Modelo predictivo del consumo de alcohol en estudiantes de la Universidad de Córdoba a partir de la minería de datos

Lila Patricia López Gaviria

Maestría en Ingeniería de Software

Universidad de Medellín

Facultad de Ingeniería

Medellín, Colombia

Año 2024

Modelo predictivo del consumo de alcohol en estudiantes de la Universidad de Córdoba a partir de la minería de datos

Lila Patricia López Gaviria

Trabajo de grado presentado como requisito para optar al título de:

Magister en Ingeniería de Software

Directora:

Phd., Doctora, Ingeniería. Emilcy Juliana Hernández Leal

Codirector:

Phd., Doctor, Ciencias Médicas. Andrés Felipe Orozco Duque

Universidad de Medellín

Facultad de Ingeniería

Medellín, Colombia

Año 2024

Dedicatoria

Doy gracias a Dios por darme la sabiduría e inteligencia para desarrollar el proyecto y también a mi esposo e hija que me apoyaron en momentos difícil. A mis padres por su compañía en toda esta aventura.

"No todo lo que se puede contar cuenta, y no todo lo que cuenta puede ser contado".
Albert Einstein

Agradecimientos

Quiero dar las gracias a los profesores Emilcy Juliana Hernández Leal y Andrés Felipe Orozco Duque por sus orientaciones durante el tiempo que estuve desarrollando el proyecto, también a mis porfesores de Maestría, en particular, la profesora María Clara Gómez Álvarez.

Resumen

El consumo de alcohol entre estudiantes universitarios es un problema cada vez más frecuente en las Instituciones de Educación Superior (IES). Esto se suma al hecho de que las bebidas alcohólicas están presentes en todo tipo de celebraciones y reuniones sociales. En este contexto, se propone una metodología para la clasificación del riesgo de consumo de alcohol basada en modelos de machine learning. Los modelos evaluados incluyen Logistic Regression, Random Forest, Perceptrón Multicapa y Support Vector Machine (SVM). Random Forest mostró el mejor desempeño general, con un F1-score de 0.45 después de la optimización de hiperparámetros y la selección de características relevantes. El modelo SVM se destacó en la métrica de recall, detectando hasta el 86 % de los casos de consumo de alcohol tras la aplicación de técnicas de balanceo como SMOTE, RUS; no obstante, esto incrementó el número de falsos positivos. Por su parte, los modelos de Regresión Logística y Perceptrón Multicapa presentaron un rendimiento moderado en comparación con los anteriores. El uso de diversas técnicas de balanceo, como SMOTE, ADASYN, RUS, Cluster Centroids, SMOTEENN y Tomek Links, contribuyó a mejorar significativamente desempeño de los modelos, especialmente en términos de recall, permitiendo así una detección de los estudiantes consumidores de alcohol.

Palabras clave: Mineía de datos, aprendizaje automático, consumo de alcohol, predictivo, algoritmos de clasificación.

Abstract

Alcohol consumption among university students is an increasingly frequent problem in Higher Education Institutions (HEIs). This is in addition to the fact that alcoholic beverages are present in all kinds of celebrations and social gatherings. In this context, a methodology for alcohol consumption risk classification based on machine learning models is proposed. The models evaluated include Logistic Regression, Random Forest, Multilayer Perceptron and Support Vector Machine (SVM). Random Forest showed the best overall performance, with an F1-score of 0.45 after hyperparameter optimization and relevant feature selection. The SVM model excelled in the recall metric, detecting up to 86 % of alcohol consumption cases after the application of balancing techniques such as SMOTE, RUS; however, this increased the number of false positives. On the other hand, the Logistic Regression and Multilayer Perceptron models presented a moderate performance compared to the previous ones. The use of various balancing techniques, such as SMOTE, ADASYN, RUS, Cluster Centroids, SMOTEENN and Tomek Links, contributed to significantly improve the performance of the models, especially in terms of recall, thus allowing the detection of student alcohol consumers.

Keywords: Data mining, alcohol drinking, machine learning, prediction, classification algorithms

Lista de Figuras

Lista de Tablas

Tabla de Contenido

1. Introducción

El problema de consumo de alcohol en estudiantes universitarios cada vez es más frecuente en Instituciones de Educación Superior - IES. Esto se suma al hecho de que las bebidas alcohólicas están presentes en todo tipo de celebraciones y reuniones sociales; acompañadas del ocio y el esparcimiento con amigos. En Colombia, cerca de siete millones de personas con edades entre 12 y 65 años son consumidores de alcohol, lo que equivale al 35 % de la población en ese rango de edades, Minsalud (2016).

Por otro lado, Kharabsheh y cols. (2019) afirman que hay pocas investigaciones que trabajan con técnicas de minería de datos para la predicción del consumo de alcohol. Ahora bien, como señala Valdiviezo-Diaz, Torres-Carrión, Bustamante-Granda, y Sánchez-Puertas (2020), rara vez se investigan factores de consumo de alcohol. Sin embargo, Hasta la fecha, no se ha desarrollado un modelo predictivo para el consumo de alcohol entre estudiantes de la Universidad de Córdoba. Ante esta necesidad, se llevó a cabo un proceso de recolección de datos mediante dos fuentes principales. En primer lugar, el instrumento utilizado en este estudio fue adaptado a partir del test, compuesto por 40 preguntas (ver Anexo 1), que incluían variables sociodemográficas, socioeconómicas, académicas y psicosociales. Las preguntas relacionadas con el consumo de alcohol se basaron en el Test de Identificación de los Trastornos Debidos al Consumo de Alcohol (AUDIT) Saunders, Aasland, Babor, De la Fuente, y Grant (1993).

En segundo lugar, se recopilaron datos complementarios de los encuestados a través de los sistemas de información académica Academusoft y PowerCampus, en formato Excel. Las variables incluidas abarcaron información demográfica (edad, sexo biológico, estado civil, factor RH, entre otros), antecedentes académicos, factores socioeconómicos y antecedentes relacionados con el consumo de alcohol. Estos datos fueron utilizados para entrenar los modelos predictivos aplicados en este estudio.

En relación con el estudio de Lamprou (2021), este investigó los factores subyacentes que provocan el fenómeno del consumo compulsivo de alcohol, recopilando y analizando datos de estudiantes de la Universidad de Linköping. En esta investigación, se aplicaron técnicas de minería de datos como árboles de decisión, bosque aleatorio y regresión logística. Los resultados mostraron que la regresión logística fue el método más confiable para predecir el consumo excesivo de alcohol, con una exactitud del 86.50 %, una precisión del 92.64 % y un recall del 90.96 %. Además, el estudio reveló que los factores de riesgo más significativos eran

el tiempo que los estudiantes dedicaban a socializar con sus amigos y su participación en actividades extracurriculares. En conclusión, la investigación sugiere que la cultura estudiantil es el factor más influyente en el consumo de alcohol entre los estudiantes universitarios, lo que resalta la importancia del entorno social en la prevalencia de este comportamiento.

En ese sentido, el proceso metodológico empleado en esta investigación constó de cuatro etapas principales: 1) Revisión del estado del arte y refinamiento del marco conceptual; 2) Caracterización y preparación de las fuentes de datos para la construcción del modelo; 3) Diseño de metodología basada en un modelo de machine learning para clasificar el riesgo del consumo de alcohol; y 4) Validación del modelo mediante la aplicación de técnicas de minería de datos. Para la fase experimental, se adoptó la metodología CRISP-DM, la cual ofreció un enfoque estructurado y flexible que facilitó el desarrollo y la implementación del modelo predictivo, asegurando una adecuada gestión y análisis de los datos.

Ahora bien, la construcción del modelo predictivo tiene como objetivo servir de apoyo a la oficina de bienestar universitario en la toma de decisiones relacionadas con la intervención y prevención en la salud física y mental de los estudiantes de la Universidad de Córdoba. Este modelo facilitará la identificación temprana de factores de consumo de alcohol, permitiendo así, implementar estrategias más efectivas y focalizadas para mejorar el bienestar integral de la comunidad estudiantil.

1.1. Planteamiento del problema

El consumo de alcohol entre estudiantes universitarios es un problema cada vez más común en las Instituciones de Educación Superior (IES). Esto se suma al hecho de que las bebidas alcohólicas están presentes en todo tipo de celebraciones y reuniones sociales, donde suelen estar asociadas al ocio y la convivencia con amigos. En Colombia, aproximadamente siete millones de personas entre 12 y 65 años consumen alcohol, lo que representa el 35 % de la población dentro de ese rango de edad Minsalud (2016).

Según la Organización Mundial de la Salud (OMS), el consumo nocivo de alcohol ocupa el tercer lugar entre los principales factores de riesgo de muerte prematura y discapacidad a nivel mundial OMS (2018). En este contexto, la Universidad de Córdoba llevó a cabo un estudio en 2022 sobre la dinámica del consumo de alcohol en 3,657 estudiantes universitarios, con el objetivo de identificar patrones de consumo, situaciones asociadas y niveles de consumo. Los resultados obtenidos en este estudio sirvieron como base para el diseño de un modelo predictivo del consumo de alcohol en estudiantes de la Universidad de Córdoba.

En referencia a diversos estudios, reportados por los investigadores sobre las características psicosociales de los consumidores de alcohol y otras sustancias. Uno de ellos, llevado a cabo en 13 universidades de Ecuador, evaluó a 3,741 estudiantes universitarios mediante 11 pruebas

psicológicas, entre ellas AUDIT-C, TFDN, ASSIST, PHQ-9, AAQ-7, UCLA-R, SLQ, BISS-11, PSS-10, TIPI-SPA, además de una encuesta sociodemográfica. Los hallazgos de esta investigación destacaron la relevancia de la información psicosocial obtenida, la cual fue clave para identificar variables asociadas al consumo de alcohol y diseñar estrategias efectivas de prevención e intervención.Reátegui y cols. (2020).

En este estudio, se reportó que las técnicas más utilizadas en minería de datos fueron las redes bayesianas y los árboles de decisión, que permiten identificar patrones sociodemográficos asociados al consumo de alcohol Rodríguez de la Cruz, González-Angulo, Salazar-Mendoza, Camacho-Martínez, y López-Cocotle (2022). Por otro lado, Kharabsheh y cols. (2019) afirman que hay pocas investigaciones que trabajan con técnicas de minería de datos para la predicción del consumo de alcohol. Ahora bien, como señala Valdiviezo-Diaz y cols. (2020). Los factores asociados al consumo de alcohol rara vez se investigan en profundidad. Sin embargo, hasta la fecha, no se ha desarrollado un modelo predictivo para el consumo de alcohol entre estudiantes de la Universidad de Córdoba. Ante esta necesidad, se realizó un proceso de recolección de datos que se llevó a cabo mediante dos fuentes principales.

1.2. Pregunta de Investigación

¿Qué técnicas de aprendizaje de máquina supervisado permite predecir el consumo de alcohol en estudiantes de la Universidad de Córdoba a partir de variables sociodemográficas, socioeconómicas, académicas y psicosociales?

1.3. Hipótesis

El uso de técnicas de aprendizaje de máquina supervisado permite predecir el consumo de alcohol en estudiantes de la Universidad de Córdoba a partir de variables sociodemográficas, socioeconómicas, académicas y psicosociales.

2. Justificación

El consumo de alcohol es un problema de salud pública de gran relevancia, ya que, según la Organización Mundial de la Salud (OMS), ocupa el tercer lugar entre los factores de riesgo de muerte prematura y discapacidad a nivel global. Cada año, aproximadamente 2.5 millones de personas, incluidas más de 300,000 jóvenes de entre 15 y 29 años, fallecen por causas relacionadas con el alcohol, lo que representa casi el 4 % de todas las muertes a nivel mundial. Además, el alcohol es un factor causal en 60 tipos de enfermedades y lesiones y está involucrado en otras 200 afecciones, lo que subraya la necesidad de abordar este problema de manera efectiva OMS (2018).

Según el Informe sobre el Consumo de Drogas en las Américas, el 54 % de los estudiantes universitarios en Colombia consume alcohol, destacándose que el 31.2 % de los hombres presenta un consumo problemático. Este panorama evidencia la urgencia de investigar los hábitos de consumo de alcohol en la población universitaria, un grupo especialmente vulnerable debido a su estilo de vida. OEA (2019). Además, el consumo de alcohol es el tercer factor de riesgo de muerte prematura y discapacidad a nivel mundial, siendo responsable del 4 % de las muertes globales. En los jóvenes, el consumo excesivo de alcohol puede tener graves consecuencias, como relaciones sexuales de alto riesgo, accidentes de tránsito y el desarrollo de enfermedades crónicas Ahumada-Cortez, Gámez-Medina, y Valdez-Montero (2017).

La presente investigación tiene como propósito desarrollar un modelo predictivo que integre variables sociodemográficas, socioeconómicas, académicas y psicosociales, permitiendo así detectar, identificar y analizar los patrones de consumo de alcohol entre los estudiantes universitarios. Particularmente, en aquellos grupos con mayor prevalencia. La clasificación de esta población según sus niveles de consumo es fundamental para comprender mejor los factores que influyen en sus hábitos. Este modelo no solo contribuirá al conocimiento académico sobre el consumo de alcohol en el entorno universitario, sino que también ofrecerá herramientas prácticas para la intervención.

Desde una perspectiva de impacto social, el desarrollo del modelo predictivo proporcionará a la oficina de bienestar universitario datos precisos para la toma de decisiones en materia de prevención e intervención, lo que apoyará significativamente la salud física y mental de los estudiantes universitarios. La viabilidad de este modelo radica en su capacidad para procesar y analizar grandes volúmenes de datos mediante técnicas de minería de datos. Además, su implementación no solo permitirá identificar de forma temprana los factores asociados al

consumo de alcohol, sino que también facilitará la creación de estrategias preventivas y de intervención más efectivas y focalizadas, mejorando el bienestar integral de la comunidad estudiantil de la Universidad de Córdoba. Este enfoque no solo aborda un problema crítico de salud pública, sino que también sienta las bases para futuras investigaciones en el ámbito de la salud pública y el comportamiento juvenil.

2.1. Objetivos

2.1.1. Objetivo General

Diseñar un modelo predictivo del consumo de alcohol en estudiantes de la Universidad de Córdoba a partir de técnicas de minería de datos

2.1.2. Objetivos Específicos

- Analizar los componentes para la construcción de un modelo predictivo del consumo de alcohol en estudiantes universitarios.
- Caracterizar las fuentes de datos y hacer el preprocesamiento y extracción de los datos que alimentarán el modelo predictivo del consumo de alcohol en estudiantes universitarios.
- Construir un modelo predictivo del consumo de alcohol en estudiantes universitarios.
- Evaluar la precisión y la efectividad del modelo predictivo del consumo de alcohol en estudiantes universitarios.

2.2. Marco Teórico

En esta investigación se abordan los conceptos clave de alcohol, alcoholismo, minería de datos, machine learning, metodología CRISP-DM aprendizaje supervisado y técnicas de desbalanceo.

2.2.1. Alcohol

El alcohol etílico o etanol, es un líquido incoloro, de sabor urente y olor fuerte, que arde fácilmente dando llama azulada y poco luminosa. Se obtiene por destilación de productos de fermentación de sustancias azucaradas o feculentas, como uva, melaza, remolacha, patata, así como de distintos cereales Pascual Pastor (2012). Además, el alcohol es una sustancia psicoactiva, clasificada como depresora del sistema nervioso central, que se encuentra comúnmente

en bebidas como cerveza, vino y licores. Su consumo puede tener efectos variados, dependiendo de la cantidad ingerida, el contexto y la tolerancia de la persona Ferrera Perera y cols. (2019).

2.2.2. Alcoholismo

Es una enfermedad crónica caracterizada por el consumo excesivo y compulsivo de alcohol, que afecta la salud física, mental y social del individuo. El alcoholismo se define por la dependencia que una persona desarrolla hacia el alcohol, interfiriendo con su capacidad para funcionar de manera normal en su vida diaria. Arteaga Yánez y cols. (2022).

2.2.3. Consumo de alcohol

los estudiantes universitarios de Colombia registraron la cifra más alta en cuanto al consumo de alcohol al menos una vez en la vida (95,8 %), en comparación con otros países latinoamericanos como Ecuador (88,7 %), Perú (87,5 %) y Bolivia (77,1 %), los cuales presentan cifras inferiores. Estos países, al igual que Colombia, se encuentran en una situación de consumo riesgoso o perjudicial Betancourth-Zambrano, Tacán-Bastidas, y Cordoba-Paz (2017).

2.2.4. Minería de datos o Data Mining

Según Frawley, Piatetsky-Shapiro, y Matheus (1992), la minería de datos se describe como la "extracción no trivial de información implícita, previamente desconocida y potencialmente útil a partir de datos". Por otro lado, Tan, Steinbach, y Kumar (2006) la definen como la "exploración y análisis, por medios automáticos o semiautomáticos, de grandes cantidades de datos para descubrir patrones significativos".

El proceso de obtención de conocimiento a partir de los datos, mediante la identificación de patrones o modelos, es descrito por Fayyad, Piatetsky-Shapiro, y Smyth (1996) como un enfoque integral que incluye la recolección y análisis de datos. Según estos autores, las técnicas de minería de datos se dividen en dos categorías: supervisadas o predictivas, y no supervisadas o descriptivas. En esta investigación, nos centraremos en las técnicas supervisadas. Como se ilustra en la Figura **2-1**.

Según García y cols. (2018), el proceso de las técnicas de minería de datos y su enfoque conceptual están orientados a la extracción de información a partir de los datos, con un énfasis particular en el desarrollo de varios algoritmos. Estos algoritmos representan la forma en que se implementa una técnica determinada, y su correcta aplicación requiere de un conocimiento técnico adecuado para seleccionar el más apropiado en función del problema de minería

Figura 2-1.: Técnicas de minería de datos
Fuente: (Ciencia de datos, García y cols. (2018), pp. 202)

de datos. Además, es esencial comprender tanto las variables como las características de los algoritmos para preparar los datos de manera efectiva antes de su análisis.

La metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) es una de las más utilizadas por los investigadores en proyectos de minería de datos. Esto se debe a que se basa en un enfoque práctico y en la experiencia de los analistas de minería de datos, quienes han contribuido significativamente a su desarrollo y adopción. Esta metodología proporciona una estructura flexible y adaptable que permite guiar el proceso de minería de datos desde la comprensión del problema hasta la implementación de los modelos predictivos. (Gironés y cols. (2017), pp. 26). En la Fig **2-2**. se esquematiza las etapas que propone CRISP-DM.

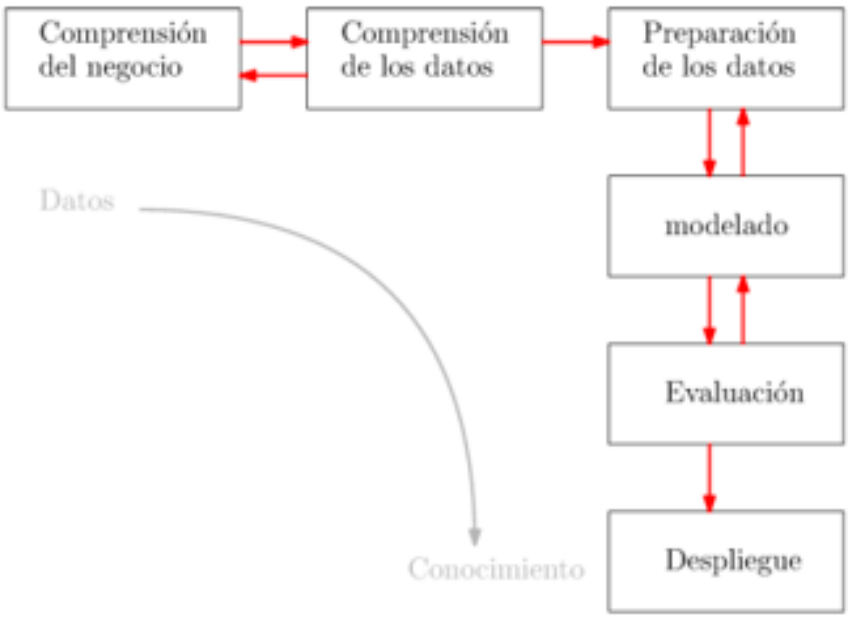


Figura 2-2.: Etapas de la metodología CRISP-DM
Fuente: (Minería de datos: Modelos y Algoritmos, Gironés y cols. (2017), pp. 28)

2.2.5. Aprendizaje Supervisado

Gerard (2021), el aprendizaje supervisado lo define como "la capacidad de encontrar patrones en los datos utilizando tanto características como etiquetas". En este contexto, cuando se utiliza un conjunto de datos, los rasgos representan las características de cada entrada, mientras que las etiquetas permiten clasificar o definir esas entradas de manera explícita.

2.2.6. Machine Learning o Aprendizaje Automático

En 1959, Arthur Samuel, pionero estadounidense en los campos de los juegos de ordenador, el aprendizaje automático y la inteligencia artificial, definió el machine learning como «un campo de estudio que otorga a las computadoras la capacidad de aprender sin estar programadas de manera explícita». Con el paso del tiempo, el machine learning ha incorporado diversos algoritmos y técnicas para el desarrollo de modelos y programas computacionales. Swamynathan (2017) lo define como «una colección de algoritmos y técnicas utilizados para crear sistemas computacionales que aprenden a partir de los datos con el fin de hacer predicciones e inferencias».

2.2.7. Técnicas de balanceo

Las técnicas de balanceo son enfoques empleados en machine learning para manejar conjuntos de datos en los que existe un desequilibrio significativo entre las clases. Estos métodos tienen como objetivo equilibrar las clases para mejorar el rendimiento de los modelos predictivos en problemas donde una clase cuenta con muchos más ejemplos que las demás, y donde la clase minoritaria es la más importante para el análisis, Cruz, González, y Rangel (2022).

