MASTERPROF UMH
UNIVERSITAS Miguel Hernández

MÁSTER UNIVERSITARIO EN FORMACIÓN DEL PROFESORADO
ESO Y BACHILLERATO, FP Y ENSEÑANZAS DE IDIOMAS

TRABAJO FIN DE MÁSTER

CREACIÓN DE MODELO DE PREDICCIÓN DEL RENDIMIENTO ACADÉMICO EN CENTROS DE EDUCACIÓN SECUNDARIA

Estudiante: Óscar Agulló Mendoza
Especialidad: Matemáticas
Tutor: David Úbeda González
Cotutor: Julio Alberto Ramos Martínez
Curso académico: 2023-24

ÍNDICE

1. Resumen y palabras clave 3
2. Introducción 4
3. Método 17
3.1. Preprocesamiento de los datos y análisis descriptivo 17
3.2. Selección y configuración de los modelos 19
3.2.1. Regresión Logística. 19
3.2.2. Random Forest 21
4. Resultados 22
4.1. Regresión Logística 22
4.2. Random Forest 24
5. Discusión y conclusiones 30
6. Contribuciones prácticas 32
7. Referencias 33
8. Anexos 35

1. Resumen y palabras clave

Resumen. La educación es un factor fundamental para el desarrollo individual y social. Sin embargo, predecir este rendimiento y actuar en consecuencia es una tarea compleja. En este contexto, la minería de datos emerge como una herramienta valiosa para comprender los factores que influyen en el éxito escolar.

El presente estudio tiene como objetivo desarrollar un modelo matemático predictivo del rendimiento académico de estudiantes de educación secundaria utilizando técnicas de minería de datos. A partir de un conjunto de datos históricos, se pretende identificar patrones y relaciones que permitan anticipar el desempeño futuro de los estudiantes.

Se analizarán variables relevantes como calificaciones, historial académico, condiciones socioeconómicas y características del entorno familiar de los estudiantes. La minería de datos ofrece un gran potencial para mejorar la comprensión del rendimiento académico y desarrollar estrategias educativas más efectivas.

El estudio espera generar información valiosa para docentes, directivos e instituciones educativas, permitiéndoles tomar decisiones más acertadas en pro del éxito académico de los estudiantes.

Palabras clave: Educación, rendimiento académico, minería de datos, secundaria, python, random forest, regresión logística

2. Introducción

Objetivo general.

El objetivo principal de este trabajo es conseguir las herramientas necesarias para mejorar la calidad de la educación que se puede ofrecer a los alumnos. Guiarlos con la mayor eficacia para mejorar, en la medida de lo posible, tanto su futuro académico como su desempeño personal.

Objetivos específicos.

Analizar la relación entre los resultados de los primeros periodos académicos y la calificación final. Este análisis nos permitirá comprender mejor la evolución del rendimiento estudiantil a lo largo del tiempo.

Desarrollar e implementar algoritmos mediante minería de datos para la predicción del rendimiento académico de los estudiantes.

Evaluar de manera crítica la efectividad de los modelos predictivos en términos de precisión y utilidad práctica. Esta evaluación nos permitirá determinar la relevancia de las predicciones generadas así como su capacidad para tomar decisiones.

Hipótesis inicial.

Partimos de la hipótesis de que hay una relación existente entre las calificaciones obtenidas en los primeros periodos y la calificación final. También creemos que las diferencias socioeconómicas de los estudiantes pueden afectar a su nivel de estudios. Tanto el tiempo que dedican a estudiar, como la implicación de los padres en su desarrollo educativo es clave para su éxito. Mediante la aplicación de algoritmos adaptados al contexto educativo, se tratarán de encontrar cuales son los factores claves para desarrollar un modelo preciso que sea capaz de predecir el rendimiento de los estudiantes.

Minería de datos - Temática general.

El motivo principal por el que la minería de datos ha atraído mucha atención en la industria de la información en los últimos años se debe a la amplia disponibilidad de grandes cantidades de datos y la necesidad de convertir esos

datos en información útil. El conocimiento obtenido puede utilizarse para aplicaciones que van desde la gestión empresarial, el control de la producción y el análisis de mercado, hasta el diseño de ingeniería y la exploración científica. La industria de las bases de datos ha experimentado una gran transformación a lo largo de las últimas décadas, motivado en parte por la necesidad de analizar y gestionar volúmenes de información cada vez mayores.

Desde la década de 1960, la tecnología de bases de información ha evolucionado desde sistemas de procesamiento de archivos primitivos hasta sistemas de bases de datos sofisticadas y potentes. El desarrollo en este campo desde la década de 1970 ha llevado al desarrollo de sistemas de bases de datos relacionales, herramientas de modelado de datos y técnicas de indexación y organización de datos. A mediados de 1980, con el aumento de actividades de investigación se popularizó la tecnología relacional. La escasez de herramientas de análisis potentes junto con la abundancia de datos, creó una situación definida como “riqueza de datos pero pobreza de información” , este rápido crecimiento superó con creces la capacidad humana para la propia comprensión.

Por lo tanto, podemos definir aquí la minería de datos como la extracción de conocimientos a partir de grandes conjuntos de datos. Un proceso por el cual se encuentran un pequeño conjunto de valiosos hallazgos partiendo de una gran cantidad de datos sin procesar (Jiawei H., & Micheline K. , 2006).

En nuestro estudio, aunque nos centramos en los datos obtenidos a partir de encuestas, es interesante considerar cómo las plataformas LMS podrían complementar nuestro análisis. La minería de datos nos permite extraer conocimiento valioso de grandes conjuntos de datos, y las técnicas de meta-learning[1], aunque no sean el foco principal de nuestro estudio, pueden ofrecer perspectivas adicionales. Estas técnicas pueden mejorar la adaptabilidad y eficiencia de los sistemas de análisis de datos, aportando un marco conceptual para comprender mejor cómo los algoritmos pueden adaptarse y generalizar a nuevas situaciones.

[1] El meta-learning se centra en automatizar el proceso de aprendizaje, buscando algoritmos que puedan aprender a aprender. Este mejora el rendimiento de la minería de datos seleccionando el algoritmo adecuado y optimizando sus parámetros.

En el mundo de la inteligencia artificial, el meta-learning se presenta como un campo de investigación crucial. El artículo "A Perspective View and Survey of Meta-Learning" de Vilalta y Drissi (2002), explora el concepto de meta-learning, sus fundamentos teóricos y su potencial para revolucionar el aprendizaje automático. Este enfoque se presenta como una estrategia prometedora para potenciar la capacidad de los sistemas de aprendizaje automático para adaptarse eficientemente a nuevas tareas y dominios.

En este contexto, se resalta la relevancia de buscar formas de invarianza en el aprendizaje continuo, como se describe en el trabajo de Thrun y Mitchell (1995). Estas invariantes sirven como guías para que el algoritmo de aprendizaje seleccione la hipótesis más adecuada al enfrentarse a un nuevo dominio, centrándose en la importancia de la adaptación y la transferencia de conocimientos en el proceso de aprendizaje.

Adicionalmente, se explora una perspectiva teórica del paradigma de aprendizaje para aprender, enmarcada en una visión empírica y bayesiana. Estas aproximaciones teóricas proporcionan un marco conceptual sólido para comprender cómo los sistemas de aprendizaje pueden mejorar su rendimiento a través de la adquisición y aplicación de metaconocimiento.

El meta-learning emerge como un campo de investigación prometedor que busca potenciar la capacidad de los algoritmos de aprendizaje automático para adaptarse a nuevas situaciones y dominios. La acumulación de metaconocimiento se presenta como una estrategia clave para mejorar la generalización y el rendimiento de los sistemas de aprendizaje automático en una amplia variedad de aplicaciones.

Minería de datos - Temática específica.

En el ámbito de la predicción del rendimiento académico, el análisis de aprendizaje se centra en medir, recopilar y analizar datos sobre el aprendizaje y su entorno, con el objetivo de optimizar los procesos educativos (Long & Siemen, 2011). Este enfoque utiliza datos detallados sobre el comportamiento del estudiante, como el número de clics, el tiempo dedicado a una página y la

retención de conceptos, para predecir el rendimiento y ofrecer intervenciones personalizadas.

Según Gámez-Granados (2023), la predicción del rendimiento académico en entornos de educación secundaria online representa un desafío debido a la falta de interacción cara a cara y el elevado número de alumnos por curso. A pesar de sus numerosas ventajas como la flexibilidad y el fácil acceso que ofrecen estos sistemas, las tasas de finalización suelen ser más bajas que en la educación tradicional. Para abordar esta problemática, el estudio propone un enfoque innovador basado en algoritmos de clasificación ordinal difusa.

El principal aporte matemático del artículo radica en la introducción de un algoritmo que utiliza clasificación ordinal difusa para predecir el rendimiento estudiantil, considerando cuatro categorías de clasificación: “Withdrawn”/”suspension” (alumnos que han abandonado el curso o han sido suspendidos del programa), “passed” (el estudiante que ha cumplido los requisitos mínimos para aprobar el curso), “approved” (alumno con buen rendimiento y que ha superado los requisitos mínimos para aprobarlo), “outstanding” (rendimiento excepcional por parte del estudiante). Este enfoque permite una predicción más precisa al penalizar los errores de predicción de manera proporcional a la jerarquía de las categorías. Además, el algoritmo ofrece resultados comprensibles para los maestros al incluir lógica difusa y un sistema basado en reglas, lo que facilita la identificación de los recursos, actividades y materiales que influyen en el rendimiento del estudiante.

