diferentes métricas reflejadas en las Tablas 86, 87 y 88 comienzan a mostrar diferencias significativas en todas las métricas, donde Random Forest claramente se comporta mejor. Es importante destacar que la clase "Positiva" sigue siendo sobreestimada en la predicción, obteniendo valores altos de recall frente a precision, mientras que las clases "Negativa" e "Indeterminada" presentan valores muy bajos en esta métrica. De lo cual se puede interpretar que las predicciones paras las clases minoritarias son muy conservadoras ya que predicen correctamente de una manera aceptable a costa de introducir muy pocas muestras en la clase de esa predicción, dejando la mayoría fuera y por ende el recall sale tan bajo.

Resulta interesante observar el F1 score para obtener una ponderación de ambas métricas y descubrir que, en el caso de SVM, a pesar de tener un 50% de precisión en la clase “Indeterminado”, su valor de F1 score es solo del 4%, en contraste con el 40% de precisión y 14% de F1 score para Random Forest en la misma clase “Indeterminado”.

Esto refleja que tener una precisión buena o un recall bueno no es significativo de tener un buen modelo, se necesitan buenos valores en ambas métricas.

PRECISION	SVM	RF
Positivo	68%	74%
Negativo	42%	59%
Indeterminado	50%	40%
Global	59%	68%

Tabla 86: Resumen precision 3 polaridades Salud Mental

RECALL	SVM	RF
Positivo	90%	92%
Negativo	19%	37%
Indeterminado	2%	8%
Global	64%	71%

Tabla 87: Resumen F1-score 3 polaridades Salud Mental

F1-SCORE	SVM	RF
Positivo	77%	82%
Negativo	26%	46%
Indeterminado	4%	14%
Global	58%	68%

Tabla 88: Resumen recall 3 polaridades Salud Mental

La evaluación de la predicción de las seis emociones se realiza mediante el análisis de las muestras reflejadas en la Tabla 62 y la Figura 24. En las Tablas 89, 90 y 91 se presenta un resumen de las métricas de precisión, recall y puntuación F1 para cada clase. Se observan diferencias significativas a favor de Random Forest, ya que SVM no logra clasificar correctamente las clases menos representadas, excepto en el caso de la clase "Indeterminado". Finalmente, al analizar los valores de F1-score, se destaca que Random Forest supera a SVM en todas las clases, logrando una exactitud del 48.5% para todas las clases. Esto demuestra que Random Forest es más versátil en todos los casos de uso, como se ha evidenciado en los escenarios anteriores, tanto para polaridades como para emociones.

PRECISION	SVM	RF
Amor/Admiración	38%	48%
Comprensión/Empatía/Identificación	40%	45%
Enfado/Desprecio/Burla	28%	51%
Gratitud	0%	87%
Indeterminado	62%	37%
Tristeza/Pena	0%	40%
Global	32%	50%

Tabla 89: Resumen precision emociones Salud Mental

RECALL	SVM	RF
Amor/Admiración	71%	56%
Comprensión/Empatía/Identificación	56%	70%
Enfado/Desprecio/Burla	5%	25%
Gratitud	0%	54%
Indeterminado	12%	27%
Tristeza/Pena	0%	10%
Global	38%	48%

Tabla 90: Resumen recall emociones Salud Mental

F1-SCORE	SVM	RF
Amor/Admiración	49%	51%
Comprensión/Empatía/Identificación	46%	55%
Enfado/Desprecio/Burla	8%	34%
Gratitud	0%	67%
Indeterminado	20%	31%
Tristeza/Pena	0%	16%
Global	31%	47%

Tabla 91: Resumen F1-score emociones Salud Mental

6

Conclusiones y líneas futuras

6.1 Conclusiones

En el presente Trabajo de Fin de Grado, se ha llevado a cabo una evaluación de la capacidad de predicción de diferentes modelos de aprendizaje automático aplicados en distintos casos de uso pertenecientes a diferentes contextos.