Las principales técnicas más utilizadas por los investigadores para abordar el problema del desbalanceo de datos incluyen:

1. **Sobremuestreo (Oversampling)**: Aumenta el número de ejemplos de la clase minoritaria, replicando ejemplos existentes o creando nuevas muestras sintéticas, Chawla, Bowyer, Hall, y Kegelmeyer (2002).
    - **SMOTE** (Synthetic Minority Over-sampling Technique): Genera ejemplos sintéticos de la clase minoritaria para equilibrar las clases. En concreto, SMOTE genera el mismo número de muestras de datos sintéticos para cada minoría original. He y Garcia (2009)
    - **ADASYN** (Adaptive Synthetic Sampling): Genera ejemplos sintéticos en áreas donde la clase minoritaria está menos representada o donde hay más riesgo de mala clasificación. He, Bai, Garcia, y Li (2008).

2. **Submuestreo (Undersampling)**: Reduce el número de ejemplos de la clase mayoritaria, eliminando algunos de ellos, lo que disminuye el sesgo hacia esa clase. He y Garcia (2009).

    - **RUS** (Random UnderSampler): Reduce el número de muestras de la clase mayoritaria, Elhassan y Aljurf (2016)

    - **Cluster Centroids**, es una técnica utilizada para manejar el desbalance de clases en minería de datos y aprendizaje automático, mediante la reducción de la clases mayoritaria al reeplazarla con los centroides de los clusters. Según Yen y Lee (2009).

3. **Técnicas Híbridas (Oversampling + Undersampling)**: Combinan el sobremuestreo y el submuestreo para equilibrar los datos, eliminando ejemplos ruidosos y generando ejemplos sintéticos para la clase minoritaria. Batista, Prati, y Monard (2004).

    - **SMOTEENN**(Systhetic Minority Over-sampling Technique combined with Edited Nearest Neighbors), es una técnica hibrida que combina el sobremuestreo sintético (SMOTE) con filtrado de datos, através del algoritmo de edición de vecinos más cercano(ENN). Según Batista y cols. (2004), afirma que una técnica eficaz para manejar conjunto de datos desbalancedaos.

    - **Tomek Links**, es una técnica empleada para limpiar los límites entre clases en conjuntos de datos desbalanceados, eliminando ejemplos ambiguos o ruidosos. De acuerdo con Tomek (1976).

2.3. Revisión de la literatura

En esta sección, se realizó el proceso de revisión de la literatura en diversas bases de datos, seleccionando estudios primarios relacionados con la aplicación de técnicas de minería de datos para la predicción de factores asociados al consumo de alcohol en estudiantes universitarios. Además, se aplicó el protocolo de revisión de literatura propuesto por Kitchenham y Charters (2007), el cual consta de las siguientes etapas: 1) Definición de las preguntas de investigación, 2) Identificación de palabras clave para la búsqueda en inglés y español, 3) Elaboración de cadenas de búsqueda, 4) Definición de los criterios de inclusión y exclusión de los estudios primarios, 5) Creación de búsquedas avanzadas en las bases de datos, 6) Extracción de información de los artículos, 7) Análisis de los artículos seleccionados mediante la aplicación de los criterios de inclusión y exclusión, y 8) Presentación de los resultados obtenidos en la revisión de la literatura. Finalmente, se presentarán algunos estudios primarios que fueron considerados en esta investigación.

2.3.1. Pregunta de investigación

La siguiente revisión de literatura tiene como propósito presentar los trabajos asociados a las técnicas de minería de datos y los algoritmos más utilizados para predecir el consumo de alcohol en universitarios. A continuación, se ha planteado la siguiente pregunta de investigación.

¿Cuáles son las técnicas de minería de datos para predecir el consumo de alcohol en estudiantes universitarios?

2.3.2. Palabras clave

Se definieron las palabras clave para realizar la búsqueda en los idiomas en inglés y español, debido a que tiene mayor posibilidad de generar resultados en las bases de datos. Como se muestra en la **Tabla2-1**.

Tabla 2-1.: Selección de las palabras clave

Minería de datos	Consumo de alcohol	Estudiantes universitarios
Data mining techniques	Consumption Alcoholic beverage	University students
Machine learning	Alcohol drinking	Universities
Classification algorithms	Alcoholism	College
Prediction	binge drinking	Higher education
Patterns		

2.3.3. Cadenas de búsqueda

Se emplearon diversas cadenas de búsqueda en las bases de datos especializadas para garantizar una revisión exhaustiva. A continuación, se presenta la cadena de búsqueda general utilizada. Además, se crearon dos cadenas de búsqueda específicas, debido a la falta de resultados en la base de datos Scopus.

- («Alcohol drinking» OR «Alcoholic beverage» OR «alcoholism» OR «binge drinking») AND («University students» OR «Higher education» OR «College») AND («Machine learning» OR «Data mining» OR «Classification algorithms» OR «Prediction» OR «Patterns»).
- («Machine learning» OR «data mining») AND («Imbalanced data») AND («categorical data» OR «qualitative data») AND («continuous data» OR «quantitate data»).

2.3.4. Selección de los criterios

Los criterios de inclusión y exclusión de la revisión de literatura se definieron de acuerdo con la aplicación de técnicas de minería de datos para predecir el consumo de alcohol en jóvenes universitarios. Como se ilustra en la **Tabla2-2**.

Tabla 2-2.: Selección de los criterios

Criterios de inclusión	Criterios Exclusión
Fecha de publicación (2018 - 2023).	Fecha de publicación superior a cinco años.
Tipos de documentos (Artículos de investigación, capítulo de libro y conference proceeding).	Tipos de documentos (Artículos de revisión, opinión y literatura gris).
Universidades	Educación básica media.
Artículos escritos en inglés y español.	Artículos escritos en otro idioma diferente al inglés y español.

2.3.5. Bases de datos

Las bases de datos seleccionadas para esta revisión fueron: Dimensions, Google Scholar, IEEE Xplore, Science Direct y Scopus. Tras ejecutar las cadenas de búsqueda en cada una de estas plataformas, se obtuvieron los siguientes resultados. Como se muestra en la **Tabla 2-3**.

Tabla 2-3.: Resultados de las cadenas de búsquedas en las bases de datos

Cadena	Dimensions	Google Scholar	IEEE Xplore	Science Direct	Scopus
A	145	26	27	18	11

2.3.6. Estudios seleccionados

Una vez aplicadas las cadenas de búsqueda en las bases de datos, se obtuvieron los siguientes hallazgos: de los 227 artículos encontrados, solo 14 cumplieron con los criterios de inclusión. En este sentido, los estudios seleccionados contribuyeron significativamente a este trabajo (ver Anexo 4).

2.3.7. Presentación resultados

Una vez finalizado el proceso de revisión de literatura en las bases de datos, se obtuvieron 14 artículos primarios relacionados con la aplicación de técnicas de minería de datos para la identificación de factores de consumo de alcohol en estudiantes universitarios. Estos

estudios seleccionados proporcionan respuestas a las preguntas de investigación planteadas inicialmente:

¿Cuáles son las técnicas de minería de datos para predecir el consumo de alcohol en estudiantes universitarios?

R1: Según Fierro, Castañeda, y Revelo-Aldás y Valdiviezo-Diaz y cols. Las técnicas de minería de datos más utilizadas para predecir el consumo de alcohol en jóvenes universitarios son los árboles de decisión, las redes neuronales y la regresión logística. Sin embargo, Marcon y cols. Menciona que las técnicas de menor uso en este ámbito incluyen las redes bayesianas y las máquinas de soporte de vectores, debido a su complejidad y menor aplicabilidad en estudios con datos limitados o desbalanceados. Además, los algoritmos más utilizados por los autores en los artículos empleados son los siguientes: árbol de decisión, red neuronal artificial, K-vecino más próximo y bosques aleatorio, y el algoritmo con menor aplicación es Naïve bayes. También, estos algoritmos son evaluados y comparados entre ellos por su precisión y efectividad en los resultados obtenidos.(Valdiviezo-Diaz y cols. (2020) Marcon y cols. (2021) Fierro y cols. (2022)).

2.3.8. Antecedentes

Lamprou (2021), realizó un estudio sobre los factores subyacentes que provocan el fenómeno del consumo compulsivo de alcohol, recopilando y analizando datos de la Universidad de Linköping. En este estudio, se aplicaron técnicas de minería de datos como árboles de decisión, bosques aleatorios y regresión logística. Los resultados mostraron que la regresión logística fue el método más confiable para predecir el consumo excesivo de alcohol, con una exactitud del 86.50 %, una precisión del 92.64 % y un recall del 90.96 %. Además, el estudio reveló que los factores de riesgo más significativos fueron el tiempo que los estudiantes pasaban con sus amigos y su participación en actividades extracurriculares. En conclusión, la investigación sugiere que la cultura estudiantil es el factor más influyente en el consumo de alcohol entre los estudiantes universitarios.

Fierro y cols. (2022), realizaron un estudio orientado a la predicción de alcoholismo en estudiantes, la información la tomaron de un repositorio de dataset Kaggle con 521 registros y seleccionaron 17 variables, obtuvieron tres modelos predictivos de los cuales se destacó el modelo de regresión lineal obteniendo una mayor precisión en comparación de los modelos K vecino más próximo y árbol de decisión. Los resultados de esta investigación identificaron que los indicadores psicosociales más relevantes para determinar la tendencia al consumo de alcohol son el estado familiar, el tipo de zona de residencia (urbana o rural), el género y la cantidad de tiempo libre.

Otro estudio orientado a la identificación de patrones de consumo de alcohol y episodios de consumo intensivo (ECI) en estudiantes universitarios de ciencias de la salud realizó un análi-

sis descriptivo mediante un cuestionario que incluía variables sociodemográficas. Además, se emplearon varios test de hipótesis paramétricos y no paramétricos, utilizando el programa SPSS. Los resultados de esta investigación evidencian asociaciones estadísticas significativas entre el consumo semanal y las variables de sexo, domicilio habitual y edad. El estudio clasificó a la población en tres patrones de consumo: riesgo bajo, riesgo moderado y riesgo elevado, destacando la alta prevalencia de episodios de consumo intensivo en estudiantes varones García-Carretero y cols. (2019).

Valdiviezo-Diaz y cols. (2020), realizaron un estudio a 7905 estudiantes universitarios. Esta investigación aplicó los tests AUDIT, TDFN y ASSITS. Obtuvieron una muestra de 478 registros, mediante la aplicación de técnicas de minería de datos para predecir patrones de poli-consumidores de tabaco y alcohol en estudiantes ecuatorianos. En este estudio evaluaron 73 variables, de las cuales están distribuidas de la siguiente manera: variables sociodemográficas, psicosociales, de salud y consumo, también emplean en esta investigación dos modelos de predicción, utilizando algoritmos de minería de datos tales como: los árboles de decisión y redes neuronales. Los resultados obtenidos en este estudio, en el que se emplearon 73 variables, destacan siete variables clave para el modelo de predicción aplicado mediante algoritmos de árboles de decisión y redes neuronales. Estos algoritmos permitieron identificar la tendencia al consumo de alcohol en función de la información recopilada en los test aplicados en esta investigación.

En general, los estudios anteriores suelen centrarse en la aplicación de pruebas como AUDIT, TDFN y ASSITS o en cuestionarios sociodemográficos. Este trabajo, sin embargo, se diferencia al incorporar datos de registros de sistemas académicos utilizados por la Universidad de Córdoba. Aunque el test AUDIT fue considerado como referencia, los modelos aquí desarrollados se alimentan directamente de la información disponible en el sistema académico. De esta manera, el modelo permite identificar a los estudiantes en riesgo de consumo de alcohol y posibilita intervenciones personalizadas, sin necesidad de incluir variables de contexto social. Esto ofrece un primer acercamiento para detectar tempranamente a la población en riesgo y facilitar intervenciones preventivas.

Por otro lado, los estudios anteriores se centran en la implementación de modelos de machine learning, sin mencionar la aplicación de técnicas de balanceo. En contraste, este trabajo se destaca por la aplicación de diversas técnicas de sobremuestreo, como SMOTE y ADASYN; de submuestreo, como RUS y Cluster Centroids; así como técnicas híbridas, como SMOTEENN y Tomek Links. Esta estrategia ha permitido mejorar significativamente el rendimiento de los modelos utilizados en este proyecto.

En conclusión, la revisión de la literatura en esta investigación permitió identificar los hallazgos más relevantes en el uso de minería de datos y los algoritmos más empleados por diversos autores para la detección del consumo de alcohol en jóvenes universitarios. Estos aportes fueron fundamentales para el desarrollo de este trabajo de grado.

3. Materiales y métodos

3.1. Proceso de entrenamiento de un modelo de Machine Learning

El proceso de entrenamiento de los modelos de Machine Learning se desarrolló en varias etapas: recolección de datos, preprocesamiento de la información, selección del modelo de clasificación, aplicación de técnicas de balanceo, entrenamiento del modelo, ajuste de hiperparámetros, evaluación del modelo y comparación de resultados. Todo esto se llevó a cabo con el objetivo de obtener el mejor clasificador para predecir el consumo de alcohol en estudiantes universitarios, como se muestra en la Figura. **3-1**.

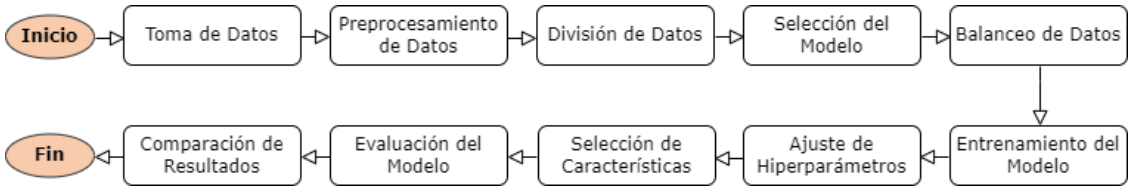


Figura 3-1.: Proceso de entrenamiento de un modelo

3.1.1. Toma de Datos (Dataset)

La recolección de datos se llevó a cabo a través de dos fuentes: en primer lugar, se utilizó una encuesta compuesta por 40 preguntas, que abarcaban variables sociodemográficas, socioeconómicas, académicas y psicosociales. Las preguntas relacionadas con el consumo de alcohol se basaron en el Test de Identificación de los Trastornos Debidos al Consumo de Alcohol - AUDIT (ver Anexo 1) Saunders y cols. (1993). En el estudio participaron 3567 estudiantes de pregrado de todos los programas académicos de la Universidad de Córdoba durante el segundo semestre del año 2022. Los datos obtenidos mediante este test no fueron procesados, ya que se utilizó solo como referencia.

Por otro lado, la información complementaria de los encuestados se obtuvo de los sistemas de información académica Academusoft y PowerCampus, en formato Excel. Los datos constaban

de 79 columnas(ver Anexo 2) y 15,639 registros correspondientes a todos los estudiantes matriculados en el segundo semestre de 2022. Esta información incluía datos relacionados variables demográficas (edad, sexo biológico, estado civil, factor RH, etc.), antecedentes escolares, factores socioeconómicos, académicos y antecedentes sobre el consumo de alcohol (variables relacionadas con el consumo de sustancias mensual, consumo en alguna etapa de la vida, tipo de sustancias, entorno social de consumo, situaciones propicias para el consumo y edad de inicio en el consumo) (Ver Anexo2). Los datos extraídos de estos sistemas fueron procesados de la siguiente forma: se depuró la información de los estudiantes matriculados en el segundo semestre de 2022. Se seleccionaron únicamente los estudiantes que participaron en la encuesta aplicada inicialmente y se incorporaron las variables sociodemográficas. Además, se eliminaron los registros vacíos y aquellos con datos incompletos, y se redujeron los atributos redundantes de cada categoría para optimizar el conjunto de datos y utilizados para entrenar los modelos aplicados en esta investigación.

3.1.2. Preprocesamiento de Datos

Una vez se recolectaron los datos a través del sistema de información académico, se procedió al preprocesamiento de los mismos. Este proceso incluyó, en primer lugar, la limpieza de los datos, lo que implicó la eliminación de registros incompletos o inconsistentes. Posteriormente, se llevó a cabo la transformación de los datos, ajustando las variables según los requerimientos del análisis, con el fin de garantizar la calidad y la coherencia de la información para los modelos de clasificación utilizados en el estudio.

3.1.3. División de los datos (Entrenamiento y prueba)

En esta investigación, los datos se dividieron en dos conjuntos: un conjunto de entrenamiento (80 %), utilizado para entrenar el modelo, y un conjunto de prueba (20 %), se emplea para evaluar el rendimiento del modelo.

3.1.4. Selección del Modelo

En relación con los modelos seleccionados para este estudio, se eligieron aquellos que mejor se ajustaran a la naturaleza de los datos y abordaran eficazmente el problema de clasificación, proporcionando un mejor desempeño en la predicción del consumo de alcohol. En ese sentido, se emplearon los siguientes modelos de clasificación: Regresión Logística, Perceptrón Multicapa, Máquinas de Soporte Vectorial (SVM) y Random Forest.

3.1.5. Balanceo de Datos

En esta investigación, se identificó un desbalance significativo en el conjunto de datos, lo que significa que las clases no están representadas de manera equitativa. En concreto, el 88.8 % de los estudiantes pertenece a la clase mayoritaria (0), que corresponde a quienes no consumen alcohol, mientras que solo el 11.2 % pertenece a la clase minoritaria (1), que incluye a los estudiantes que sí consumen alcohol. Este desequilibrio en la distribución de las clases puede distorsionar el rendimiento de los modelos de Machine Learning.
Para abordar este problema, se emplearon las siguientes técnicas:

1. **Sobremuestreo (Oversampling)**: Aumenta el número de ejemplos de la clase minoritaria, replicando ejemplos existentes o creando nuevas muestras sintéticas, como en las técnicas **SMOTE** (Synthetic Minority Over-sampling Technique) afirma Chawla y cols. (2002) y **ADASYN** (Adaptive Synthetic Sampling) He y cols. (2008).

2. **Submuestreo (Undersampling)**: Reduce el número de ejemplos de la clase mayoritaria, eliminando algunos de ellos, lo que disminuye el sesgo hacia esa clase. Se aplicaron técnicas como **RUS** (Random UnderSampler) Elhassan y Aljurf (2016) y **Cluster Centroids**, es una técnica utilizada para manejar el desbalance de clases en minería de datos y aprendizaje automático, mediante la reducción de la clases mayoritaria al reeplazarla con los centroides de los clusters. Según Yen y Lee (2009), afirma que el uso de centroides de clusters como una forma de submuestreo para mejorar el rendimineto de modelos en datos desbalanceados.

3. **Técnicas híbridas**: Combinan el sobremuestreo y el submuestreo para equilibrar los datos, eliminando ejemplos ruidosos y generando ejemplos sintéticos para la clase minoritaria. Ejemplos de estas técnicas incluyen **SMOTEENN**(Systhetic Minority Oversampling Technique combined with Edited Nearest Neighbors), es una técnica hibrida que combina el sobremuestreo sintético (SMOTE) con filtrado de datos, através del algoritmo de edición de vecinos más cercano(ENN). Conforme a Batista y cols. (2004), afirma que esta técnica es eficaz para manejar conjunto de datos desbalancedaos y **Tomek Links**, es una técnica empleada para limpiar los límites entre clases en conjuntos de datos desbalanceados, eliminando ejemplos ambiguos o ruidosos. De acuerdo con Tomek (1976), mejora la calidad del conjunto de datos al hacer que los límites entre clases sean más claros, especialmente en problemas de clasificación supervisada con desbalanceo de clases.

3.1.6. Entrenamiento del Modelo

Una vez que los datos han sido balanceados, se ajusta el modelo a los datos de entrenamiento para realizar predicciones precisas sobre el consumo de alcohol en estudiantes universitarios. A continuación, se describe este proceso utilizando el modelo Random Forest:

```
model = RandomForestClassifier(random_state=42)
model.fit(X_train_smote, y_train_smote)
```

- **Creación del modelo**: Inicialmente, se crea una instancia del modelo **RandomForestClassifier** y se ajusta a los datos de entrenamiento.
- **random_state=42**: Se utiliza este parámetro para garantizar la reproducibilidad de los resultados. Al establecer este valor, se asegura que la división aleatoria de los datos sea la misma cada vez que se entrene el modelo.
- **Entrenamiento del modelo**: El método fit(X_train_smote, y_train_smote) se utiliza para entrenar el modelo con el conjunto de datos balanceados. Aquí, X_train_smote contiene las características de los datos y y_train_smote las etiquetas. El modelo aprenderá la relación entre las características y las etiquetas para realizar predicciones precisas.

3.1.7. Ajuste de Hiperparámetros

Para abordar el problema del desbalance de los datos, se utilizaron técnicas de ajuste de hiperparámetros y validación cruzada con el objetivo de mejorar el rendimiento del modelo. El ajuste de hiperparámetros consiste en encontrar los mejores valores para los parámetros que controlan el comportamiento de los algoritmos de Machine Learning.
En ese sentido, se aplicó la técnica sistemática GridSearch, para ajustar los hiperpárametros del modelo. Consiste en:

1. Definir un conjunto de posibles valores para cada hiperparámetro.
2. Entrenar el modelo con todas las combinaciones posibles de estos valores.
3. Evaluar el rendimiento de cada combinación utilizando una métrica de evaluación adecuada.
4. Seleccionar la combinación de hiperparámetros que maximice o minimice la métrica de rendimiento.