El algoritmo propuesto, denominado NSLVOrd, emplea un enfoque de aprendizaje de reglas iterativo junto con un algoritmo genético (GA) para aprender reglas a partir de un conjunto de datos de entrenamiento. Este conjunto de reglas se crea utilizando una estrategia de cobertura secuencial que se detalla en el artículo. La función de aptitud para evaluar la calidad de las reglas aprendidas es una función multicriterio que incluye medidas como la tasa de clasificación correcta, el error medio absoluto ordinal y la simplicidad de las reglas.

El estudio lleva a cabo un exhaustivo análisis experimental utilizando el conjunto de datos Open University Learning Analytics Dataset (OULAD), que se considera

un referente en el campo. Este conjunto de datos proporciona una muestra grande y variada de estudiantes y cursos, lo que permite validar de manera significativa el enfoque propuesto.

Los resultados experimentales indican que el enfoque propuesto ha demostrado una efectividad prometedora en la predicción del rendimiento académico de los estudiantes en entornos de educación en línea. El algoritmo NSLVOrd logra una mejora significativa en la precisión de la predicción en comparación con enfoques convencionales. Además, la capacidad de explicación de las reglas generadas por el algoritmo ofrece información valiosa para comprender los factores que influyen en el rendimiento estudiantil.

En resumen, el artículo proporciona un marco teórico y práctico para utilizar algoritmos de clasificación ordinal difusa en la predicción del rendimiento académico, con aplicaciones potenciales en la optimización de estrategias de enseñanza en entornos de educación en línea y la identificación temprana de estudiantes en riesgo.

En contraposición a mi estudio, que se centra en la enseñanza presencial mediante el análisis de datos obtenidos de encuestas, he buscado estudios sobre enseñanza virtual para complementar y contrastar los hallazgos. En el ámbito de la predicción del rendimiento académico en entornos de educación secundaria online, se presentan desafíos únicos debido a la falta de interacción cara a cara y el elevado número de alumnos por curso (Gámez-Granados, 2023). Estos estudios virtuales, como el trabajo de García Saiz (2016), proponen diversas estrategias y algoritmos de minería de datos para abordar estos retos, como el uso de algoritmos de clasificación y recomendadores basados en meta-características[2]. Estos enfoques innovadores en la enseñanza virtual ofrecen perspectivas adicionales que pueden enriquecer nuestro entendimiento y

[2] Atributos que describen las propiedades de otras características o conjuntos de datos. Es decir, son características que se calculan a partir de otras características. En la minería de datos se utilizan para la selección de características, la reducción de la dimensionalidad, la mejora de la precisión de los modelos y la interpretación de los modelos. Han permitido a los investigadores desarrollar métodos más efectivos para analizar grandes conjuntos de datos. Sus grandes ventajas son la mayor capacidad de abstracción, la mejor comprensión de los datos y modelos más robustos y precisos.

optimización de los procesos educativos, aunque nuestro estudio se enfoque principalmente en datos de encuestas en entornos presenciales.

García Saiz, D. (2016) realiza cinco estudios. Esta tesis aborda el desafío de la enseñanza virtual y propone nuevas estrategias utilizando técnicas de minería de datos para construir modelos predictivos de rendimiento estudiantil.

1. Comparación de algoritmos de clasificación:
    - Objetivo: identificar el mejor algoritmo de clasificación para predecir el rendimiento estudiantil.
    - Resultados: Los algoritmos bayesianos (*Naïve Bayes* y *BayesNetwork*) y los árboles de decisión mostraron un buen rendimiento en diferentes conjuntos de datos.

2. Uso de meta-características simples:
    - Objetivo: Evaluar la efectividad de meta-características simples para la recomendación de clasificadores.
    - Resultados: Las meta-características simples demostraron ser útiles para predecir el rendimiento de los clasificadores y construir recomendaciones efectivas.

3. Recomendaciones basadas en meta-características:
    - Objetivo: Desarrollar recomendaciones para seleccionar el algoritmo de clasificación con mejor rendimiento.
    - Resultados: Las recomendaciones basadas en meta-características simples lograron reducir el número de clasificadores considerados y mejorar la precisión.

4. Meta-características de complejidad y contexto:
    - Objetivo: Incorporar meta-características de complejidad y contexto para mejorar la capacidad predictiva de las recomendaciones.
    - Resultados: La inclusión de estas meta-características mejoró significativamente la capacidad predictiva de las recomendaciones.

5. Modelos de regresión para predecir el rendimiento:
    - Objetivo: Proponer un enfoque novedoso para predecir el rendimiento de los clasificadores en nuevos conjuntos de datos.
    - Resultados: Se propone un enfoque basado en modelos de regresión para predecir el rendimiento de los clasificadores, aunque aún presenta desafíos.

En resumen, estos cinco estudios, que forman parte de la tesis doctoral de Diego García Saiz, ofrecen una amplia exploración sobre el uso de técnicas de minería de datos para mejorar la enseñanza, desde la comparación de algoritmos de clasificación hasta la construcción de recomendadores y la aplicación de modelos de regresión para predecir el rendimiento de los clasificadores. Estos hallazgos proporcionan una base sólida para futuras investigaciones en este campo en constante evolución. (García Saiz, 2016).

El análisis de Big Data está emergiendo como una herramienta poderosa en la educación superior para mejorar la toma de decisiones y optimizar los resultados de los estudiantes. En particular, el uso de algoritmos y técnicas de minería de datos se ha centrado en predecir el rendimiento académico de los estudiantes. Hrabowski y Suess (2010) destacan la capacidad del análisis de Big Data para identificar obstáculos en el acceso y la usabilidad de los estudiantes, así como para evaluar intervenciones. Utilizando modelos descriptivos, correlacionales y predictivos, el análisis académico combina grandes conjuntos de datos con técnicas estadísticas para mejorar la toma de decisiones y proporcionar información detallada sobre el rendimiento académico.

Si bien el análisis de Big Data ofrece numerosas oportunidades para mejorar la eficacia educativa, también nos encontramos con desafíos significativos. Estos incluyen la aceptación institucional del análisis, la facilidad del software, la integración de datos de múltiples fuentes y la garantía de la calidad y algo que ha cobrado gran relevancia los últimos años, la seguridad de los datos junto con la cuestión ética en la recopilación y el uso de los datos, así como la responsabilidad institucional en la toma de decisiones basada en dicha información (Jones, 2012; Slade & Prinsloo, 2013).

En conclusión, el uso de algoritmos y técnicas de minería de datos para predecir el rendimiento académico ofrece una gran oportunidad para mejorar la educación superior. Sin embargo, es necesario abordar los desafíos y consideraciones éticas para garantizar una puesta en marcha efectiva y responsable. Futuras investigaciones se centrarán en el desarrollo de estructuras de gestión de datos, políticas de gobierno y estándares de calidad para maximizar los beneficios de esta tecnología en el ámbito educativo.

Descripción de la base de datos empleada.

Este conjunto de datos proviene de dos escuelas portuguesas y aborda el rendimiento estudiantil en dos materias distintas: Matemáticas (mat) y Lengua Portuguesa (por). Los datos recopilados incluyen las calificaciones de los estudiantes, así como características demográficas, sociales y relacionadas con la escuela. Estos datos fueron obtenidos a través de informes escolares y cuestionarios. Tenemos 33 atributos entre los que podemos en tres tipos de datos:

Numéricos:

- El atributo age (edad) cuya media es de 16.74 años y una desviación estándar de 1.22. Los estudiantes tienen edades que van desde los 15 hasta los 22 años, con la mayoría concentrada en el rango de 16 a 17 años.
- Medu (nivel de educación de la madre), con valores del 0 al 4, donde 0 es que carece de educación escolar, 1 indica educación primaria (hasta cuarto grado), 2 corresponde a haber cursado de 5º a 9º grado, 3 significa haber terminado secundaria y 4 un nivel de educación superior. La media es de 2.51 y la desviación estándar es de 1.13. Esto sugiere que la mayoría de las madres tienen algún nivel de educación, con un promedio cercano a la educación secundaria.
- En cuanto al atributo Fedu (nivel de educación del padre), con los mismos parámetros de medida, la media es de 2.31 con una desviación estándar

de 1.10. Similar a Medu, la mayoría de los padres tienen algún nivel de educación, con un promedio cercano a la educación secundaria.

- Traveltime indica el tiempo de viaje al colegio, donde 1 representa una duración menor a 15 minutos, 2 es entre 15 y 30 minutos, 3 corresponde desde 30 minutos a una hora y 4 cualquier tiempo superior a una hora. La media es de 1.57 y una desviación estándar de 0.75. La mayoría de los estudiantes tienen un corto tiempo de viaje al colegio, sugiriendo que viven cerca de la institución educativa.
- El atributo studytime (tiempo de estudio semanal) también se representa con valores del 1 al 4, donde 1 indica menos de 2 horas semanales, 2 es para quienes estudian entre 2 y 5 horas a la semana, 3 para aquellos estudiantes que dedican entre 5 y 10 horas semanales y 4 para quienes superan las 10. La media es de 1.93 con una desviación estándar de 0.83. Esto nos dice que los estudiantes dedican alrededor de 2-3 horas por semana al estudio fuera del horario escolar.
- En relación al atributo failures (número de fracasos en el pasado), la media es de 0.22 y la desviación estándar es de 0.59. La mayoría de los estudiantes no tienen fracasos previos, lo que sugiere un buen rendimiento académico general.
- El atributo famrel (calidad de las relaciones familiares), escala que va de 1 para las relaciones muy malas con la famila hasta 5 para aquellas que son excelentes, tiene una media de 3.93 y una desviación estándar de 0.96. La mayoría de los estudiantes reportan tener una buena o muy buena relación familiar.
- En cuanto al atributo freetime (tiempo libre después de la escuela), también con una escala del 1 (muy poco tiempo libre) al 5 (mucho tiempo libre). La media es de 3.18 con una desviación estándar de 1.05. Los estudiantes tienen un tiempo libre moderado después de la escuela.
- El atributo goout (salir con amigos), con la misma escala que los dos atributos anteriores (1 muy poco, 5 para mucho), tiene una media de 3.18

y una desviación estándar de 1.18. Esto sugiere que los estudiantes tienen un nivel medio de actividad social.

