UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Informática

Modelización de los factores que inciden en el rendimiento académico de los estudiantes universitarios con técnicas de estadística multivariante y de machine learning

Trabajo Fin de Grado

Grado en Ciencia de Datos

AUTOR/A: Carles Vega, Josep

Tutor/a: Carot Sierra, José Miguel

Cotutor/a: Conchado Peiró, Andrea

CURSO ACADÉMICO: 2022/2023

Modelización de los factores que inciden en el rendimiento académico de los estudiantes universitarios con técnicas de estadística multivariante y de aprendizaje automático

Resum

En aquest treball final de grau es van a elaborar diferents models d'aprenentatge automàtic per poder predir tant el rendiment acadèmic com l'eficiència dels estudiants universitaris i conèixer si hi ha relacions amb les variables com el sexe o el grau que estudien.

L'objectiu d'aquesta investigació és intentar determinar en quina mesura l'aprenentatge automàtic pot ser capaç de predir el rendiment acadèmic universitari amb precisió, sensibilitat i especificitat. En aquest estudi es va utilitzar una població de 5035 alumnes de grau amb diferents característiques recollides en 44 variables, 19 de caràcter demogràfic, 19 més de caràcter acadèmic i finalment unes 6 de caràcter personal.

Per donar resultats es van utilitzar eines d'aprenentatge automàtic per poder predir tant la nota mitjana com la taxa d'eficiència, entre aquestes eines hi ha els algoritmes d'aprenentatge automàtic anomenats: Màquines de Suport Vectorial (SVM), Arbres de decisió, Random Forest, KNN veïns i mètodes de regressió PLS.

Paraules Clau: Educació, taxa d'eficiència, machine learning, aprenentatge supervisat

Resumen

En este trabajo final de grado se elaboraron distintos modelos de aprendizaje automático para poder predecir tanto el rendimiento académico como la eficiencia de los estudiantes universitarios y conocer si existen relaciones con las variables como sexo o el grado que estudian.

El objetivo de esta investigación es tratar de determinar en qué medida el aprendizaje automático puede ser capaz de predecir el rendimiento académico universitario con precisión, sensibilidad y especificidad. En este estudio se utilizó una población de 5035 alumnos de grado con distintas características recogidas en 44 variables, 19 de ellas de carácter demográfico, otras 19 de carácter académico y por último unas 6 de carácter personal.

Para dar resultados se utilizaron herramientas de aprendizaje automático para poder predecir tanto la nota media como la tasa de eficiencia, entre estas herramientas se encuentran los algoritmos de aprendizaje automático llamados: Máquinas de Soporte Vectorial (SVM), Árboles de decisión, Random Forest, KNN vecinos y métodos de regresión PLS.

Palabras Clave: Educación, tasa de eficiencia, machine learning, aprendizaje supervisado.

Abstract

In this final undergraduate work, different machine learning models were developed to predict both the academic performance and the efficiency of university students and to find out if there are relationships with variables such as gender or the grade they study.

The objective of this research is to try to determine to what extent machine learning can be able to predict university academic performance with accuracy, sensitivity and specificity. This study used a population of 5035 undergraduate students with different characteristics collected in 44 variables, 19 of which were of a demographic nature, 19 of an academic nature and 6 of a personal nature.

To give results, machine learning tools were used to predict both the average grade and the efficiency rate, among these tools are machine learning algorithms called: Support Vector Machines (SVM), Decision Trees, Random Forest, KNN neighbors and PLS regression methods.

Keywords: Education, efficiency rate, machine learning, supervised learning.

Agradecimientos

A todos mis profesores y profesoras y en especial a mis tutores, José Miguel y Andrea por su ayuda, paciencia y dedicación y por cada pregunta respondida. Agradecerle también a toda mi familia por darme ánimos durante todas las fases del proyecto y por apoyarme hasta el final.

Índice de contenido

Índice de figuras

Índice de Tablas

1. Introducción

Uno de los temas más controvertidos durante mucho tiempo, en el campo de la educación es el rendimiento académico. La educación es un logro académico y por tanto nos referimos a él como un éxito o un fracaso de los estudiantes. Teniendo en cuenta que los estudiantes son parte fundamental de las instituciones educativas y sobre todo del país, se crea la necesidad de ser capaz de identificar de antemano cuáles serán los estudiantes que tendrán un rendimiento bueno o un rendimiento malo con la finalidad de que las universidades puedan brindar más apoyo o proporcionar alternativas mejoradas a los alumnos.

Para situarnos en contexto podemos decir que el Sistema Universitario Valenciano presenta para el curso 21/22 un total de 129.000 estudiantes, frente al 1.340.632 estudiantes del conjunto de universidades españolas. En el Sistema Universitario Valenciano la tasa de rendimiento de los grados se sitúa en el 88% mientras que el SUE lo fija en el 84,6%, en ambos casos el dato publicado en el último informe corresponde con el curso 19/20.

El rendimiento académico de los estudiantes universitarios es un tema de gran interés en el ámbito educativo. Hay múltiples factores que pueden influir en el desempeño de los estudiantes, como su nivel socioeconómico, su motivación, sus hábitos de estudio, entre otros.

La aplicación de técnicas de estadística multivariante y de aprendizaje automático puede ayudar a identificar los factores más relevantes que inciden en el rendimiento académico de los estudiantes universitarios. Estas técnicas permiten analizar conjuntamente múltiples variables y encontrar patrones y relaciones entre ellas.

Algunas de las variables que se suelen analizar son: el historial académico previo, el nivel socioeconómico del estudiante, la dedicación al estudio, la asistencia a clase, el uso de recursos didácticos, el interés por la materia, entre otras.

El análisis de estos factores puede ayudar a las instituciones educativas a tomar decisiones más informadas para mejorar el desempeño académico de los estudiantes. Por ejemplo, se pueden identificar grupos de estudiantes que necesitan una atención más personalizada o diseñar programas específicos para mejorar la motivación y el compromiso de los estudiantes con su formación.

En resumen, el uso de técnicas de estadística multivariante y de aprendizaje automático para analizar los factores que inciden en el rendimiento académico de los estudiantes universitarios puede proporcionar valiosa información para la toma de decisiones educativas y mejorar la calidad de la enseñanza.

Debido a este panorama presentado, consideramos la siguiente pregunta general: ¿Qué tan bien el aprendizaje automático puede predecir el rendimiento académico de los estudiantes universitarios?

En este estudio se comparan algoritmos de aprendizaje automático para predecir la nota media y tasa de eficiencia con la más alta precisión posible, de tal forma que se identifique qué alumnos son los que tendrán un rendimiento académico más bajo. También podemos decir que contribuirá de manera social a los alumnos ya que ayudará al desarrollo profesional y en el futuro de los alumnos, así estos tendrán una mejor calidad de vida.

Por todo esto se va a tratar de aplicar aprendizaje automático para predecir la nota media y la tasa de eficiencia de los estudiantes de grado universitario. Este objetivo nos permite plantearnos la hipótesis: El aprendizaje automático puede predecir la nota y la tasa de eficiencia de los alumnos universitarios. Basada en esta también nos surgen otras hipótesis o cuestiones cómo puede el aprendizaje automático predecir con precisión la nota media y la tasa de eficiencia o predecir con sensibilidad y especificidad el rendimiento académico de los alumnos universitarios.

1.1 Motivación

La educación es un aspecto fundamental en la vida de las personas, tanto para ellas misma, para poder tener un futuro mejor y más próspero, como para un país, ya que consigue mejorar el nivel de vida de todas las personas que viven en él. Las dificultades académicas hacen que los alumnos soporten una profunda insatisfacción personal, desmotivación e incluso puede hacer que la autoestima de los alumnos baje.

El bajo rendimiento se convierte en un proceso en el que no se avanza y se mueve en círculos a ningún destino, la desmotivación hace que no se tenga un buen rendimiento y eso hace a su vez que los alumnos sigan sin motivación. Unas malas notas pueden aparecer por distintas causas que pueden depender o no del alumno, por ello si se consigue identificar esas causas y poder predecir qué alumnos tendrán ese mal rendimiento respecto a los demás compañeros, se puede ayudar y apoyar a estos alumnos para que consigan alcanzar ese rendimiento académico y poder así optar a un futuro mejor.

Por todo ello la realización de este tipo de investigaciones puede aportar información valiosa tanto para la comunidad académica como para la sociedad en general, ya que permiten identificar los factores que influyen en el rendimiento académico de los estudiantes universitarios y, por tanto, desarrollar estrategias y programas que contribuyan a mejorar la calidad de la educación y el rendimiento de los estudiantes.

Además, la aplicación de técnicas de estadística multivariante y aprendizaje automático en la investigación permite obtener una perspectiva más completa y detallada de los datos, lo que permite identificar patrones y relaciones entre los diferentes factores que pueden afectar al rendimiento académico.

1.2 Objetivos

1.2.1 Objetivo general

El objetivo principal de este trabajo es desarrollar un modelo de predicción del rendimiento académico en estudiantes universitarios y por tanto determinar la capacidad de los algoritmos de aprendizaje automático aplicados a un conjunto de datos de estudiantes de grado de la Universitat Politécnica de València para predecir su rendimiento académico. A partir de un análisis exhaustivo de las variables disponibles en el conjunto de datos, se busca identificar aquellas que tienen una mayor influencia en la nota media de los estudiantes y, posteriormente, aplicar los algoritmos de aprendizaje automático para construir modelos predictivos precisos. Se pretende evaluar la capacidad de estos modelos para predecir la nota media de los estudiantes con un alto grado de precisión, lo que permitiría tomar medidas preventivas y proactivas para mejorar el rendimiento académico de los estudiantes.

1.2.2 Objetivos específicos

Para lograr este objetivo general, se plantearán una serie de objetivos específicos que permitirán guiar el proceso de investigación y análisis de datos.

Con la realización de estos objetivos específicos, se espera obtener un modelo de predicción del rendimiento académico en estudiantes universitarios que permita identificar factores críticos y tomar medidas preventivas para mejorar el desempeño académico de los estudiantes. Estos objetivos son:

1. Revisión científica de los factores que influyen en rendimiento**.** Realizar una revisión científica de los factores que influyen en el rendimiento académico tiene como propósito conocer y comprender los diferentes factores que pueden influir en el desempeño de los estudiantes universitarios. Para llevar a cabo este objetivo, se realizará una búsqueda exhaustiva de la literatura científica relacionada con el rendimiento académico en estudiantes universitarios. Se identificarán y recopilarán artículos científicos, tesis y otros estudios que hayan investigado los diferentes factores que influyen en el rendimiento académico, tales como el nivel socioeconómico y los factores relevantes.

Una vez recopilada la información, se llevará a cabo una revisión crítica y análisis de los estudios seleccionados, con el fin de identificar los factores más significativos y determinar cómo se relacionan entre sí.

Al finalizar este objetivo**,** se espera haber obtenido una visión completa y detallada de los diferentes factores que influyen en el rendimiento académico en estudiantes universitarios, lo cual servirá como base para el desarrollo del modelo de predicción del rendimiento académico y la identificación de estrategias efectivas para mejorar el desempeño académico de los estudiantes.

2. Especificar y estimar modelos predictivos basado en la revisión científica y relaciones observadas entre variables a nivel descriptivo, tiene como propósito desarrollar un modelo de predicción del rendimiento académico en estudiantes universitarios utilizando técnicas

de aprendizaje automático. Para lograr este objetivo, se utilizará la información recopilada en la revisión científica y en el análisis exploratorio de los datos, para identificar las variables más relevantes que influyen en el rendimiento académico de los estudiantes.
A partir de la selección de las variables más relevantes, se definirán y especificarán modelos predictivos utilizando técnicas de aprendizaje automático, como los algoritmos de máquinas de vectores de soporte, k vecinos más cercanos, PLS y árboles de decisión, entre otras.

3. Seleccionar el modelo que mejor ajusta y es coherente con investigaciones anteriores supone tener como propósito elegir el modelo de predicción del rendimiento académico en estudiantes universitarios que mejor se ajusta a los datos y es coherente con los resultados de investigaciones anteriores. Para lograr este objetivo, se llevará a cabo una evaluación de los diferentes modelos desarrollados en el proyecto, utilizando técnicas de validación y comparación de modelos, como la validación cruzada y el análisis de las métricas de precisión del modelo.

   Además, se comparará el modelo seleccionado con los resultados de investigaciones anteriores, para determinar si el modelo se ajusta a los hallazgos de la literatura científica y es coherente con los factores que se sabe que influyen en el rendimiento académico de los estudiantes universitarios.

   Al finalizar este objetivo, se espera haber seleccionado el modelo de predicción del rendimiento académico en estudiantes universitarios que mejor se ajusta a los datos y es coherente con los hallazgos de investigaciones anteriores, lo que permitirá identificar con antelación a los estudiantes en riesgo de bajo rendimiento académico y tomar medidas preventivas para mejorar su desempeño académico.

2. Marco teórico

En esta sección, se aborda el primer objetivo del trabajo, que consiste en desarrollar un marco teórico sólido para el estudio de los factores que influyen en el rendimiento académico de los estudiantes universitarios y en la predicción de dicho rendimiento.

El marco teórico se centra en la revisión de la literatura científica relevante, con el fin de identificar los factores que han sido identificados como influyentes en el rendimiento académico de los estudiantes universitarios, y de esta manera, proporcionar una base sólida para la selección de las variables que se utilizarán en el modelo de predicción.

En conjunto, el marco teórico proporciona una visión amplia y detallada de los factores que influyen en el rendimiento académico de los estudiantes universitarios, lo que permite una selección adecuada de las variables a utilizar en el modelo de predicción y una mejor comprensión de los factores que afectan el éxito académico de los estudiantes universitarios.

2.1 El Espacio Europeo de Educación Superior y el Sistema Universitario Español

La convergencia de sistemas educativos universitarios supuso el comienzo del Espacio Europeo de Educación Superior -EEES-, desencadenando una serie de reformas a nivel curricular, organizativo y estructural (Díez et al., 2011).

En nuestro país, los estudios de grado y máster se regían por el Real Decreto 1393/2007 por el que se establece la ordenación de las enseñanzas universitarias oficiales y por el Real Decreto 43/2015 que lo modifica. Actualmente está vigente el Real Decreto 822/2021 de organización de las enseñanzas universitarias y procedimiento de aseguramiento de la calidad.

La creación de este EEES establece dos objetivos estratégicos que son: aumentar el nivel de empleo en la Unión Europea y un sistema de educación de formación superior que suponga un polo de atracción para estudiantes y profesorado de todo el mundo. En nuestro país, la implantación se concretó en su totalidad en el curso 2010/2011.