Por ejemplo, para un modelo de Random Forest, se realiza un Grid Search sobre los hiperparámetros n_estimators (número de árboles) y max_depth (profundidad máxima de los árboles) de la siguiente manera:

Validación Cruzada

La validación cruzada es una técnica utilizada para evaluar el rendimiento del modelo mientras se evita el sobreajuste (overfitting). En lugar de entrenar y probar el modelo en un solo conjunto de datos, el conjunto completo se divide en varias particiones o "folds". Esto

```python
# Definir el modelo
modelrfc = RandomForestClassifier(random_state=42)

# Definir los parámetros para GridSearchCV
params = {
    'n_estimators': [50, 100, 200],
    'max_depth': [None, 10, 20, 30],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
    'bootstrap': [True, False]
}
```

Figura 3-2.: Ejemplo hiperpárametros

permite que el modelo se entrene y evalúe en diferentes subconjuntos de los datos.
En este estudio, se utilizó cv=5, lo que significa que el conjunto de datos se dividió en 5 partes. El proceso de validación cruzada funciona de la siguiente manera:

1. El conjunto de datos se divide en 5 partes o "folds".
2. Se entrena el modelo 5 veces, donde en cada iteración, 4 partes se utilizan para el entrenamiento y 1 parte para la evaluación.
3. Al final, se calcula el promedio de las 5 evaluaciones como la métrica final de rendimiento del modelo.

Grid Search con Validación Cruzada

El Grid Search con validación cruzada combina ambas técnicas de manera eficiente. En este proceso, cada combinación de hiperparámetros se entrena y evalúa utilizando validación cruzada, lo que permite obtener una estimación más precisa del rendimiento del modelo para cada configuración de hiperparámetros. De esta forma, se selecciona la mejor combinación que optimiza el rendimiento general del modelo. Como lo muestra la Figura **3-3**.

En cuanto a las técnicas de Grid Search y validación cruzada, el método **search.fit()** realiza la búsqueda de hiperparámetros y ajusta el modelo utilizando validación cruzada. La métrica resultante, como **search_best_score_**, representa el promedio de las puntuaciones obtenidas en las diferentes iteraciones de la validación cruzada. Por otro lado, **search_best_params_** proporciona la combinación de hiperparámetros que resultó en el mejor rendimiento promedio.

En resumen, Grid Search busca la mejor combinación de hiperparámetros y validación cruzada garantiza que el rendimiento del modelo sea evaluado de manera justa, lo que resulta

```
# Configurar y Ejecutar GridSearchCV con Validación Cruzada
search = GridSearchCV(modelrfc, params, cv=5, scoring='accuracy', n_jobs=-1)

# Ajustar el modelo
search.fit(X_train_smote, y_train_smote)

# Resultados
print("Best parameter (CV score=%0.3f):" % search.best_score_)
print(search.best_params_)
```

Figura 3-3.: Ejemplo de Grid Search con validación Cruzada

en un modelo más robusto y preciso.

3.1.8. Selección de Características

En esta investigación, se empleó la selección de características para reducir la dimensionalidad del conjunto de datos, seleccionando un subconjunto óptimo de características que mejoraran el rendimiento del modelo. Se utilizó la técnica Sequential Feature Selection (SFS), un método de búsqueda heurística que selecciona características de manera secuencial, ya sea añadiendo o eliminando características. Este proceso evalúa iterativamente un subconjunto de características y su impacto en la precisión del modelo.

Además, se aplicó la variante Forward Selection (Selección hacia adelante), que funciona de la siguiente manera:

- Comienza con un modelo vacío, sin características seleccionadas.
- En cada iteración, se añade una característica que, al combinarse con las seleccionadas previamente, proporciona la mayor mejora en el rendimiento del modelo.
- El proceso continúa hasta alcanzar un criterio de parada, como un número máximo de características seleccionadas o un umbral de mejora en el rendimiento.

A continuación, se presenta un ejemplo de la selección de características hacia adelante:
En relación con la Selección Secuencial de Características (SFS), se utilizaron los siguientes parámetros clave:

- forward=True: Indica que se está aplicando la Selección Hacia Adelante.
- k_features=20: Este parámetro especifica que se desea seleccionar un total de 20 características.
- floating=False: Selección secuencial normal sin flotación.

```
# Defino las características secuencial
sfs = SFS(best_modelrf,
          k_features=20,
          forward=True,
          floating=False,
          scoring='f1',
          cv=5)

# Ajustar y seleccionador las características de los datos
sfs = sfs.fit(X_train_smote, y_train_smote)

# Obtener las características seleccionadas
selected_features = sfs.k_feature_idx_
```

Figura 3-4.: Ejemplo de Selección Secuencial de Características (SFS)

- scoring='f1': Define que se utilizará la la media armónica entre precisión y recall, como métrica de evaluación.
- cv=5: Realiza una validación cruzada con 5 particiones para asegurar que la selección de características sea robusta y generalizable.

La SFS es una técnica eficaz para identificar las características más relevantes en un conjunto de datos, lo que contribuye a reducir la complejidad del modelo y mejorar su rendimiento, Whitney (1971). Aunque este método puede ser costoso en términos de tiempo computacional, resulta muy valioso en escenarios donde la interpretabilidad y la selección óptima de características son prioritarias.

3.1.9. Evaluación del Modelo

La evaluación de un modelo de machine learning implica medir y comprender su desempeño una vez que ha sido entrenado. Es crucial que el modelo no solo memorice los datos de entrenamiento, sino que también sea capaz de generalizar bien a datos nuevos. Para abordar este aspecto, se emplearon diversas métricas de evaluación adecuadas para problemas de clasificación, tales como:

1. **Precisión (Precision)**: Mide la proporción de verdaderos positivos entre los casos que el modelo ha predicho como positivos. En otras palabras, indica cuántos de los casos predichos como consumidores de alcohol realmente lo son.

La fórmula de la Precisión es:

$$\text{Precisión} = \frac{\text{Verdaderos Positivos (TP)}}{\text{Verdaderos Positivos (TP)} + \text{Falsos Positivos (FP)}}$$

2. **Recall (Sensibilidad)**: Evalúa la proporción de verdaderos positivos entre todos los casos que realmente son positivos. Es decir, mide cuántos de los verdaderos consumidores de alcohol fueron correctamente identificados por el modelo.
   La fórmula de la Recall es:

$$\text{Recall} = \frac{\text{Verdaderos Positivos (TP)}}{\text{Verdaderos Positivos (TP)} + \text{Falsos Negativos (FN)}}$$

3. **F1-Score**: Es el promedio armónico de precisión y recall. Esta métrica es especialmente útil en problemas de desbalance de clases, ya que proporciona un equilibrio entre la precisión y el recall, considerando ambos aspectos de manera conjunta.
   La fórmula de la F1-Score es:

$$\text{F1-Score} = 2 \times \frac{\text{Precisión} \times \text{Recall}}{\text{Precisión} + \text{Recall}}$$

La evaluación de un modelo es crucial para entender su comportamiento con datos no vistos y garantizar que esté generalizando adecuadamente. Al utilizar métricas de evaluación apropiadas, se obtiene una visión integral del rendimiento del modelo, lo que permite identificar áreas de mejora y realizar ajustes necesarios para optimizar su capacidad predictiva.

3.1.10. Comparación de Resultados

La comparación de resultados es una etapa crucial en el proceso de evaluación de modelos de machine learning. Permite identificar cuál modelo ofrece el mejor rendimiento y es más adecuado para abordar el problema de clasificación en cuestión. Además, esta comparación facilita la selección del modelo óptimo al evaluar métricas de desempeño y considerar las características prácticas de cada modelo, lo que ayuda a tomar decisiones informadas sobre cuál es el más adecuado para el problema específico.

4. Resultados

A continuación, se presentan los resultados de los experimentos realizados en esta investigación. En primer lugar, se llevó a cabo un análisis exploratorio de los datos para comprender su distribución y características generales. Posteriormente, se entrenaron modelos de clasificación sin aplicar ninguna técnica de balanceo de clases. En segundo lugar, se implementaron diversas técnicas de balanceo de clases, tales como el sobremuestreo (SMOTE y ADASYN), el submuestreo (RUS, Cluster Centroids) y técnicas híbridas (SMOTEENN y Tomek Links), con el objetivo de abordar el desbalance de clases. En tercer lugar, se ajustaron los hiperparámetros y se realizó validación cruzada para optimizar el rendimiento de los modelos. Finalmente, se aplicó la selección de características a los modelos, con el propósito de identificar el modelo más adecuado para clasificar el riesgo de consumo de alcohol en estudiantes universitarios.

4.1. Análisis Exploratorio de Datos

Este proceso se desarrolló a través de un análisis exploratorio de los datos. Inicialmente, se realizó la depuración y limpieza de los registros nulos o vacíos, así como la normalización o estandarización de las características. Además, se llevó a cabo la codificación de las variables categóricas, utilizando la técnica de One Hot Encoding para cada una de las variables. Como se muestra en la Tabla 4-1.

Tabla 4-1.: Preparación de los datos sin aplicar One Hot Encoding

Sexo	Estado_Civil	Edad	Estrato	Consumo
M	Soltero	25	1	Perjudicial
M	Soltero	21	1	No consume
M	Soltero	22	1	No consume
F	Soltera	22	1	No consume
M	Soltero	21	1	Riesgo

La tabla anterior muestra algunas variables del conjunto de datos (ver anexo 2). En este sentido, se presentan los primeros cinco registros sin aplicar la técnica de One Hot Encoding, en la cual cada categoría se convierte en una columna binaria con valores 0 o 1. Además, se

incluyen algunos datos relacionados con el consumo de alcohol, como: sexo (M, F), estado civil (soltero(a), casado(a), divorciado(a), unión libre, viudo(a)), edad, estrato (1 al 6) y la variable de salida (consumo).

Tabla 4-2.: Preparación de los datos con One Hot Encoding

Sexo	Estado_Civil	Edad	Estrato	Consumo
0	1	25	1	1
0	1	21	1	0
0	1	22	1	0
1	1	22	1	0
0	1	21	1	1

Posteriormente, se aplicó la técnica de One Hot Encoding a las variables Sexo, Estado Civil y Consumo. Esto permitió preparar los datos para aplicar modelos de Machine Learning con el objetivo de predecir el consumo de alcohol en estudiantes universitarios. Como se muestra en la Tabla 4-2.
Ahora bien, después de realizar la encuesta a 2954 estudiantes de la Universidad de Córdoba, estos fueron los resultados obtenidos para la variable Sexo biológico:

- Femenino: 1541 estudiantes, lo que representa el 52.17 %.
- Masculino: 1413 estudiantes, lo que representa el 47.83 %.

La diferencia mayoritaria es de 2 %, con una mayor representación el sexo femenino. Como se muestra en la Figura. **4-1**.

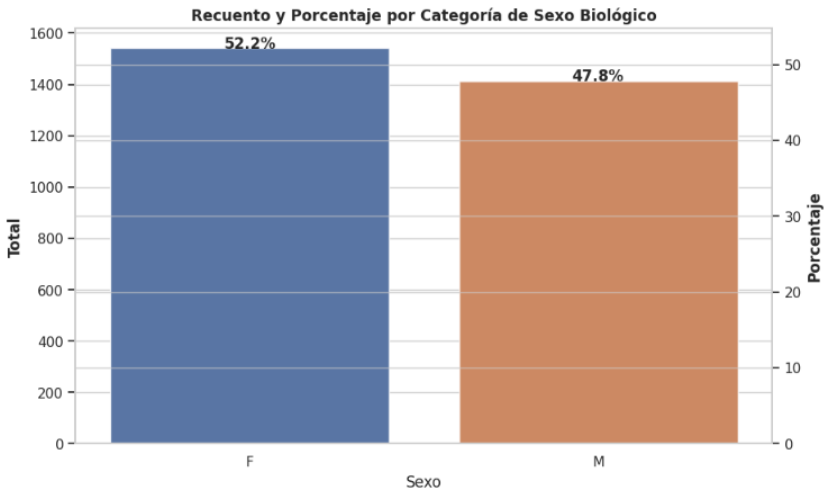


Figura 4-1.: Sexo biológico

Con respecto a los resultados obtenidos en el analisis exploratorio, se observó que el 93.7 %

de las mujeres encuestadas no consume alcohol, mientras que el 6.3 % sí lo hace. En cuanto a los hombres, el 83.5 % no consume alcohol y el 16.5 % sí lo hace. Esto evidencia un desbalance significativo en los datos. Como se ilustra en la Figura **4-2**.

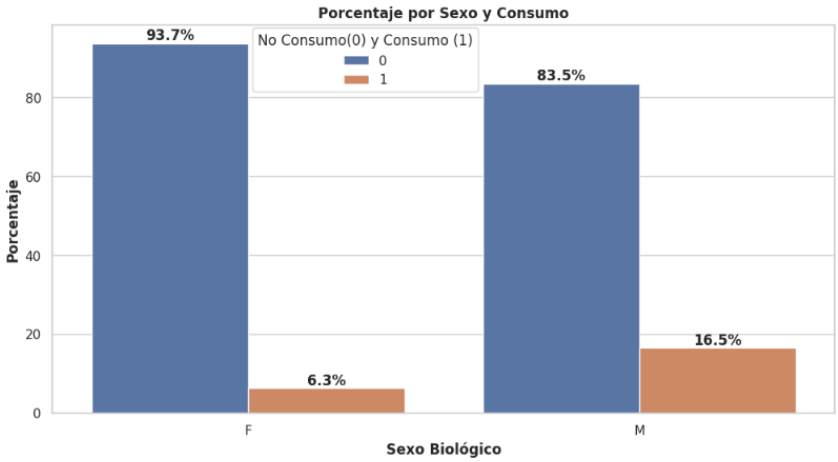


Figura 4-2.: Consumo de alcohol por sexo biológico

En el análisis exploratorio de los datos, se identificó que el 88.8 % de los encuestados no consume alcohol, mientras que el 11.2 % sí lo hace. Este desbalance significativo en los datos implica que una clase tiene muchas más instancias que la otra, lo que podría afectar más adelante la precisión de los modelos de Machine Learning al sesgarse hacia la clase mayoritaria. Como resultado, estos modelos podrían mostrar un rendimiento deficiente en la clasificación de la clase minoritaria. A continuación, se muestra la distribución de los estudiantes que consumen(1) y los que no lo consumen(0) alcohol. Como se muestra en la Figura **4-3**.

Tras realizar el análisis exploratorio de las variables, los resultados indican una relación entre el estrato socioeconómico y el consumo de alcohol, como se observa en la Figura **4-4**.

En ese sentido, los datos sugieren que la mayoría de las personas pertenece al estrato socioeconómico bajo (estrato 1), y dentro de este grupo, hay una tendencia más elevada al consumo de alcohol en comparación con los estratos 2 y 3.

En relación con los resultados obtenidos entre las variables «estado civil» y «consumo», se presenta un desglose detallado en la Figura **4-5**.

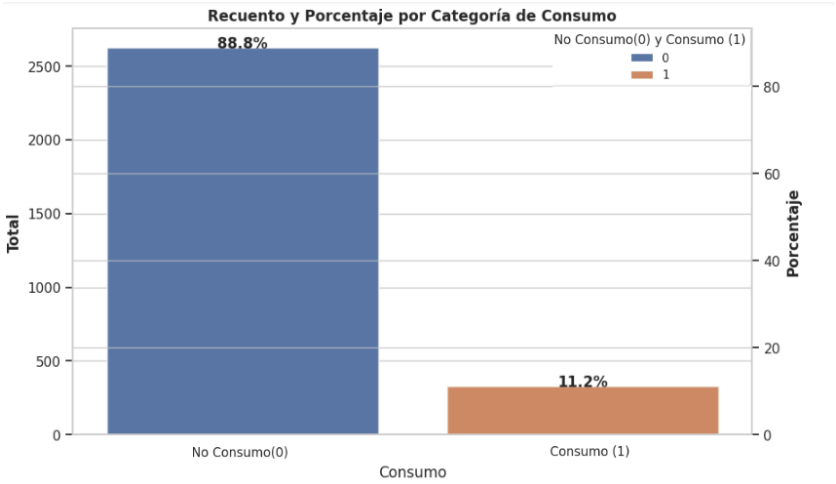


Figura 4-3.: Consumo de alcohol en estudiantes universitarios

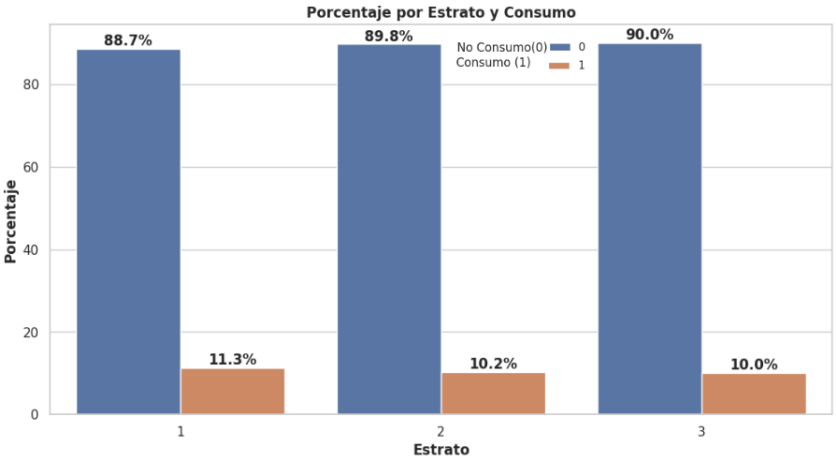


Figura 4-4.: Consumo de alcohol por estrato socioeconómico

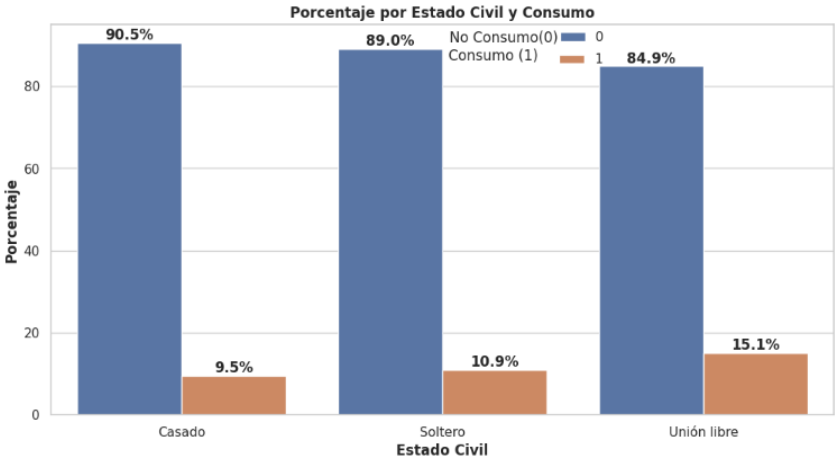


Figura 4-5.: Consumo de alcohol por estado civil

En general, los datos indican que el consumo de alcohol es menor entre las personas casadas, mientras que es más elevado entre aquellos que viven en unión libre, situándose los solteros en un punto intermedio.

A continuación, se presentan los resultados del análisis de los datos en relación con las variables rango de edad y consumo de alcohol. Como se muestra en Figura **4-6**.

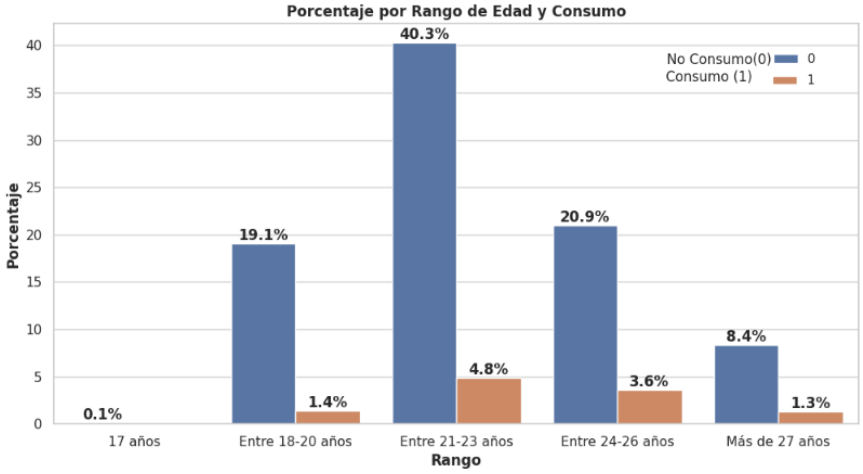


Figura 4-6.: Consumo de alcohol por rango de edad

En general, los resultados indican que el patrón de consumo de alcohol sigue un ciclo caracterizado por un aumento en la juventud temprana, alcanzando un pico en la adultez joven (entre los 21 y 23 años), y un descenso en la adultez media y posterior.