- Para el atributo Dalc (consumo de alcohol en días laborales, con una escala del 1 al 5), la media es de 1.50 con una desviación estándar de 0.92. El consumo de alcohol en días laborales es bajo en promedio.
- En relación al atributo Walc (consumo de alcohol durante el fin de semana, con una escala del 1 al 5), la media es de 2.28 y la desviación estándar es de 1.28. El consumo de alcohol durante el fin de semana es moderado en promedio, superior al de días laborables.
- El atributo health (estado de salud actual) tiene una media de 3.54 y una desviación estándar de 1.45. La mayoría de los estudiantes tienen un estado de salud bueno.
- En cuanto al atributo absences (número de ausencias escolares), la media es de 3.66 y la desviación estándar es de 4.64. El número de ausencias escolares es bajo en promedio, pero con una variabilidad considerable.
- El atributo G1 (nota del primer periodo, con valores del 0 al 20) tiene una media de 11.40 y una desviación estándar de 2.75. La nota promedio del primer periodo es aproximadamente 11-12, con cierta consistencia en el rendimiento académico.
- Para el atributo G2 (nota del segundo periodo), la media es de 11.57 con una desviación estándar de 2.91. La nota promedio del segundo periodo es similar a la del primer periodo.
- En relación al atributo G3 (nota final), la media es de 11.91 y la desviación estándar es de 3.23. La nota final promedio es ligeramente más alta que las notas del primer y segundo periodo, lo que sugiere una mejora en el rendimiento hacia el final del año escolar.

En la figura 1 podemos apreciar de manera gráfica la distribución de los atributos que acabamos de describir.

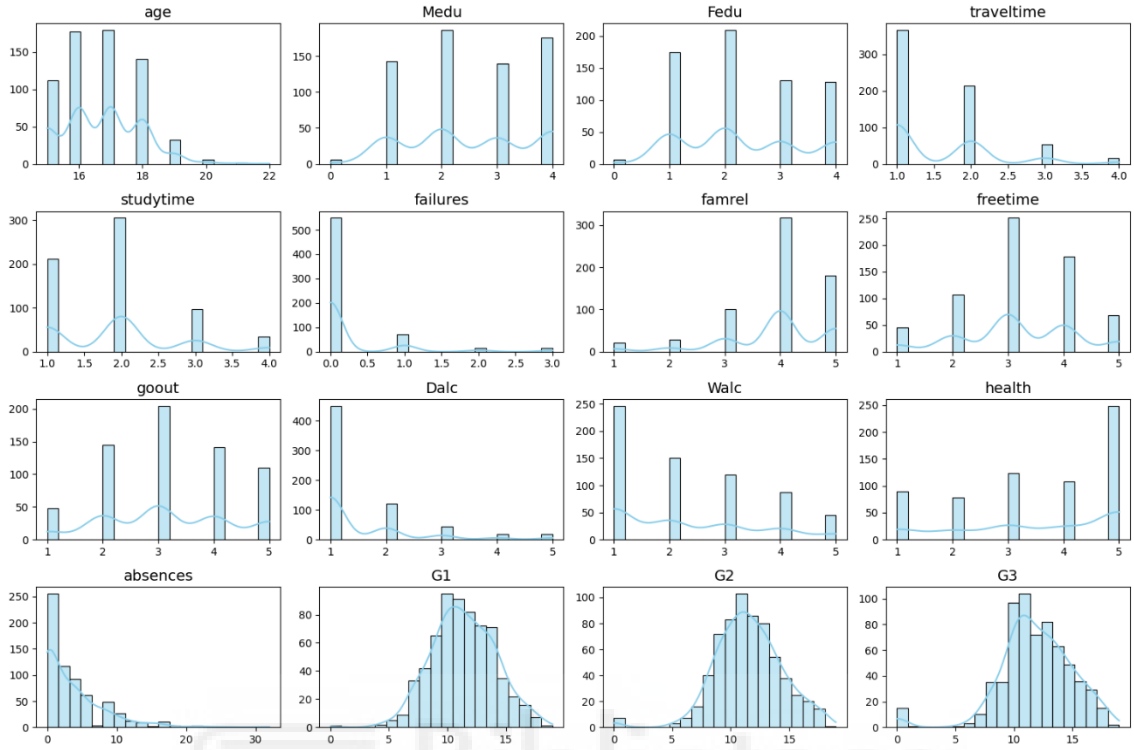

Figura 1. Atributos numéricos.

Booleanos. En la Figura 2 se presentan diagramas circulares que muestran la distribución de las variables. Estos gráficos proporcionan una representación visual de la proporción de casos en cada categoría de las siguientes variables:

- schoolsup: Apoyo educativo adicional.
- famsup: Apoyo educativo familiar.
- paid: Clases pagas adicionales dentro del tema del curso (Matemáticas o Portugués).
- activities: Actividades extracurriculares.
- nursery: Asistió a la guardería.
- higher: Quiere seguir educación superior.
- internet: Acceso a Internet en casa.
- romantic: Con una relación romántica.

Figura 2. Variables booleanas.

Categóricas. La figura 3 muestra diagramas de barras que nos proporcionan una visualización clara de la distribución en diferentes categorías de las siguientes variables:

- school: La escuela a la que asiste el estudiante, con los valores 'GP' (Gabriel Pereira) o 'MS' (Mousinho da Silveira).
- sex: El género del estudiante, con los valores binarios 'F' (femenino) o 'M' (masculino).
- address: El tipo de dirección del hogar del estudiante, con los valores 'U' (urbana) o 'R' (rural).
- famsize: El tamaño de la familia, con los valores 'LE3' (igual o menor a 3) o 'GT3' (mayor a 3).

- Pstatus: El estado de convivencia de los padres, con los valores 'T' (viviendo juntos) o 'A' (separados).
- Mjob: El trabajo de la madre, con opciones como 'teacher', 'health', 'services', 'at_home' u 'other'.
- Fjob: El trabajo del padre, con las mismas opciones que Mjob.
- reason: Razón para elegir esta escuela, con opciones como 'close to home', 'school reputation', 'course preference' u 'other'.
- guardian: Tutor del estudiante, con opciones como 'mother', 'father' u 'other'.

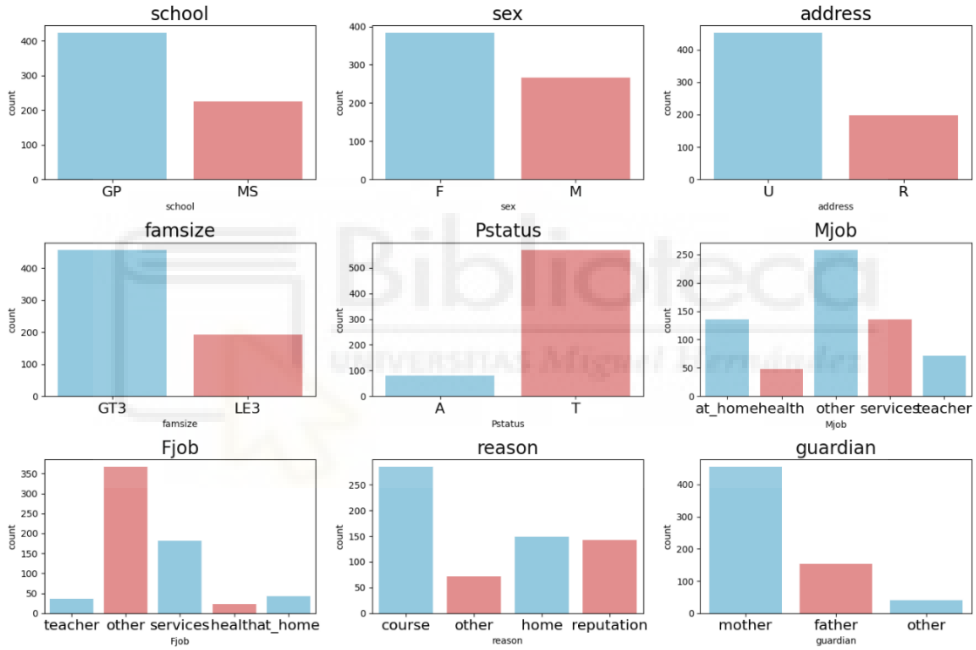

Figura 3. Variables catergóricas.

3. Método

En este apartado, se presenta el diseño y la ejecución de las simulaciones realizadas para abordar el desafío de ser capaces de predecir el rendimiento académico final de los estudiantes en función de sus características personales. El objetivo principal de este estudio es identificar el modelo de aprendizaje automático más adecuado para esta tarea, considerando la precisión, la robustez y la generalización. Para ello, han adoptado dos algoritmos de clasificación: Regresión Logística y Random Forest, complementados por otras técnicas avanzadas.

3.1. Preprocesamiento de los datos y análisis descriptivo

Se llevó a cabo para garantizar la calidad del conjunto de datos utilizado en las simulaciones. Se realizó la eliminación de datos vacíos: Se identificaron y eliminaron filas con valores faltantes en las variables relevantes, asegurando que solo se utilizaran datos completos para el entrenamiento de los modelos.