Sin embargo, en la creación del EEES el paso definitivo viene marcado por la Declaración de Bolonia de 1999, aunque como paso previo hay que citar la reunión celebrada el 25 de mayo de 1998 por los ministros de educación de Francia, Alemania, Italia y Reino Unido que firmaron en la Sorbona una Declaración que buscaba el desarrollo de un EEES.

En la Declaración de Bolonia de 1999, en la que participaron 30 estados europeos se estableció el año 2010 como plazo para la realización del EEES y en esta declaración donde además se sentaron las bases para poder construir el EEES en base a unos principios de evaluación, acreditación, movilidad, diversidad y competitividad y unos objetivos que son:

- Lograr un sistema fácilmente legible y comparable de titulaciones.
- La adopción de un sistema basado, fundamentalmente, en dos ciclos principales (grado y postgrado).
- El establecimiento de un sistema común de créditos, como el sistema europeo de transferencia de créditos, ECTS.
- La promoción de la cooperación europea para asegurar niveles de calidad y metodologías comparables.
- La promoción de la movilidad para los estudiantes, profesores y personal.

A continuación, se citan las distintas reuniones que marcan la cronología del Espacio Europeo de Educación Superior:

- Declaración de la Sorbona (mayo de 1998)
- Declaración de Bolonia (junio de 1999)
- Comunicado de Praga (mayo de 2001)
- Comunicado de Berlín (octubre de 2003)
- Comunicado de Bergen (mayo de 2005)
- Comunicado de Londres (mayo de 2007)
- Comunicado de Leuven/Louvain-la-Neuve (abril de 2009)
- Comunicado de Budapest y Viena (marzo de 2010)
- Comunicado de Bucarest (abril de 2012)
- Comunicado de Ereván (mayo de 2015)
- Comunicado de París (mayo de 2018)

Actualmente, 48 países integran el EEES. El Proceso de Bolonia está en consonancia con los objetivos de la UE de crear un Espacio Europeo de Educación a alcanzar en 2025 para promover la movilidad y el reconocimiento académico de los ciudadanos europeos.

Así pues, el cambio supone un nuevo reto en el proceso de enseñanza aprendizaje de los estudiantes universitarios, donde es necesario cambiar los roles de docentes y estudiantes para

alcanzar estos objetivos y plantear nuevos retos para el futuro de la enseñanza superior. De esta manera, resultará necesario crear un sistema de becas y ayudas al estudio suficiente y potente, que garantice el acceso a los estudios universitarios de grado y posgrado a todos los ciudadanos en igualdad de oportunidades, y que facilite la movilidad.

Todo ello buscando las nuevas metodologías que nos permitan la transformación de nuestro sistema educativo, pasando de un sistema basado en la "enseñanza" a otro basado en el "aprendizaje" y siempre implicando al estudiante, con metodologías más activas y con un cambio de rol del profesorado. En este cambio las universidades y estudiantes deben ser socios activos e implicarse en la docencia e investigación.

El sistema universitario español es un conjunto de instituciones académicas que ofrecen estudios universitarios en todo el territorio nacional. Está regulado por el Ministerio de Universidades y se compone de universidades públicas y privadas, así como de centros de investigación y de formación superior no universitaria.

En particular el sistema universitario español está compuesto por un total de 83 universidades de las cuales 50 son públicas y 33 son privadas distribuidas por toda la geografía del país. Estas universidades ofrecen una amplia variedad de estudios, desde grados y posgrados hasta programas de doctorado e investigación.

El sistema universitario español se rige por la Ley Orgánica de Universidades (LOU) y, posteriormente, por la Ley de la Ciencia, la Tecnología y la Innovación (LCTI), ambas diseñadas para garantizar la calidad, la excelencia y la competitividad de las universidades españolas en el contexto europeo e internacional.

Las universidades españolas se organizan en torno a facultades, escuelas y departamentos, y ofrecen estudios en todas las áreas de conocimiento, desde humanidades y ciencias sociales hasta ciencias, tecnología e ingeniería, pasando por ciencias de la salud y arte y diseño.

Para acceder a la universidad en España, los estudiantes deben superar una prueba de acceso, conocida como Selectividad o EBAU (Evaluación de Bachillerato para el Acceso a la Universidad), que evalúa los conocimientos adquiridos durante la educación secundaria y el bachillerato. También existen otras vías de acceso a la universidad, como la acreditación de experiencia laboral o la realización de estudios de formación profesional.

En cuanto a la financiación de la educación universitaria, en España se aplica un sistema mixto, en el que las universidades públicas reciben financiación del Estado y las comunidades autónomas, mientras que las universidades privadas se financian principalmente mediante el pago de matrículas y donaciones.

En general, el sistema universitario español cuenta con una buena reputación internacional y atrae a estudiantes de todo el mundo que buscan una educación de calidad en un ambiente multicultural y enriquecedor.

Por tanto, cabe destacar que las universidades juegan un papel muy importante en el desarrollo socioeconómico de un país, ya que son entidades que se ocupan de transmitir el conocimiento y fomentar la investigación.

Los datos reflejan que desde los años 80, se ha producido un aumento de la demanda en educación superior, y, además, la mayoría de los estudiantes optan por la enseñanza pública. La competitividad también existe en el sistema universitario, y por ello el nivel de calidad y eficiencia de los estudios cursados en muchos casos resulta determinante a la hora de optar al mercado laboral. Y es notable que en términos de eficacia las universidades públicas españolas llevan la delantera a las privadas.

2.2 Oferta educativa Universitaria

A partir de la información disponible en el Informe Datos y Cifras del sistema Universitario Español, publicación 2021-2022, podemos ver a través de los datos en los puntos siguientes cuál es la situación del Sistema Universitario en España.

2.2.1 Universidades y centros

Según datos del Ministerio de Universidades, actual responsable de la de educación superior en España, el Sistema Universitario Español (SUE) en el curso 2020-2021 se contó con 84 universidades con actividad, 50 públicas y 34 privadas. Las primeras, han sido creadas por una entidad pública, y, por tanto, su financiación viene dada principalmente por el Estado y de acuerdo con la Comunidad Autónoma donde vaya a establecerse y las privadas son creadas por instituciones a título personal, y por tanto, no se ven afectadas por cambios políticos ni recortes en los presupuestos, ya que no están reguladas por el Gobierno. A su vez, las universidades públicas y privadas se dividen en universidades presenciales y no presenciales. Éstas se distribuyen en un total de 168 municipios, de los cuales en 145 está presente la universidad pública, mientras que la privada está en unos 49 municipios.

Aparte hay un total de 1.067 centros universitarios, contando escuelas y facultades, 544 institutos universitarios, 52 escuelas de doctorado, 56 hospitales universitarios y 77 fundaciones.

2.2.2 Oferta de estudios universitarios

La estructura de las enseñanzas oficiales se recoge en el artículo 37 de la Ley Orgánica de Universidades, que actualmente se encuentra en revisión, y se dividen en tres ciclos: Grado, Máster y Doctorado. En el caso de grado, este tiene una duración mínima de 240 créditos ECTS (4 años) y su finalidad es una formación general del estudiante, orientada al ejercicio de actividades profesionales. Normalmente se dividen en cinco áreas o ramas de conocimiento: Ciencias Sociales y Jurídicas, Ingeniería y Arquitectura, Artes y Humanidades, Ciencias de la Salud y Ciencias. Para el Máster la duración de estos títulos están entre 60 y 120 créditos ECTS, equivalente a uno o dos cursos académicos, y su finalidad es la formación avanzada del estudiante, orientada a la especialización académica o profesional, o bien a promover tareas investigadoras. Finalmente, los estudios de Doctorado tienen como finalidad la formación avanzada del estudiante

en técnicas de investigación, incorporando cursos, seminarios o cualquier otra actividad enfocada a la formación investigadora.

En lo referente a la oferta de estudios universitarios, en el curso de 2020-2021 se presentaron 3.062 titulaciones de grado, de las que un 73,3 % eran de universidades públicas. Desde el año 2012 este número de titulaciones impartidas ha ido creciendo a un ritmo asombroso.

La rama de Ciencias sociales y Jurídicas fue la rama con un número más grande de titulaciones de Grado impartidas con un total de 1.093 titulaciones de Grado. Por otro lado, la rama de Ciencias cuenta con un número bastante menor de titulaciones impartidas con un total de 258.

La proporción de grados impartidos en universidades públicas es muy distinta según la rama que se imparte, esta variación va del 91% en Ciencias a poco más del 60 % en Ciencias Sociales y Jurídicas y Ciencias de Salud.

2.2.3 Acceso a la Universidad

En el año 2020 en las pruebas de Acceso a la universidad (PAU) se registraron 322.823 estudiantes, lo que supone un ascenso del 9,5 % respecto al año anterior. La convocatoria de 2020 fue la convocatoria con mayor número de matriculados de los últimos años. De los estudiantes registrados, 306.820 se presentaron y finalmente de esos, aprobaron 274.278. En el resto de las pruebas (para mayores de 25 años, mayores de 45 años y mayores de 40 con experiencia laboral) se matricularon 28.177 estudiantes de los cuales 17.262 se presentaron y finalmente 10.190 aprobaron.

En estas pruebas de acceso a la educación superior, las mujeres son mayoría, la proporción de mujeres matriculadas respecto del total de los matriculados fue de un 56,9%. Pero en las pruebas de mayores de 25 y 45 son los hombres los que tienen la mayoría de los matriculados.

Si miramos las PAU de forma genérica, observamos que el 57,8 % de los matriculados eran mujeres, el 5,2% tenía nacionalidad extranjera y el 7,8 % era mayor de 20 años.

Respecto a las comunidades autónomas, las que obtuvieron una proporción de aprobados más altas fueron el País Vasco y la Comunidad Valenciana con un 97% y 96% respectivamente, mientras que los porcentajes más bajos de aprobados se dieron en Galicia, Canarias y Extremadura con 88%, 89% y 90% respectivamente.

2.3 Estudiantes en el Sistema Universitario Español

A partir de la información disponible en el Informe Datos y Cifras del sistema Universitario Español, publicación 2021-2022, en este caso podemos observar los siguientes datos.

2.3.1 Matriculación y Tasa de Escolarización

Durante el curso 2020-2021 los matriculados en universidades de España fueron 1.679.518 repartiéndose en un 80% en grado, un 15% en máster y un 5% en Doctorado.

Las universidades no presenciales tuvieron 291.165 matriculados lo que supone el 17,4% del total de alumnos.

La tasa neta de escolarización en Educación Universitaria es una tasa que mide la proporción de la población de entre 18 y 24 años que está cursando estudios de Grado o Máster. Para el curso 2020-2021 esta tasa se sitúa en 32%, lo que quiere decir que uno de cada tres personas de entre 18 y 24 años realiza o por lo menos está matriculado en una titulación universitaria. Como en el Acceso a la Universidad encontramos diferencias importantes en este tipo de tasas de escolarización entre las distintas Comunidades Autónomas, esto se debe a los alumnos que estudian fuera de su comunidad autónoma de origen. Haciendo así que las comunidades autónomas que reciben más alumnos sean las que incrementen sus tasas y por el contrario las comunidades autónomas que más alumnos se marchan disminuyen sus tasas de escolarización.

2.3.2 Matriculados y Egresados

De los matriculados del curso 2020-2021, un 55,6% eran mujeres y si miramos el nivel académico. El porcentaje de las mujeres es distinto según las ramas de estudio. Para alumnos matriculados en el curso 2020/2021 hubo una gran proporción (71,4%) que escogió Ciencias de la Salud y por otro lado hubo una proporción mucho menor en Ingenierías y Arquitectura (25,7%).

2.3.3 Estudiantes de Grado

2.3.3.1 Movilidad Interna de los Estudiantes de Grado

En torno al 90% de los estudiantes que hacen las PAU en Madrid, Valencia y Cataluña no se van fuera de su comunidad autónoma para matricularse en una universidad. Pero en el caso de Castilla-La Mancha, Baleares y Extremadura son solo el 50 % de los estudiantes los que se quedan para matricularse en una universidad de su comunidad autónoma.

Por otro lado, Navarra, Castilla y León y Madrid son las comunidades autónomas que más estudiantes reciben de fuera de su comunidad autónoma, con un 30 % de los estudiantes con la residencia habitual fuera de la comunidad.

Hablando en términos generales, el 16,8% de los alumnos matriculados en universidades españolas tiene su residencia habitual fuera de la comunidad autónoma donde se encuentra su universidad. También son el 29,6% de alumnos los que estudian en una provincia distinta de la suya dentro de la misma comunidad autónoma.

2.3.3.2 Preinscripción en Estudios de Grado

La tasa de estudiantes que pudieron matricularse en su primera opción de titulación fue de 73,9% en el curso de 2019-2020.

La tasa de ocupación de las titulaciones, séase las plazas ofertadas cubiertas, fue distinto dependiendo de la rama de enseñanza, en Ciencias de la Salud fue del 100,6% y en Artes y Humanidades fue del 86,9%.

Respecto a la tasa de preferencia que recoge las solicitudes en primera opción por cada plaza ofertada destaca la rama de las Ciencias de la Salud con un 350%. (más de tres alumnos por plaza ofertada).

En general, a lo largo de los últimos años las matrículas de nuevo ingreso y la oferta de plazas se mantienen estables. Sin embargo, entre los cursos 2010-2011 y 2019-2020 se observan variaciones por rama de enseñanza, si en Ingeniería y Arquitectura cae un 7,9% la matrícula de nuevo ingreso en Ciencias aumenta un 10,5%.

En estas ramas de enseñanza y en los ámbitos de estudio se observan grandes diferencias entre notas medias de admisión y notas de corte. Los ámbitos en los que la nota media de admisión fue mucho mayor que la nota de corte fueron los relacionados con las ramas de las Ciencias y las Ciencias de la Salud.

2.3.3.3 Matriculados y Egresados en Grado

Con la información disponible referente al curso 2020-2021 se puede destacar el aumento en el número de personas matriculadas en Grado, aunque ya venía creciendo en los anteriores años el curso 2020-2021 tiene matriculados a 44.253 estudiantes más que el anterior.

Un estudiante egresado se conoce como una persona que ha cursado y aprobado satisfactoriamente la totalidad del plan de estudios reglamentado para un programa o una carrera, pero que aún no ha recibido el título académico. El número de estudiantes egresados en el curso de 2019-2020 en grado fue de 208.345.

El 51% de los estudiantes tienen entre 18 y 21 años, sin embargo, esta cifra baja en la rama del Arte y Humanidades (47,2%) y sube considerablemente en la rama de ciencias (63,5%). En el caso de los estudiantes egresados, se destaca que el 4% tienen más de 40 años y en Ciencias únicamente encontramos el 0.8% de estudiantes de este tipo.