4.1.1. Matriz de correlación

La matriz de correlación muestra las relaciones entre las diferentes variables de un conjunto de datos. Cada celda contiene un coeficiente de correlación, que indica tanto la fuerza como la dirección de la relación lineal entre dos variables. A continuación, se destacan las observaciones más relevantes:

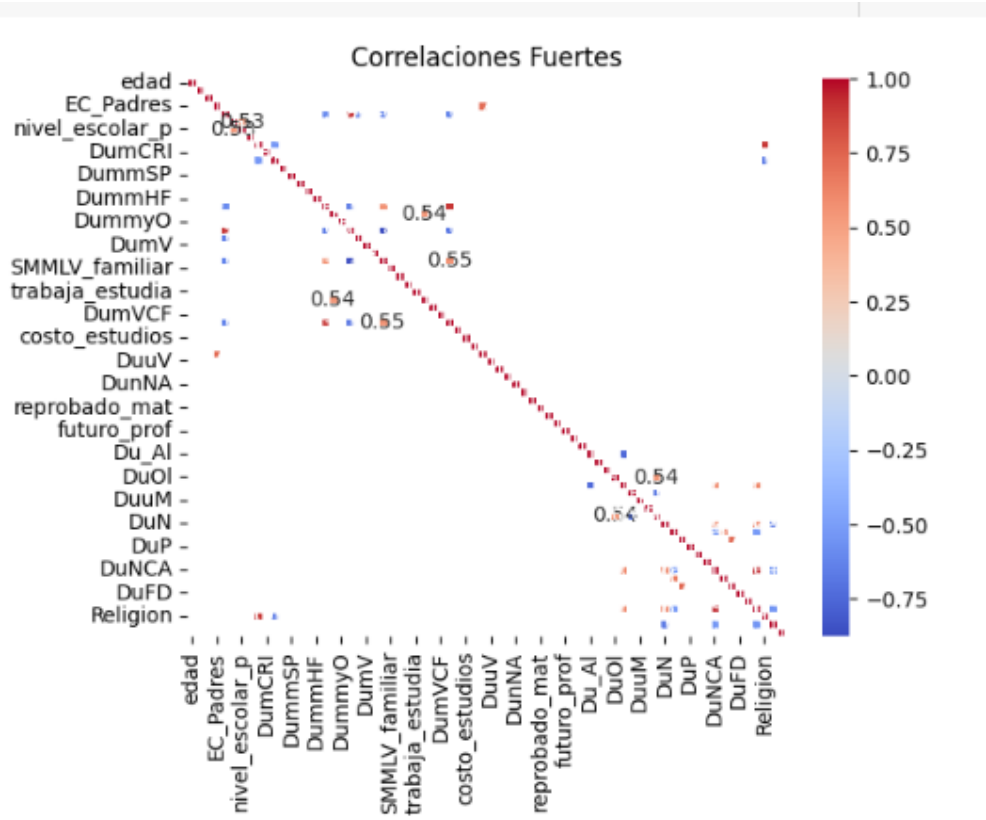


Figura 4-7.: Matriz de correlación fuertes

Observaciones relevantes de la matriz de correlación

1. **Correlaciones positivas fuertes**: Existen varias correlaciones positivas, cercanas a 0.54 o 0.55, entre ciertas variables. Por ejemplo, la correlación entre: **nivel_escolar_p** y **EC_Padres** (0.53)sugiere que el nivel escolar podría estar relacionado con el entorno educativo familiar, lo que podría influir en las características del consumo de alcohol en los estudiantes. Asimismo, variables como **DumHF** y **DumYO** presentan correlaciones similares, lo que indica que podrían representar categorías relacionadas entre sí y compartir factores comunes.

2. **Correlaciones negativas**: No parece haber fuertes correlaciones negativas en las variables visibles en este gráfico, ya que los valores están más concentrados en la zona positiva.

3. **Relaciones débiles o nulas**: La mayoría de las celdas tienen valores cercanos a 0, lo que indica una falta de relación lineal entre muchas variables.

4.2. Correlaciones de variables

Una vez, se realizó el análisis de correlación entre la variable objetivo (Consumo) y otras variables relevantes. La siguiente figura **4-8** presenta las correlaciones positivas y negativas identificadas:

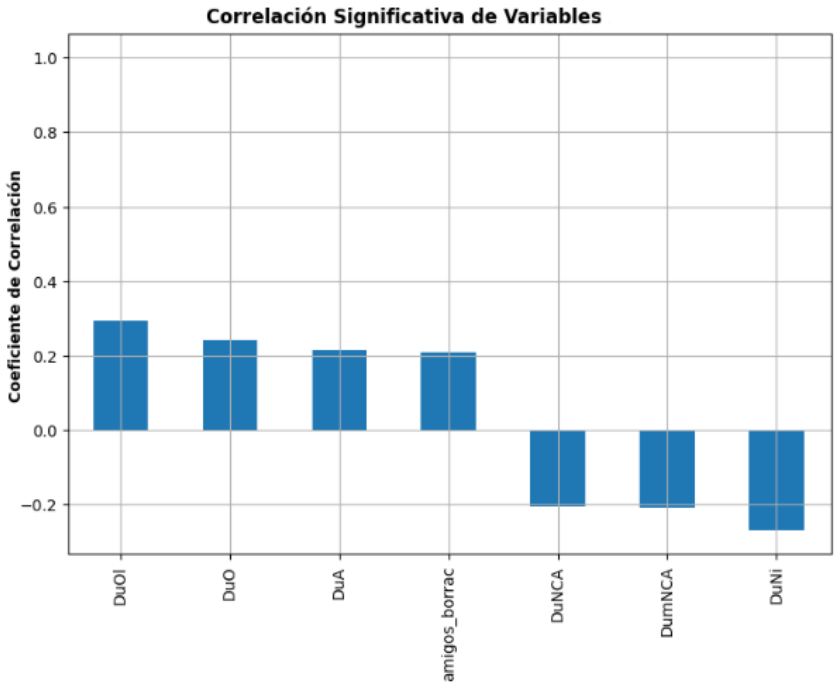


Figura 4-8.: Correlaciones de variables

En relación con los resultados previos, se identificaron varias variables con una correlación positiva significativa:

- **DuOl** (Consumo de alcohol mensual), esta variable está relacionada con un aspecto de la frecuencia de consumo del último mes, y su valor positivo indica que está directamente relacionado con un aumento en el «Consumo». Aunque, la correlación es moderada (0.29), sigue siendo relevante.
- **DuO** (Consumo de sustancias en el último mes) y **DuA** (entorno social), estas variables relacionadas al consumo sustancias y en entorno social. Sin embargo, con una correlación menor que DuOl. Esto indica que un aumento en estos factores se asocia con un aumento en la variable de «Consumo».
- **amigos_borrac** (Influencia de amigos consumidores), la correlación positiva indica que el consumo de alcohol aumenta si los amigos del individuo también tienden a emborracharse, lo cual puede deberse a la influencia social.

No obstante, las variables con correlaciones significativas indican que cuando su valor es alto, el consumo de alcohol tiende a ser bajo:

- **DuNCA** (no consume alcohol en entornos sociales) y **DumNCA** (no consume alcohol en celebraciones o fiestas), estas variables muestran una correlación negativa moderada con el consumo de alcohol, lo que sugiere que podrían representar factores o actividades que inhiben esta conducta. En este sentido, están asociadas con la abstinencia de alcohol en contextos sociales o en situaciones propicias para el consumo, como las celebraciones de fin de año.
- **DuNi** (no consume ninguna sustancia), esta es la variable con la correlación negativa más fuerte con el consumo, lo que indica un vínculo importante. En ese sentido, representa la actividad incompatible con el consumo de sustancias por lo menos alguna vez en el último mes. Este factor puede ser un predictor importante en la reducción del consumo.

En resumen, los resultados de correlación proporcionan una visión sobre las variables que ejercen mayor influencia en el consumo de alcohol. La influencia de correlaciones positivas, como en el caso de DuOl y amigos_borrac, destaca el papel del consumo de alcohol mensual y la influencia de amigos consumidores. Por otro lado, factores con correlación negativa, como DuNi, resaltan la relevancia de no consumir ningunas sustancias, al menos una vez en el último mes, puede ser un factor relevante en la disminución del consumo.

4.3. Descripción de resultados de los modelos de clasificación

En este contexto, los modelos de clasificación se emplearon sin utilizar técnicas para abordar el desbalanceo de clases, lo que llevó a que los algoritmos se inclinaran hacia la clase mayoritaria. Este sesgo afecta negativamente la capacidad del modelo para predecir de manera correcta la clase minoritaria. En particular, los modelos como RandomForest, LogisticRegression, Perceptron multicapa y SVM se entrenaron sin corregir el desbalance de los datos, lo que resultó en un menor rendimiento en la predicción de los estudiantes que consumen alcohol. Aunque las métricas de precisión pueden ser altas para la clase mayoritaria, métricas importantes como el Recall y el F1-score para la clase minoritaria se ven afectadas negativamente.

Los resultados obtenidos evidencian cómo diferentes modelos de clasificación se comportan al ser entrenados y evaluados en un conjunto de datos desbalanceado, donde una clase es mayoritaria y otra es minoritaria, sin aplicar técnicas de manejo de desbalance. A continuación, se presenta un desglose de cada modelo y las métricas de evaluación correspondientes(ver

Tabla 4-3):

Tabla 4-3.: Modelos de clasificación sin utilizar técnica de balanceo

Modelos	Precision	Recall	F1-score
LogisticRegression	0.55	0.17	0.26
Perceptron multicapa	0.53	0.18	0.27
SVM	0.00	0.00	0.00
RandomForest	0.56	0.05	0.09

Estos resultados reflejan que ninguno de los modelos está desempeñándose adecuadamente debido al desbalance de clases. En particular:

- Logistic Regression, Perceptrón Multicapa, y Random Forest muestran una precisión moderada, pero con un recall muy bajo, lo que sugiere que estos modelos están fallando en identificar correctamente los ejemplos de la clase minoritaria (consumidores de alcohol). Además, su desempeño en la clase de interés es deficiente.
- El SVM (Support Vector Machine) ha colapsado por completo, ya que no logra predecir ningún ejemplo de la clase positiva (consumidores de alcohol), lo que indica una grave incapacidad para abordar el desbalance de clases.

En resumen, los resultados resaltan la necesidad de aplicar técnicas para manejar el desbalance de clases con el fin de mejorar la capacidad de los modelos para detectar correctamente a la clase minoritaria.

Por otro lado, los resultados obtenidos al aplicar la técnica de sobremuestreo SMOTE, muestran una mejora significativa en la capacidad de los modelos para detectar la clase minoritaria, especialmente en los valores de recall. Como se muestra en la **Tabla 4-4**.

Tabla 4-4.: Modelos de clasificación con técnica de balanceo SMOTE(Synthetic Minority Over-sampling Technique)

SMOTE

Modelos	Precision	Recall	F1-score
LogisticRegression	0.47	0.22	0.30
Perceptron multicapa	0.36	0.40	0.38
SVM	0.41	0.64	0.50
RandomForest	0.49	0.25	0.33

En los apartados anteriores, se han presentado los resultados de los diferentes modelos, acompañados de sus respectivas métricas de evaluación. En particular, se ha descrito el comportamiento de cada modelo en relación con el tratamiento del desbalance de clases,

evaluando su rendimiento mediante las métricas de precisión, recall y F1-score. Estas métricas permiten una comprensión integral de la efectividad de cada modelo al enfrentar el problema del desbalance de clases, proporcionando una visión comparativa del compromiso entre la minimización de los falsos positivos y falsos negativos. A continuación, se realiza un análisis exhaustivo de los resultados obtenidos, destacando las fortalezas y limitaciones de cada enfoque.

1. **Logistic Regression**
   La precisión de 0.47 indica que, de todas las predicciones positivas realizadas por el modelo, solo el 47 % fueron correctas. Aunque es un valor moderado, está lejos de ser ideal. El recall de 0.22 revela que el modelo solo identificó correctamente el 22 % de las instancias positivas reales, lo que sigue siendo bajo, sugiriendo que el modelo está fallando en detectar muchos casos de la clase minoritaria, a pesar de la aplicación de SMOTE. Finalmente, el F1-score de 0.30 refleja la dificultad del modelo para equilibrar precisión y recall. Aunque SMOTE mejoró ligeramente el rendimiento, el modelo aún tiene dificultades para identificar correctamente la clase minoritaria.

2. **Perceptrón Multicapa**
   La precisión de 0.36 es relativamente baja, lo que indica que solo el 36 % de las predicciones positivas del modelo son correctas, lo que refleja que el modelo comete muchos errores al predecir la clase minoritaria. El recall de 0.40 es superior al de la regresión logística, lo que significa que el Perceptrón Multicapa es capaz de detectar el 40 % de los casos positivos reales. Esto sugiere que SMOTE ayudó a mejorar la capacidad del modelo para identificar más instancias de la clase minoritaria. El F1-score de 0.38 muestra un mejor equilibrio entre precisión y recall en comparación con la regresión logística, aunque sigue siendo un rendimiento no óptimo. Esto implica que el modelo aún no alcanza un nivel satisfactorio de predicción en la clase minoritaria.

3. **SVM (Support Vector Machine)**
   La precisión de 0.41 indica que el 41 % de las predicciones positivas realizadas por el modelo fueron correctas, lo que es una mejora moderada. El recall de 0.64 es el más alto entre los modelos probados, lo que significa que el SVM detecta el 64 % de los casos positivos reales. Esto sugiere que la aplicación de SMOTE ha sido efectiva para mejorar la capacidad del modelo en la identificación de la clase minoritaria. El F1-score de 0.50 refleja un mejor equilibrio entre precisión y recall en comparación con otros modelos evaluados. Este valor muestra que el modelo SVM, después de aplicar SMOTE, ha logrado un rendimiento adecuado en la detección de los consumidores de alcohol, aunque aún puede mejorar en precisión.

4. **Random Forest**
   La precisión de 0.49 indica que el 49 % de las predicciones positivas realizadas por el modelo fueron correctas, lo que representa un valor moderado. El recall de 0.25

muestra que el modelo solo fue capaz de identificar el 25 % de las instancias positivas reales, lo que sigue siendo un valor bajo. Aunque la aplicación de SMOTE ha mejorado ligeramente la capacidad del modelo para detectar casos positivos, este sigue sin captar la mayoría de los ejemplos de la clase minoritaria. El F1-score de 0.33 refleja que el modelo aún enfrenta dificultades para equilibrar precisión y recall. Aunque SMOTE ha contribuido a una mejora en el rendimiento, el modelo aún necesita ajustes adicionales para mejorar su capacidad de predicción, especialmente en la detección de la clase minoritaria.

En relación con los resultados obtenidos, se ha observado una mejora significativa en el recall tras la aplicación de SMOTE. El SVM ha sido el modelo con el mejor desempeño en esta métrica, alcanzando un recall de 0.64, lo que indica que es más eficaz para detectar la clase minoritaria. A pesar de esta mejora en la detección de la clase minoritaria, algunos modelos han presentado una reducción en la precisión, lo que puede explicarse por la generación de ejemplos sintéticos a través de SMOTE, lo que a su vez puede aumentar los falsos positivos.

El SVM también ha obtenido el mejor F1-score (0.50), reflejando que ha logrado el equilibrio más adecuado entre precisión y recall. Otros modelos, como el Random Forest y el Perceptrón Multicapa, han mostrado un rendimiento inferior en comparación, lo que sugiere que, aunque SMOTE ha mejorado el desempeño general, sigue habiendo margen para ajustes y optimización en estos algoritmos.

Ahora bien, tras la aplicación de la técnica de sobremuestreo ADASYN a los modelos, se observaron resultados significativamente diferentes en comparación con el uso de SMOTE. ADASYN, al generar ejemplos sintéticos enfocados en las regiones donde la clasificación de la clase minoritaria presenta mayores dificultades, mostró un impacto variable en el rendimiento de los modelos. Esta variabilidad se refleja en las métricas de evaluación, indicando que ADASYN puede ser más eficaz en algunos casos al abordar la complejidad intrínseca de las muestras minoritarias. A continuación, se analizará cómo esta técnica afectó el comportamiento de cada modelo en términos de precisión, recall y F1-score, proporcionando una comparación detallada con los resultados obtenidos mediante SMOTE. Como se ilustra en la **Tabla 4-5**.

Tabla 4-5.: Modelos de clasificación con técnica de desbalanceo ADASYD

ADASYD

Modelos	Precision	Recall	F1-score
LogisticRegression	0.45	0.22	0.30
Perceptron multicapa	0.47	0.15	0.23
SVM	0.33	0.55	0.42
RandomForest	0.47	0.34	0.39

En relación a los resultados obtenidos de los modelos aplicados, se pueden observar los siguientes comportamientos:

1. **Logistic Regression**
   En este caso, ADASYN no mostró una mejora significativa, con una precisión de 0.45 indica que solo el 45 % de las instancias positivas son reales. Aunque es un valor moderado y un recall de 0.22 muestra que el modelo solo identificó correctamente el 22 % de las instancias positivas reales. Aun así, sigue bajo, teniendo problemas para ofrecer predicciones precisas, lo que resultó en un F1-score de 0.30 relativamente bajo. Indicando que este modelo aún enfrenta dificultades para manejar el desbalance de clases incluso con sobremuestreo.
2. **Perceptrón Multicapa**
   Este modelo mostró un recall de 0.15 bajo, lo que implica una mayor capacidad para detectar la clase minoritaria, pero la precisión de 0.47 se vio afectada, incrementando el número de falsos positivos. De hecho el F1-score de 0.30 bajo. A pesar de la mejora, su desempeño global no superó a otros modelos en términos de balance entre precisión y recall.
3. **SVM (Support Vector Machine)**
   El SVM fue el modelo que mejor aprovechó las técnicas de sobremuestreo, con un recall de 0.55 superior a otros modelos, especialmente tras la aplicación de SMOTE y ADASYN. Esto implica que fue más eficaz en la detección de la clase minoritaria. Además, su precisión de 0.33 y F1-score de 0.49, lo que refleja un buen equilibrio entre precisión y recall.
4. **Random Forest**
   Aunque Random Forest mejoró en algunas métricas, no logró superar al modelo SVM en términos de recall de 0.34. Sin embargo, su precisión de 0.47 fue más estable en comparación con otros modelos, lo que indica que logró un mejor control sobre los falsos positivos. Aun así, su F1-score de 0.39 no fue tan alto como el del SVM, lo que indica que aún tiene espacio para mejorar en la detección de la clase minoritaria.

En general, el SVM fue el modelo que mejor se comportó tras la aplicación de técnicas de sobremuestreo, mostrando un equilibrio adecuado entre las métricas de evaluación. Aunque otros modelos como el Perceptrón Multicapa y Random Forest también mejoraron, no alcanzaron el mismo nivel de desempeño en términos de balance entre precisión y recall.

Por otro lado, tras la aplicación de la técnica de submuestreo aleatorio (RUS, por sus siglas en inglés) a los modelos, se obtuvieron resultados que evidencian el impacto de la reducción de la clase mayoritaria en el rendimiento de los mismos. El submuestreo aleatorio, al eliminar instancias de la clase mayoritaria, introduce un cambio notable en el balance del conjunto de datos, lo que se reflejó en un ajuste de las métricas de precisión, recall y F1-score. Aunque esta

técnica es efectiva para equilibrar las clases, su implementación también conlleva la pérdida de información, lo que puede afectar el desempeño general del modelo, especialmente en términos de precisión y capacidad predictiva. A continuación en la **Tabla 4-6**, se presentan los resultados detallados del impacto del submuestreo en cada modelo.

Tabla 4-6.: Modelos de clasificación con técnica de desbalanceo RandomUnderSampler(Submuestreo Aleatorio)

RUS			
Modelos	**Precision**	**Recall**	**F1-score**
LogisticRegression	0.23	0.66	0.34
Perceptron multicapa	0.24	0.70	0.36
SVM	0.25	0.79	0.38
RandomForest	0.26	0.82	0.39

En relación a los resultados obtenidos al aplicar RUS muestran un cambio notable en el desempeño de los modelos:

1. **Logistic Regression**
   Aunque el recall de este modelo mejoró significativamente, alcanzando un valor de 0.66, su precisión sigue siendo baja con un 0.23. Esto indica que el modelo está detectando más instancias de la clase minoritaria, pero con un alto número de falsos positivos, lo que afecta negativamente su F1-score (0.34).

2. **Perceptrón Multicapa**
   El Perceptrón Multicapa también vio mejoras en el recall (0.70), aunque su precisión sigue siendo baja (0.24). El F1-score es mejor que el de la regresión logística, pero sigue siendo modesto con un 0.36, mostrando que el modelo está equilibrando de manera limitada en la precisión y el recall.

3. **SVM (Support Vector Machine)**
   El SVM mostró una mejora considerable en el recall (0.79), siendo uno de los modelos más efectivos en detectar instancias de la clase minoritaria. Sin embargo, su precisión sigue siendo baja (0.25), lo que afecta ligeramente su F1-score (0.38).

4. **Random Forest**
   Este modelo obtuvo los mejores resultados tras aplicar RUS, con un recall de 0.82 y la precisión más alta entre los modelos probados (0.26). Su F1-score de 0.39 refleja que, si bien sigue habiendo un número considerable de falsos positivos, el modelo está logrando un mejor equilibrio entre precisión y recall que los otros.

En resumen, el Random Forest mostró el mejor desempeño con RUS en términos de recall y F1-score, lo que sugiere que este modelo es más eficaz para detectar la clase minoritaria después del submuestreo, aunque la precisión sigue siendo baja para todos los modelos.

Una vez obtenidos los resultados tras la aplicación de la técnica de submuestreo Cluster Centroid, los modelos mostraron comportamientos diferenciados. Esta técnica, que reduce la clase mayoritaria mediante la generación de centroides que representan grupos de muestras, busca mejorar el balance de clases manteniendo la información representativa de la distribución original. Como se muestra en la **Tabla 4-7**:

Tabla 4-7.: Modelos de clasificación con técnica de desbalanceo Cluster Centroids

Cluster Centroids

Modelos	Precision	Recall	F1-score
LogisticRegression	0.24	0.76	0.37
Perceptron multicapa	0.25	0.85	0.39
SVM	0.20	0.89	0.32
RandomForest	0.28	0.84	0.41

Los resultados anteriores reflejan cómo los modelos se adaptaron a esta reducción estructurada, con un impacto notable en las métricas de precisión, recall y F1-score. El uso de centroides permitió preservar cierta información clave, a diferencia del submuestreo aleatorio, aunque las limitaciones en la representación de la clase mayoritaria podrían haber influido en la capacidad predictiva de los modelos. A continuación, se detallan los resultados específicos para cada modelo tras la aplicación de Cluster Centroid.

1. **Logistic Regression**
   Con una precisión de 0.24, el modelo sigue mostrando un alto número de falsos positivos, aunque su recall aumentó notablemente hasta 0.76. El F1-score de 0.37 refleja un balance moderado entre las métricas.

2. **Perceptrón Multicapa**
   Este modelo mejoró su recall hasta 0.85, lo que indica que está identificando la mayoría de los ejemplos de la clase minoritaria, aunque con una precisión de 0.25. El F1-score de 0.39 muestra que el modelo equilibra las métricas ligeramente mejor que la regresión logística.

3. **SVM (Support Vector Machine)**
   El Support Vector Machine tuvo el recall más alto (0.89), detectando la mayoría de los casos positivos, pero su precisión es la más baja (0.20), lo que resulta en un F1-score de 0.32, el más bajo entre los modelos. Esto sugiere que, aunque detecta más instancias de la clase minoritaria, también produce una cantidad significativa de falsos positivos.