Realización de análisis descriptivo para una buena comprensión de los datos y poder llevar a cabo un estudio más efectivo. En las figuras 1 y 2 se muestra una breve descripción gráfica de las variables booleanas y categóricas.

El análisis proporciona información muy útil sobre cómo las variables están relacionadas entre sí. Para ello creamos una matriz de correlaciones, calculando el coeficiente de correlación de Pearson entre todas las parejas de variables numéricas.

En la figura 4 podemos observar el resultado, donde vemos como se cumple una de nuestras hipótesis de partida. Existe una fuerte relación entre las calificaciones obtenidas en los primeros periodos y la calificación final. Sin embargo, dado que nuestro objetivo principal es entender cómo podemos ayudar a los alumnos según sus características personales y no basarnos únicamente en sus calificaciones anteriores, hemos decidido eliminar las variables correspondientes (G1 y G2) para mejorar la eficacia de nuestro modelo.

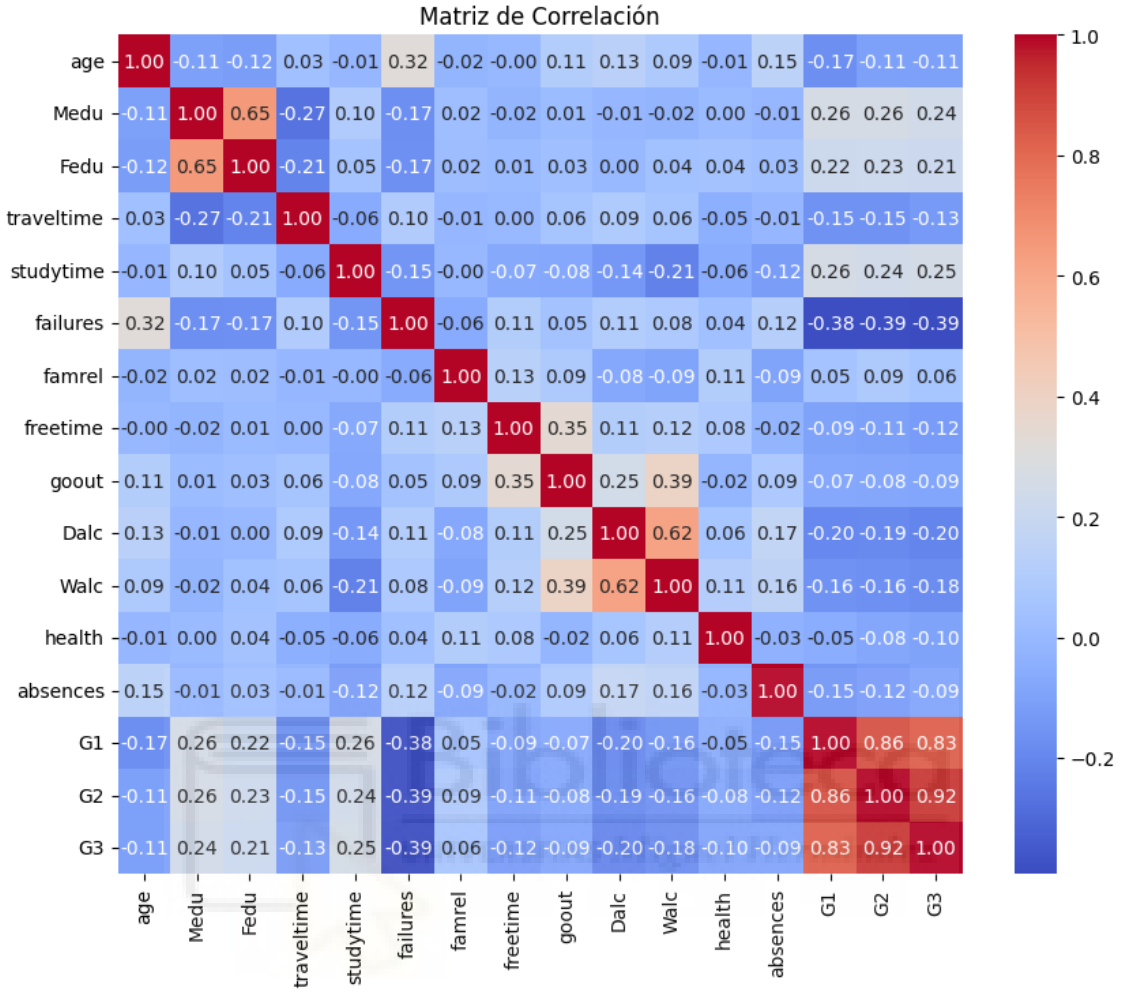

Figura 4. Matriz de correlaciones.

En la figura 5, hacemos un análisis de los valores atípicos, comprobamos que existen unos 16 datos atípicos, que representan menos de un 3% del conjunto de datos. Son datos extremadamente bajos, 0 casi todos y alguno cerca del 1. Pueden tener un impacto significativo en el análisis y los resultados del modelo por lo que haremos el análisis dos veces, con y sin dichos valores. Podría ser interesante explorar la naturaleza de estos valores, es probable que se trate de estudiantado que deja de acudir a clase, y quizás existan patrones comunes entre ellos, con una base más grande podríamos hacer un análisis específico de aquellos estudiantes con calificaciones tan bajas.

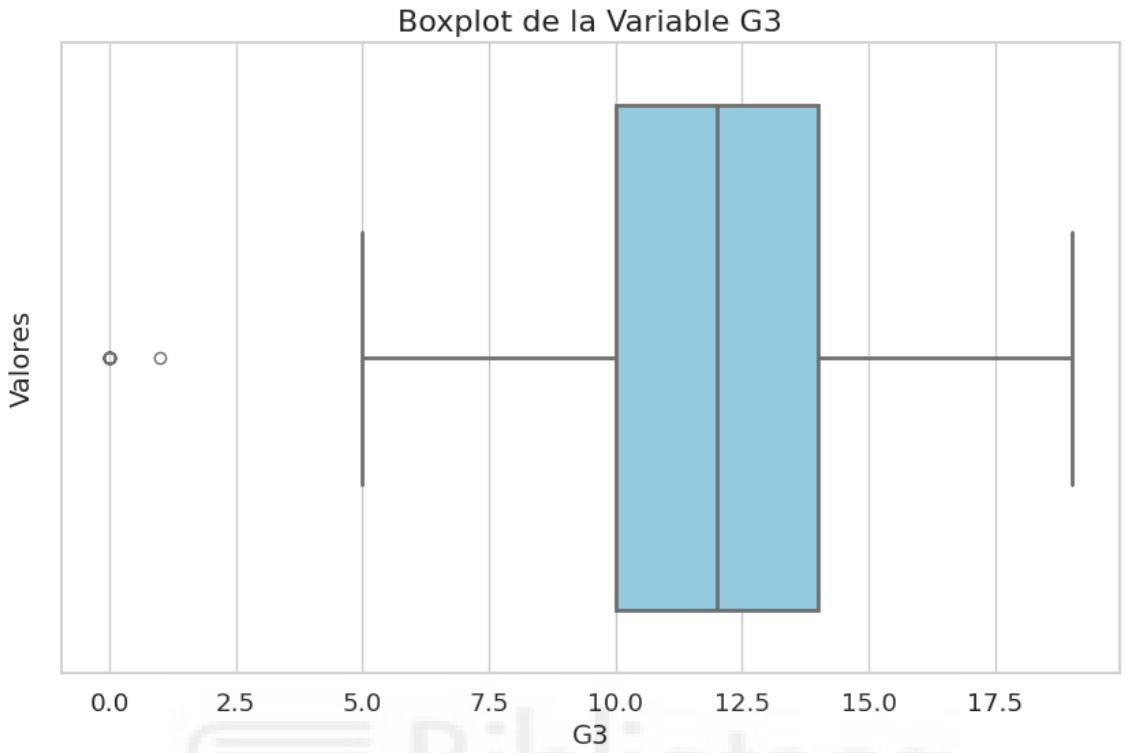

Figura 5. Boxplot variable objetivo.

3.2. Selección y configuración de los modelos.

Se evaluaron cuidadosamente diversos algoritmos de aprendizaje automático para seleccionar los más adecuados para el problema de clasificación binaria. Se consideraron la regresión logística y el random forest.

3.2.1. Regresión Logística.

La regresión logística es un algoritmo de aprendizaje supervisado utilizado para la clasificación binaria, es decir, predice la probabilidad de que una observación pertenezca a una de las dos categorías. Es un algoritmo muy versátil y utilizado en el campo del aprendizaje automático.

Preparación de los datos.

Después de haber preparado los datos, eliminamos las columnas “G1” y “G2” de los datos originales. Separamos las características de la variable objetivo que en este caso será “G3”. A su vez convertimos dicha variable en binaria, 1 para los valores de “G3” que son mayores o iguales que 10 (aprobado), 0 de lo contrario.

Realizaremos una codificación "one-hot"[3] de las características categóricas como el género o la institución educativa. El último proceso de preparación de los datos será dividirlos en conjunto de entrenamiento (80%) y prueba (20%).

Entrenamiento del modelo.
Se inicia y entrena un modelo de regresión logística sin escalado de características. Para ello utilizamos la librería scikit-learn, que nos proporciona todas las herramientas necesarias para este modelo (*train_test_split*, *LogisticRegression*, *classification_report*, *accuracy_score*). Con ello evaluamos el modelo utilizando el conjunto de prueba.

Entrenamiento del modelo con datos escalados.
Se escalan las características utilizando *StandardScaler*[4], se aplica a ambos conjuntos de dato y se vuelve a iniciar el modelo poniendo un máximo de iteraciones (1000) y volviendo a comprobar los resultados.