2.4 Rendimiento académico en la universidad española

El rendimiento académico del estudiantado universitario constituye un factor imprescindible en el abordaje del tema de la calidad de la educación superior, debido a que es un indicador que permite una aproximación a la realidad educativa (de Miguel Díaz et al, 2002).

El análisis del rendimiento académico además permite predecir posibles resultados académicos lo que hace que sea una herramienta de gran utilidad para la toma de decisiones que permitan mejorar los resultados académicos.

Existen trabajos académicos previos que tratan de determinar qué factores pueden afectar al rendimiento. Así pues, tal y como define Garbanzo Vargas (2007), existen múltiples determinantes internos y externos al estudiante que están asociados a su rendimiento académico.

Y de esta manera establece tres grupos de determinantes, personales, sociales e institucionales. El estudio trata de determinar que los factores que se recogen en cada uno de los grupos inciden en el rendimiento académico del estudiante y puede presentar interrelaciones que se producen entre sí.

En primer lugar, en los determinantes personales, encontramos factores de índole personal tales como: competencia cognitiva, motivación, condiciones cognitivas, auto eficiencia percibida, bienestar psicológico, satisfacción y abandono respecto a los estudios, asistencia a clases, inteligencia, aptitudes, sexo, formación académica previa a la Universidad y nota de acceso a la universidad

Como segundo grupo se citan los determinantes sociales, que son factores asociados al rendimiento académico de índole social que interactúan con la vida académica del estudiante y cuyas interrelaciones se pueden producir entre sí y entre variables personales e institucionales. En los determinantes sociales encontramos los siguientes factores: las diferencias sociales, el entorno familiar, el nivel educativo de los progenitores o adultos responsables del estudiante, nivel educativo de la madre, contexto socioeconómico y variables demográficas.

En último lugar tenemos los determinantes institucionales: esta categoría es definida por Carrión (2002), como componentes no personales que intervienen en el proceso educativo, donde al interactuar con los componentes personales influye en el rendimiento académico alcanzado. Estos factores se refieren a normas, condiciones, requisitos de ingreso y requisitos entre materias, entre otros factores que se hayan definido en la institución educativa. Estos factores institucionales asociados al rendimiento académico en estudiantes universitarios son: elección de estudios de interés del estudiante, complejidad en los estudios Institucionales, servicios Institucionales de apoyo, ambiente estudiantil, relación estudiante-profesor y pruebas específicas de ingreso a la carrera.

Otro estudio realizado por Esparza-Paz et al. (2020), establece que existen cuatro grupos de factores que afectan al rendimiento académico de los estudiantes. Estos factores estarían agrupados en factores socioeconómicos, factores escolares-académicos, factores familiares y factores personales. Y según refleja sus análisis los factores de mayor importancia e incidencia en el rendimiento académico de los estudiantes sobre los que se realizó la investigación son los factores socioeconómicos seguidos de los familiares, observando en último lugar con un efecto casi similar a los factores escolares o académicos y los factores personales.

En otros estudios previos tratan de determinar por qué la cantidad de estudio por parte de los estudiantes universitarios tiene o no un impacto mínimo en su rendimiento académico. El estudio se basa en las teorías de la práctica deliberada y el aprendizaje académico autorregulado. En el estudio resultó que las notas obtenidas en secundaria y en las pruebas de acceso a la universidad estaban positivamente relacionadas con la nota media de la universidad y explicaron la variabilidad independiente. La calidad del entorno en el que se estudia, cómo estudiar solo en un entorno tranquilo, aparece como un predictor significativo del rendimiento. El grado de planificación a largo plazo, asistir a clases y usar recursos como tutores, bibliotecas y grupos de estudio también mostraron una correlación positiva con la nota media de los estudiantes. El estudio mostró que la cantidad total de tiempo de estudio no era un predictor significativo del rendimiento, pero sí lo fue cuando se combinó con la calidad del entorno de estudio y con la puntuación obtenida en los exámenes de acceso. Por ello se sugiere que los estudiantes con notas de acceso más altas y que estudian en un entorno tranquilo pueden estudiar de manera más efectiva y obtener las mismas o mejores calificaciones con menos tiempo de estudio (Plant, et al, 2004).

Otra investigación en esta línea tiene por objetivo dar a conocer cuáles son las variables académicas y la importancia de éstas en el rendimiento académico de los estudiantes universitarios. A nivel nacional, se cuestiona si estos cambios de la educación superior en el último siglo han favorecido la calidad de la educación universitaria. Y se describen y analizan las variables académicas que influyen en el rendimiento de los estudiantes universitarios. Estas variables han sido validadas en diferentes investigaciones y su conocimiento y estudio puede contribuir en la mejora de la educación superior universitaria (Fernández, 2011).

Otros autores se centran en ver si existen relaciones positivas entre el apoyo social percibido, la inteligencia emocional y el rendimiento académico. En este caso los resultados muestran que la familia y las amistades son las principales fuentes de apoyo emocional y que el apoyo del profesorado y de los iguales es importante para mejorar el rendimiento académico. Por ello se recomienda fortalecer las relaciones de apoyo por parte del profesorado y entre el propio alumnado para mejorar el rendimiento académico (Lasarte et al, 2019).

2.5 El rendimiento de los estudiantes

En el conjunto de indicadores académicos a la hora de determinar el rendimiento de los estudiantes y la eficiencia de la universidad, destaca claramente la tasa de eficiencia de los alumnos matriculados. Aunque junto a estos indicadores también se podrían destacar la tasa de abandono del título y la tasa de graduación del título.

2.5.1 La tasa de rendimiento de Grado

La tasa de rendimiento se define como la relación porcentual entre el número total de créditos ordinarios superados por los estudiantes en un determinado curso académico y el número total de créditos ordinarios matriculados por los mismos.

Con relación a las tasas de rendimiento, según indica el informe de cifras y datos de la Universidad Española del curso 2021-2022, encontramos los datos comentados a continuación.

La tasa de rendimiento de los alumnos en el curso de 2019-2020, último dato disponible, fue de 84,6%, diferenciando por titulaciones el mínimo rendimiento dado fue en la rama de ingeniería y arquitectura con un 74,4% y el máximo rendimiento dado fue en la rama de las ciencias de la salud con un 90,6%.

Esta tasa de rendimiento se ha visto incrementada en este curso debido probablemente a la pandemia por el COVID 19 y la docencia online.

Hacer notar la relación de la nota de admisión al título de grado y la tasa de rendimiento. Estamos pasando de un 71% de tasa de rendimiento de los estudiantes que entraron con notas más bajas a un 95 % de tasa de rendimiento de estudiantes que entraron con las mayores notas de admisión.

2.5.2 La tasa de eficiencia de los alumnos de grado

Esta tasa se refiere a la relación porcentual entre el número total de créditos teóricos del plan de estudios a los que debieron haberse matriculado a lo largo de sus estudios el conjunto de estudiantes graduados en un determinado curso académico y el número total de créditos en los que realmente han tenido que matricularse.

La tasa de eficiencia en total de todas las instituciones universitarias de España es de un 89,7%, siendo en las universidades privadas esta tasa superior a la de las universidades públicas con un 93,5% y un 89% respectivamente. En cuanto a la diferencia entre ramas, la rama con mayor tasa de eficiencia es Ciencias de la Salud mientras que la menor tasa de eficiencia está en las titulaciones de Ingeniería y Arquitectura.

2.5.3 La nota media de alumnos de grado

La media del expediente académico de cada alumno será el resultado de aplicar la suma de los créditos obtenidos por los alumnos multiplicados cada uno de ellos por el valor de las calificaciones que correspondan y dividida por el número de créditos totales por el alumno.

En España, la nota media de los alumnos del curso 2019-2020 es de 7,27, diferenciando entre universidad públicas y privadas existe una pequeña superioridad de la nota media por parte de las universidades privadas frente a las públicas con un 7,35 y un 7,24 respectivamente. Diferenciando por ramas, las Ciencias de la Salud siguen siendo la rama con más nota media con un 7,44 y por otro lado la Ingeniería y Arquitectura tiene la nota media más baja con un 6,86.

2.6 Papel de la inteligencia artificial en el análisis de datos

La inteligencia artificial (IA) se está utilizando cada vez más en el análisis de datos educativos para ayudar a mejorar el proceso educativo y el rendimiento académico de los estudiantes. La IA puede procesar grandes cantidades de datos de forma rápida y eficiente, lo que permite a los educadores tomar decisiones informadas sobre cómo mejorar la enseñanza y el aprendizaje.

Algunas de las formas en que la IA se está utilizando en el análisis de datos educativos son:

1. *Predicción del rendimiento académico*: La IA puede analizar datos históricos de los estudiantes, como las calificaciones, el progreso del aprendizaje y la asistencia, para predecir su rendimiento académico futuro. Esto permite a los educadores intervenir tempranamente si un estudiante está en riesgo de no alcanzar sus objetivos.

2. *Personalización del aprendizaje:* La IA puede analizar los datos de los estudiantes para identificar patrones en su aprendizaje y ofrecer recomendaciones personalizadas. Esto puede incluir materiales de aprendizaje adaptados a su nivel, sugerencias de actividades complementarias, y recomendaciones de tutorías y asesoramiento.

3. *Análisis de datos en tiempo real:* La IA puede analizar los datos en tiempo real, lo que permite a los educadores detectar patrones y tendencias en el proceso educativo de los estudiantes de forma inmediata. Esto les permite intervenir en tiempo real y ajustar la enseñanza para mejorar el proceso educativo.

4. *Detección de plagio*: La IA puede analizar los trabajos y las tareas de los estudiantes para detectar el plagio y prevenir el fraude académico.

En resumen, la IA se está convirtiendo en una herramienta cada vez más valiosa para el análisis de datos en el proceso educativo. La IA puede ayudar a los educadores a personalizar el aprendizaje, intervenir tempranamente y mejorar el rendimiento académico de los estudiantes.

Por lo que se observa que además la inteligencia artificial está teniendo un impacto significativo en la educación. Resaltando los siguientes ítems:

1. *Personalización de la enseñanza*: la IA permite adaptar el aprendizaje a los estudiantes individuales y proporcionar una experiencia de aprendizaje personalizada.

2. *Evaluación automatizada:* la IA puede realizar evaluaciones más rápidas y precisas, lo que permite una retroalimentación más oportuna y precisa para los estudiantes.

3. *Accesibilidad a materiales educativos:* la IA ayuda a expandir el acceso a materiales educativos de alta calidad a través de plataformas en línea.

4. *Mejora de la eficiencia*: la IA permite un uso más eficiente del tiempo y los recursos, lo que significa que los profesores pueden enfocarse en tareas más valiosas como la mentoría y el desarrollo de habilidades sociales y emocionales de los estudiantes.

Sin embargo, es importante señalar que la IA no es una solución completa y única y que su uso en la educación debe ser cuidadosamente considerado. La IA debe ser utilizada para complementar y mejorar la enseñanza humana, no para reemplazarla.

3. Metodología

3.1 Conjunto de datos

3.1.1 Recopilación de los datos

La recopilación de datos es un paso fundamental en cualquier proyecto de investigación. En concreto, estos datos fueron proporcionados por la Universidad Politécnica de Valencia con el fin especifico de que sean tratados y utilizados únicamente para el fin de este trabajo final de grado.

3.1.2 Descripción de los datos

El conjunto de datos está formado por una población de 5035 alumnos de los cuales se conocen 43 características de carácter demográfico, económico y social. Entre estas variables tenemos:

- DNI: documento nacional de identidad del alumno.
- Título: Grado
- TIT: Titulación del grado específico
- FECHA_TF_YYYYMMDD: fecha de finalización del grado por parte del alumno.
- Cohorte: promoción del alumno.
- MEDIA_CON_PFC: media con su proyecto final de carrera.
- MEDIA_SUS: media contando asignaturas suspendidas.
- MEDIA_SUS_NP: media contando asignaturas suspendidas y no presentadas.
- ORDEN_PROMOCION: orden por nota de toda la promoción.
- MEDIA_OFICIAL: media oficial del alumno en el grado.
- DUR_REAL: tiempo que tardó el alumno en completar el grado.
- CRE_MAT: créditos matriculados por el alumno.
- CRE_SUP: créditos superados por el alumno.
- T_EFI: tasa de eficiencia del alumno.
- PAIS_FAMILIA: País de la familia del alumno.

- FamExtrenjera: Si la familia es de nacionalidad española o extranjera.
- PROVINCIA_FAMILIA: código de la provincia donde vive la familia del alumno.
- FueraProvValencia: Si la familia vive en la provincia de Valencia o no.
- POSTAL_FAMILIA: código postal de la familia del alumno.
- POSTAL_ALUMNO: código postal del alumno únicamente.
- IVEL_ESTUDIOS_PA: nivel de estudios del Padre del alumno.
- NIVEL_ESTUDIOS_MA: nivel de estudios de la madre del alumno.
- TRABAJO_A: el nivel de cualificación del alumno para el trabajo.
- TrabajoAlumno: si el alumno trabaja o no trabaja.
- TRABAJO_A2: que tipo de trabajo ejerce.
- VOLUNTARIADO_A: si el alumno ha realizado algún tipo de voluntariado o no.
- TRABAJO_P: tipo de trabajo del padre del alumno.
- TRABAJO_M: tipo de trabajo de la madre del alumno.
- Sexo: hombre o mujer
- ING_INGRESO: vías de ingreso del alumno al grado.
- ING_CUPO: vía de acceso al grado del alumno.
- FP: si accedió desde FP o no.
- ING_EST: estudios previos
- ING_NOTA: nota de acceso al grado.
- ANY_ING: año de ingreso en la universidad.
- UPV_DESDE: Desde cuando el alumno pertenece a la UPV.
- EDAD: edad del alumno al presentar el TFG.

- DOMICILIO: donde vive el alumno
- TUTORIZADO _SN: si el alumno disponía de un tutor que favoreció su integración en la universidad.
- TutorPATU: si ha sido tutor de PATU
- ISEC_Padre (índice socioeconómico): índice socioeconómico del padre.
- ISEC_Madre: índice socioeconómico de la madre.
- ISEC_Ambos: índice socioeconómico de ambos padres.

El ISEC (Índice Socioeconómico de la Familia) es una medida utilizada en investigación social y de salud para evaluar el nivel socioeconómico de una familia. El ISEC combina información sobre la educación, el empleo y el ingreso de los padres de una familia para proporcionar una medida global del estatus socioeconómico.

El cálculo del ISEC se basa en la posición de los padres en la estructura social, y se calcula utilizando información sobre su educación, empleo y nivel de ingresos. Por lo general, se utiliza una fórmula que asigna puntos a cada uno de estos tres factores y los suma para obtener una puntuación total de ISEC. La fórmula específica utilizada para calcular el ISEC puede variar dependiendo del contexto y del país.