4. **Random Forest**
   El Random Forest muestra el mejor equilibrio con una precisión de 0.28, un recall de 0.84, y el mejor F1-score de 0.41. Este modelo mantiene un mejor balance entre

identificar correctamente la clase minoritaria y minimizar los falsos positivos.

En conclusión, Cluster Centroid mejoró considerablemente el recall en todos los modelos, con el SVM obteniendo el mayor valor en esta métrica. Sin embargo, el Random Forest nuevamente se destaca por lograr el mejor equilibrio general entre precisión y recall, reflejado en su superior F1-score.

En relación a los resultados obtenidos tras la aplicación de la técnica híbrida SMOTEENN a los modelos, se observó un comportamiento notablemente distinto en comparación con otras técnicas de balanceo de clases. SMOTEENN, al combinar el sobremuestreo sintético de SMOTE con el submuestreo basado en la eliminación de vecinos por la técnica de edición ENN (Edited Nearest Neighbors), permitió un enfoque más refinado tanto en la generación de nuevas instancias de la clase minoritaria como en la eliminación de instancias ruidosas o redundantes de la clase mayoritaria. Como se muestra en la **Tabla 4-8**.

Tabla 4-8.: Modelos de clasificación con técnica de desbalanceo SMOTENN

SMOTENN

Modelos	Precision	Recall	F1-score
LogisticRegression	0.26	0.83	0.40
Perceptron multicapa	0.31	0.84	0.45
SVM	0.21	0.96	0.35
RandomForest	0.33	0.74	0.46

Los modelos mostraron un rendimiento variable en términos de precisión, recall y F1-score, reflejando cómo esta técnica híbrida puede mejorar la capacidad predictiva en escenarios donde el desbalance de clases y el ruido de los datos son problemáticos. A continuación, se presentan los resultados detallados para cada modelo.

1. **Logistic Regression**
   Con una precisión de 0.26 indica que el 26 % de las instancias positivas son reales, lo que sigue siendo un valor bajo. El recall de 0.83 indica que el 83 % de las predicciones positivas realizadas por el modelo fueron correctas. La regresión logística tiene un buen desempeño en detectar la clase minoritaria, pero con algunos falsos positivos. Su F1-score de 0.40 muestra un equilibrio moderado entre ambas métricas.

2. **Perceptrón Multicapa**
   Este modelo presenta una precisión mejorada de 0.31 y un recall de 0.84, lo que significa que detecta más casos positivos con una menor tasa de falsos positivos. Su F1-score de 0.45 es el segundo mejor, mostrando un buen balance entre las métricas.

3. **SVM (Support Vector Machine)**
   Aunque el recall es el más alto (0.96), indicando que el SVM está capturando casi todos los ejemplos de la clase minoritaria, su precisión es baja (0.21), lo que resulta en

un alto número de falsos positivos. El F1-score de 0.35 refleja este desequilibrio entre precisión y recall.

4. **Random Forest**
   Este modelo muestra el mejor desempeño en general, con una precisión de 0.33, un recall de 0.74, y el F1-score más alto (0.46). Aunque su recall es ligeramente inferior en comparación con otros modelos, logra el mejor equilibrio entre la capacidad para detectar la clase minoritaria y evitar falsos positivos.

En resumen, SMOTEENN ayudó a mejorar el rendimiento general de los modelos, con Random Forest destacándose por su mejor F1-score, lo que indica que es el modelo más equilibrado. SVM, aunque tiene un recall excelente, sufre de baja precisión, lo que lo hace menos confiable en términos de clasificación general.

En relción a la aplicación de la técnica de Tomek Links, los resultados reflejaron una tendencia clara en los modelos, caracterizada por una mejora en la precisión general, pero un rendimiento deficiente en la clasificación de la clase minoritaria. Esta técnica, que elimina pares de ejemplos cercanos de diferentes clases, tiene como objetivo limpiar el límite de decisión entre las clases, lo que puede favorecer la precisión en la clase mayoritaria a expensas de la clase minoritaria. Este comportamiento genera un desequilibrio en las métricas, afectando significativamente el balance entre precisión, recall y F1-score. Si bien se observó un incremento en la precisión global, la capacidad de los modelos para identificar correctamente instancias de la clase minoritaria se redujo, lo que indica una pérdida de sensibilidad hacia estas muestras. Como se describe en la **Tabla 4-9**.

Tabla 4-9.: Modelos de clasificación con técnica de desbalanceo Tomek Links

Tomek Links

Modelos	Precision	Recall	F1-score
LogisticRegression	0.50	0.20	0.29
Perceptron multicapa	0.62	0.20	0.30
SVM	0.00	0.00	0.00
RandomForest	0.47	0.09	0.15

En relación a, los resultados obtenidos, tras aplicar Tomek Links muestran una mejora general de la precision. Aunque, el recall y F1-Score su rendimiento es bajo.

1. **Logistic Regression**
   Con una precisión de 0.50 indica que el 50 % de las predicciones positivas fueron correctas. Sin embargo, su recall es bastante bajo (0.20), lo que significa que solo detecta el 20 % de los ejemplos de la clase minoritaria. Finalmente, el F1-score de 0.29 refleja que el modelo no está logrando un buen equilibrio entre precisión y recall, especialmente

en la clase minoritaria.

2. **Perceptrón Multicapa**
   El modelo tiene la precisión más alta (0.62), lo que indica que está haciendo mejores predicciones positivas en comparación con los demás modelos. Sin embargo, al igual que la regresión logística, su recall es bajo (0.20), lo que limita su capacidad para detectar correctamente ejemplos de la clase minoritaria. Su F1-score de 0.30 muestra que, aunque tiene una buena precisión, su capacidad para encontrar un equilibrio entre precisión y recall es limitada.

3. **SVM (Support Vector Machine)**
   Este modelo ha colapsado completamente con Tomek Links, mostrando una precisión, recall y F1-score de 0.00 en todos los casos. Esto sugiere que el SVM no ha logrado detectar ningún caso positivo en la clase minoritaria, posiblemente debido a la eliminación excesiva de ejemplos clave por parte de Tomek Links.

4. **Random Forest**
   Este modelo muestra la precisión de 0.47, lo que sugiere que el 47 % de las predicciones positivas fueron correctas. No obstante, su recall es extremadamente bajo (0.09), lo que indica que está identificando muy pocos ejemplos de la clase minoritaria. El F1-score de 0.15 es el más bajo, lo que refleja un desequilibrio muy fuerte entre precisión y recall.

La aplicación de Tomek Links ha mejorado la precisión general en algunos casos, pero ha empeorado gravemente la capacidad de los modelos para detectar la clase minoritaria, especialmente en el caso de los modelos SVM y Random Forest. Si bien Perceptrón Multicapa ha mostrado la mejor precisión, su bajo recall sigue siendo un problema importante, lo que se evidencia es que Tomek Links puede haber eliminado demasiados ejemplos minoritarios críticos, comprometiendo el rendimiento de los modelos.

Después de aplicar técnicas avanzadas como el ajuste de hiperparámetros mediante Grid Search, validación cruzada, y selección de características utilizando Sequential Feature Selection (SFS) a los modelos de clasificación, se obtuvieron resultados que permitieron una evaluación exhaustiva de su rendimiento. Estas técnicas fueron implementadas con el objetivo de optimizar el desempeño predictivo de los modelos y mejorar la precisión en la identificación de patrones relevantes para predecir el consumo de alcohol en estudiantes universitarios. El ajuste de hiperparámetros, mediante la búsqueda sistemática de combinaciones óptimas, junto con la validación cruzada para asegurar la generalización, y la selección secuencial de características para reducir la dimensionalidad y eliminar variables irrelevantes, contribuyeron a afinar la capacidad de los modelos. Se presentan en la **Tabla** 4-10.

Tabla 4-10.: Modelos de clasificación con ajuste hiperparárametro, validación cruzada y selección de características

Modelos	Precision	Recall	F1-score
LogisticRegression	0.26	0.78	0.39
Perceptron multicapa	0.25	0.79	0.37
SVM	0.23	0.86	0.36
RandomForest	0.32	0.79	0.45

Los resultados obtenidos reflejan una mejora significativa en las métricas de evaluación, destacando el modelo con el mejor rendimiento en términos de precisión, recall y F1-score. A continuación, se describen en detalle los resultados de cada modelo, destacando cuál de ellos se posiciona como el más adecuado para la predicción del consumo de alcohol en la población universitaria.

1. **Logistic Regression**
   Con la precisión de 0.26, lo que significa que el 26 % de las predicciones positivas fueron correctas. Sin embargo, su recall de 0.78 es relativamente alto, lo que indica que el modelo detectó el 78 % de los ejemplos de la clase minoritaria. El F1-score de 0.39 refleja un mejor equilibrio entre precisión y recall en comparación a los resultados anteriores de los modelos sin aplicación de ajuste de hiperparámetros, lo que sugiere una mejora moderada.

2. **Perceptrón Multicapa**
   Este modelo presenta una precisión similar a la de la regresión logística (0.25), lo que indica un rendimiento comparable en términos de predicción de la clase positiva. Su recall de 0.79 sugiere que detecta un 79 % de los ejemplos de la clase minoritaria, lo que es positivo. Finalmente, el F1-score de 0.37 indica que el modelo está obteniendo un rendimiento aceptable, pero no óptimo.

3. **SVM (Support Vector Machine)**
   El SVM presenta la precisión más baja (0.23), lo que implica que el 23 % de las predicciones positivas fueron correctas. Sin embargo, su recall de 0.86 es el más alto de todos los modelos, lo que significa que este modelo detectó la mayor parte de los ejemplos de la clase minoritaria. El F1-score de 0.36 refleja que, aunque tiene el mejor recall, el desequilibrio con la precisión afecta el rendimiento general del modelo.

4. **Random Forest**
   El modelo con mejor rendimiento en términos de precisión (0.32), lo que significa que predice con mayor exactitud los ejemplos positivos. No obstante, su recall de 0.79 es comparable al del Perceptrón Multicapa y la regresión logística, indicando una buena capacidad para detectar la clase minoritaria. El F1-score de 0.45 es el más alto de todos los modelos, lo que indica que Random Forest logró el mejor equilibrio entre precisión

y recall, siendo el modelo con el mejor rendimiento general.

En general, los modelos de clasificación aplicados permiten identificar factores asociados al consumo de alcohol, facilitando la detección de estudiantes en riesgo de desarrollar conductas problemáticas relacionadas con este hábito. Este enfoque representa un avance significativo, ya que proporciona una base sólida para implementar estrategias de prevención y atención temprana, contribuyendo a mitigar las posibles consecuencias negativas en la salud física, mental y académica de los estudiantes.

En conclusión, los resultados obtenidos no solo tienen un impacto directo en la detección y prevención del consumo de alcohol, sino que también contribuyen al bienestar integral de los estudiantes, fortalecen la gestión institucional y generan nuevas oportunidades para el avance en este campo.

4.3.1. Matriz de confusión de los modelos de clasificación

Logistic Regression

La matriz de confusión generada por el modelo de Regresión Logística aplicado al problema de clasificación de estudiantes universitarios consumidores de alcohol proporciona una representación detallada del desempeño del modelo al clasificar correctamente o incorrectamente a los estudiantes en las categorías de «consume» y «no consume».

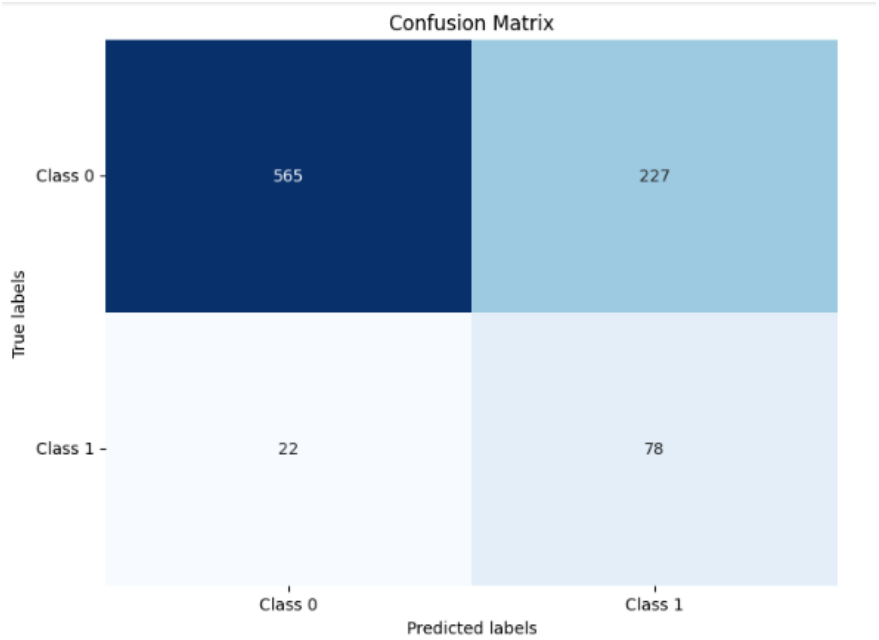


Figura 4-9.: Matriz de Confusión del Modelo Logistic Regression

- **565 (True Negatives - TN)**: El modelo predijo correctamente que 565 estudiantes

no consumen alcohol.

- **227 (False Positives - FP)**: Hubo 227 casos en los que el modelo predijo erróneamente que los estudiantes consumen alcohol, aunque en realidad no lo era.
- **22 (False Negatives - FN)**: El modelo no identificó a 22 estudiantes que realmente consumen alcohol, clasificándolos incorrectamente como no consumidores.
- **78 (True Positives - TP)**: El modelo predijo correctamente que 78 estudiantes consumen alcohol.

El modelo tiene una buena capacidad para identificar a los estudiantes que consumen alcohol (recall alto); sin embargo, su baja precisión indica que comete un número significativo de errores al clasificar como consumidores a estudiantes que en realidad no lo son.

Perceptrón Multicapa

La matriz de confusión generada por el modelo Perceptrón Multicapa ofrece una visión detallada de su desempeño, al mostrar tanto las clasificaciones correctas como los errores al identificar a los estudiantes en cada categoría.

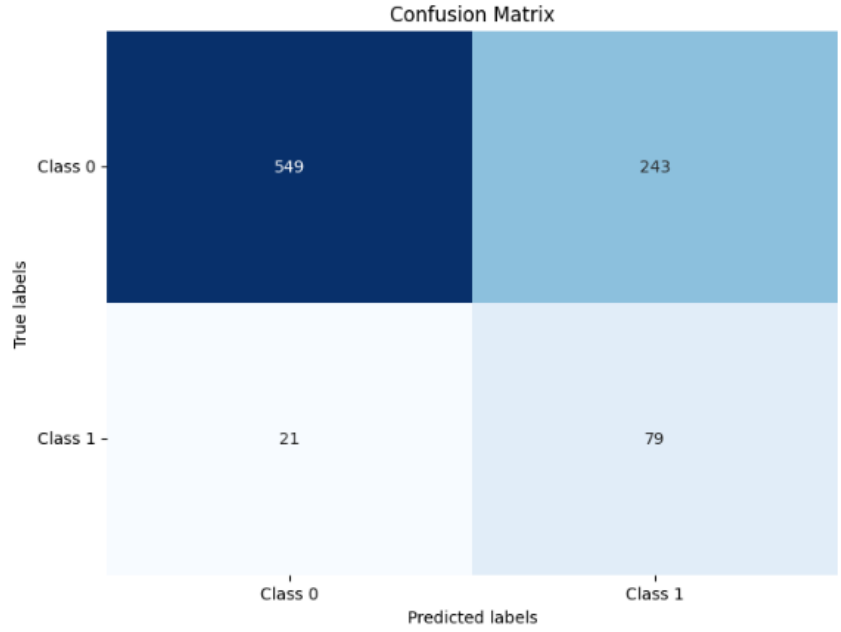


Figura 4-10.: Matriz de Confusión del Modelo Perceptrón Multicapa

- **549 (True Negatives - TN)**: El modelo predijo correctamente que 549 estudiantes no son consumidores de alcohol.
- **243 (False Positives - FP)**: En 243 casos, el modelo clasificó erróneamente a estu-

diantes no consumidores como consumidores.

- **21 (False Negatives - FN)**: El modelo no detectó correctamente a 21 estudiantes que realmente eran consumidores y los clasificó como no consumidores.
- **79 (True Positives - TP)**: El modelo identificó correctamente a 79 estudiantes como consumidores de alcohol.

El modelo Perceptrón Multicapa muestra una buena capacidad para identificar a los consumidores de alcohol (recall alto), lo que resulta especialmente útil en contextos donde es prioritario minimizar los falsos negativos y no pasar por alto casos de consumo. Sin embargo, su baja precisión indica una alta proporción de falsos positivos, clasificando incorrectamente a estudiantes no consumidores como consumidores.

SVM (Support Vector Machine)

La matriz de confusión obtenida con el modelo SVM (Support Vector Machine) es la siguiente:

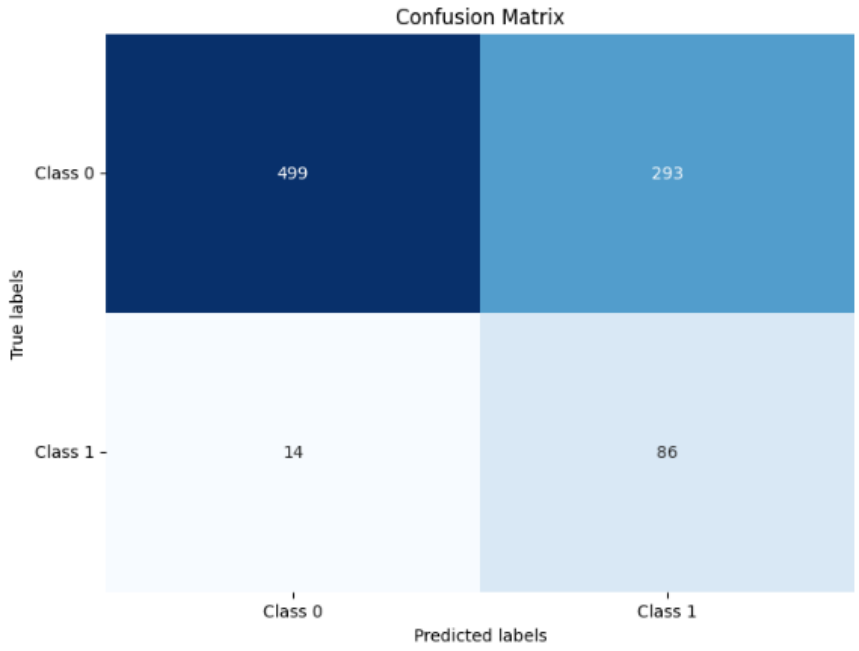


Figura 4-11.: Matriz de Confusión del Modelo SVM (Support Vector Machine)

- **499 (True Negatives - TN)**: El modelo predijo correctamente a 499 estudiantes como no consumidores de alcohol, lo que coincide con sus valores reales.
- **293 (False Positives - FP)**: En 293 casos, el modelo clasificó erróneamente a estudiantes no consumidores como consumidores.

- **14 (False Negatives - FN)**: El modelo no detectó correctamente a 14 estudiantes que eran consumidores, clasificándolos como no consumidores.
- **86 (True Positives - TP)**: El modelo predijo correctamente que 86 estudiantes consumen alcohol.

El modelo SVM presenta un alto recall (86 %), lo que indica su efectividad para identificar a los consumidores de alcohol. Sin embargo, su baja precisión (23 %) señala un elevado número de falsos positivos, clasificando erróneamente a muchos estudiantes no consumidores como consumidores. Este aspecto puede ser problemático si las predicciones incorrectas tienen un alto costo o consecuencias negativas.

Random Forest

La matriz de confusión obtenida con el modelo Random Forest, identificado como el mejor modelo tras el proceso de ajuste y validación, ofrece una visión detallada de su rendimiento en la clasificación del consumo de alcohol entre estudiantes universitarios. Como afirman Chawla y cols. (2002), el desempeño de los algoritmos de aprendizaje automático se evalúa comúnmente a través de la matriz de confusión, la cual resume los resultados de clasificación de manera efectiva, tal como se ilustra en la Figura **4-12** (para un problema de dos clases). En esta matriz, las columnas representan las clases predichas, mientras que las filas corresponden a las clases reales, permitiendo una comparación clara entre las predicciones del modelo y las observaciones verdaderas.

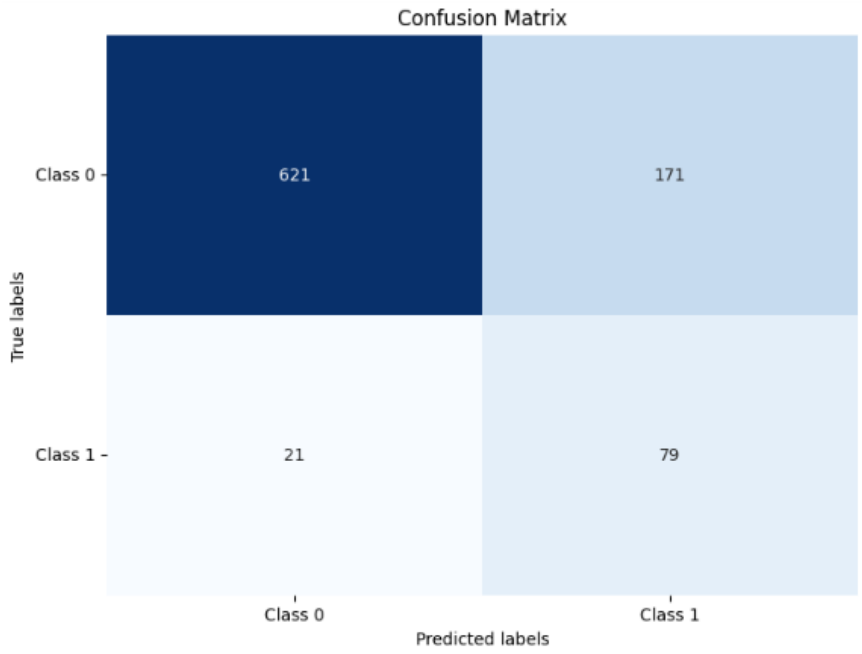


Figura 4-12.: Matriz de Confusión del Modelo Random Forest

Interpretación de la Matriz de confusión

- **621 (True Negatives - TN)**: El modelo predijo correctamente que 621 estudiantes no consumen alcohol.
- **171 (False Positives - FP)**: El modelo predijo erróneamente que 171 estudiantes consumen alcohol cuando en realidad no lo hacen.
- **21 (False Negatives - FN)**: El modelo no detectó a 21 estudiantes que realmente consumen alcohol, clasificándolos incorrectamente como no consumidores.
- **79 (True Positives - TP)**: El modelo predijo correctamente que 79 estudiantes consumen alcohol.