En el primer caso de uso se procedió a la clasificación de vinos tintos y blancos utilizando parámetros químicos. Se logró una precisión del 99.5% mediante la implementación de un modelo de Random Forest. A continuación, se exploró la predicción de la calidad del vino en función de dichos parámetros químicos, empleando diferentes modelos de aprendizaje automático. Se empezó con un modelo de regresión basado en capas neuronales donde los resultados obtenidos mostraron una precisión del 59% para el vino blanco y del 62% para el vino tinto. Al aplicar el algoritmo de Support Vector Machine se elevó la precisión a un 68% tanto para el vino tinto como para el blanco. Por último, utilizando el modelo de Random Forest se obtuvieron resultados de precisión más elevados, llegado a niveles del 71% para el vino blanco y del 75% para el vino tinto. Al comparar las distintas métricas de rendimiento (precision, recall, f1-score) se llegó a la conclusión de que, en este problema en particular, el modelo de Random Forest mostró los mejores resultados.

En la segunda parte de este trabajo se aplicaron algunos de estos modelos de aprendizaje automático aplicados al procesamiento de lenguaje natural para el análisis de sentimientos en diferentes redes sociales. Para ello, se realizaron pruebas de clasificación utilizando los modelos de SVM y Random Forest en dos bases de datos diferentes. En ambas bases de datos, se tenía una variable independiente que correspondía a texto, así como dos variables objetivo: polaridad (positiva, negativa e indeterminada) y emociones. La variable texto debía ser procesada mediante técnicas de procesamiento del lenguaje natural para posteriormente ser evaluada en los distintos modelos predictivos.

En la primera base de datos, se recopilaron diversos comentarios de usuarios en el ámbito de los videojuegos dentro de la aplicación Twitch categorizados en polaridad y emociones. Esta base de datos presentaba cierto desequilibrio y un posible problema de etiquetado en las variables relacionadas con las emociones, puesto que algunos comentarios contenían varias emociones y se producía una mezcla de dichas emociones. Para predecir las polaridades, se implementaron dos modelos diferentes: uno considerando únicamente las clases positivo y negativo, y otro que también incluía la clase indeterminado. Utilizando el modelo SVM, se logró una precisión del 73.15% en la

predicción de dos polaridades categorías, y del 55% para las tres polaridades. Por otro lado, el modelo basado en Random Forest alcanzó una precisión del 80% para las dos polaridades y del 70% para las tres polaridades, mostrando ser más versátil en todas las métricas a pesar del bajo número de muestras en la clase indeterminado. Posteriormente, se procedió al estudio de las emociones, donde se contaba con siete categorías. Al utilizar el modelo SVM, se obtuvo una precisión tan solo del 31%, mientras que con Random Forest se logró una precisión del 44.2%. El modelo SVM presentó dificultades en la clasificación de las emociones, ya que tendía a agrupar la mayoría de sus predicciones en la clase mayoritaria (la clase que presentaba mayor desbalanceo con mayor número de muestras), mientras que Random Forest mostró índices de precisión entre el 40% y el 50% para todas las categorías siendo más fuerte a la hora de afrontar los diferentes desbalances de las bases de datos. Sin embargo, es posible que existiera un conflicto en el etiquetado de las emociones.

En relación con la última base de datos, vinculada a la salud mental en redes sociales, la variable independiente consistía en texto que requería ser procesado utilizando técnicas de procesamiento del lenguaje natural, como en el anterior modelo. Esta base de datos mostraba un desequilibrio considerable en la variable objetivo de polaridad, donde la clase positiva estaba representada en gran medida. Al aplicar el modelo SVM al caso de dos polaridades, se obtuvo una precisión del 73%, mientras que para las tres polaridades se redujo su precisión al 65%. Por otro lado, el modelo Random Forest logró una precisión del 74.35% para el caso de dos polaridades y del 71.4% para el caso de tres polaridades. En el escenario de dos polaridades, ambos modelos presentaron un rendimiento muy similar, con métricas equiparables. Sin embargo, al expandir a tres polaridades, Random Forest demostró una mejor capacidad de predicción para la nueva clase y una menor confusión con las demás clases en comparación con SVM. El último análisis realizado para este corpus de salud mental se enfocó en la predicción emocional. En este caso, el corpus presentaba la categorización en seis emociones diferentes con un desequilibrio existente, aunque no tan significativo como en la base de datos de Twitch. El modelo SVM logró una precisión del 38.5%, mientras que Random Forest obtuvo una precisión del 48.5%. SVM mostró los mismos problemas que en la base de datos anterior, al clasificar únicamente las clases predominantes y proporcionar predicciones en su mayoría hacia esas clases. Por otro lado, Random Forest demostró una capacidad más sólida para clasificar todas las clases, con precisiones cercanas al 50% para todas ellas.