En algunos casos, el ISEC se calcula utilizando información sobre el hogar en su conjunto (por ejemplo, el ingreso familiar total), mientras que en otros casos se utiliza información específica sobre los padres (por ejemplo, la educación de cada uno de ellos). En el caso del ISEC combinado de ambos padres, se utiliza la información de ambos padres para calcular una puntuación conjunta que refleje el estatus socioeconómico de la familia en su conjunto.

Es importante tener en cuenta que el ISEC es una medida simplificada y limitada del estatus socioeconómico de una familia, y no debe utilizarse como una medida definitiva o única de la desigualdad socioeconómica. Sin embargo, puede ser útil para comparar el estatus socioeconómico de diferentes familias en un contexto determinado.

3.1.3 Perfil del participante

En el presente conjunto de datos de alumnos de las cohortes de 2013 y 2014, esto quiere decir que los datos que se recogen es de alumnos que comenzaron sus estudios universitarios en los años 2013 y 2014, se incluyen alumnos graduados con edades comprendidas entre los 21 y los 66 años, lo que sugiere que se trata de un grupo bastante heterogéneo en cuanto a la edad se refiere. Además, es importante destacar que el conjunto de datos se incluye tanto a estudiantes que residen

en la provincia como a aquellos que no lo hacen, lo que a su vez puede influir en factores como el nivel socioeconómico o la distancia a la universidad, entre otros.

Otro aspecto interesante que considerar es la presencia de estudiantes de familia extranjera, lo que añade una dimensión multicultural y diversa al perfil de los participantes en el conjunto de datos. En general, el perfil de los participantes en este conjunto de datos es bastante amplio y diverso en cuanto a edad, lugar de residencia y origen familiar, lo que sugiere que se pueden encontrar resultados y patrones muy diversos al analizar las variables del conjunto de datos.

3.2 Preprocesado de los datos

Esta etapa de preprocesamiento de datos será esencial, ya que los datos brutos raramente están en una forma que se pueda usar directamente para su análisis. Durante el preprocesamiento, se llevarán a cabo una serie de pasos para limpiar, transformar y normalizar los datos, de manera que sean adecuados para su uso en modelos de aprendizaje automático o análisis estadísticos. Algunos de los pasos que se incluirán en este preprocesamiento de datos serán la limpieza de datos incompletos o duplicados, la eliminación de valores atípicos o errores, la normalización y la estandarización de los datos, así como la selección de características relevantes.

3.2.1 Limpieza de los datos

3.2.1.1 Datos redundantes

El primer paso en el preprocesamiento de datos es la identificación y eliminación de datos redundantes. Los datos redundantes son aquellos que se repiten en el conjunto de datos y no aportan información adicional. La presencia de datos redundantes puede aumentar la complejidad del modelo, disminuir la eficiencia del algoritmo de aprendizaje automático y llevar a resultados engañosos.

Tras verificar si existen filas o columnas duplicadas, se observa que no existen ni filas ni columnas duplicadas.

3.2.1.2 Datos faltantes

El análisis y tratamiento de los datos faltantes es uno de los pasos fundamentales en el preprocesamiento de datos. Los datos faltantes son aquellos que no están disponibles en una o varias variables de un conjunto de datos. La presencia de datos faltantes puede ser problemática, ya que pueden causar sesgos en el análisis y la interpretación de los datos, y pueden afectar la eficacia de los modelos predictivos y descriptivos. Por lo tanto, el tratamiento adecuado de los datos faltantes es crucial para garantizar la calidad y la validez de los análisis de datos. En esta sección, se presentará un análisis exhaustivo de los diferentes tipos de datos faltantes y las técnicas para tratarlos de manera eficaz.

Tabla 1: Valores faltantes por variable

Variable	Recuento de valores
DNI	0
Título	0
TIT	0
FECHA_TF_YYYYMMDD	0
Cohorte	0
MEDIA_CON_PFC	0
MEDIA_SUS	0
MEDIAS_SUS_NP	0
ORDEN_PROMOCION	0
MEDIA_OFICIAL	0
DUR_REAL	0
CRE_MAT	0
CRE_SUP	0
T_EFI	0
PAIS_FAMILIA	0
FamExtranjera	0
PROVINCIA_FAMILIA	35
FueraProvValencia	35
POSTAL_FAMILIA	18
POSTAL_ALUMNO	10
NIVEL_ESTUDIOS_PA	17

NIVEL_ESTUDIOS_MA	17
TRABAJO_A	22
TrabajoAlumno	22
TRABAJO_A2	17
VOLUNTARIADO_A	59
TRABAJO_P	20
TRABAJO_M	18
Mujer	0
ING_INGRESO	0
ING_CUPO	1294
FP	1294
ING_EST	1296
ING_NOTA	1295
ANY_ING	0
UPV_DESDE	0
EDAD	0
DOMICILIO	0
TUTORIZADO_SN	0
TutorPATU	0
ISEC_Padre	0
ISEC_Madre	0
ISEC_Ambos	0

En esta tabla se muestra la cantidad de datos faltantes por cada variable del conjunto de datos, resaltar las variables ING_CUPO, FP, ING_EST y ING_NOTA que tiene un gran número de valores faltantes de entorno a los 1295. Como estas 4 variables tienen un número similar de datos faltantes, sería bueno investigar si es posible que tengan una relación y así tratarlos de manera adecuada.

Ahora bien, que ya han sido identificados, se analizara caso por caso para tratar de perder el menor número de información posible.

Para las variables con relativamente pocos datos faltantes en comparación con el tamaño conjunto de los datos, se ha decidido que la mejor forma de tratarlos es reemplazar los datos faltantes por la moda de esas variables evitando así la perdida innecesaria de información que sería eliminar todos los casos. Estas son las variables que han sido tratadas de esta manera:

- FueraProvValencia
- Trabajo_alumno,
- Provincia_familia,
- Postal_familia,
- Postal_alumno,
- trabajo_A,
- trabajo_A2,
- trabajo_padre,
- trabajo_madre,
- nivel_estudios_pa,
- Voluntariado_A,
- nivel_estudios_ma

De esta manera, se ha asegurado que los datos faltantes no afecten los resultados del análisis y se han obtenido valores más precisos y confiables para estas variables.

Para la variable ING_NOTA tenemos muchos valores faltantes, unos 1295, son muchos registros como para eliminarlos todos, así que para ING_NOTA que es la nota de ingreso podemos imputar usando KNN vecinos y así que escoja la media de los k vecinos más cercanos para sustituir al valor faltante o imputarse por la mediana pero aquí tendremos muchos valores con el mismo valor numérico y podría crear una distribución totalmente distinta a la que teníamos en los datos, por ello vamos a escoger la opción de imputar mediante KNN vecinos usando los 5 vecinos más cercanos.

Finalmente, tras analizar el caso de los valores faltantes en ING_CUPO, FP y ING_EST se percató de que las tres variables hacen referencia al modo de acceso a la universidad y que cuando existen valores faltantes en las tres variables es debido a que entraron por el cupo de FP, ya que cuando existe el valor faltante en FP, sí que hay un valor en ING_CUPO de FP por lo tanto deducimos que no se quiso redundar la información y que solo se ponía en una de las tres variables. Por ello simplemente tenemos que rellenar los datos faltantes con modo de ingreso FP y así obtendremos unos datos coherentes.

En conclusión, el tratamiento de los datos faltantes es esencial para garantizar la calidad de los datos y obtener resultados precisos y confiables en el análisis posterior.

Como vemos una vez aplicadas distintas maneras de tratar los datos faltantes nuestro conjunto de datos queda completo y sin huecos vacíos. Ahora ya que no queda ni un valor faltante en nuestro conjunto de datos podemos pasar al siguiente paso de la preparación de los datos para nuestros algoritmos.

3.2.2 Tratamiento de variables categóricas

El tratamiento de variables categóricas es un paso fundamental en cualquier análisis de datos que involucre información cualitativa. Las variables categóricas representan características que no pueden ser medidas numéricamente, sino que se describen en términos de etiquetas o categorías. En este caso, estas variables son utilizadas para codificar información sobre grupos o características de los individuos, como su género, nivel educativo nivel de trabajo o modo de acceso a la universidad y el domicilio. Sin embargo, para que estas variables sean útiles en el análisis de datos, es necesario transformarlas en una forma numérica, ya que muchos algoritmos y modelos de aprendizaje automático requieren datos numéricos como entrada. En este apartado se presentarán las técnicas más comunes para el tratamiento de variables categóricas, incluyendo la codificación one-hot, la codificación ordinal. Además, se discutirán las mejores prácticas para el tratamiento de variables categóricas en función del tipo de datos y los objetivos del análisis.

- FamExtranjera que es una variable que recoge si la familia es española o de nacionalidad extranjera, ya que solo hay dos categorías podemos usar una codificación one-hot creando una nueva variable binaria. En este caso una nueva variable que indicará con un 1 si la familia es española o con 0 si no lo es.

- FueraProvValencia que es una variable que recoge si el alumno reside fuera de la provincia de Valencia o en la propia provincia por lo que igual que antes usaremos una codificación one-hot como en el anterior caso.

- ISEC_Ambos una variable que recoge si el nivel socioeconómico de los padres del alumno es alto o es bajo y por tanto usaremos una codificación one-hot creado una variable binaria en la que 1 sea que el nivel es alto y 0 sea que el nivel es bajo.

- trabajaAlumno una variable que recoge si el alumno tiene un trabajo mientras estudia o no lo tiene y por tanto usaremos una codificación one-hot creado una variable binaria en la que 1 sea que no tiene trabajo y 0 sea que si tiene trabajo.

- Las variables DOMICILIO y ING_CUPO tienen más de dos categorías por lo que para esta cada categoría se asocia a un número que se habrá codificado de manera ordinal.

Se ha realizado una adecuada transformación y codificación de las variables categóricas para poder ser utilizadas en futuros análisis y modelos de aprendizaje automático. Se han utilizado técnicas de codificación one-hot y codificación ordinal para transformar las variables categóricas

en variables numéricas y así poder ser procesadas por los algoritmos de aprendizaje automático. También se ha realizado una exploración y análisis detallado de cada una de las variables categóricas para entender su impacto en la variable objetivo y poder determinar cuáles son las más relevantes para su inclusión en los modelos. En general, el tratamiento de variables categóricas ha sido fundamental para poder llevar a cabo un análisis riguroso y preciso de los datos.

3.3 Análisis de los datos

En este trabajo, se pretende llevar a cabo la creación y evaluación de distintos modelos predictivos basados en técnicas de aprendizaje automático. Concretamente, se explorarán los algoritmos SVR, KNN, árboles de decisión, Random Forest y PLS, con el objetivo de seleccionar el que presente mejor precisión en la predicción del rendimiento académico de los estudiantes universitarios. Además, se llevará a cabo la evaluación de la importancia de los predictores en cada modelo, lo que permitirá conocer qué variables tienen una mayor influencia en el resultado final.

En el desarrollo de este proyecto se utilizaron diversas herramientas y librerías de Python para el preprocesado, análisis de datos e implementación de modelos predictivos. A continuación, se presentan las principales librerías utilizadas:

- **Pandas**: para la manipulación y limpieza de datos, así como para la exploración inicial de los mismos.
- **NumPy**: para el manejo de arrays y operaciones matemáticas con los datos.
- **Matplotlib** y **Seaborn**: para la visualización y generación de gráficos descriptivos de los datos.
- **Statsmodels**: para realización de análisis estadísticos como ANOVA.
- **Scikit-learn**: proporciona una amplia selección de algoritmos de aprendizaje automático supervisados y no supervisados para problemas de clasificación, regresión, agrupamiento y reducción de la dimensionalidad, así como herramientas para la preparación y transformación de datos, validación de modelos y selección de modelos.

Estas librerías son ampliamente utilizadas en el análisis de datos y en la implementación de modelos predictivos en Python, y su utilización en este proyecto permitió una exploración detallada de los daos y la construcción de modelos robustos y precisos.

3.4 Algoritmos de aprendizaje supervisado

3.4.1 Fases de un proyecto de aprendizaje automático

Un proyecto de aprendizaje automático sigue ciertas fases esenciales para poder crear un modelo de aprendizaje automático efectivo. Estas fases están diseñadas para garantizar que los datos sean procesados adecuadamente, que los algoritmos de aprendizaje automático sean seleccionados y entrenados apropiadamente, y que los resultados finales sean evaluados y mejorados. En general, se pueden dividir en las siguientes fases:

- Recopilación y preparación de datos
- Selección del modelo y entrenamiento
- Evaluación del modelo
- Ajuste del modelo y mejora de su desempeño
- Implementación del modelo

Cada una de estas fases es fundamental para el éxito del proyecto y requiere atención y esfuerzo por parte de los profesionales involucrados. En la siguiente sección se explicará con más detalle cada una de estas fases.

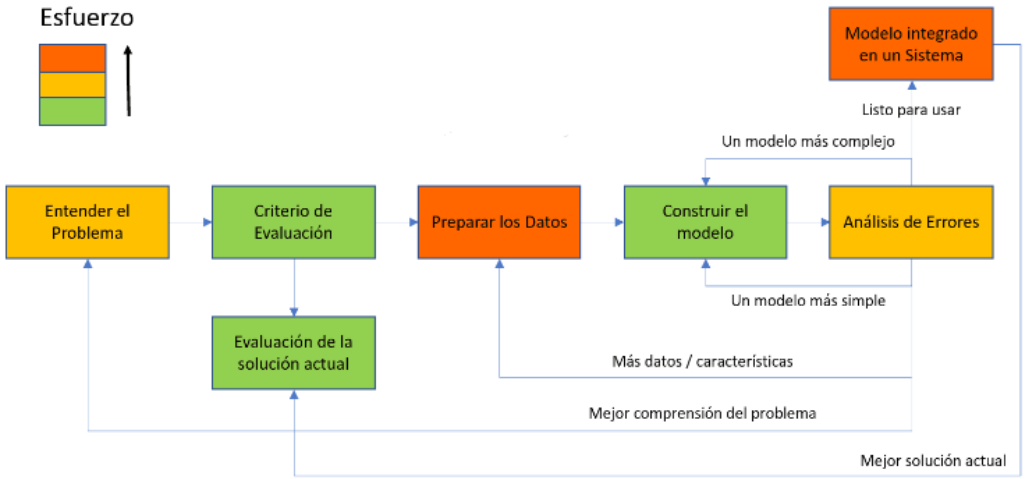

Figura 1: Proyecto de aprendizaje automático.