El modelo tiene un buen desempeño identificando a los estudiantes que no consumen alcohol (621 verdaderos negativos), pero todavía se producen errores en la clasificación de los estudiantes que consumen (solo 79 verdaderos positivos frente a 21 falsos negativos). Sin embargo, los 171 falsos positivos indica que el modelo tiende a clasificar incorrectamente a algunos estudiantes como consumidores de alcohol, lo que afecta la precisión.

Aunque todos los modelos muestran un alto recall y son efectivos para identificar la mayoría de los consumidores de alcohol, enfrentan desafíos en cuanto a precisión, ya que clasifican erróneamente a un número considerable de no consumidores como consumidores. Esto sugiere que los modelos requieren ajustes adicionales, especialmente en lo relacionado con el equilibrio de clases y los umbrales de clasificación, para mejorar la diferenciación entre las clases y reducir los falsos positivos. Random Forest parece ser el modelo más prometedor, dado su rendimiento equilibrado entre precisión y recall, aunque aún se necesita una optimización adicional para maximizar su efectividad.

En conclusión, el modelo Random Forest muestra un buen equilibrio, pero podría mejorarse aún más para reducir los falsos positivos y falsos negativos, lo que mejoraría tanto la precisión como el recall. A pesar de ello, es el mejor modelo en términos de rendimiento global en este estudio.

4.4. Diseño final del esquema de clasificación

El diseño de un esquema de clasificación basado en Random Forest para clasificar el riesgo del consumo de alcohol en estudiantes de la Universidad de Córdoba se basa en un enfoque estructurado que abarca todas las etapas del proceso, desde la preparación del conjunto de datos hasta los resultados del modelo. A continuación, se describen los pasos clave que conforman este esquema, con el propósito de contextualizar su objetivo y funcionamiento:

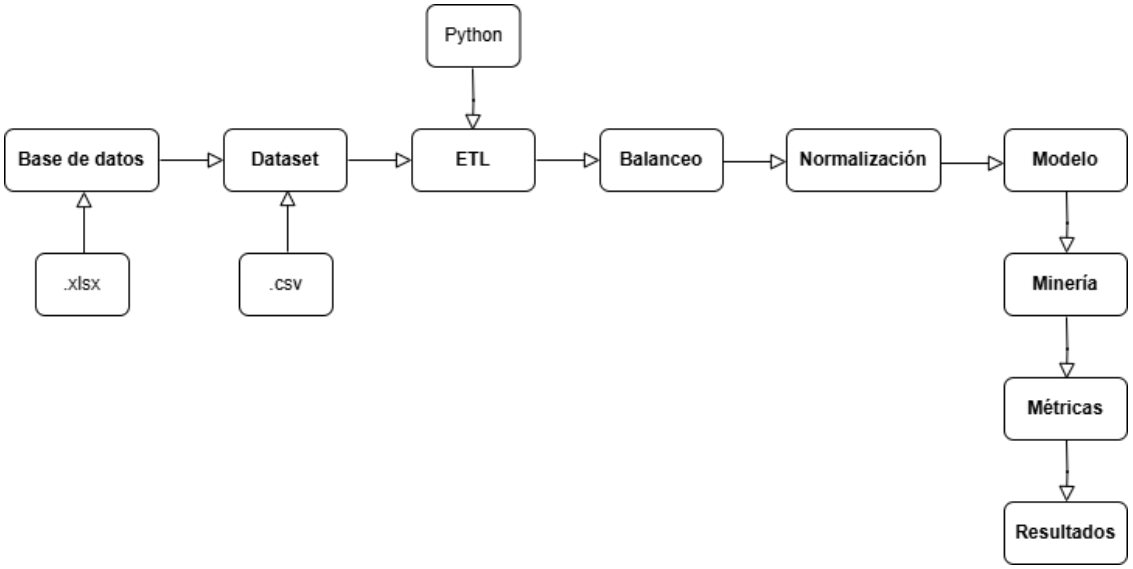


Figura 4-13.: Diseño del modelo para clasificar el consumo de alcohol

La Figura **4**-13, describe el proceso completo del modelo Random Forest diseñado para predecir el consumo de alcohol en estudiantes universitarios. Este flujo proporciona una visión estructurada de cada etapa del análisis:

1. **Base de datos**: La base de datos de los encuestados se obtuvo a través de los sistemas de información académica Academusoft y PowerCampus, exportada en formato Excel. Las variables incluidas abarcaron datos demográficos (edad, sexo biológico, estado civil, factor RH, entre otros), antecedentes académicos, factores socioeconómicos y antecedentes relacionados con el consumo de alcohol.
2. **Dataset(Datos originales)**: Los datos originales consisten en 71 variables que incluyen información sociodemográfica, socioeconómicas, académica, y psicosocial de los estudiantes. Estas variables son los insumos iniciales para la construcción del modelo.
3. **ETL**: El preprocesamiento se centra en mejorar la calidad de los datos y prepararlos para el modelo. Incluye:
   - **Balanceo de clases con SMOTEENN**: Dado que el consumo de alcohol es una clase desbalanceada (pocos estudiantes «Consumidores» en comparación con

los «No consumidores»), se utiliza SMOTEENN para combinar técnicas de sobremuestreo de la clase minoritaria (SMOTE) y submuestreo de la clase mayoritaria (ENN), logrando un conjunto de datos equilibrado.

- **Normalización y estandarización**: MinMaxScaler escala las variables al rango [0, 1], asegurando que no dominen características de mayor escala.
  StandardScaler, transforma las características para que tengan media 0 y desviación estándar 1, mejorando la estabilidad del modelo.

4. **División de los datos**, el conjunto de datos se divide en dos partes principales:
   - **Datos de entrenamiento (80 %)**: Utilizados para entrenar el modelo y ajustar sus parámetros.
   - **Datos de prueba (20 %)**: Se emplea para evaluar el desempeño del modelo en datos no vistos, asegurando que su rendimiento sea generalizable.
5. **Modelo de Random Forest**: Se seleccionó el modelo Random Forest debido a su destacado desempeño en la predicción del consumo de alcohol en estudiantes de la Universidad de Córdoba. Este algoritmo es reconocido por su capacidad para minimizar el sobreajuste en los datos de entrenamiento y mejorar la generalización en datos no vistos. Singh, Singh, Gourisaria, y Sharma (2022); Ishaq y cols. (2021); Afifi y cols. (2024).
6. **Ajuste de Hiperparámetros (GridSearch)**: Se realizó una búsqueda exhaustiva de los valores óptimos para los siguientes hiperparámetros del modelo Random Forest utilizando la técnica de GridSearch:
   - **n_estimators**: Número de árboles en el bosque (50, 100, 200).
   - **max_depth**: Profundidad máxima de cada árbol (None, 10, 20, 30).
   - **min_samples_split**: Número mínimo de muestras requeridas para dividir un nodo (2, 5, 10).
   - **min_samples_leaf**: Número mínimo de muestras necesarias en cada hoja (1, 2, 4).
   - **bootstrap**: Indica si se utiliza muestreo con reemplazo para construir los árboles (True o False).

   Este proceso permitió optimizar el modelo al identificar la combinación de hiperparámetros que maximiza su desempeño. La evaluación se realizó utilizando métricas como el F1-score, garantizando un balance adecuado entre precisión y recall. A continuación, se presentan los resultados obtenidos:

Tabla 4-11.: Mejores parámetros encontrados por GridSearchCV

n_estimators	max_depth	min_samples_split	min_samples_leaf	bootstrap
50	None	10	1	False

En relación con los resultados obtenidos, A continuación, se describe la mejor combinación de hiperparámetros:

- **Fitting 5 folds for each of 216 candidates**: El proceso de ajuste implicó un total de 1080 evaluaciones realizadas, resultado de evaluar las 216 combinaciones de hiperparámetros definidas en la rejilla, cada una a través de 5 particiones (folds) de validación cruzada. Este enfoque exhaustivo asegura que el modelo sea evaluado, identificando la configuración óptima para maximizar el desempeño.
- **n_estimators: 50**, el modelo estará compuesto por un conjunto de 50 árboles de decisión, lo que permite aprovechar la diversidad entre ellos para reducir el sobreajuste y mejorar la capacidad de generalización del modelo.
- **max_depth: None**, los árboles en el modelo crecerán de manera completa, deteniéndose únicamente cuando todos los nodos hoja sean puros (es decir, contengan datos de una sola clase) o cuando no sea posible realizar más divisiones significativas en los datos.
- **min_samples_split: 10**, un nodo deberá contener al menos 10 muestras para que pueda dividirse. Este requisito ayuda a prevenir divisiones excesivas que podrían llevar al sobreajuste, mejorando así la capacidad del modelo para generalizar a nuevos datos.
- **min_samples_leaf: 1**, en cada árbol del modelo, las hojas contendrán al menos una muestra, que es el valor mínimo permitido. Esto asegura que cada partición final del árbol tenga al menos un dato representativo, permitiendo una máxima profundidad en los árboles.
- **bootstrap: False**, el modelo fue configurado para no utilizar la técnica bootstrap al construir los árboles, lo que significa que no se generarán subconjuntos de datos mediante muestreo con reemplazo. En su lugar, cada árbol se entrenará utilizando la totalidad de los datos disponibles, lo que puede resultar en una representación más completa de las características del conjunto de datos en cada árbol.

7. **Selección de Características**: Mediante el método de Selección de Características Secuencial (SFS), se identificaron las 20 características más relevantes para maximizar el desempeño del modelo durante la validación cruzada. Las características seleccionadas son: **(1, 2, 4, 5, 6, 7, 13, 30, 37, 39, 47, 51, 52, 53, 58, 64, 65, 66, 67, 68)**(Ver anexo 4).
   Este proceso permitió reducir la dimensionalidad del problema, eliminando característi-

cas irrelevantes o redundantes. Esto no solo mejora la eficiencia computacional, sino que también incrementa la capacidad predictiva y generalización del modelo.

8. **Evaluación de Random Forest**: Se llevó a cabo una evaluación exhaustiva mediante validación cruzada con 5 particiones (k-folds = 5). Este enfoque divide los datos en cinco subconjuntos, alternando entre entrenamiento y validación en cada iteración. Esto reduce el riesgo de obtener resultados sesgados o dependientes de un solo subconjunto de datos. Ahora bien, las métricas de evaluación utilizadas fueron:
   - **Precisión**: Proporción de predicciones correctas sobre el total de predicciones realizadas.
   - **Recall**: Capacidad del modelo para identificar correctamente los casos positivos (estudiantes consumidores de alcohol).
   - **F1-score**: Métrica que combina precisión y recall en un único valor armónico, especialmente útil en problemas con clases desbalanceadas.
9. **Resultados**: La salida del modelo Random Forest, diseñada para predicir el consumo de alcohol, es una etiqueta binaria que clasifica a los estudiantes en una de las dos categorías:
   - **0 (No consume)**: El estudiante no es clasificado como consumidor de alcohol.
   - **1 (Consume)**: El estudiante es clasificado como consumidor de alcohol.
   - **Predicciones individuales de los árboles**: El modelo Random Forest está compuesto por múltiples árboles de decisión. Cada árbol realiza una predicción individual basándose en las características seleccionadas y el conjunto de datos de entrenamiento. Para cada registro de entrada (estudiante), un árbol predice si pertenece a la clase 0 (No consume) o 1 (Consume).
   - **Voto mayoritario**: El resultado final del modelo (0 o 1) se determina a través de un proceso de votación. Cada árbol «vota» por una de las dos clases, y la clase con más votos es seleccionada como la predicción del modelo.

El modelo desarrollado, basado en técnicas avanzadas de minería de datos como Random Forest, incorpora métodos de preprocesamiento, balanceo de datos (SMOTEENN), selección de características (SFS), ajuste de hiperparámetros (GridSearch) y evaluación del modelo. El objetivo general de diseñar un modelo predictivo para el consumo de alcohol en estudiantes de la Universidad de Córdoba se cumple parcialmente, ya que el modelo es capaz de detectar y clasificar correctamente a una parte significativa de los estudiantes que consumen alcohol, como lo muestra el recall (79 %) para la clase 1 (ver tabla 4-10). Sin embargo, se identifican áreas de mejora, principalmente en la precisión (32 %) para la clase 1, lo que indica que el modelo podría estar clasificando incorrectamente a algunos estudiantes no consumidores como consumidores.

Este desafío está relacionado con la desigualdad en la distribución de las clases (792 no consumidores y 100 consumidores), lo cual ha afectado significativamente el desempeño del modelo, como se refleja en las métricas de precisión y recall. Es común que los modelos tiendan a predecir con mayor frecuencia la clase mayoritaria (en este caso, «no consume»), lo que explica el recall para la clase 1, pero con una baja precisión.

Con los resultados obtenidos en el diseño del modelo, se propone utilizarlo como una herramienta tamizable *SAMHSA* (1994) en una prueba inicial, cuyo objetivo sea identificar a los estudiantes con mayor riesgo de consumo de alcohol en la Universidad de Córdoba. Este modelo permitiría detectar rápidamente a los estudiantes en riesgo, para posteriormente perfilarlos y aplicar una segunda prueba diagnóstica más específica. Esta segunda prueba estaría diseñada para reducir los falsos positivos y proporcionar un diagnóstico más preciso de los estudiantes que realmente consumen alcohol en el contexto universitario. De esta manera, se optimizarían los recursos de intervención y se garantizaría una atención más focalizada y efectiva para los estudiantes en riesgo.

4.5. Discusión de Resultados

Tras analizar los resultados obtenidos con los modelos de clasificación —Regresión Logística, Perceptrón Multicapa, SVM y Random Forest—, se identificaron hallazgos relevantes que permiten reflexionar sobre el desempeño y aplicabilidad de estos modelos en la clasificación del consumo de alcohol en estudiantes universitarios. En esta sección, se discutirá el significado e implicaciones de estos hallazgos, así como las posibles estrategias para abordar las limitaciones observadas.

En relación, al modelo Random Forest, tras la optimización de hiperparámetros mediante Grid Search y la selección de características, alcanzó un balance adecuado entre recall (0.74) y F1-score (0.46), lo que evidencia su solidez en la identificación de estudiantes en riesgo de consumo de alcohol. Este resultado resalta la importancia de priorizar el recall en problemas de esta naturaleza, donde el objetivo principal es emitir alertas tempranas. Aunque la precisión fue moderada, la capacidad del modelo para identificar un mayor número de casos positivos es crucial para evitar la exclusión de estudiantes en riesgo. En este sentido, la intervención temprana, aun con la presencia de falsos positivos, puede tener un impacto preventivo significativo en el bienestar de la comunidad estudiantil.

Por otro lado, el modelo SVM mostró un desempeño destacado en términos de recall, alcanzando hasta un 86 % con técnicas de balanceo como SMOTE y RUS. Sin embargo, su baja precisión limita la utilidad práctica del modelo, ya que el alto número de falsos positivos podría generar intervenciones innecesarias. Este resultado enfatiza la necesidad de un enfoque más equilibrado que permita mejorar la precisión sin sacrificar el recall.

Las técnicas de balanceo de clases (SMOTE, ADASYN, RUS, entre otras) mejoraron la detección de la clase minoritaria (estudiantes que consumen alcohol), pero también evidenciaron que el desbalanceo inicial afecta significativamente la capacidad predictiva de los modelos. Aunque el recall mejoró, la precisión general se redujo, lo que sugiere que estas técnicas, si bien son útiles, no solucionan completamente el problema. Este hallazgo refuerza la importancia de recopilar datos adicionales para la clase minoritaria, lo que permitiría entrenar modelos más robustos y mejorar la capacidad de generalización.

Un aspecto importante a discutir es la limitación de los datos utilizados, que provienen exclusivamente de una institución de educación superior. Esto podría afectar la generalización de los modelos a otros contextos. Los patrones de consumo de alcohol pueden variar significativamente entre universidades y regiones debido a diferencias socioeconómicas, culturales y educativas. Por ello, para mejorar la validez externa de los modelos, se requiere ampliar la base de datos con información proveniente de diversas instituciones y contextos.

Los datos utilizados en este estudio incluyeron variables sociodemográficas, económicas y académicas; sin embargo, factores psicosociales como el estrés, la familia o el entorno social no fueron analizados en profundidad. Incluir estos factores podría haber enriquecido el análisis y mejorado la precisión de las predicciones, dado que el comportamiento de consumo de alcohol en estudiantes universitarios está influenciado por diversos elementos psicosociales que no fueron completamente capturados en este trabajo. Fierro y cols. (2022).

Las correlaciones identificadas en el conjunto de datos sugieren que variables como el consumo mensual de alcohol, el consumo de sustancias y la influencia de amigos consumidores son predictores clave del consumo de alcohol. Estas correlaciones positivas indican que el entorno social y las conductas de consumo en el entorno cercano aumentan significativamente la probabilidad de consumo en los estudiantes. Por otro lado, las variables con correlación negativa sugieren que ciertos comportamientos o características personales pueden reducir la probabilidad de consumo de alcohol. Por ejemplo, se observó que realizar actividades incompatibles con el consumo de sustancias, al menos una vez en el último mes, puede ser un factor relevante en la disminución del consumo.Este factor podría ser un predictor importante para la reducción del riesgo de consumo.

En conclusión, los resultados obtenidos ofrecen una base sólida para comprender los factores que influyen en el consumo de alcohol en estudiantes universitarios. No obstante, el análisis también destaca áreas clave de mejora, especialmente en términos de generalización y la inclusión de variables psicosociales. El equilibrio entre recall y precisión continúa siendo un desafío importante; sin embargo, con las estrategias adecuadas, es posible desarrollar modelos más robustos y eficaces, que sean útiles para implementar intervenciones tempranas y efectivas.

5. Conclusiones, recomendaciones y trabajos futuros

En esta sección se presentan las conclusiones, recomendaciones y propuestas para trabajos futuros de este trabajo de grado.

5.1. Conclusiones

Los resultados obtenidos en este estudio destacan que el consumo de alcohol en estudiantes universitarios está influenciado por una variedad de factores, entre ellos el consumo mensual de alcohol, el uso de sustancias, el entorno social y la influencia de amigos consumidores. Estos hallazgos refuerzan la importancia del entorno social como un factor determinante en los patrones de consumo, lo que podría orientar futuras intervenciones preventivas. Sin embargo, el desbalance en los datos y la necesidad de aplicar técnicas de balanceo limitan la capacidad de los modelos para generalizar estos resultados a otras poblaciones.

Dado este contexto, las intervenciones futuras deberían focalizarse en los grupos con mayor riesgo de consumo problemático, ajustando las estrategias según las características identificadas en los datos. Esto requiere superar las limitaciones inherentes al desbalance de clases mediante la recolección de un volumen mayor de datos representativos, especialmente de la clase minoritaria. Una base de datos más equilibrada permitiría entrenar modelos más robustos, mejorando la precisión de las predicciones y optimizando las estrategias de intervención.

Los modelos entrenados inicialmente, sin aplicar técnicas de balanceo, mostraron un desempeño deficiente al clasificar la clase minoritaria. El bajo valor de recall evidenció dificultades para identificar correctamente a los estudiantes en riesgo de consumo de alcohol, lo que refleja un sesgo hacia la clase mayoritaria. Esta situación compromete la utilidad práctica del modelo, ya que no logra detectar adecuadamente los casos relevantes, limitando su capacidad para apoyar la implementación de alertas tempranas.

La aplicación de técnicas de balanceo como SMOTE, ADASYN, RUS, Cluster Centroids, SMOTEENN y Tomek Links mejoró considerablemente el rendimiento de los modelos en términos de recall, permitiendo identificar una mayor proporción de casos pertenecientes a la clase minoritaria. Sin embargo, esta mejora en recall a menudo se logró a expensas de una

disminución en la precisión, lo que generó un aumento en los falsos positivos. Este compromiso entre precisión y recall subraya la importancia de ajustar cuidadosamente las técnicas de balanceo para maximizar la detección de casos positivos sin sacrificar excesivamente la precisión del modelo.

En ese sentido, los modelos probados, Random Forest mostró consistentemente el mejor desempeño global, con el mayor F1-score (0.45 tras aplicar Grid Search y selección de características), lo que refleja un buen equilibrio entre precisión y recall. No obstante, SVM también presentó un alto recall después de aplicar técnicas como SMOTE y RUS, llegando a detectar hasta un 86 % de los casos de consumo, aunque su precisión fue menor, lo que resultó en falsos positivos. A demás, Logistic Regression y el Perceptrón Multicapa tuvieron un rendimiento moderado en general, pero mostraron mejoras significativas en el recall con las técnicas de desbalanceo.

Una vez, se empleó Grid Search y validación cruzada mejoró considerablemente el rendimiento de los modelos. Esto permitió optimizar los parámetros clave, mejorando tanto la capacidad predictiva como la generalización de los modelos. Sin embargo, Random Forest, después de ajustar hiperparámetros como el número de estimadores y la profundidad de los árboles, se destacó como el modelo más robusto, con buenos resultados tanto en precisión como en recall.