Selección de características utilizando RFE.
RFE es el acrónimo de Recursive Feature Elimination. Técnica utilizada en el aprendizaje automático para la selección de características, que consiste en seleccionar las características de manera recursiva mediante la eliminación de las menos importantes en cada iteración del proceso de entrenamiento del modelo. Importando la clase RFE de scikit-learn, seleccionamos las 10 características más relevantes para el modelo. Una vez hecho, volvemos a probar el modelo para comprobar si existe una mejora en la precisión.

Visualización de los resultados.
Para ello calculamos la curva ROC, que nos muestra el rendimiento de un modelo de clasificación binaria en diferentes umbrales de discriminación. Representa la tasa de verdaderos positivos en el eje Y y la tasa de falsos

[3] Técnica utilizada para convertir variables categóricas en un formato propicio para los algoritmos de aprendizaje automático. Crea una nueva columna para cada categoría única en la variable original y le asigna un valor binario.
[4] Técnica de preprocesamiento de datos utilizada para estandarizar las características numéricas. Las características tendrán una media igual a cero y una desviación estándar igual a uno. Es útil cuando las características tienen escalas diferentes.

positivos en el eje X. Un área bajo la curva más cercana a 1 indica un mejor rendimiento del modelo.

Añadimos también la matriz de confusión, para ver los verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos. Con ello ayudamos en la visualización de la precisión, sensibilidad y especificidad del modelo.

Por último, en una gráfica de barras mostramos la importancia de las 10 características que hemos calificado como las más importantes para el modelo, veremos en qué medida afectan cada una de ellas al resultado global del modelo.

3.2.2. Random Forest.

Preparación de los datos.

Como hicimos anteriormente, después de haber preparado los datos, eliminamos las columnas “G1” y “G2” de los datos originales, separamos las características de la variable objetivo. A su vez convertimos dicha variable en binaria. Realizamos la codificación de las características para convertir las variables categóricas en variables “dummy”, eliminando las columnas redundantes para evitar la multicolinealidad y así garantizar la estabilidad y la interpretación adecuada del modelo. Después de haber definido los conjuntos de prueba y entrenamiento, escalamos todos los datos para mejorar el modelo.

Mejora del modelo mediante SMOTE.

SMOTE (Técnica de sobre muestreo de minorías sintéticas) es una técnica utilizada cuando existe un desequilibrio significativo entre las clases objetivo. Con ella generamos muestras sintéticas seleccionando aleatoriamente una observación de la clase minoritaria y encontrando sus vecinos más cercanos dentro de la misma clase. Así aumentamos el conjunto de datos y el entrenamiento del modelo suele tener un mejor rendimiento.

Visualización de los resultados.

Como hemos hecho en la regresión logística, visualizamos los resultados mediante el cálculo de la curva ROC, la matriz de confusión y viendo mediante un gráfico de barras la importancia de cada variable en el funcionamiento del modelo.

4. Resultados

4.1. Regresión Logística

En la tabla 1 vemos una tabla detallada describiendo y comparando las diferencias en el proceso de mejora del modelo de regresión logística. Tanto el modelo con los datos escalados como sin escalar muestra un *accuracy* [5] idéntico, lo que nos indica, que el escalado no tuvo un impacto significativo. La precisión muestra que el 78.46% de las predicciones fueron correctas. Sin embargo, para los estudiantes que no aprobaron (clase 0), la precisión, el *recall*[6] y *F1-score*[7] son más bajos en comparación a los que aprobaron (clase 1). Esto nos sugiere que el desempeño es muy superior para predecir a los estudiantes que aprobaron versus a los que no.

Después de aplicar la reducción de variables con el método RFE, el *accuracy* aumenta hasta un 80.77% de precisión. También aumentaron ligeramente las métricas de la clase 0, pero sigue siendo un soporte muy desequilibrado.

Modelo	Accuracy Score	Precisión (Clase 0)	Recall (Clase 0)	F1-score (Clase 0)	Precisión (Clase 1)	Recall (Clase 1)	F1-score (Clase 1)	Support (Clase 0)	Support (Clase 1)
Sin Escalado	0,7846	0.57	0.50	0.53	0.84	0.88	0.86	32	98
Con Escalado	0,7846	0.57	0.53	0.55	0.85	0.87	0.86	32	98
RFE	0,8077	0.64	0.50	0.56	0.85	0.91	0.88	32	98

Tabla 1. Comparación proceso Reg. Log.

En la tabla 2 vemos el resultado de realizar el mismo proceso, pero habiendo eliminado previamente los valores atípicos, el modelo presenta una leve mejora después de haber seleccionado las variables más importantes pero viendo los datos sigue existiendo esa dificultad para predecir la clase 0.

Modelo	Accuracy Score	Precisión (Clase 0)	Recall (Clase 0)	F1-score (Clase 0)	Precisión (Clase 1)	Recall (Clase 1)	F1-score (Clase 1)	Support (Clase 0)	Support (Clase 1)
Sin Escalado	0,7874	0.59	0.53	0.56	0.85	0.87	0.86	32	95
Con Escalado	0,7795	0.56	0.56	0.56	0.85	0.85	0.85	32	95
RFE	0,8110	0.65	0.53	0.59	0.85	0.91	0.88	32	95

Tabla 2. Comparación proceso Reg. Log. sin outliers.

[5] Precisión de un modelo de aprendizaje automático en la clasificación de datos.
[6] Mide la proporción de los casos positivos reales que fueron correctamente identificados por el modelo.
[7] Estimador de la capacidad de clasificación de una prueba.

En la figura 6 vemos la curva ROC para evaluar el rendimiento del modelo. En el caso presentado, el AUC de 0.82 sugiere un buen rendimiento general del modelo. Sin embargo, la sensibilidad del 88.8% en el umbral de clasificación de 0.5, junto con la especificidad del 50%, revela un desequilibrio en la capacidad del modelo para diferenciar entre instancias positivas y negativas.

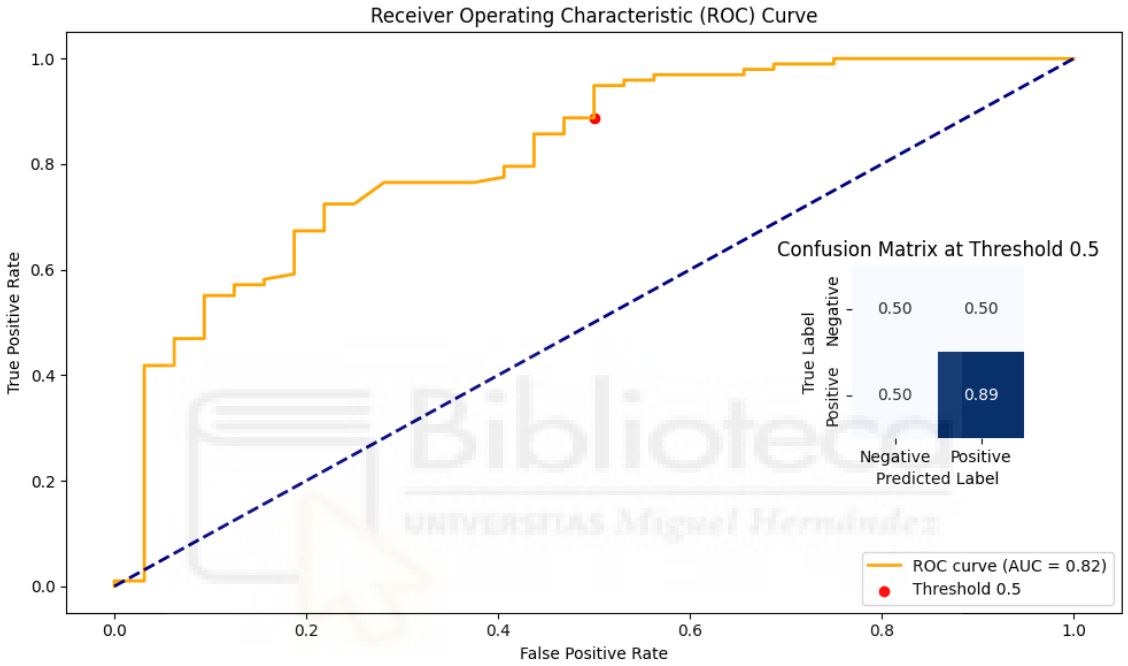

Figura 6. Curva ROC modelo Reg. Log.

En la figura 7 vemos en qué medida afecta cada una de las características más relevantes seleccionadas para el modelo. La característica "failures" muestra la mayor importancia con un valor de 1.171656, indicando que el número de fracasos en el pasado tiene un impacto significativo en la predicción del modelo. A continuación, "school_MS" tiene una importancia de 0.722752, sugiriendo que asistir a la escuela Mousinho da Silveira es un factor relevante. Todas las características combinadas proporcionan una visión comprensiva de los factores que más afectan al resultad del modelo.