Fuente: IArtificial.net

3.4.2 Máquinas de vectores de soporte (SVM)

SVM (Support Vector Machines) es un algoritmo de aprendizaje supervisado utilizado principalmente para problemas de clasificación y regresión. El objetivo principal de SVM es encontrar un hiperplano que separe los datos en dos o más clases.

Supongamos que tienes un conjunto de datos de entrenamiento etiquetados, es decir, cada punto de datos está asociado con una etiqueta de clase conocida. El objetivo de SVM es encontrar un hiperplano en el espacio de características que separe los puntos de datos en diferentes clases.

El hiperplano es una superficie de decisión que se utiliza para clasificar nuevos puntos de datos. Para entender el concepto de hiperplano, considera el caso más simple de clasificación binaria donde tienes dos clases etiquetadas como +1 y -1. En este caso, el hiperplano es una línea

recta en el espacio de dos dimensiones que separa los puntos de datos positivos y negativos. El hiperplano se puede definir por la siguiente ecuación:

$$w*x + b = 0$$

Donde w es un vector de pesos que define la dirección del hiperplano y b es un término de sesgo que desplaza el hiperplano desde el origen. La variable x es un vector de características que representa un punto de datos de entrada.

Para entrenar el modelo SVM, primero se selecciona un tipo de kernel que se utiliza para transformar los datos a un espacio de mayor dimensión. Luego, se seleccionan los parámetros del modelo, como C y gamma, que afectan la forma del hiperplano.

El algoritmo SVM busca el hiperplano que maximiza la distancia entre los puntos de datos de las diferentes clases. Esta distancia se llama margen y se mide perpendicularmente al hiperplano. El margen se define como la distancia entre el hiperplano y el punto de datos más cercano de cada clase.

El objetivo del algoritmo SVM es maximizar el margen, lo que se traduce en minimizar la función de costo regularizada. Esta función de costo incluye dos términos: uno que penaliza la clasificación incorrecta de los puntos de datos de entrenamiento y otro que penaliza la complejidad del modelo. El parámetro C controla el equilibrio entre estos dos términos.

Por último, se calcula la distancia del punto de datos al hiperplano y se clasifica el punto en la clase correspondiente según en qué lado del hiperplano caiga.

En el caso de este proyecto, se utilizará SVR, una variante de SVM (Support Vector Machines) que se utiliza para resolver problemas de regresión en lugar de problemas de clasificación. En lugar de predecir la pertenencia de un punto de datos a una de dos clases, como hace SVM, SVR se utiliza para predecir un valor continuo.

El funcionamiento de SVR es similar al de SVM, pero en lugar de buscar un hiperplano que separe los datos en diferentes clases, busca una línea que mejor se ajuste a los datos y permita predecir el valor numérico objetivo.

3.4.3 K - vecinos más cercanos

KNN es un algoritmo de aprendizaje supervisado utilizado para la clasificación y la regresión. La sigla KNN significa "K-Nearest Neighbors", que en español significa "K-Vecinos Más Cercanos".

El funcionamiento del algoritmo KNN es relativamente simple. En el caso de clasificación, se trata de asignar una etiqueta de clase a un objeto a partir de las etiquetas de los objetos más cercanos en un conjunto de entrenamiento. En el caso de regresión, se trata de predecir el valor numérico de un objeto a partir de los valores numéricos de los objetos más cercanos.

Para utilizar el algoritmo KNN, se debe elegir un valor de K, que representa el número de vecinos más cercanos que se deben considerar para la clasificación o regresión. A continuación, se calcula la distancia entre el objeto de prueba y todos los objetos del conjunto de entrenamiento. Los K vecinos más cercanos se seleccionan en función de la distancia y se utiliza su etiqueta de clase (en el caso de clasificación) o su valor numérico (en el caso de regresión) para predecir la etiqueta de clase o el valor numérico del objeto de prueba.

La distancia se puede calcular utilizando diferentes métricas, como la distancia euclidiana o la distancia de Manhattan. En general, se utiliza la distancia euclidiana para datos continuos y la distancia de Hamming para datos discretos.

El algoritmo KNN tiene algunas ventajas, como su simplicidad y su capacidad para manejar datos no lineales. Sin embargo, también tiene algunas limitaciones, como la sensibilidad a los valores atípicos y la necesidad de almacenar todos los datos de entrenamiento en la memoria.

En resumen, KNN es un algoritmo de aprendizaje supervisado utilizado para la clasificación y la regresión. El algoritmo utiliza los K vecinos más cercanos en un conjunto de entrenamiento para predecir la etiqueta de clase o el valor numérico de un objeto de prueba. La distancia se puede calcular utilizando diferentes métricas y el valor de K se elige previamente. El algoritmo tiene algunas ventajas, como su simplicidad y su capacidad para manejar datos no lineales, pero también tiene algunas limitaciones, como la sensibilidad a los valores atípicos y la necesidad de almacenar todos los datos de entrenamiento en la memoria.

3.4.4 Árboles de decisión

Los árboles de decisión son un algoritmo de aprendizaje supervisado utilizado para la clasificación y la regresión. El algoritmo construye un árbol de decisiones a partir de un conjunto de datos de entrenamiento, en el que cada nodo del árbol representa una pregunta sobre los datos y las ramas del árbol representan las posibles respuestas a la pregunta.

Cuando se utiliza el árbol de decisión para la clasificación, cada hoja del árbol representa una clase o etiqueta de clasificación. Para la regresión, cada hoja del árbol representa un valor numérico.

Existen varios tipos de árboles de decisión, como los árboles de clasificación y regresión (CART) y los árboles de regresión y clasificación múltiple (MART).

Los árboles CART son un tipo de árbol de decisión utilizado tanto para la clasificación como para la regresión. En el caso de la clasificación, el árbol CART se construye dividiendo repetidamente el conjunto de datos de entrenamiento en subconjuntos más pequeños en función de las características que mejor separan las clases. En el caso de la regresión, el árbol CART se construye dividiendo el conjunto de datos en subconjuntos más pequeños en función de las características que mejor explican la variabilidad en la variable objetivo.

Los árboles MART, por otro lado, se utilizan principalmente para la regresión y son una variante de los árboles de regresión y clasificación múltiple. A diferencia de los árboles CART, los árboles MART utilizan una técnica llamada boosting para mejorar la precisión de las predicciones. Boosting implica la construcción de varios árboles de decisión, cada uno de los cuales se construye en función de los errores de los árboles anteriores.

En resumen, los árboles de decisión son un algoritmo de aprendizaje supervisado utilizado para la clasificación y la regresión. Los árboles CART se utilizan tanto para la clasificación como para la regresión y se construyen dividiendo repetidamente el conjunto de datos de entrenamiento en subconjuntos más pequeños en función de las características que mejor separan las clases o explican la variabilidad en la variable objetivo. Los árboles MART, por otro lado, se utilizan principalmente para la regresión y utilizan una técnica llamada boosting para mejorar la precisión de las predicciones.

3.4.5 Random Forest

Random Forest es un algoritmo de aprendizaje supervisado utilizado tanto para la clasificación como para la regresión. Se basa en la combinación de varios árboles de decisión, donde cada árbol se construye utilizando una muestra aleatoria del conjunto de datos de entrenamiento y un subconjunto aleatorio de características.

En el caso de la clasificación, la predicción final de Random Forest se determina por mayoría de votos de los árboles individuales, es decir, se selecciona la clase que es elegida por la mayoría de los árboles. En el caso de la regresión, la predicción final se determina por la media de las predicciones de los árboles individuales.

La idea detrás de Random Forest es que la combinación de varios árboles de decisión con diferentes subconjuntos de datos y características puede mejorar la precisión general del modelo, ya que reduce el sobreajuste y aumenta la generalización.

La combinación de los resultados de los árboles en Random Forest se realiza de manera diferente para la clasificación y la regresión.

En el caso de la regresión, la predicción final se determina por la media de las predicciones de los árboles individuales. Cada árbol en el bosque genera una predicción para una entrada dada, y la predicción final es el promedio de estas predicciones. Por ejemplo, si ha ay 100 árboles en el bosque y cada uno predice que una entrada tiene un valor objetivo de 5, la predicción final será 5.

En este caso, la combinación de los resultados de los árboles ayuda a mejorar la precisión y la estabilidad del modelo en comparación con el uso de un solo árbol de decisión.

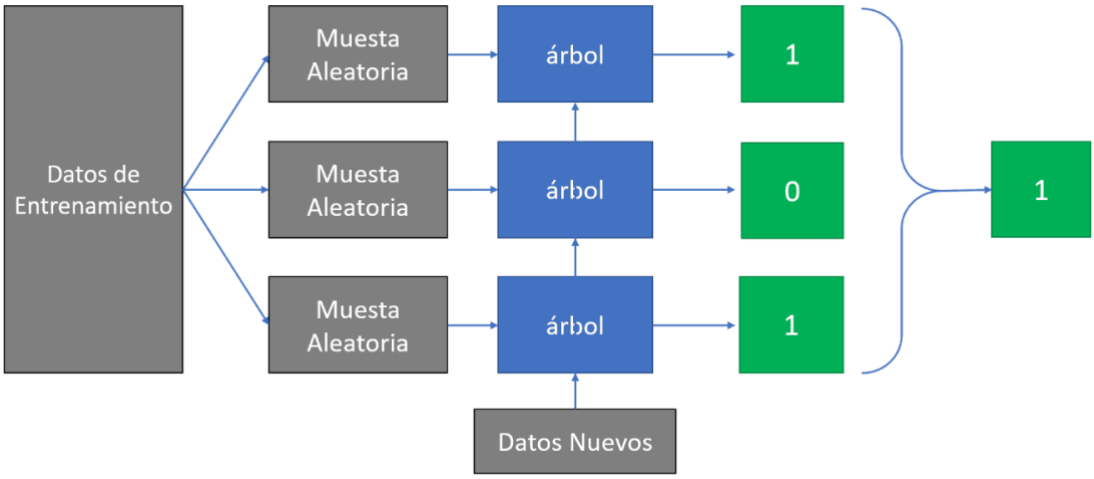

Figura 2: Esquema del algoritmo de Random Forest.

Fuente: IArtificial.net

Random Forest tiene varios hiperparámetros que pueden ser ajustados para mejorar el rendimiento del modelo, como el número de árboles, la profundidad máxima de los árboles y el número máximo de características a considerar en cada división de un árbol.

En resumen, Random Forest es un algoritmo de aprendizaje supervisado utilizado tanto para la clasificación como para la regresión. Combina varios árboles de decisión con diferentes subconjuntos de datos y características para mejorar la precisión general del modelo. La predicción final se determina por mayoría de votos en el caso de la clasificación y por la media de las predicciones en el caso de la regresión.

3.4.6 Regresión de mínimos Cuadrados Parciales

Partial Least Squares (PLS) es un algoritmo de aprendizaje supervisado utilizado para la regresión y el análisis de datos multivariantes. El objetivo principal de PLS es encontrar una relación lineal entre un conjunto de variables predictoras y una variable de respuesta continua.

La idea detrás de PLS es similar a la de la regresión lineal, pero en lugar de ajustar un modelo lineal a todas las variables predictoras al mismo tiempo, PLS descompone las variables predictoras en una serie de componentes latentes que maximizan la covarianza entre las variables predictoras y la variable de respuesta.

El proceso de construcción del modelo de PLS se divide en dos fases. En la primera fase, PLS calcula una serie de componentes latentes que describen la relación entre las variables predictoras y la variable de respuesta. Cada componente latente se construye mediante la combinación lineal

de las variables predictoras originales y se selecciona de tal manera que maximiza la covarianza entre las variables predictoras y la variable de respuesta. En la segunda fase, se utiliza la información obtenida de la primera fase para construir un modelo de regresión lineal que puede predecir la variable de respuesta.

PLS tiene varios hiperparámetros que pueden ser ajustados para mejorar el rendimiento del modelo, como el número de componentes latentes y la forma en que se manejan los datos faltantes.

En resumen, PLS es un algoritmo de aprendizaje supervisado utilizado para la regresión y el análisis de datos multivariantes. Descompone las variables predictoras en una serie de componentes latentes que maximizan la covarianza entre las variables predictoras y la variable de respuesta, y utiliza esta información para construir un modelo de regresión lineal que puede predecir la variable de respuesta.

3.5 Optimización de hiper parámetros para modelos de aprendizaje automático

Los hiperparámetros son parámetros que no se aprenden automáticamente a partir de los datos de entrenamiento, sino que deben ser ajustados por el usuario antes de la fase de entrenamiento. Estos parámetros controlan la complejidad del modelo, el proceso de entrenamiento y la capacidad de generalización del modelo. Algunos ejemplos de hiperparámetros comunes incluyen la profundidad del árbol en un algoritmo de árbol de decisión, el número de vecinos en un algoritmo de vecinos más cercanos y el número de componentes latentes.

La elección adecuada de los hiperparámetros es crucial para obtener un modelo de aprendizaje automático que se ajuste bien a los datos y generalice bien a nuevos datos. Existen varias estrategias para encontrar los valores óptimos de los hiperparámetros:

- *Búsqueda exhaustiva de cuadrícula:* Consiste en definir un rango de valores posibles para cada hiperparámetro y luego evaluar el rendimiento del modelo para todas las combinaciones posibles de valores de hiperparámetros. Este enfoque es simple y garantiza encontrar el mejor conjunto de hiperparámetros dentro del espacio de búsqueda especificado, pero puede ser muy costoso computacionalmente en caso de que el espacio de búsqueda sea grande.
- *Búsqueda aleatoria:* Consiste en muestrear los valores de los hiperparámetros de forma aleatoria dentro de un rango específico de valores. En lugar de explorar todo el espacio de búsqueda, la búsqueda aleatoria busca puntos aleatorios dentro de ese espacio. Este enfoque puede ser más eficiente que la búsqueda exhaustiva de cuadrícula si el espacio de búsqueda es grande, pero también puede requerir más iteraciones para encontrar los mejores valores de hiperparámetros.
- *Optimización bayesiana:* Este enfoque utiliza el teorema de Bayes para encontrar los valores óptimos de los hiperparámetros. En lugar de explorar todo el espacio de búsqueda, el algoritmo optimiza una función objetivo que estima la calidad del modelo para diferentes conjuntos de hiperparámetros. Este enfoque puede ser más eficiente que la búsqueda exhaustiva de cuadrícula y la búsqueda aleatoria, especialmente para espacios de búsqueda de alta dimensionalidad.

- *Ajuste en línea*: Consiste en ajustar los hiperparámetros del modelo durante el entrenamiento en función del rendimiento del modelo en un conjunto de validación. Este enfoque puede ser más rápido que los métodos anteriores, pero también es menos sistemático y puede llevar a sobreajuste en conjuntos de datos pequeños.