En contexto, la Selección Secuencial de Características (SFS) ayudó a reducir la dimensionalidad del conjunto de datos, seleccionando un subconjunto de características relevantes que mejoraron el rendimiento de los modelos. Esta técnica permitió que los modelos se concentraran en las variables más informativas, mejorando la generalización y reduciendo el riesgo de sobreajuste.

General, el modelo Random Forest ajustado con técnicas de balanceo, ajuste de hiperparámetros y selección de características fue el mejor modelo de clasificar el riesgo de consumo de alcohol en estudiantes universitarios. Este modelo mostró un buen equilibrio entre la capacidad para identificar los consumidores de alcohol (recall) y la precisión en sus predicciones.

En conclusión, aunque los modelos aplicados fueron útiles para obtener información relevante, aún hay margen para mejorar en la predicción precisa del consumo de alcohol, especialmente para los consumidores moderados o frecuentes, debido al desbalance de clases y la complejidad del fenómeno del consumo de alcohol, limitan su capacidad para ser considerados una solución completa.

En general, uno de los principales logros de este trabajo de grado fue mejorar significativamente las métricas de evaluación para la clase minoritaria a los modelos, incrementándolas desde valores muy bajos hasta casi un 50 %.

5.2. Recomendaciones

Es fundamental garantizar que los datos utilizados para entrenar los modelos estén debidamente balanceados y limpios. Si bien las técnicas de balanceo, como SMOTE, RUS y las híbridas, son útiles para mitigar el problema de la distribución desigual de las clases, obtener un mayor volumen de datos pertenecientes a la clase minoritaria o realizar una recolección más exhaustiva, enfocándose en aumentar el volumen de datos de la clase minoritaria (estudiantes consumidores de alcohol), puede contribuir significativamente a mejorar el rendimiento del modelo.

Continuar afinando la optimización de los hiperparámetros es crucial para mejorar el rendimiento de los modelos. Aunque ya se ha implementado Grid Search, se recomienda complementar este enfoque con técnicas más avanzadas como Random Search, Bayesian Optimization y AutoML (Auto Machine Learning). Estas metodologías permiten explorar de manera más eficiente el espacio de hiperparámetros, lo que resulta fundamental para optimizar el equilibrio entre precisión y recall. Una evaluación exhaustiva de diferentes configuraciones contribuirá a obtener resultados más robustos y adecuados para el problema de clasificación.

Se recomienda explorar el uso de modelos adicionales, como XGBoost, LightGBM, CatBoost, así como enfoques basados en bagging y boosting, los cuales son ampliamente reconocidos por su robustez y eficacia en problemas de clasificación. La incorporación de estos algoritmos podría mejorar significativamente los resultados obtenidos, proporcionando alternativas sólidas y precisas para enfrentar el desafío de la predicción en este contexto.

Además, se podría implementar un meta-modelo con StackingClassifier, utilizando como base modelos como RandomForest, GradientBoosting y SVC, optimizando sus hiperparámetros y ajustando los pesos para maximizar la efectividad general.

Se sugiere la participación de expertos en psicología o salud para interpretar y validar las predicciones generadas por el modelo. Esta colaboración permitiría garantizar que las intervenciones derivadas de dichas predicciones sean éticamente apropiadas y efectivas. Además, contribuiría a aumentar la pertinencia y el impacto de las estrategias propuestas, fortaleciendo su aplicación en el ámbito de la salud universitaria.

Finalmente, es crucial realizar una evaluación y ajuste periódico del modelo. Esto incluye la implementación de validación cruzada, la actualización constante del conjunto de datos y un análisis exhaustivo del impacto de sus predicciones. Estas acciones permitirán garantizar que el modelo permanezca relevante y útil en el contexto dinámico de la salud universitaria.

5.3. Trabajos futuros

Se sugiere realizar análisis adicionales para explorar cómo los factores psicosociales, como el estrés, el ambiente familiar y el desempeño académico, influyen en el consumo de alcohol.

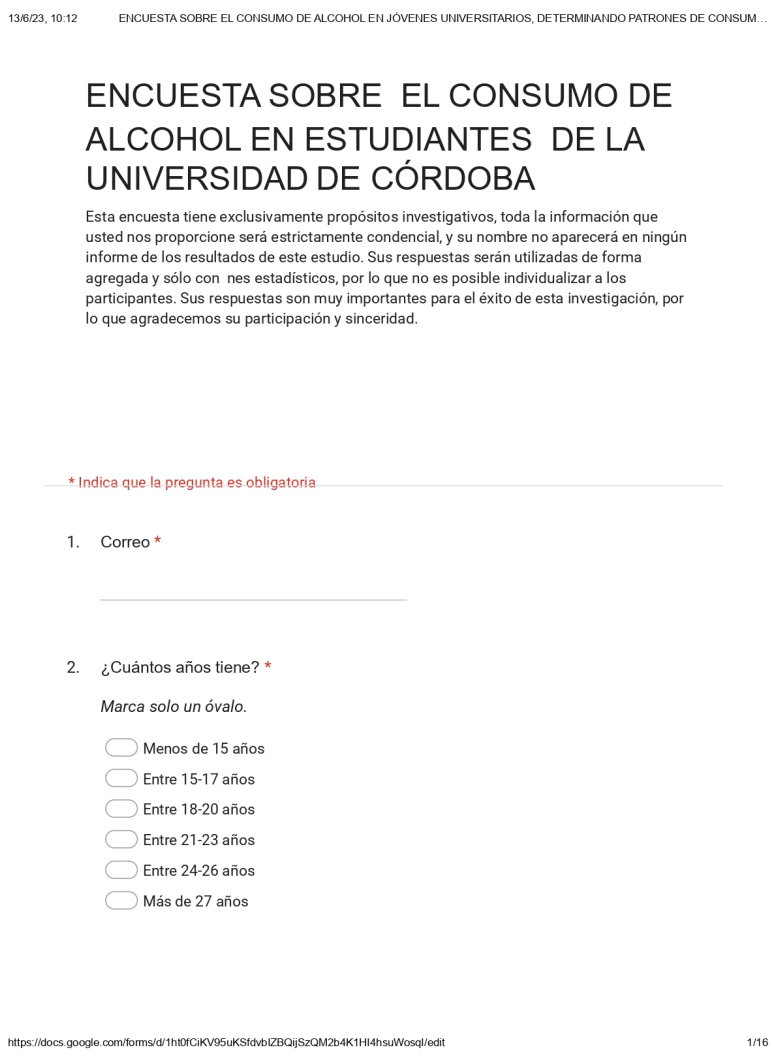
Integrar estos factores en los modelos predictivos podría mejorar significativamente la precisión de las predicciones, proporcionando una comprensión más profunda de los patrones de consumo en la población estudiantil. Para ello, se sugiere diseñar encuestas o instrumentos específicos que evalúen estos aspectos de forma rigurosa, lo que enriquecería y fortalecería el conjunto de datos disponible.

La recolección de datos de estudiantes provenientes de otras universidades o de distintos países podría enriquecer significativamente un modelo genérico. Esta ampliación de la base de datos permitiría evaluar la aplicabilidad del modelo en diversos contextos culturales y educativos, fortaleciendo su validez y robustez. Además, este enfoque facilitaría la identificación de patrones comunes y divergencias entre diferentes poblaciones, proporcionando una perspectiva más integral y matizada del fenómeno estudiado.

Realizar un estudio comparativo entre diversas regiones podría facilitar la identificación de patrones específicos que influyen en el consumo de alcohol en distintas poblaciones. Esta aproximación permitiría una comprensión más profunda de las variables contextuales y culturales que afectan los hábitos de consumo, enriqueciendo tanto el análisis como las intervenciones propuestas. Además, sería relevante investigar cómo factores económicos, sociales y de acceso a programas de prevención y tratamiento impactan los resultados obtenidos, lo cual proporcionaría información valiosa para diseñar estrategias más efectivas y personalizadas.

Finalmente, se podría desarrollar una aplicación interactiva que permita a las universidades monitorear el consumo de alcohol en sus estudiantes y aplicar intervenciones personalizadas basadas en los resultados del modelo predictivo. Esta herramienta podría incluir recomendaciones específicas, materiales educativos y acceso a recursos de apoyo, lo que contribuiría a reducir los riesgos asociados al consumo de alcohol y mejorar el bienestar general de la comunidad estudiantil.

A. Anexo 1: Encuesta sobre el consumo de alcohol en estudiantes de la Universidad de Córdoba



13/6/23, 10:12 ENCUESTA SOBRE EL CONSUMO DE ALCOHOL EN JÓVENES UNIVERSITARIOS, DETERMINANDO PATRONES DE CONSUM...

ENCUESTA SOBRE EL CONSUMO DE ALCOHOL EN ESTUDIANTES DE LA UNIVERSIDAD DE CÓRDOBA

Esta encuesta tiene exclusivamente propósitos investigativos, toda la información que usted nos proporcione será estrictamente condencial, y su nombre no aparecerá en ningún informe de los resultados de este estudio. Sus respuestas serán utilizadas de forma agregada y sólo con nes estadísticos, por lo que no es posible individualizar a los participantes. Sus respuestas son muy importantes para el éxito de esta investigación, por lo que agradecemos su participación y sinceridad.

* Indica que la pregunta es obligatoria

1. Correo *

2. ¿Cuántos años tiene? *

Marca solo un óvalo.

- Menos de 15 años
- Entre 15-17 años
- Entre 18-20 años
- Entre 21-23 años
- Entre 24-26 años
- Más de 27 años

https://docs.google.com/forms/d/1ht0fCiKV95uKSfdvbIZBQijSzQM2b4K1HI4hsuWosqI/edit 1/16

Figura A-1.: Preguntas de la encuesta del consumo de alcohol

3. Indique su género *

Marca solo un óvalo.

- Masculino
- Femenino

4. Indique la facultad a la que pertenece *

Marca solo un óvalo.

- Facultad Ciencias Básicas
- Facultad Ciencias Agrícolas
- Facultad Medicina Veterinaria y Zootecnia
- Facultad de Educación y Ciencias Humanas
- Facultad Ciencias de la Salud
- Facultad de Ingenierías
- Facultad Ciencias Económicas, Jurídicas y Administrativas

5. Indique su programa académico *

Marca solo un óvalo.

- Biología
- Química
- Geografía
- Matemáticas
- Física
- Estadística
- Ingeniería Agronómica
- Medicina Veterinaria y Zootecnia
- Acuicultura
- Bacteriología
- Enfermería
- Tecnología en Regencia y Farmacia
- Administración en Salud
- Licenciatura en Ciencias Sociales
- Licenciatura Educación Física, Recreación y Deporte
- Licenciatura en Literatura y Lengua Castellana
- Licenciatura en Informática
- Licenciatura en Lengua Extranjera con Énfasis en Inglés
- Licenciatura en Ciencias Naturales y Educación Ambiental
- Licenciatura en Educación Infantil
- Licenciatura en artística
- Ingeniería Mecánica
- Ingeniería Ambiental
- Ingeniería Industrial
- Ingeniería de Alimentos
- Ingeniería de Sistemas
- Derecho
- Administración en Finanzas y Negocios Internacionales

6. ¿En qué semestre académico se encuentra actualmente? *

Marca solo un óvalo.

- I
- II
- III
- IV
- V
- VI
- VII
- VIII
- IX
- X

7. ¿Cuál es el nivel escolar de su madre? *

Marca solo un óvalo.

- Primaria incompleta
- Primaria completa
- Bachillerato incompleto
- Bachillerato completo
- Técnico
- Tecnóloga
- Profesional
- Especialista
- Profesional con maestría
- Profesional con doctorado

8. ¿Cuál es el nivel escolar de su padre? *

Marca solo un óvalo.

- Primaria incompleta
- Primaria completa
- Bachillerato incompleto
- Bachillerato completo
- Técnico
- Tecnólogo
- Profesional
- Especialista
- Profesional con maestría
- Profesional con doctorado

9. Indique el número de hermanos *

Marca solo un óvalo.

- No tiene
- Uno
- dos
- tres
- Más de tres

10. ¿Cuál es su religión? *

Marca solo un óvalo.

- Católica
- Cristiana
- Evangélica
- Adventista
- Pentecostal
- No profesa ninguna religión
- Otra

11. ¿Con quién vive? *

Marca solo un óvalo.

- Sólo con padres
- Padres y hermanos
- Sólo abuelos
- Hogar familiar (padres, hermanos, tíos, abuelos)
- Compañero sentimental
- Vivo sólo
- Otro

12. ¿Cuál es su estado civil? *

Marca solo un óvalo.

- Soltero
- Casado
- Viudo
- Divorciado
- Unión libre

13. ¿Cuánto es aproximadamente el ingreso en su familia en salarios mínimos mensuales legales vigentes (SMMLV)? *

Marca solo un óvalo.

- Menos de un SMMLV
- Entre uno y dos SMMLV
- Entre dos y tres SMMLV
- Entre tres y cuatro SMMLV
- Entre cuatro y cinco SMMLV
- Más de cinco SMMLV

14. Estrato socioeconómico de la vivienda donde reside *

Marca solo un óvalo.

- 0
- 1
- 2
- 3
- 4
- 5
- 6

15. Actualmente, ¿usted trabaja además de estudiar? *

Marca solo un óvalo.

- Sí
- No

16. ¿Cuál es su situación de vivienda mientras está estudiando? *

Marca solo un óvalo.

- Vivo solo
- Vivo con mis padres
- Vivo en casa de un familiar
- Vivo con mi pareja
- Otra

17. ¿Le resulta fácil o difícil asumir el costo de sus estudios? *

Marca solo un óvalo.

- Muy difícil
- Difícil
- Ni fácil ni difícil
- Fácil
- Muy fácil

18. ¿Cuál es el estado civil de sus padres? *

Marca solo un óvalo.

- Casado(a)
- Divorciado(a)
- Separado(a)
- Viudo(a)
- Unión libre
- Soltero(a)
- No aplica

19. ¿Tuvo problemas académicos durante sus estudios secundarios? *

Marca solo un óvalo.

- Nunca o rara vez
- Varias veces
- Con frecuencia

20. ¿Qué tan satisfecho se siente con lo que está estudiando? *

Marca solo un óvalo.

- Muy satisfecho
- Satisfecho
- Regular
- Nada satisfecho
- No sé

21. ¿Cuántas asignaturas ha reprobado en su vida universitaria? *

Marca solo un óvalo.

- Ninguna
- Una
- Entre 2 y 3
- Entre 3 y 4
- Más de 5
- No aplica, está en primer semestre o año

22. ¿Ha pensado alguna vez abandonar los estudios que está realizando? *

Marca solo un óvalo.

- Nunca
- Alguna vez
- Varias veces

23. ¿Cree que logrará terminar fácilmente sus estudios y graduarse? *

Marca solo un óvalo.

- Sí, fácilmente
- Sí, pero con ciertas dificultades
- Sí, pero con muchas dificultades
- No creo que lo lograré

24. ¿Cuál es su percepción sobre su futuro profesional? *

Marca solo un óvalo.

- Muy optimista
- Optimista
- Pesimista
- Muy pesimista
- No lo tengo claro

25. ¿Tiene familiares que se emborrachan frecuentemente? *

Marca solo un óvalo.

- ◯ Ninguno
- ◯ Uno
- ◯ Dos o más
- ◯ No sabe

26. ¿Tiene amigos que se emborrachan frecuentemente? *

Marca solo un óvalo.

- ◯ Ninguno
- ◯ Uno
- ◯ Dos o más
- ◯ No sabe

27. Marque cual o cuales de las siguientes sustancias ha consumido por lo menos una vez en el último mes *

Selecciona todos los que correspondan.

- ☐ Alcohol
- ☐ Marihuana
- ☐ Cocaína
- ☐ Tabaco
- ☐ Éxtasis
- ☐ Anfetaminas
- ☐ Metanfetaminas
- ☐ Tranquilizantes
- ☐ Inhalantes
- ☐ Crack
- ☐ Analgésicos opiodes
- ☐ Cigarrillos electrónicos
- ☐ Ninguna

28. Marque cual o cuales de las siguientes sustancias ha consumido por lo menos una vez en la vida *

Selecciona todos los que correspondan.

- [ ] Alcohol
- [ ] Marihuana
- [ ] Cocaína
- [ ] Tabaco
- [ ] Éxtasis
- [ ] Anfetaminas
- [ ] Metanfetaminas
- [ ] Tranquilizantes
- [ ] Inhalantes
- [ ] Crack
- [ ] Analgésicos opiodes
- [ ] Cigarrillos electrónicos
- [ ] Ninguna

29. ¿Con qué tipo de personas acostumbra a consumir alcohol? *

Marca solo un óvalo.

- ◯ Amigos
- ◯ Familia
- ◯ Pareja
- ◯ Compañeros de trabajo
- ◯ Solo
- ◯ No consume alcohol

30. ¿En qué situaciones mayormente acostumbra a consumir alcohol? *

Marca solo un óvalo.

- Celebraciones con amigos
- Celebraciones con familiares
- Fiestas decembrinas
- Salidas a bailar
- No consume alcohol

31. ¿A qué edad consumió alcohol por primera vez? *

Marca solo un óvalo.

- Menos de 10 años
- Entre 10 y 14 años
- Entre 14 y 16 años
- Entre 16 y 18 años
- Mayor de 18 años
- Nunca ha consumido alcohol

32. ¿Con qué frecuencia consume alguna bebida alcohólica? *

Marca solo un óvalo.

- Nunca
- Una o menos veces al mes
- De 2 a 4 veces al mes
- De 2 a 3 veces a la semana
- 4 o más veces a la semana

33. ¿Cuántas consumiciones de bebidas alcohólicas suele realizar en un día de consumo normal? *

Marca solo un óvalo.

- Ninguna
- 1 o 2
- 3 o 4
- 5 o 6
- 7, 8 o 9
- 10 o más

34. ¿Con qué frecuencia toma 6 o más tragos en un sólo día? *

Marca solo un óvalo.

- Nunca
- Menos de una vez al mes
- Mensualmente
- Semanalmente
- A diario o casi a diario

35. ¿Con qué frecuencia en el curso del último año ha sido incapaz de parar de beber una vez había empezado? *

Marca solo un óvalo.

- Nunca
- Menos de una vez al mes
- Mensualmente
- Semanalmente
- A diario o casi a diario

36. ¿Con qué frecuencia en el curso del último año no pudo hacer lo que se esperaba de usted porque había bebido? *

Marca solo un óvalo.

- Nunca
- Menos de una vez al mes
- Mensualmente
- Semanalmente
- A diario o casi a diario

37. ¿Con qué frecuencia en el curso del último año ha necesitado beber en ayunas para recuperarse después de haber bebido mucho el día anterior? *

Marca solo un óvalo.

- Nunca
- Menos de una vez al mes
- Mensualmente
- Semanalmente
- A diario o casi a diario

38. ¿Con qué frecuencia en el curso del último año ha tenido remordimientos o sentimientos de culpa después de haber bebido? *

Marca solo un óvalo.

- Nunca
- Menos de una vez al mes
- Mensualmente
- Semanalmente
- A diario o casi a diario

39. ¿Con qué frecuencia en el curso del último año no ha podido recordar lo que sucedió la noche anterior porque había estado bebiendo? *

Marca solo un óvalo.

- Nunca
- Menos de una vez al mes
- Mensualmente
- Semanalmente
- A diario o casi a diario

40. ¿Usted o alguna otra persona ha resultado herido porque usted había bebido? *

Marca solo un óvalo.

- No
- Sí, pero no en el curso del último año
- Sí, el último año

41. ¿Algún familiar, amigo, médico o profesional sanitario ha mostrado preocupación por su consumo de bebidas alcohólicas o le han sugerido que deje de beber ? *

Marca solo un óvalo.