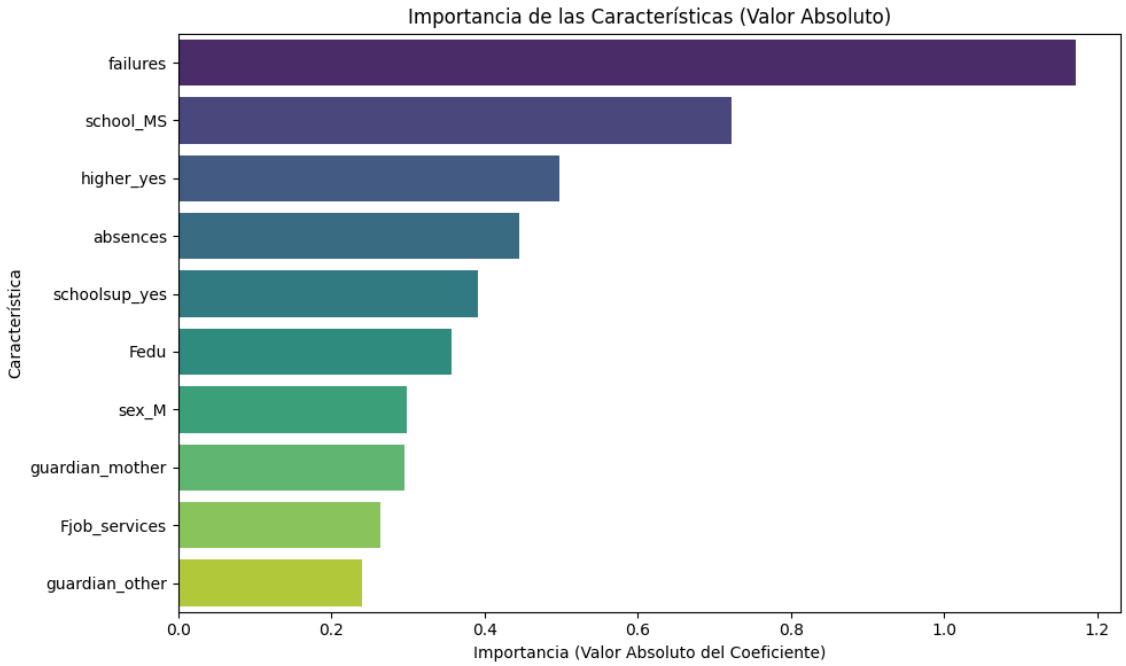

Figura 7. Importancia de las características.

4.2. Random Forest

En la tabla 3 podemos ver una comparativa del modelo random forest, antes y después de aplicar la técnica SMOTE para abordar el desequilibrio de clases, el modelo ha mejorado significativamente su rendimiento en términos de precisión, *recall* y *f1-score* para ambas clases.

El *accuracy* pasa de predecir un 82.31% a un 86.22%, además el reporte de clasificación muestra que el modelo tiene una buena precisión tanto para los suspendidos (clase 0), como los aprobados (clase 1). Esto nos deja entrever que a diferencia de la regresión logística, este modelo sí es capaz de identificar correctamente las instancias positivas y negativas con una precisión y un *recall* razonables.

Modelo	Accuracy Score	Precisión (Clase 0)	Recall (Clase 0)	F1-score (Clase 0)	Precisión (Clase 1)	Recall (Clase 1)	F1-score (Clase 1)	Support (Clase 0)	Support (Clase 1)
sin SMOTE	0,8231	0.7	0.5	0.58	0.85	0.93	0.89	32	98
con SMOTE	0,8622	0.92	0.8	0.85	0.82	0.93	0.87	98	98

Tabla 3. Random Forest con vs. sin SMOTE

En ambos conjuntos de datos observamos que después de aplicar la técnica SMOTE, mejora el *accuracy* y las métricas de precisión, *recall* y el *f1-score* para ambas clases en comparación con el modelo sin SMOTE.

En la tabla 4, donde nos encontramos los resultados de haber eliminado los valores atípicos se puede observar que el *accuracy* baja de predecir el 86.22% al 82.63%. Esta diferencia en la precisión puede deberse a varios factores. Por ejemplo, la presencia de outliers en el conjunto original puede haber proporcionado alguna información adicional al modelo. Ya que los valores eliminados en su totalidad eran de estudiantes suspendidos, y a su vez con una calificación extremadamente baja puede que mostrasen algunas características en común que ayudan a predecir este tipo de resultados

Modelo	Accuracy Score	Precisión (Clase 0)	Recall (Clase 0)	F1-score (Clase 0)	Precisión (Clase 1)	Recall (Clase 1)	F1-score (Clase 1)	Support (Clase 0)	Support (Clase 1)
sin SMOTE	0,7953	0.6	0.56	0.58	0.86	0.87	0.86	32	95
con SMOTE	0,8263	0.84	0.81	0.82	0.82	0.84	0.83	95	95

Tabla 4. Random Forest con vs. sin SMOTE sin outliers

Para optimizar el análisis hemos realizado pruebas con las diferentes posibilidades, en la tabla 5 podemos ver las siguientes pruebas:

- RF Normal: El modelo realizado con todos sus atributos y sin aplicar ninguna técnica de reducción de variables, ni generando nuevas instancias.
- RF 10 var: Modelo reduciendo variables y entrenándolo solo con las 10 consideradas de mayor relevancia.
- RF SMOTE doble: Modelo generando nuevas instancias tanto en los datos de entrenamiento como en los de testeo.
- RF SMOTE simple: Modelo generando nuevas instancias solo para el conjunto de entrenamiento.
- RF SMOTE simple 10 var: Modelo reduciendo a los atributos más importantes y aplicando SMOTE al conjunto de datos de entrenamiento.
- RF SMOTE doble 10 var: Modelo reducido y aplicando SMOTE a ambos conjuntos de datos.

En las figuras 8,9,10,11,12 y 13 vemos una comparación de los diferentes resultados que otorga cada modelo, analizando su sensibilidad y especificidad de mediante un gráfico de curva ROC y calculando el rendimiento de cada modelo mediante una matriz de confusión.

	Accuracy	Precisión Suspenso	Precisión Aprobado	Recall Suspenso	Recall Suspenso	f1-score Suspenso	f1-score Aprobado	Muestra
RF Normal	82.3%	0.7	0.85	0.5	0.93	0.58	0.89	130
RF 10 var	79.2%	0.59	0.84	0.5	0.89	0.54	0.87	130
RF SMOTE doble	86.2%	0.92	0.82	0.8	0.93	0.85	0.87	196
RF SMOTE simple	83.1%	0.71	0.86	0.53	0.93	0.61	0.89	130
RF SMOTE simple 10 var	75.4%	0.5	0.84	0.53	0.83	0.52	0.84	130
RF SMOTE doble 10 var	72.9%	0.78	0.69	0.63	0.83	0.7	0.75	196

Tabla 5. Comparación de modelos

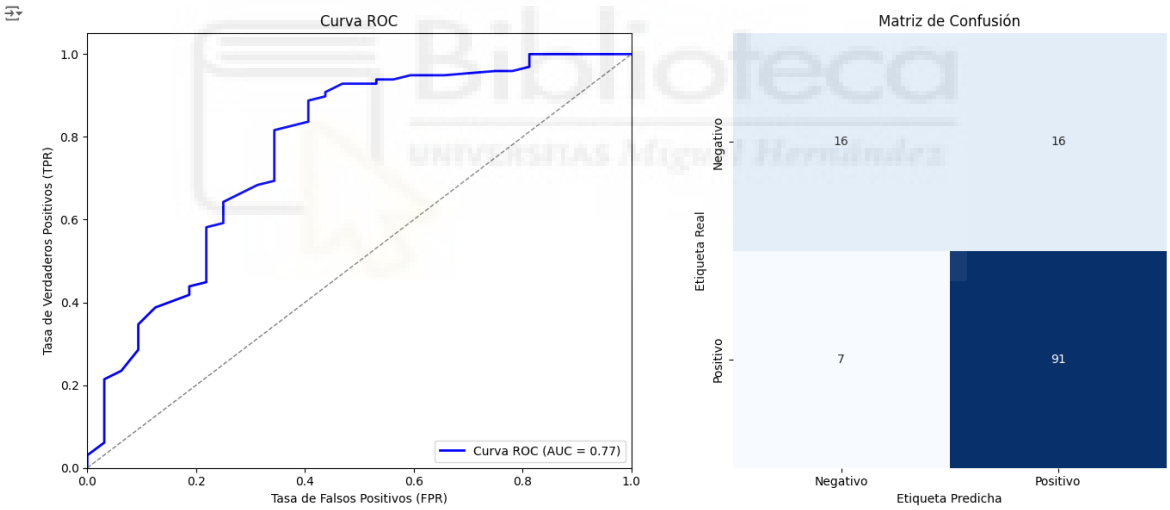

Figura 8. RF Normal.

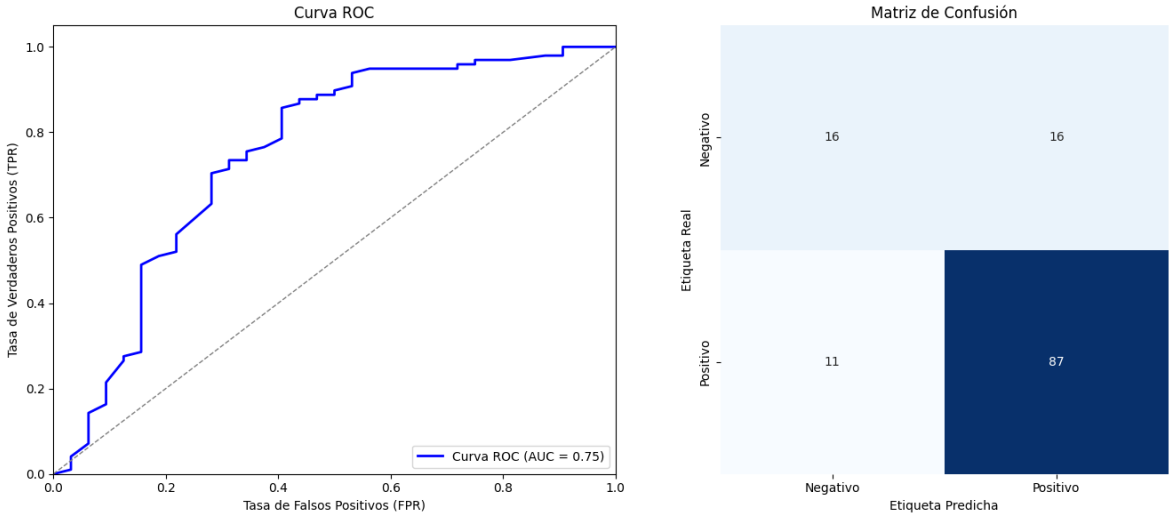

Figura 9. RF 10 var.