Model	Overview	Hyperparameters
C4.5	J48 Decision Tree	c = {0.005,0.10,0.20,0.30,0.40,0.50,0.60,0.70}
NNET	3-layer Neural Network	Size = {4,…,28}, decay = {0.10,0.20}
KNN	K- Nearest Neighbor	c = {2*(0,...,7) +1}
RF	Random Forest	Mtry = {10,50,100,200,250, 500, 1000}
SVM	Support Vector Machine	c = $\{2^{-6},\ldots,2^{10}\}$

Figura 3: Ejemplo de hiperparámetros

Es importante tener en cuenta que la elección del método adecuado para optimizar los hiperparámetros dependerá de la complejidad del modelo, el tamaño del conjunto de datos y los recursos computacionales disponibles.

3.6 Métricas de evaluación de modelos de aprendizaje automático

Un aspecto importante en la concepción de modelos válidos es la validación del modelo predictivo. Para ello existen métricas que explican diferentes aspectos del modelo. Para este proyecto usamos 4 métricas usadas usualmente para evaluar los problemas de regresión.

Desde el punto de vista de la interpretación del resultado un modelo con un valor más alto de MAE generalmente es considerado peor que un modelo con un valor más bajo de MAE. El MAE es una métrica que mide la magnitud del error absoluto entre las predicciones y los valores reales. Un valor más bajo de MAE indica una menor magnitud del error y, por lo tanto, una mayor precisión en las predicciones del modelo. Por lo tanto, un modelo con un valor más alto de MAE se considera menos preciso que un modelo con un valor más bajo. Sin embargo, es importante tener en cuenta que el MAE solo mide la magnitud del error y no tiene en cuenta la dirección del error.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}i|$$

Desde el punto de vista de la interpretación del resultado un modelo con un valor más alto de MSE generalmente es considerado peor que un modelo con un valor más bajo de MSE. El MSE es una métrica que mide la magnitud del error cuadrático entre las predicciones y los valores reales. Un valor más bajo de MSE indica una menor magnitud del error cuadrático y, por lo tanto, una mayor precisión en las predicciones del modelo. Por lo tanto, un modelo con un valor más alto de MSE se considera menos preciso que un modelo con un valor más bajo. Además, el MSE da más peso a los errores más grandes y es más sensible a los outliers, lo que significa que un modelo con un valor más alto de MSE puede ser afectado por un solo punto de datos con un error muy grande.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Desde el punto de vista de la interpretación del resultado un modelo con un valor más alto de RMSE generalmente es considerado peor que un modelo con un valor más bajo de RMSE. El RMSE es una métrica que mide la magnitud del error cuadrático medio entre las predicciones y los valores reales. Se calcula como la raíz cuadrada del MSE. Un valor más bajo de RMSE indica una menor magnitud del error cuadrático medio y, por lo tanto, una mayor precisión en las predicciones del modelo. Por lo tanto, un modelo con un valor más alto de RMSE se considera menos preciso que un modelo con un valor más bajo. Al igual que el MSE, el RMSE da más peso a los errores más grandes y es más sensible a los outliers.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2}$$

Desde el punto de vista de la interpretación del resultado, un modelo con un valor más alto de R^2 generalmente es considerado mejor que un modelo con un valor más bajo de R^2. El R^2 es una métrica que mide qué tan bien el modelo explica la variabilidad de los datos de respuesta. Un valor más cercano a 1 indica que el modelo explica una mayor cantidad de la variabilidad de los datos de respuesta y, por lo tanto, una mayor precisión en las predicciones del modelo. Por lo tanto, un modelo con un valor más alto de R^2 se considera más preciso que un modelo con un valor más bajo. Sin embargo, es importante tener en cuenta que un valor alto de R^2 no siempre garantiza un buen modelo. Hay otros factores que deben ser considerados al evaluar un modelo, como la sensibilidad a los datos fuera de muestra y la capacidad de generalización.

$$R^2 = \frac{\sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2}$$

3.7 Aplicación de técnicas de Machine a estudios sobre rendimiento académico

En esta parte se estudiarán los resultados de trabajos similares con objetivos similares al de este trabajo.

En el estudio llamado "Prediction of students Performance Using Machine Learning" por Chahuan (2019), cuyo objetivo era el de predecir la nota media de los estudiantes del curso de Ciencias de la computación que comprenden las ramas de la ingeniería de software, el estudio de algoritmos y disciplinas como el big data y la ciberseguridad entre el año 2015 y 2018, como variables explicativas eligió las notas de teoría, las notas de prácticas y a estas aplicó los algoritmos de regresión: KNN, SVM, árboles de decisión, Random Forest y de regresión lineal.

Después de aplicar todos estos algoritmos, concluyó que la técnica de regresión lineal múltiple era la que mejor resultados obtiene, con un error cuadrático de 0.040, un error cuadrático medio de 0.2, error absoluto 0.149 y R-cuadrado igual a 0.940. Lo interesante de este estudio es la comparación que se realizó para llegar a un modelo con mejor precisión.

Otro estudio llamado "A Machine Learning Model to Predict the Performance of University Students" por Canagareddy (2019), este trabajo fue llevado adelante debido a que se identificó que la mayoría de los estudiantes repiten curso, por ello se establece como objetivo intentar predecir el rendimiento para poder ayudar a los alumnos a mejorar. Los algoritmos que se escogieron para

las predicciones del rendimiento fueron algoritmos de predicción SVM, Random Forests y Logic Regression. Se utilizaron datos de 2000 estudiantes de una facultad de la Universidad de Mauricio. Se concluyó que, como algoritmo de predicción, Random Forest era el más eficiente.

El siguiente estudio titulado "Predicting Students Academic Performance Using Support Vector Machine" de Burman (2019), surgió debido al problema del bajo rendimiento de los estudiantes por ello el objetivo de este estudio era mejorar ese rendimiento académico. Para ello, se creó un modelo de SVM (Support Vector Machine) para clasificar a los alumnos en 3 categorías: alto, medio y bajo. Los datos se extrajeron de un cuestionario, obteniendo así parámetros psicológicos motivacionales y socio económicos con alrededor de 1000 registros para el análisis. Se obtuvieron mejores resultados con la función de base radial con un 90 % de precisión.

También hay una investigación con el título de "Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis" por Hamoud (2018). El objetivo de esta investigación fue buscar lo que determina el éxito o el fracaso de los estudiantes. En la investigación se utilizaron técnicas de minería de datos específicamente el árbol de decisión. Los datos se obtuvieron de una encuesta de 60 preguntas acerca de la salud, la vida social, las relaciones y el rendimiento económico del estudiante a 161 alumnos. Se utilizaron algoritmos como J48, Random Forest y RepTree, J48 obtuvo la mejor precisión de 0.629 y una Tasa de verdadero positivo del 0.634 mientras que RECALL (casos negativos) igual a 0.409.

De manera parecida al trabajo de Hamoud (2018), también Segura (2018) en su trabajo titulado "Using Decision Trees For Predicting Academic Performance Based On Socio-Economic Factors" que fijó el objetivo en determinar la forma esta vez en la que los factores socioeconómicos afectan a la educación. Por supuesto considero datos socioeconómicos combinados con datos académicos, y a estos datos se le aplicaron algoritmos de clasificación y técnicas de aprendizaje automático. Acerca de los datos un total de 15.445 eran registros acerca de las notas de los alumnos, registros académicos y unos 19.070 datos era registros socioeconómicos como los índices socioeconómicos de los padres, estudios de los mismos, etc… Como resultado se obtuvo que los algoritmos basados en árboles de decisión mostraron la mayor precisión de entorno al 60% y que los árboles con degradado una precisión cerca del 68 %, en ambos casos se observó que existían indicadores que influyen de mayor forma para la predicción como las becas académicas, la edad en el momento y la edad cuando empezó el proyecto. Como conclusión Segura consideró que los factores socioeconómicos de los estudiantes no guardaban una relación ni influye en las notas de los estudiantes. Además, Segura apunta que sería interesante para en futuras investigaciones tomar en cuenta más los factores psicológicos y emocionales.

Otro trabajo muy parecido a los anteriores es la investigación de Chiheb (2017) llamada: "Predicting Students performance using decision trees: Case of an Algerian University", Cuyo objetivo principal era identificar a los estudiantes de buen desempeño y ayudar a los graduados a elegir un máster adecuado según los resultados. Con datos recolectados de los departamentos de matemáticas e informática, Ciencias y Ciencias de la computación de la Universidad de Jijel de estudiantes de las promociones de 2009-2010 y 2014-2015. Después de metódico proceso de limpieza, transformación y preparación de los datos, usó como algoritmo predictivo el árbol de decisión que dio como resultado una tasa de acierto del 80 % para la variable éxito en la Licenciatura, un 55% para el Máster SIAD y un 100% para el máster R&S.

En un trabajo de investigación titulado "Uso de técnicas de Machine Learning para predecir el rendimiento académico de los estudiantes de la carrera de ingeniería Civil en informática de la Universidad del BIO BIO, Chillan" por Soto (2015) se presentó una encuesta con 33 preguntas realizada desde el año 2013 hasta el 2015 y se obtuvieron datos de los estudiantes acerca de sus motivaciones, su manera de estudiar, factores emocionales, autoestima, índices económicos, intereses y expectativas laborales así como donde estudiaron previamente, se buscó la manera de predecir su rendimiento. Se utilizó un algoritmo K vecinos cercanos dando como resultado una efectividad del 41% y un error de 1.24 para la asignatura de programación y un 60 % de efectividad con un error de 0.4 para la programación orientada a objetos.

4. Resultados

4.1 Análisis exploratorio de los datos

En esta sección, se llevará a cabo un análisis exploratorio de los datos (EDA) con el objetivo de comprender mejor la distribución y las relaciones entre las variables que se utilizarán para la construcción de modelos de aprendizaje automático. En particular, se utilizarán los algoritmos de SVM, PLS, árboles de decisión y KNN para desarrollar modelos de predicción.

En esta sección, se explorará la distribución de las variables, las correlaciones entre ellas y la presencia de valores atípicos, con el objetivo de garantizar que los modelos de aprendizaje automático sean sólidos y precisos. Una vez completado el análisis exploratorio, se construirán y evaluarán los modelos de SVM, PLS, árboles de decisión y KNN.

La exploración de variables numéricas nos permitirá entender cómo se distribuyen las variables, si existen valores extremos o valores faltantes, y si hay alguna relación entre ellas. Comencemos con la estimación de varios estadísticos descriptivos.

Tabla 2: Estadísticos descriptivos

	count	mean	std	min	25%	50%	75%	max
MEDIA_SUS	5005	7.193882	0.909486	4.45	6.54	7.18	7.82	10
MEDIAS_SUS_NP	5005	7.112599	0.990874	3.4	6.44	7.14	7.8	10
MEDIA_OFICIAL	5005	7.241738	0.787521	5.3	6.6	7.2	7.8	9.7
DUR_REAL	5005	3.736464	1.228768	1	4	4	4	8
CRE_MAT	5005	199.5463	76.1853	6	181.5	238.5	244.5	364.5
CRE_SUP	5005	189.5958	72.9171	6	166.5	230	240	273
T_EFI	5005	0.951397	0.080127	0.21	0.92	1	1	1
ING_NOTA	5005	9.084096	1.790893	5	8	9.1	10.1	14
EDAD	5005	25.77682	6.716506	21	22	23	26	66

Los resultados indican que la nota media en el grupo de datos es de 7.2, con una desviación estándar de 0.7, lo que sugiere que los datos están bastante concentrados en torno a la media. El rango de notas se encuentra entre 5.3 y 9.7, lo que indica que hay una variedad de notas en el conjunto de datos. El percentil 25 (25% de los datos) está en 6.6, lo que significa que el 25% de los estudiantes obtuvo una nota inferior a 6.6. El percentil 50 (50% de los datos) se encuentra en 7.2, lo que significa que el 50% de los estudiantes obtuvo una nota inferior a 7.2. Finalmente, el percentil 75 (75% de los datos) está en 7.8, lo que indica que el 75% de los estudiantes obtuvo una nota inferior a 7.8.

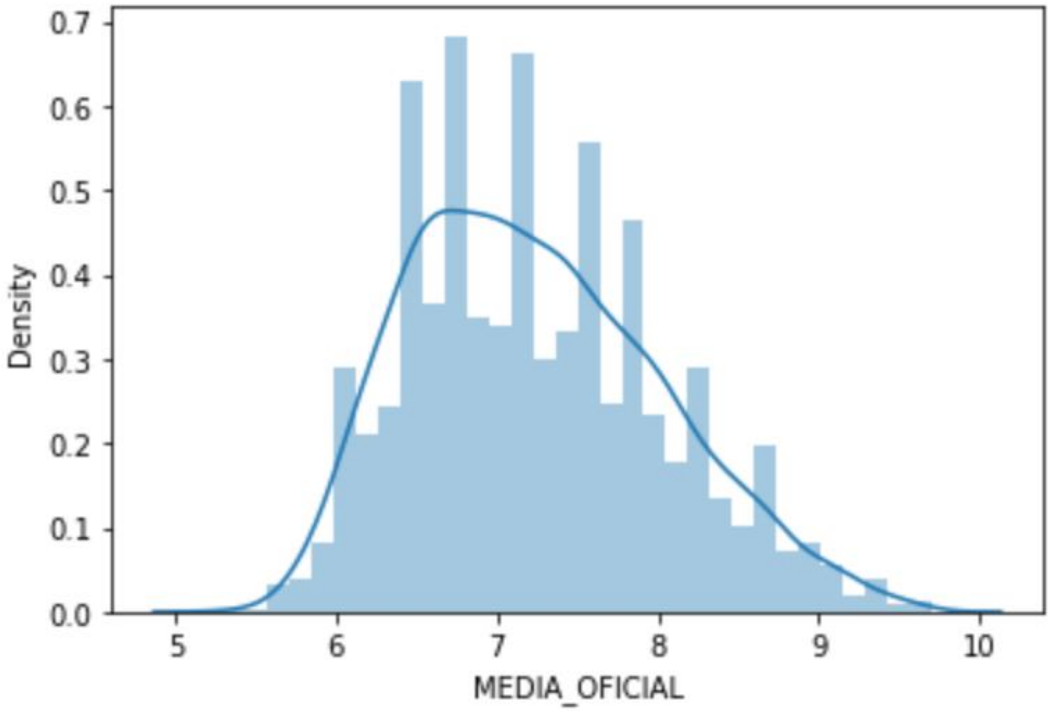

Figura 4: Histograma de la Media oficial

En general, estos resultados proporcionan una idea general del rango y la distribución de la nota media en el conjunto de datos. Además, con el histograma notamos distintos picos, esto indica que hay ciertos valores que son más comunes que otros. En este caso, los valores comunes son alrededor de 6.5, 6.75, 7.2 y 7.5.