- No
- Sí, pero no en el curso del último año
- Sí, el último año

B. Anexo 2: Registros de las bases de datos de información de Academusoft y Powercampus

REGISTROS DE LAS BASES DE DATOS DE INFORMACIÓN DE ACADEMUSOFT Y POWERCAMPUS

CORREO
ID
ID_ESTUDIANTE
FACULTAD
DEPARTAMENTO
PROGRAMA
CODIGO_SNIES
PERIODO_ADMISION
PERIODOS_CURSADOS
SEDE
ID_TIPO_DOCUMENTO
NUM_DOCUMENTO
FECHA_EXPEDICION
LUGAR_EXPEDICION
PRIMER_NOMBRE
SEGUNDO_NOMBRE
APELLIDOS
PRIMER_APELLIDO
SEGUNDO_APELLIDO
SEXO_BIOLOGICO
ESTADO_CIVIL
FACTOR_RH
FECHA_NACIMIENTO
ESTRATO
PAIS_NACIMIENTO
DEPARTAMENTO_DE_NACIMIENTO
CIUDAD_NACIMIENTO
ESTATURA
EPS
SISBEN
DESCRIPCION_SISBEN
PAIS_RESIDENCIA
DEPARTAMENTO_RESIDENCIA
CIUDAD_RESIDENCIA
DIRECCION
BARRIOVEREDA
TELEFONO_FIJO
CELULAR
PAIS_ESTUDIO
DEPARTAMENTO_ESTUDIO
CIUDAD_ESTUDIO
INST_GRADUACION
CARACTER
ENFASIS
FORMA_OBTENCION
NO_SNP
CIUDAD_PRESENTACION
FECHA_PRESENTACION
TIPO_PRUEBA
UBICACION_SEMESTRAL
PROMEDIO_ACUMULADO
PROMEDIO_SEMESTRE
PERIODO_MATRICULA
MATERIAS_TOMADAS
MATERIAS_APROBADAS
CATEGORIA
CIRCUNSCRIPCION
MODALIDAD
ACCESO_INTERNET
CUENTA_CON_PORTATIL_O_PC_ESCRITORIO
CUENTA_CELULARINTELIGENTE_O_TABLE
GRUPO_ETNICO
PERSONA_CON_DISCAPACIDAD
CAPACIDAD_EXCEPCIONAL
PERTENECE_POBLACION_VULNERABLE
POBLACION_VULNERABLE
GENERO
ORIENTACIÓN_SEXUAL
DESVINCULACIÓN_GRUPOS_ARMADOS
INMIGRANTES
TIPO_DISCPACIDAD
ESTADO_DE_EMBARAZO
TIENE_HIJOS
CUANTOS_HIJOS
SOLTERA_CABEZA_FAMILIA
CONSUME_SUSTANCIAS_PSICOACTIVA1
SUSTANCIAS_PSICOACTIVA
FRECUENCIA_CONSUME

Figura B-1.: Registros de base de datos académica

REGISTROS UTILIZADOS PARA ENTRENAR A LOS MODELOS

# ATRIBUTOS	DESCRIPCIÓN
0 EDAD	17, 18-20, 21-23, 24-26 y 27+.
1 SEXO	Masculino o Femenino.
2 SEMESTRE_ACADEMICO_ACTUAL	1-10.
3 EC_PADRES	Estado civil de los padres.
4 ESTADO_CIVIL	Estado civil de los estudiantes.
5 NIVEL_ESCOLAR_M	Nivel escolaridad de la Madre.
6 NIVEL_ESCOLAR_P	Nivel escolaridad del Padre.
7 N_HERMANOS	Número de hermanos.
8 DUMMC	Profesa la religión católica.
9 DUMCRI	Profesa la religión cristiana.
10 DUMNPNR	No profesa ninguna religión
11 DUMMO	Profesa otra religión.
12 DUMMSP	Vive en el hogar de los papás.
13 DUMMPH	Vive en el Papá y hermanos.
14 DUMMSA	Vive con los abuelos.
15 DUMMHF	Hogar familiar.
16 DUMMCS	Hogar del compañero sentimental.
17 DUMMVS	Vive solo.
18 DUMMYO	Vive otro.
19 DUMS	Estado Civil (Soltero(a))
20 DUMC	Estado Civil (Casado(a))
21 DUMV	Estado Civil (Viudo(a))
22 DUMD	Estado Civil (Divorciado(a))
23 DUMUL	Estado Civil (Unión Libre)
24SMMLV_FAMILIAR	Ingreso familiar
25 ESTRATO_SOCIAL_VI	Estrato social de la vivienda
26 ESTRATO	Estrato socioeconómico
27 TRABAJA_ESTUDIA	Si trabaja y estudia.
28 DUMVS	Condiciones de vivienda (Vivo solo).
29 DUMVMP	Condiciones de vivienda (Padres).
30 DUMVCF	Condiciones de vivienda (Casa familiar).
31 DUMVP	Condiciones de vivienda (Pareja)
32 DUOO	Condiciones de vivienda (Otros)
33 COSTO_ESTUDIOS	Costos de los estudios.
34 DUMMYC	Estado civil de los padres (Casado).
35 DUDU	Estado civil de los padres (Unión libre).
36 DUUV	Estado civil de los padres (Soltero).
37 DUUS	Estado civil de los padres (Unión libre).
38 DUSS	Estado civil de los padres (Divorciado).
39 DUNNA	Estado civil de los padres (N/A).
40 ANTECEDENTES_ESCOLAR	Antecedentes escolares
41 SATISFECHO_EST	Satisfecho de estudiar
42 REPROBADO_MAT	Reprobado asignaturas
43 DESERTAR_EST	Abandonar los estudios

Figura B-2.: Registros de base de datos académica

44 ESTUDIOS_GRADUARSE	Logrará terminar fácilmente sus estudios y graduarse
45 FUTURO_PROF	Percepción sobre su futuro profesional
46 FAMILIARES_BORRAC	Familiares que consumen alcohol
47 AMIGOS_BORRAC	Entorno social.
48 DU_AL	Sustancias que ha consumido (Alcohol).
49 DUMI	Sustancias que ha consumido (Marihuana).
50 DUTI	Sustancias que ha consumido (Tabaco).
51 DUOL	Sustancias que ha consumido (Otras).
52 DUNI	Sustancias que ha consumido (Ninguna).
53 DUAL	Sustancias alguna vez en la vida (Alcohol).
54 DUUM	Sustancias alguna vez en la vida (Marihuana).
55 DUT	Sustancias alguna vez en la vida (Tabaco).
56 DUO	Sustancias alguna vez en la vida (Otras).
57 DUN	Sustancias alguna vez en la vida (Ninguna).
58 DUA	Personas que acostumbra a consumir (Amigos).
59 DUF	Personas que acostumbra a consumir (Familia).
60 DUP	Personas que acostumbra a consumir (Pareja).
61 DUCT	Personas que acostumbra a consumir (Compañero de trabajo).
62 DUMMS	Personas que acostumbra a consumir (Solo).
63 DUNCA	Personas que acostumbra a consumir (No consume alcohol).
64 DUCA	Situaciones a consumir (Celebraciones con amigos).
65 DUCF	Situaciones a consumir (Celebraciones familiares).
66 DUFD	Situaciones a consumir (Fiesta de decembrinas).
67 DUSB	Situaciones a consumir (Salidas a bailar).
68 DUMNCA	Situaciones a consumir (No consume alcohol).
69 RELIGION	Profesa alguna religión
70 EDAD_CONSUMO	Edad cuando inicio el consumo
71 CONSUMO	No consume (0) y Consume (1).

Figura B-3.: Registros de base de datos académica

C. Anexo 3: Selección de Características del Modelo Random Forest

Los índices (1, 2, 4, 5, 6, 7, 13, 30, 37, 39, 47, 51, 52, 53, 58, 64, 65, 66, 67, 68) corresponden a las características seleccionadas mediante la técnica de Selección de Características Secuencial (SFS), aplicando el modelo Random Forest. Esto implica que, de todas las variables disponibles, estas 20 fueron identificadas como las más relevantes para maximizar el F1-score del modelo durante la validación cruzada, optimizando su capacidad predictiva.

```
[ ]  from mlxtend.feature_selection import SequentialFeatureSelector as SFS
     from sklearn.model_selection import GridSearchCV

     # Definir características secuencial
     sfs_0 = SFS(best_model,
                 k_features=20,  # Puedes ajustar el número de características que deseas seleccionar
                 forward=True,
                 floating=False,
                 scoring='f1',
                 cv=5)

     # Ajustar características a los datos
     sfs_0 = sfs_0.fit(X_train , y_train )

     # Obtener las características seleccionadas
     selected_features = sfs_0.k_feature_idx_
```

```
[ ]  sfs_0.k_feature_idx_

     (1, 2, 4, 5, 6, 7, 13, 30, 37, 39, 47, 51, 52, 53, 58, 64, 65, 66, 67, 68)
```

Figura C-1.: Selección de características secuencial (SFS), más relevantes del modelo

C.1. Selección de Características de clasificación de Random Forest

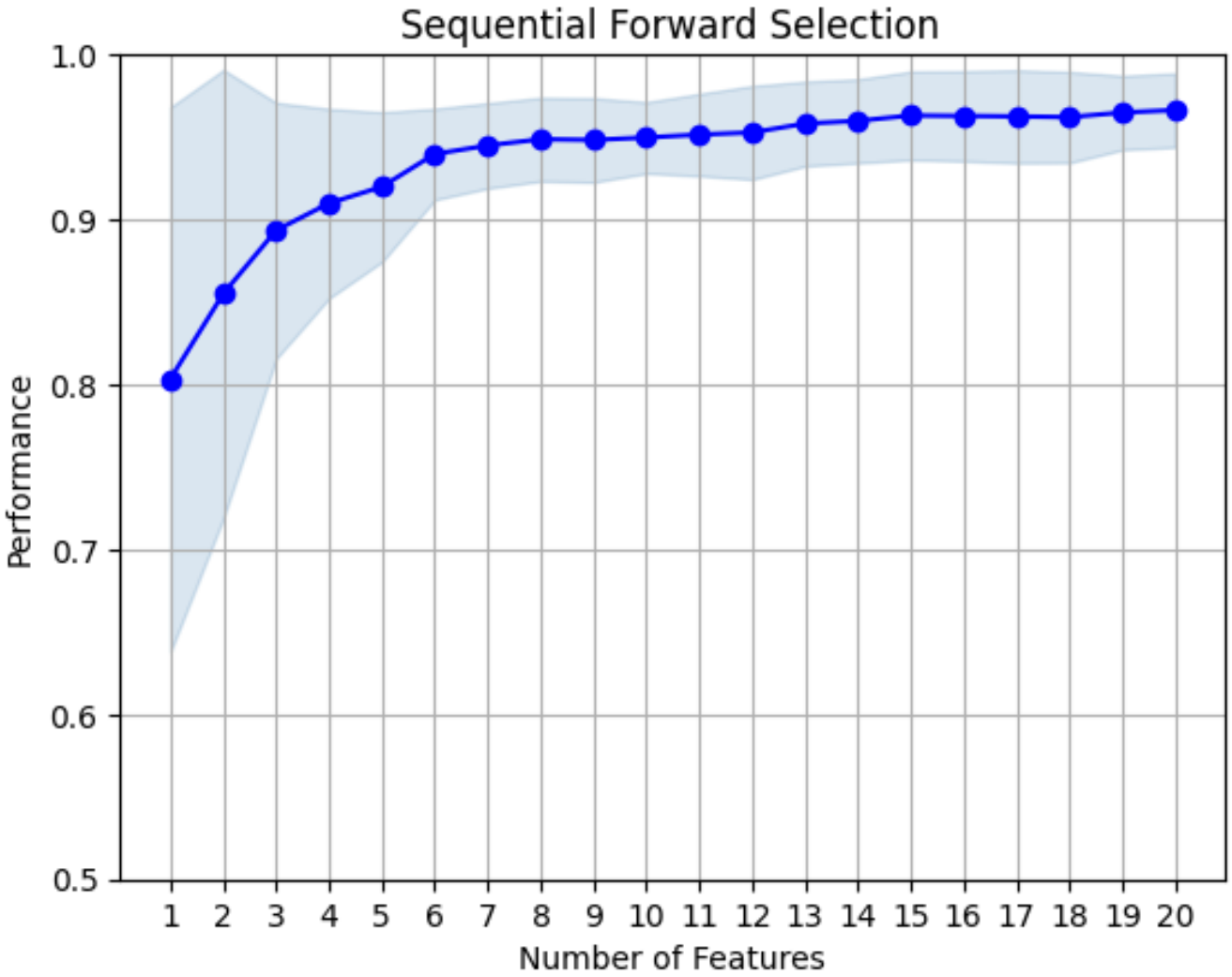


Figura C-2.: resultados de la Selección Secuencial Adelante (Sequential Forward Selection, SFS).

CARACTERÍSTICAS SELECCIONADAS PARA ENTRENAR AL MODELO RANDOM FOREST

# ATRIBUTOS	DESCRIPCIÓN
1 SEXO	Masculino o Femenino.
2 SEMESTRE_ACADEMICO_ACTUAL	1-10.
4 ESTADO_CIVIL	Estado civil de los estudiantes.
5 NIVEL_ESCOLAR_M	Nivel escolaridad de la Madre.
6 NIVEL_ESCOLAR_P	Nivel escolaridad del Padre.
7 N_HERMANOS	Número de hermanos.
13 DUMMPH	Vive en el Papá y hermanos.
30 DUMVCF	Condiciones de vivienda (Casa familiar).
37 DUUS	Estado civil de los padres (Unión libre).
39 DUNNA	Estado civil de los padres (N/A).
47 AMIGOS_BORRAC	Entorno social.
51 DUOL	Sustancias que ha consumido (Otras).
52 DUNI	Sustancias que ha consumido (Ninguna).
53 DUAL	Sustancias alguna vez en la vida (Alcohol).
58 DUA	Personas que acostumbra a consumir (Amigos).
64 DUCA	Situaciones a consumir (Celebraciones con amigos).
65 DUCF	Situaciones a consumir (Celebraciones familiares).
66 DUFD	Situaciones a consumir (Fiesta de decembrinas).
67 DUSB	Situaciones a consumir (Salidas a bailar).
68 DUMNCA	Situaciones a consumir (No consume alcohol).

Figura C-3.: Características seleccionadas para entrenar al modelo Random Forest

D. Anexo 4: Resultados de la Revisión Sistemática de Literatura

RESULTADOS DE LA REVISIÓN SISTEMÁTICA DE LITERATURA

No	Título del artículo	Autores
1	Machine learning with computer networks: Techniques, datasets and models.	Afifi., *et al. 2024*
2	El consume de alcohol como problema de salud pública.	Ahumada-Cortez., *et al 2017.*
3	Alcohol Consumption Rate Prediction using Machine Learning Algorithms	Singh *et al.*
4	Binge drinking in early adulthood: A machine learning approach	Dell *et al.*
5	Percepción de riesgo de consumo de alcohol y tabaco en universitarios del área de salud	Rodríguez *et al.*
6	Predicting binge drinking among university students: Application of integrated behavioral model	Gutema, *et al.*
7	Patterns of high-risk drinking among medical students: A web-based survey with machine learning	Marcon *et al.*
8	Patterns of Alcohol Consumption and Use of Health Services in Spanish University Students: UniHcos Project	Romero *et al.*
9	Perceived social support from significant others among binge drinking and polyconsuming Spanish university students	Tinajero *et al.*
10	Influence of attitudes and alcohol consumption on tobacco use among university students in Ecuador: an explanatory model with SEM	Moreta *et al.*
11	Aplicación de técnicas de minería de datos para la predicción del consumo de tabaco y alcohol en estudiantes universitarios	Valdiviezo *et al*
12	Modelos predictivos para la estimación de adolescentes con tendencia al alcoholismo	Salazar *et al.*
13	The relationship between demographic variables and substance use in undergraduates	Rogowska, 2019
14	Alcohol consumption patterns of university students of health sciences.	García-Carretero., *et al. 2019*

Figura D-1.: Prsentación de los Resultados de RSL

Referencias

Afifi, H., Pochaba, S., Boltres, A., Laniewski, D., Haberer, J., Leonard, P., ... others (2024). Machine learning with computer networks: Techniques, datasets and models. *IEEE access*.

Ahumada-Cortez, J. G., Gámez-Medina, M. E., y Valdez-Montero, C. (2017). El consumo de alcohol como problema de salud pública. *Ra Ximhai*, *13*(2), 13–24.

Arteaga Yánez, Y. L., Peraza de Aparicio, C. X., Ortega Guevara, N. M., Luna Álvarez, H. E., Zurita Barrios, N. Y., López Gamboa, Y., ... others (2022). *Cuidados de enfermería en la salud mental.* Quito, Universidad Metropolitana.

Batista, G. E., Prati, R. C., y Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, *6*(1), 20–29.

Betancourth-Zambrano, S., Tacán-Bastidas, L., y Cordoba-Paz, E. G. (2017). Consumo de alcohol en estudiantes universitarios colombianos. *Universidad y salud*, *19*(1), 37–50.

Chawla, N. V., Bowyer, K. W., Hall, L. O., y Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

Cruz, E., González, M., y Rangel, J. C. (2022). Técnicas de machine learning aplicadas a la evaluación del rendimiento ya la predicción de la deserción de estudiantes universitarios, una revisión. *Prisma Tecnológico*, *13*(1), 77–87.

Elhassan, T., y Aljurf, M. (2016). Classification of imbalance data using tomek link (t-link) combined with random under-sampling (rus) as a data reduction method. *Global J Technol Optim S*, *1*, 2016.

Fayyad, U., Piatetsky-Shapiro, G., y Smyth, P. (1996, Mar.). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3), 37. Descargado de `https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1230` doi: 10.1609/aimag.v17i3.1230

Ferrera Perera, M. Y., y cols. (2019). Opinión de los/as empresarios/as de la hostelería sobre posibles medidas preventivas en relación al abuso y dependencia al alcohol.

Fierro, F. S., Castañeda, J., y Revelo-Aldás, M. (2022, 6). Modelos predictivos para la estimación de adolescentes con tendencia al alcoholismo. *AXIOMA*, *1*, 74-79. doi: 10.26621/ra.v1i26.779

Frawley, W. J., Piatetsky-Shapiro, G., y Matheus, C. J. (1992). Knowledge discovery in

databases: An overview. *AI magazine*, *13*(3), 57–57.

García, J., Molina, J., Berlanga, A., Patricio, M., Bustamante, A., y Padilla, W. (2018). *Ciencia de datos : técnicas analíticas y aprendizaje estadístico. bogotá, colombia. publicaciones altaria, sl*.

García-Carretero, M. A., Moreno-Hierro, L., Martínez, M. R., de los Ángeles Jordán-Quintero, M., Morales-García, N., y O'Ferrall-González, C. (2019, 9). Alcohol consumption patterns of university students of health sciences. *Enfermeria Clinica*, *29*, 291-296. doi: 10.1016/j.enfcli.2019.01.003

Gerard, C. (2021). *Practical machine learning in javascript: Tensorflow. js for web developers*. Springer.

Gironés, J., Quiles, R. C., Roma, J. C., Alfonso, J. M., Casas, J., y Minguillón, J. (2017). *Minería de datos: modelos y algoritmos*. Editorial UOC.

He, H., Bai, Y., Garcia, E. A., y Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. En *2008 ieee international joint conference on neural networks (ieee world congress on computational intelligence)* (pp. 1322–1328).

He, H., y Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, *21*(9), 1263–1284.

Ishaq, A., Sadiq, S., Umer, M., Ullah, S., Mirjalili, S., Rupapara, V., y Nappi, M. (2021). Improving the prediction of heart failure patients' survival using smote and effective data mining techniques. *IEEE access*, *9*, 39707–39716.

Kharabsheh, M., Meqdadi, O., Alabed, M., Veeranki, S., Abbadi, A., y Alzyoud, S. (2019). A machine learning approach for predicting nicotine dependence. *International Journal of Advanced Computer Science and Applications*, *10*(3).

Kitchenham, B., y Charters, S. M. (2007). *Guidelines for performing systematic literature reviews in software engineering.* Descargado de `https://www.researchgate.net/publication/302924724`

Lamprou, S. (2021). *A study in alcohol: A comparison of data mining methods for identifying binge drinking risk factors in university students.*

Marcon, G., de Ávila Pereira, F., Zimerman, A., da Silva, B. C., von Diemen, L., Passos, I. C., y Recamonde-Mendoza, M. (2021, 9). Patterns of high-risk drinking among medical students: A web-based survey with machine learning. *Computers in Biology and Medicine*, *136*. doi: 10.1016/j.compbiomed.2021.104747

Minsalud, M. d. J. y. d. D. O. d. D. d. C. y. M. d. E. N. (2016). Estudio nacional de consumo de sustancias psicoactivas en población escolar colombia. *SD [cited 2016 Noviembre 16. Available from: https://www.minjusticia.gov.co/programasco/ODC/Publicaciones/Publicaciones/CO03142016*$_{e}$*studio*$_{c}$*o*

OEA, C. I. p. e. C. d. A. d. D. C., Organización de los Estados Americanos. (2019). *Informe sobre el consumo de drogas en las américas.*

OMS, O. M. d. l. S. (2018). *El consumo nocivo de alcohol mata a más de 3 millones de personas al año, en su mayoría hombres.* OMS Ginebra.

Pascual Pastor, F. (2012). 3. conceptos y diagnóstico del alcoholismo. *MONOGRAFÍA SObRE*, 121.

Reátegui, R., Torres-Carrión, P., López, V., Galárraga, A., Grondona, G., y Nuñez, C. L. (2020). Cluster analysis base on psychosocial information for alcohol, tobacco and other drugs consumers. En *Communications in computer and information science* (Vol. 1194 CCIS, p. 269-283). Springer. doi: 10.1007/978-3-030-42520-$3_2$2

Rodríguez de la Cruz, P. J., González-Angulo, P., Salazar-Mendoza, J., Camacho-Martínez, J. U., y López-Cocotle, J. J. (2022). Percepción de riesgo de consumo de alcohol y tabaco en universitarios del área de salud. *Sanus*, *7*.

Samhsa. (1994, octubre). `https://www.bvscolombia.org/tamizaje-y-evaluacion-spa/`. BVS Colombia. (Accessed: 2025-1-9)

Saunders, J. B., Aasland, O. G., Babor, T. F., De la Fuente, J. R., y Grant, M. (1993). Development of the alcohol use disorders identification test (audit): Who collaborative project on early detection of persons with harmful alcohol consumption-ii. *Addiction*, *88*(6), 791–804.

Singh, A., Singh, V., Gourisaria, M. K., y Sharma, A. (2022). Alcohol consumption rate prediction using machine learning algorithms. En *2022 oits international conference on information technology (ocit)* (p. 85-90). doi: 10.1109/OCIT56763.2022.00026

Swamynathan, M. (2017). *Mastering machine learning with python in six steps: A practical implementation guide to predictive data analytics using python*. Springer.

Tan, P.-N., Steinbach, M., y Kumar, V. (2006). Data mining introduction. *People's Posts and Telecommunications Publishing House, Beijing*.

Tomek, I. (1976). Two modifications of cnn. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-6*(11), 769-772. doi: 10.1109/TSMC.1976.4309452

Valdiviezo-Diaz, P., Torres-Carrión, P., Bustamante-Granda, B. F., y Sánchez-Puertas, R. N. (2020, 11). Aplicación de técnicas de minería de datos para la predicción del consumo de tabaco y alcohol en estudiantes universitarios. *Revista Ibérica de Sistemas e Tecnologias de Informação Iberian Journal of Information Systems and Technologies*, *32*, 242-255.

Whitney, A. W. (1971). A direct method of nonparametric measurement selection. *IEEE transactions on computers*, *100*(9), 1100–1103.

Yen, S.-J., y Lee, Y.-S. (2009). Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Systems with Applications*, *36*(3), 5718–5727.