En conclusión, Random Forest ha mostrado ser el algoritmo de aprendizaje automático que presenta un mejor comportamiento en todos los casos de estudio, logrando resultados superiores en todas las métricas analizadas y obteniendo resultados satisfactorios de manera destacada.

6.2 Líneas futuras

Después de analizar los diferentes modelos, sería beneficioso considerar varias líneas de investigación para mejorar los resultados. En primer lugar, es fundamental aumentar significativamente el número de muestras en las diferentes bases de datos. Esto permitiría obtener modelos más precisos y optimizar la selección de hiperparámetros. Además, se debe prestar especial atención a equilibrar las bases de datos, intentando que haya un número similar de muestras en las diferentes clases. De esta manera, todas las clases podrían contribuir adecuadamente al entrenamiento y la evaluación del modelo.

Por otro lado, la clasificación de vinos en la base de datos portuguesa arrojó buenos resultados. Sin embargo, sería interesante poner a prueba esos modelos con una base de datos de vinos de otro origen, por ejemplo, español. Esto permitiría evaluar cómo se desempeñan los modelos en una muestra con características diferentes. Al probar con una base de datos de distinta procedencia, se podría obtener una visión más amplia y generalizable del rendimiento de los modelos de clasificación de vinos.

También sería interesante realizar una nueva evaluación de la base de datos de Twitch después de realizar el etiquetado adecuado en las diversas clases de la variable emoción. Esto nos permitiría determinar si los resultados mejoran de manera significativa con la información correctamente clasificada. Además, sería recomendable realizar un análisis más profundo del procesamiento del lenguaje natural, explorando la posibilidad de asignar diferentes pesos a las palabras para lograr una mejor separación entre las distintas clases. También sería interesante probar algoritmos más avanzados en esta área, lo que podría brindar mejores resultados en la clasificación de las emociones en los datos de Twitch.

7

Bibliografía

[1] «Página web de Google Colaboratory». https://colab.research.google.com/ (accedido 28 de junio de 2023).

[2] «Página web de Python». https://www.python.org/ (accedido 28 de junio de 2023).

[3] «Página Web de scikit-learn». https://scikit-learn.org/stable/ (accedido 3 de julio de 2023).

[4] Y. Er y A. Atasoy, «The Classification of White Wine and Red Wine According to Their Physicochemical Qualities», *IJISAE*, n.º 4, pp. 23-26, 2016, doi: 10.1039/b000000x.

[5] S. Kumar, K. Agrawal, y N. Mandan, «Red wine quality prediction using machine learning techniques», en *2020 International Conference on Computer Communication and Informatics, ICCCI 2020*, Institute of Electrical and Electronics Engineers Inc., ene. 2020. doi: 10.1109/ICCCI48352.2020.9104095.

[6] T. Hui-Ye Chiu, C.-W. Wu, y C.-H. Chen, «A HYBRID WINE CLASSIFICATION MODEL FOR QUALITY PREDICTION».

[7] R. S. Akanksha Trivedi, *Wine Quality Detection through Machine Learning Algorithms*. IEEE, 2018.

[8] F. Izaurieta y C. Saavedra, «Redes Neuronales Artificiales».

[9] W. S. Noble, «What is a support vector machine?», 2006. [En línea]. Disponible en: http://www.nature.com/naturebiotechnology

[10] F. Chollet, «Deep Learning with Python», 2018.

[11] P. Probst, M. N. Wright, y A. L. Boulesteix, «Hyperparameters and tuning strategies for random forest», *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, n.º 3. Wiley-Blackwell, 1 de mayo de 2019. doi: 10.1002/widm.1301.