La cola del histograma está desplazada hacia la derecha significa que hay valores altos o extremos en la distribución de la nota media, lo cual puede deberse a varios factores como, por ejemplo, que existan un número reducido de estudiantes con notas muy altas que estén estirando el promedio hacia arriba.

En el presente análisis también se han examinado las notas medias de los estudiantes por facultades de la universidad. Para ello, se han utilizado gráficos de histogramas para representar la distribución de las notas en cada una de las facultades.

A partir de los histogramas generados, se puede realizar un análisis comparativo de las notas medias entre las diferentes facultades y observar si existen patrones o tendencias comunes.

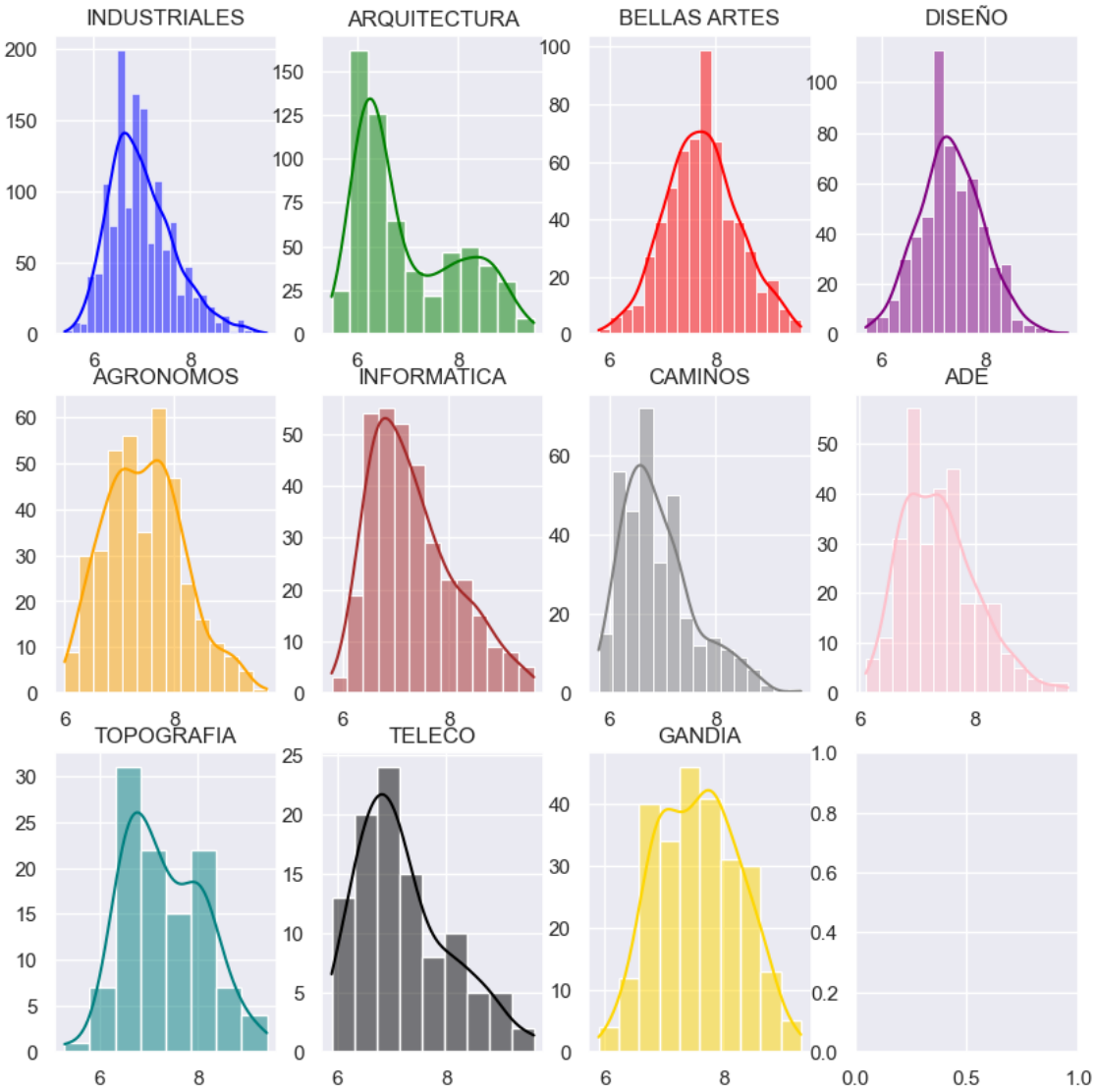

Figura 5: Histograma de la nota media por Facultades

Para la ETSIIndustriales con picos en torno a 6.5 y 7, sugiere que hay una cantidad significativa de observaciones con valores cercanos a estas dos cifras. La cola hacia la derecha indica que hay algunas observaciones con valores muy altos que están aumentando ligeramente la asimetría de la distribución. Además, se observa que alrededor de los picos en valores como 6.5, 7 y 7.5 hay valores que no llegan a la línea de base lo que indica que hay muy pocas observaciones entorno a esos valores.

Con una cola muy desplazada hacia la derecha vemos que hay algunas observaciones con valores muy altos por encima del 8.

Para las Facultad de Bellas Artes y ETSIDiseño, sus distribuciones son casi simétricas, se puede decir que la distribución de las notas es equilibrada y que hay una cantidad similar de

estudiantes en ambos lados de la media. Sin embargo, para la Facultad de Bellas Artes vemos un pico justo en el 8, es decir que hay un gran grupo de estudiantes con notas más altas que el resto, para la ETSID sucede algo muy parecido con un gran número de estudiantes con notas de entorno al 7.

Para las Escuelas TSI Informática, Caminos, Canales y Puertos y Telecomunicación su distribución está desplazada hacia la izquierda, lo que significa que la mayoría de las observaciones se encuentran en valores más bajos que el promedio. En el caso de una distribución de notas, esto podría indicar que la mayoría de los estudiantes obtuvieron calificaciones más bajas de lo esperado.

El resto de las escuelas y facultades, ETSI Agronómica y del Medio Natural, Administración y Dirección de empresas (ADE), ETSIGeodésica, Cartográfica y Topográfica y la Escuela Politécnica Superior de Gandía parece que siguen una distribución de notas bastante equilibrada, lo que indican que no hay tendencia a notas muy altas ni muy bajas. Sin embargo, si se observan gran cantidad de observaciones entorno a la nota de 7 y 8 en la Facultad de ADE y ETSIAgronómica respectivamente.

Cabe destacar que la facultad de arquitectura sigue una distribución que no es completamente normal y podría ser bimodal no simétrica ya que la cola del histograma esta desplazada hacia la derecha. Con varios picos alrededor de 6 y después una mayor cantidad de observaciones entorno al 8 indica que hay dos grupos o subpoblaciones distintas en la muestra, con diferentes niveles de notas. Las barras muy pequeñas en el medio podrían indicar que hay una menor cantidad de estudiantes con notas en esa zona en particular.

El análisis de datos en el ámbito educativo es una herramienta fundamental para entender el rendimiento y desempeño de los estudiantes. En este caso, se utilizó la función. groupby() para agrupar a los estudiantes según el nivel de estudios de sus padres y obtener así la media de su desempeño académico. Por lo tanto, obtenemos el desempeño de los estudiantes en cuanto a nota media según la pareja de nivel de estudios que formen el padre y la madre.

Tabla 3: Recuento estudios de parejas de padres

Nivel de estudios padre	Nivel de estudios madre	Recuento
Estudios Primarios (Básicos)	Estudios Primarios (Básicos)	1144
	Estudios Secundarios (Cou, Bachiller,	346
	Estudios Terciarios (Universitarios)	175
	Sin estudios (Si sabe leer y escribir)	34
Estudios Secundarios (Cou, Bachiller, Fp2)	Estudios Primarios (Básicos)	442
	Estudios Secundarios (Cou, Bachiller,	669
	Estudios Terciarios (Universitarios)	305
	Sin estudios (Si sabe leer y escribir)	26
Estudios Terciarios (Universitarios)	Estudios Primarios (Básicos)	182
	Estudios Secundarios (Cou, Bachiller,	366
	Estudios Terciarios (Universitarios)	1108
	Sin estudios (Si sabe leer y escribir)	4
Sin estudios (Si sabe leer y escribir)	Estudios Primarios (Básicos)	65
	Estudios Secundarios (Cou, Bachiller,	16
	Estudios Terciarios (Universitarios)	15
	Sin estudios (Si sabe leer y escribir)	99

En esta tabla, se muestra un recuento del número de veces que aparecen esas parejas de padre y madre en cuanto a nivel de estudios en los datos. Las parejas más comunes entre los padres son ambos padres con estudios primarios con 1144 casos seguidos de ambos padres con estudios primarios con 1108 casos.

Véase ahora una comparativa de la nota media promedio de los estudiantes en función de la pareja a nivel de estudios que formen los padres:

Figura 6: Gráfico de barras de nota media en función de estudios de los padres

Para las parejas mencionadas anteriormente, ambos padres con estudios terciarios y ambos padres con estudios primarios la nota media es de 7.2 y 7.3 respectivamente. Cabe destacar la nota media de los estudiantes cuyo padre alcanzó los estudios secundarios y cuya madre no tiene estudios, es de 7.4, la más alta de todas las parejas. Por otro lado, la más baja es la nota media de los estudiantes cuyo padre tiene estudios terciarios y su madre no tiene estudios, con un 6,7.

Para continuar con el estudio de la influencia de los padres en la nota media de los estudiantes, se analiza la relación entre el trabajo de los padres y la nota media de los estudiantes, con el fin de identificar posibles correlaciones que puedan explicar los resultados académicos de los jóvenes. Para ello, se utilizaron varios diagramas de caja y bigotes que representan la distribución de la nota media según el tipo de trabajo de los padres y las madres.

Primero, veamos unas tablas que indican la nota media de los estudiantes según los distintos trabajos que realizan padres y madres.

Tabla 4: Nota media del estudiante en función trabajo del padre

Trabajo del padre	Nota
Artesanos y trabajadores cualificados	7,2
Dirección de las empresas y de las administraciones	7,1
Empleados de tipo administrativo	7,3
Fuerzas armadas	7,3
Inactivo, Desocupado o Jubilado	7,2
Operadores de instalaciones y maquinaria, y montadores	7,4
Trabajadores cualificados en la agricultura y en la pesca	7,2
Trabajadores de los servicios de restauración, personales,	7,2
Trabajadores no cualificados	7,3
Técnicos y profesionales científicos e intelectuales	7,2
Técnicos y profesionales de apoyo	7,3

En el caso del trabajo del padre son los estudiantes en los que el padre trabaja en operadores de instalaciones y maquinaria y montadores los que obtienen mejor media, pero por lo general, no hay diferencias significativas entre unos y otros.

Tabla 5: Nota media del estudiante en función del trabajo de la madre

Trabajo de la madre	Nota
Artesanos y trabajadores cualificados	7,3
Dirección de las empresas y de las administraciones	7,2
Empleados de tipo administrativo	7,2
Fuerzas armadas	7,3
Inactivo, Desocupado o Jubilado	7,3
Operadores de instalaciones y maquinaria, y montadores	7,4
Trabajadores cualificados en la agricultura y en la pesca	7,5
Trabajadores de los servicios de restauración, personales,	7,2
Trabajadores no cualificados	7,2
Técnicos y profesionales científicos e intelectuales	7,2
Técnicos y profesionales de apoyo	7,2

En el caso del trabajo de la madre son los estudiantes con madre que trabaja en Trabajadores cualificados en la agricultura y en la pesca los que obtienen las mejores calificaciones, pero también observamos que no hay diferencias significativas entre unos y otros.

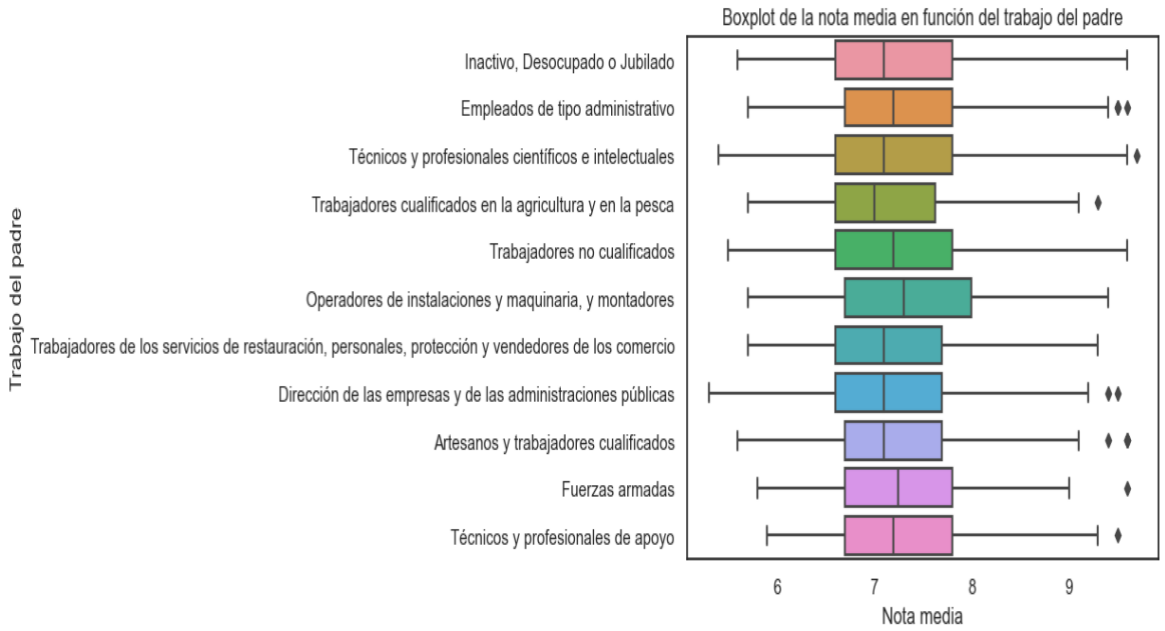

Figura 7. Diagrama de cajas y bigotes en función del trabajo del padre

Si nos fijamos en la mediana, por lo general, para todos los tipos de trabajo del padre se sitúa en el centro por lo que la distribución es simétrica y la media, mediana y moda coinciden. Se ven diversos valores atípicos para los trabajos de empleados de tipo administrativo, dirección de empresas y administraciones públicas y artesanos y trabajadores cualificados. En resumen, no se encuentran diferencias significativas en la nota media en función del trabajo del padre.