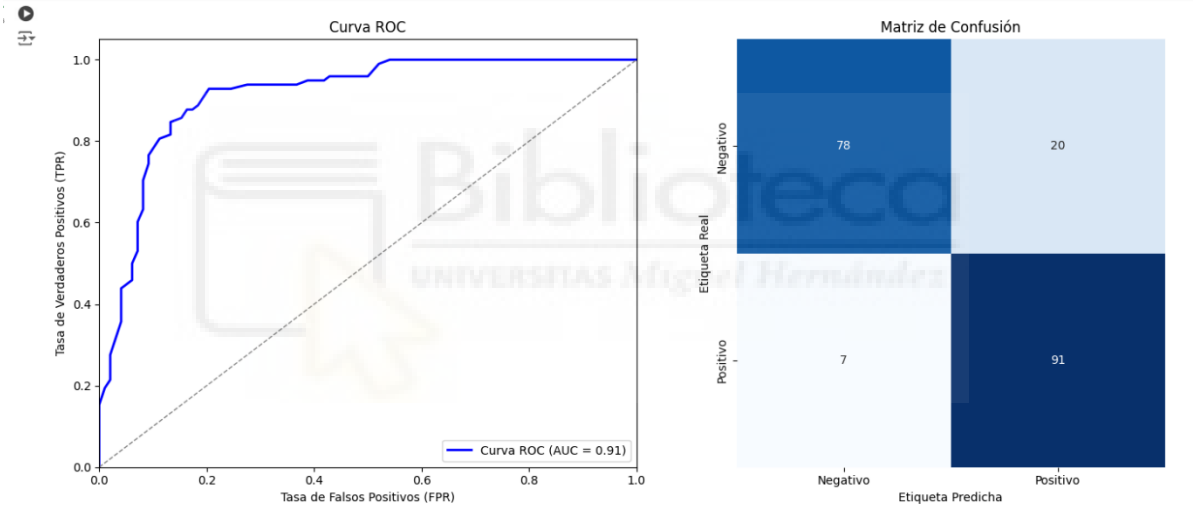

Figura 10. RF SMOTE doble.

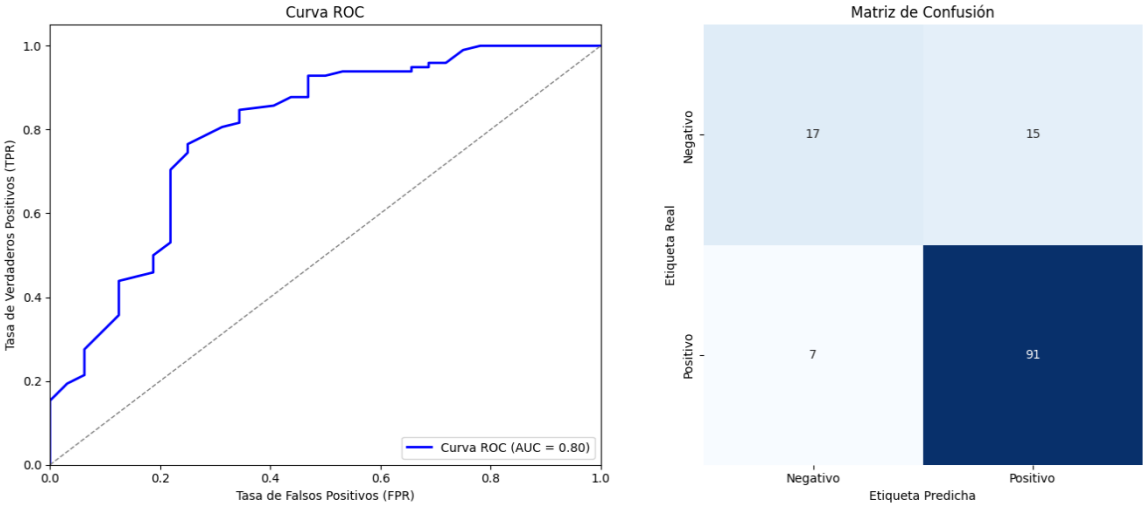

Figura 11. RF SMOTE simple.

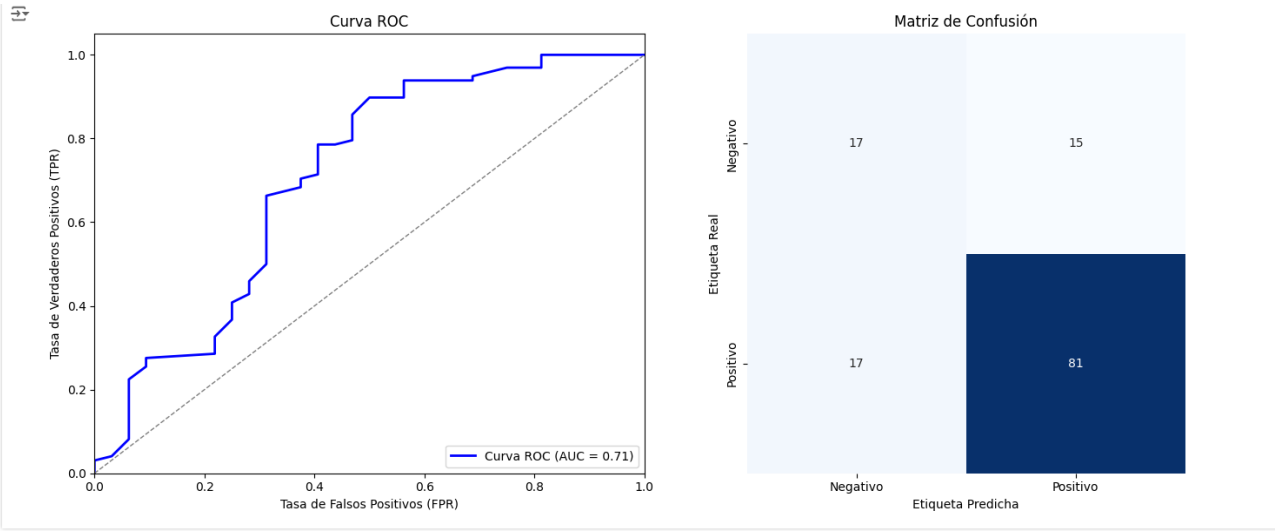

Figura 12. RF SMOTE simple 10 var.

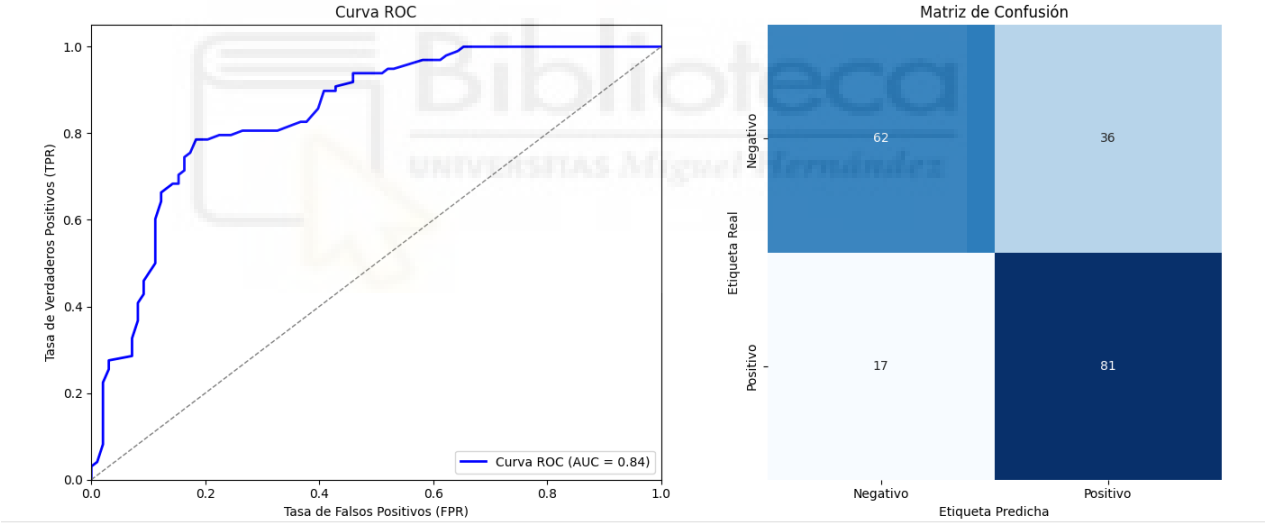

Figura 13. RF SMOTE doble 10 var.

Para poder visualizar la importancia de las características en el modelo, nos hemos quedado con el que hemos generado nuevas instancias de la clase minoritaria tanto en la variable de prueba como en la de entrenamiento (RF SMOTE doble), sacrificando así un poco la sensibilidad en pro de que la especificidad sea lo más elevada posible. En la figura 14 podemos ver como la característica 'failures' (número de fracasos en cursos anteriores) tiene la mayor importancia en la predicción del rendimiento académico de los estudiantes,

seguida de 'school_MS' (tipo de escuela), 'Fedu' (nivel educativo del padre), 'absences' (número de ausencias) y 'Medu' (nivel educativo de la madre).

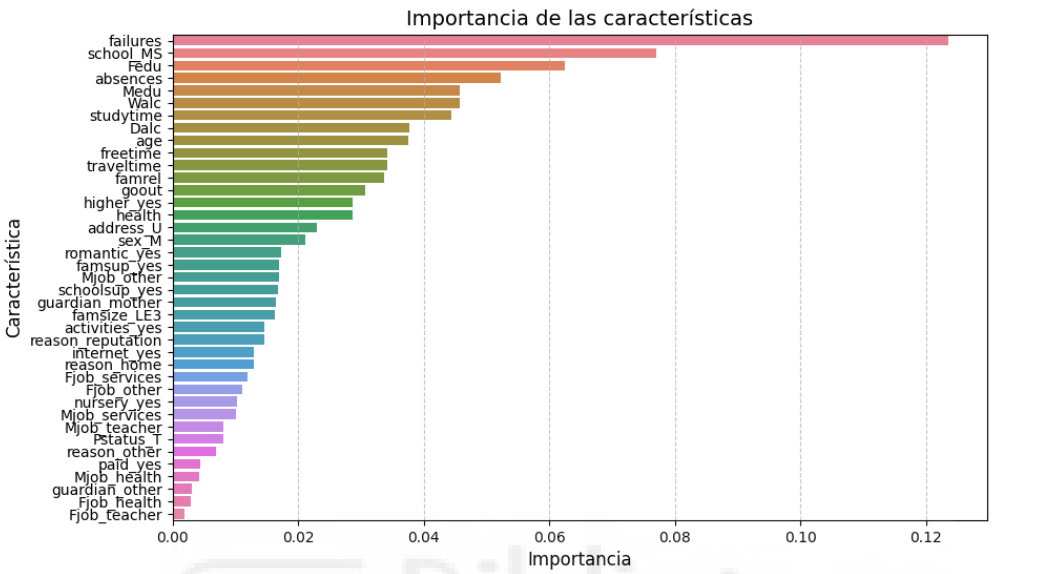

Figura 14. Importancia de las características.

5. Discusión y conclusiones

La minería de datos ha mostrado ser una herramienta muy útil para mejorar la educación y predecir el rendimiento académico de los estudiantes, coincidiendo con la perspectiva presentada por Hrabowski y Suess (2010). La importancia de analizar grandes conjuntos de datos para comprender y optimizar los procesos educativos se refleja en los estudios de Long & Siemen (2011) y Jones (2012), lo que respalda la elección del enfoque de minería de datos en este estudio.

Aunque el enfoque propuesto en este estudio muestra resultados prometedores en términos de precisión en la predicción del rendimiento académico, enfrenta desafíos similares a los discutidos por Slade & Prinsloo (2013), como la garantía de la calidad de los datos y las consideraciones éticas en la recopilación y el uso de la información, sobre todo al estar tratando datos de estudiantes menores de edad.

El estudio proporciona una sólida base teórica y práctica para utilizar la minería de datos en la predicción del rendimiento académico de los estudiantes, respaldado por numerosas investigaciones previas en el campo. A pesar de los desafíos señalados, el enfoque propuesto muestra resultados prometedores y proporciona una contribución muy valiosa para poder mejorar la calidad de la educación y ser capaces de guiar a los estudiantes hacia el éxito académico. A pesar de los buenos resultados, no hay que olvidar la necesidad de abordar las deficiencias y consideraciones éticas para maximizar los beneficios de la minería de datos en la educación.

En cuanto a los objetivos y las hipótesis del estudio, se lograron con éxito. Desde el inicio, el propósito era desarrollar un modelo predictivo eficiente para anticipar el rendimiento académico final de los estudiantes. Para ello, se llevó a cabo una serie de etapas analíticas que permitieron comprender la relación entre las calificaciones obtenidas en los primeros períodos académicos (G1 y G2) y la calificación final (G3). La implementación de algoritmos de minería de datos, específicamente random forest y regresión logística permitió realizar predicciones basadas en los datos recopilados. Estos algoritmos se seleccionaron por su capacidad para modelar relaciones complejas y su robustez, aspectos cruciales dada la naturaleza compleja de los datos de los

estudiantes. Los resultados obtenidos de estos experimentos demostraron una capacidad predictiva prometedora de ambos modelos, con puntajes de precisión que superaron el 80% en términos de precisión global. Demostrando, con el modelo Random Forest que somos capaces de predecir con antelación cuándo un estudiante podría llegar a fracasar en sus estudios, con la finalidad de que el centro pueda intervenir y realizar un estudio de la problemática asociada a esta persona y ayudarla en la medida de lo posible a encauzar sus estudios.

Las conclusiones extraídas del estudio resultan de gran utilidad para llegar a comprender el impacto y la utilidad práctica. Se confirmó la importancia de la relación entre las calificaciones de los primeros trimestres y las calificaciones finales. Destacando así la necesidad de analizar el progreso académico de los estudiantes desde etapas tempranas. Además, se validó la eficacia de los modelos predictivos desarrollados, donde el modelo hecho mediante Random Forest propició unos resultados superiores al hecho con regresión logística. La diferencia en el rendimiento de los modelos resaltó la importancia de seleccionar el algoritmo adecuado según las características específicas del conjunto de datos y los objetivos del estudio.

6. Contribuciones prácticas

Los resultados obtenidos en este estudio ofrecen una base sólida para ayudar a los centros educativos en la predicción del rendimiento académico de los estudiantes. Los modelos desarrollados pueden identificar a los estudiantes en riesgo y permitir intervenciones para mejorar su rendimiento académico. Al utilizar estos modelos, los centros educativos pueden prever el rendimiento futuro de los estudiantes utilizando información sobre sus condiciones familiares, socioeconómicas y el entorno estudiantil. Esto les permite intervenir para garantizar los recursos y apoyo necesario a aquellos estudiantes en riesgo de fracaso académico, antes de que sea demasiado tarde.

Continuar la investigación a través de nuestra línea implica explorar nuevas variables, técnicas y estrategias para mejorar la precisión y la utilidad práctica de los modelos de predicción realizados. La incorporación de datos adicionales, más detallados sobre los estudiantes, información sanitaria o información sobre las actividades extraescolares sería una posible área de investigación futura. Con estas nuevas variables, y una muestra más grande se podría generar modelos de mayor precisión. Al mismo tiempo que la ampliación de muestra y de variables, la evaluación del impacto de las intervenciones realizadas en los estudiantes, basadas en los modelos predictivos debería ser otra línea de investigación interesante. Esto implicaría crear estrategias específicas de apoyo académico para los estudiantes identificados como en riesgo por los modelos y evaluar si estas intervenciones conducen a mejoras notables en su rendimiento académico.

En resumen, continuar la investigación a través de nuestra línea implica explorar nuevas variables, técnicas y estrategias para mejorar la precisión y la utilidad práctica de los modelos de predicción del rendimiento académico, con el objetivo final de contribuir al éxito estudiantil y mejorar los resultados educativos.

7. Referencias

Han, J., & Kamber, M. (2006). Data mining: Concepts and techniques. Morgan Kaufmann.

Vilalta, R., & Drissi, Y. (2002). A Perspective View and Survey of Meta-Learning. Artificial Intelligence Review, 18, 77–95.

Thrun, S., & Mitchell, T. M. (1995). Lifelong robot learning. Robotics and Autonomous Systems, 15(1–2), 25-46.

Gámez-Granados, J. C., Esteban, A., Rodriguez-Lozano, F. J., & Zafra, A. (2023). An algorithm based on fuzzy ordinal classification to predict students' academic performance. Applied Intelligence, 53, 27537–27559.

García Saiz, Diego. "Minería de datos aplicada a la enseñanza virtual: nuevas propuestas para la construcción de modelos y su integración en un entorno amigable para el usuario no experto." Tesis Doctoral, dirigida por Marta Elena Zorrilla Pantaleón, 5 de abril de 2016.

Hrabowski, F. A. III, & Suess, J. (2010). Reclaiming the lead: higher education's future and implications for technology. EDUCAUSE Review, 45(6).

Long, P., & Siemen, G. (2011). Penetrating the fog: analytics in learning and education. EDUCAUSE Review, 46(5), 30–40.

Jones, S. (2012). Technology review: the possibilities of learning analytics to improve learner-centered decision-making. Community College Enterprise, 18(1), 89–92.

Slade, S., & Prinsloo, P. (2013). Learning analytics: ethical issues and dilemmas. American Behavioral Scientist, 57(10), 1509–1528.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830. Recuperado de https://scikit-learn.org

McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, 51-56. Recuperado de https://pandas.pydata.org

van der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy Array: A Structure for Efficient Numerical Computation. Computing in Science & Engineering, 13(2), 22-30. Recuperado de https://numpy.org

Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Ostblom, J., Lukauskas, S., Gemperline, D., Augspurger, T., Halchenko, Y., Cole, J., Warmenhoven, J., Ruiter, J., Vanderplas, J., Hoyer, S., Villalba, S., Quintero, E., Bachant, P., Martin, M., Meyer, K., Miles, A., ... Seaborn Contributors. (2022). Seaborn: Statistical Data Visualization. Journal of Open Source Software, 7(77), 3021. Recuperado de https://seaborn.pydata.org

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90-95. Recuperado de https://matplotlib.org

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. Journal of Machine Learning Research, 18(17), 1-5. Recuperado de https://imbalanced-learn.org

8. Anexos

Enlace al código python

Csv con los datos recopilados de los institutos portugueses