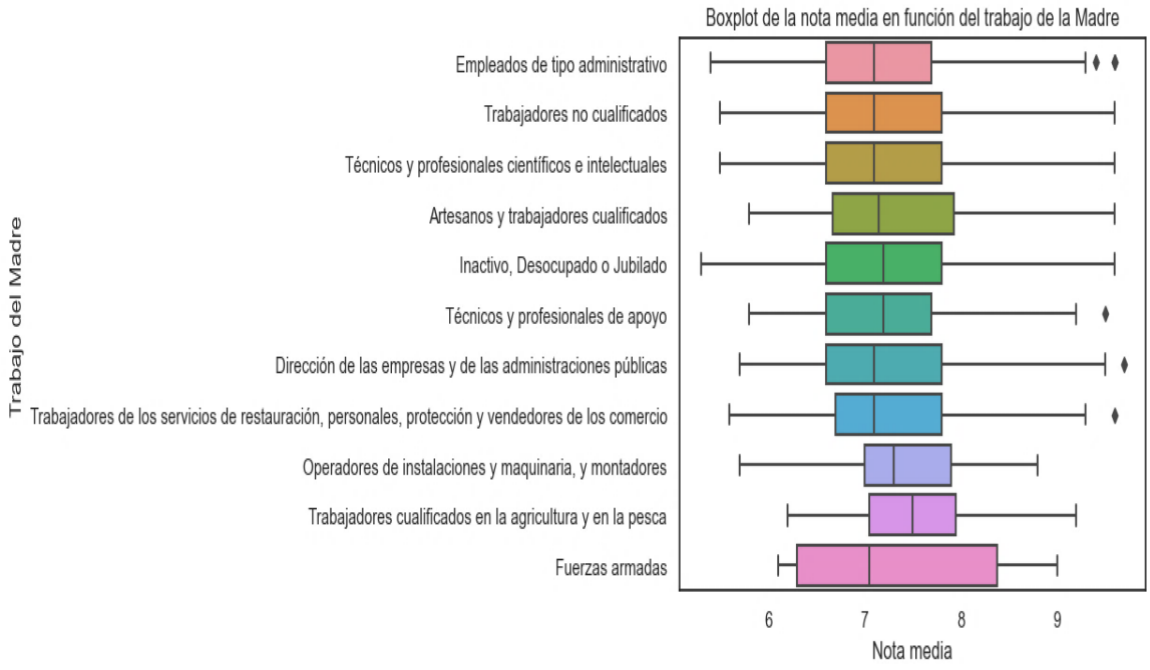

Figura 8: Diagrama de caja y bigotes en función del trabajo de la madre

En este caso la mediana no se sitúa en el centro para la mayoría de los tipos de trabajo si no que vemos por ejemplo para las fuerzas armadas y operadores de instalaciones y maquinaria y montadores que existe una asimetría positiva ya que se encuentra más sesgada a la derecha. Los datos se concentran en la parte inferior de la distribución, la media por tanto será mayor que la mediana. Hay que destacar que para los trabajadores cualificados en la agricultura y en la pesca y operadores de instalaciones y maquinaria y montadores solo el 25 % de los estudiantes obtuvo notas por debajo del 7.2. Igualmente, no se encuentran diferencias significativas en la nota media de los estudiantes en función del trabajo de la madre.

En este gráfico de barras se presenta la nota media de los estudiantes universitarios en función de su modo de acceso a la universidad. Cada barra representa un modo de acceso diferente y su altura indica la nota media obtenida por los estudiantes que accedieron a la universidad de esa forma. Este tipo de gráfico permite visualizar de forma clara y sencilla las diferencias en las notas medias entre los distintos modos de acceso a la universidad, lo que puede ser útil para analizar posibles factores que influyen en el rendimiento académico de los estudiantes universitarios.

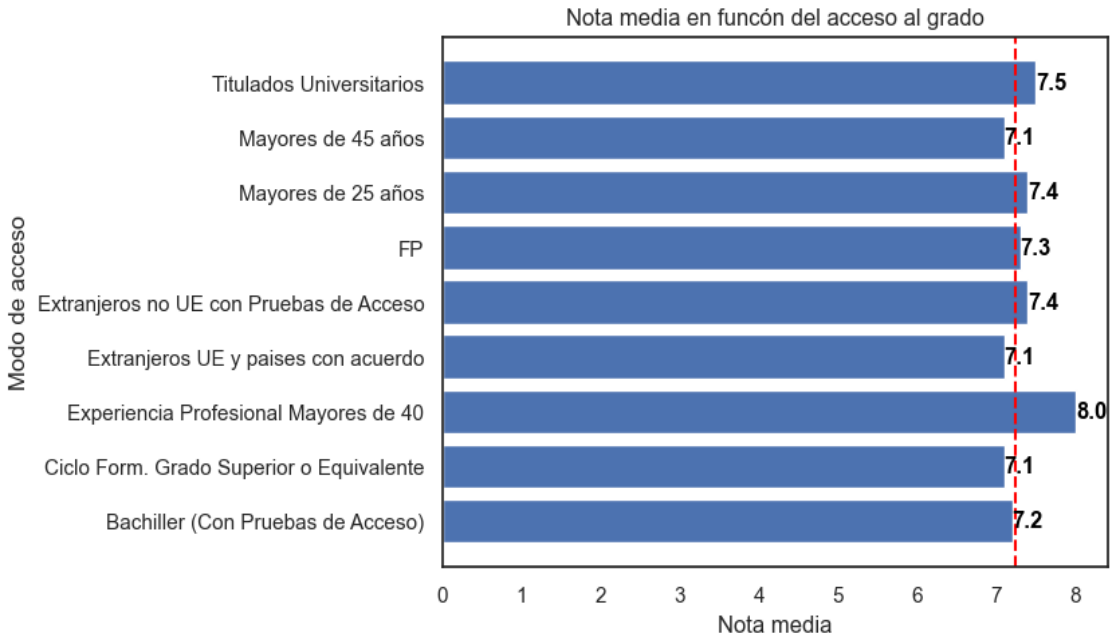

Figura 9: Gráfico de barras nota media en función modo de acceso

Los estudiantes que entraron con Experiencia profesional Mayores de 40 años son los que tienen una nota media superior al resto, esto puede deberse a que son un grupo muy reducido, sin embargo, si nos fijamos en los grupos más grandes como son el acceso por Bachiller y por FP la nota media es superior de 7.3 a 7.2 en estudiantes que accedieron mediante FP, quizá pueda deberse a la preparación más técnica que se recibe en la formación profesional.

Con el fin de determinar si que el alumno trabaje o no durante los años que tarda en acabar su carrera universitaria, influye en el rendimiento, se realiza un análisis de varianza (ANOVA) que será útil para determinar si existe una relación significativa. En este caso se desea explorar si el hecho de que un estudiante trabaje o no tienen alguna relación con su rendimiento académico en términos de la nota media. Específicamente, se busca determinar si existe una diferencia significativa en las notas medias de los estudiantes que trabajan y los que no trabajan.

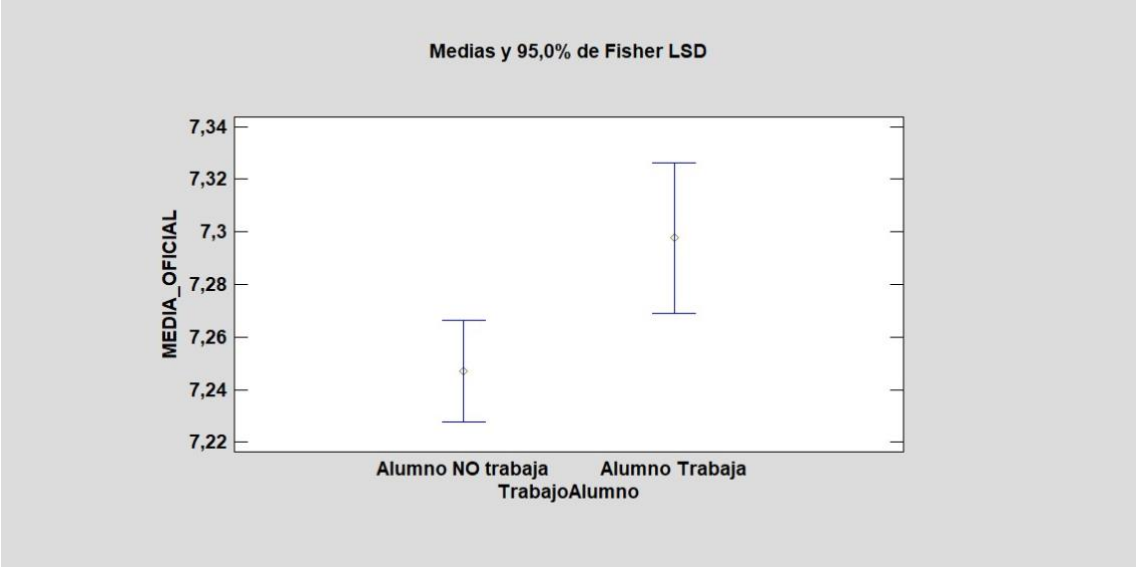

Figura 10: ANOVA simple trabajoAlumno-Nota media

Parece ser que, si existe una diferencia, aunque es una diferencia muy pequeña habrá que tenerlo en cuenta para la elaboración del modelo. Además, se observa que el estudiante que SI trabaja obtiene mejores calificaciones.

Sabiendo esto, se ha aplicado un ANOVA de varios factores para evaluar si la nota media de los estudiantes universitarios esta influenciada por la combinación de varios factores, tales como el género, y si el estudiante trabaja o no.

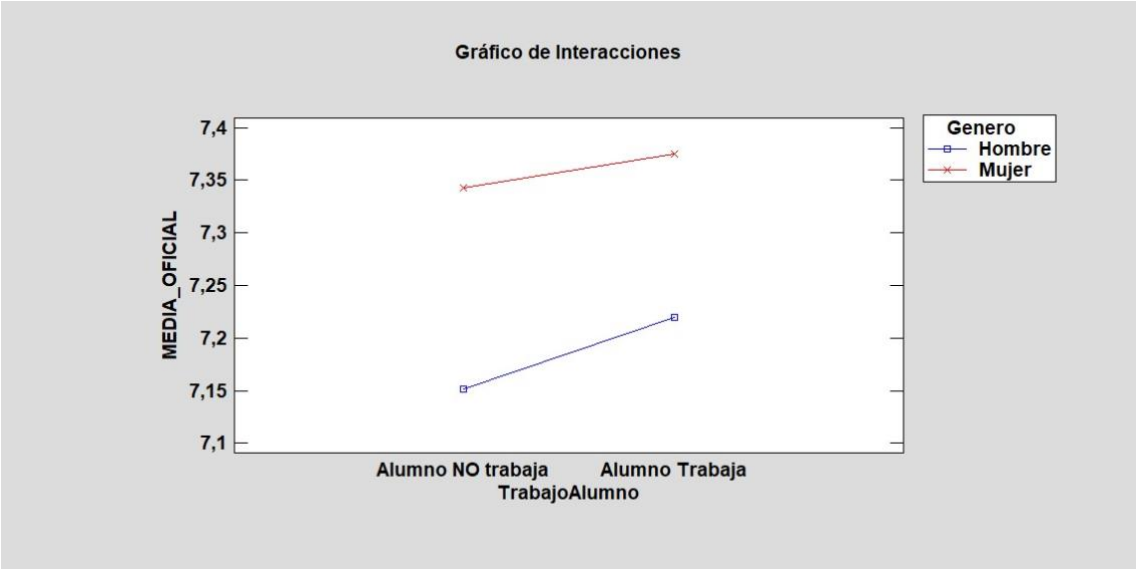

Figura 11: ANOVA de trabajoAlumno, genero - Nota media

Véase que si hay una diferencia entre los dos grupos, hombre y mujer. Para las mujeres no existe gran diferencia si el estudiante trabaja o no, pero para los hombres sí que existe una diferencia más notable si trabaja o no trabaja el estudiante

Para poder explorar la relación entre las variables para entender mejor los datos y, en última instancia, construir modelos precisos. Una herramienta útil para examinar la relación entre variables numéricas es el gráfico de correlación. En este caso particular, nos enfocamos en la relación de varias variables numéricas con la variable objetivo nota media oficial del estudiante. El objetivo es identificar aquellas variables que estén fuertemente correlacionadas con la nota media y, por lo tanto, pueden ser útiles para predecir su valor. Para que el gráfico de correlación se vea más completo se han creado 4 nuevas variables:

- CRE_MATxAÑO: créditos matriculados por año
- CRE_SUPxAÑO: créditos suspendidos por año
- CRE_SUSPENDIDOS: créditos suspendidos
- CRE_SUSPENDIDOSxAÑO: créditos suspendidos por año

	MEDIA_SUS	MEDIAS_SUS_NP	ORDEN_PROMOCION	MEDIA_OFICIAL	DUR_REAL	CRE_MAT	CRE_SUP	T_EFI
MEDIA_SUS	1.000000	0.969266	-0.548588	0.935919	-0.378376	-0.315891	-0.186177	0.531237
MEDIAS_SUS_NP	0.969266	1.000000	-0.548874	0.910637	-0.367591	-0.275660	-0.132602	0.628375
ORDEN_PROMOCION	-0.548588	-0.548874	1.000000	-0.545211	0.204034	0.150176	0.058632	-0.386341
MEDIA_OFICIAL	0.935919	0.910637	-0.545211	1.000000	-0.252412	-0.190206	-0.082645	0.451350
DUR_REAL	-0.378376	-0.367591	0.204034	-0.252412	1.000000	0.898028	0.846361	-0.213542
CRE_MAT	-0.315891	-0.275660	0.150176	-0.190206	0.898028	1.000000	0.977354	-0.042068
CRE_SUP	-0.186177	-0.132602	0.058632	-0.082645	0.846361	0.977354	1.000000	0.146403
T_EFI	0.531237	0.628375	-0.386341	0.451350	-0.213542	-0.042068	0.146403	1.000000
ING_NOTA	0.400088	0.409432	-0.236705	0.374938	-0.131957	-0.014975	0.069695	0.311575
ANY_ING	0.323178	0.313018	-0.207864	0.212024	-0.906173	-0.847183	-0.812463	0.138830
UPV_DESDE	-0.156485	-0.112035	0.037203	-0.082445	0.534236	0.635041	0.656613	0.114572
EDAD	0.200685	0.152729	-0.069209	0.142708	-0.513349	-0.622063	-0.639535	-0.089495
CRE_MATxAÑO	-0.098859	-0.033965	0.021854	-0.024775	0.320702	0.675547	0.712045	0.246043
CRE_SUPxAÑO	0.098009	0.177600	-0.113422	0.135080	0.211331	0.573013	0.668738	0.512319
CRE_SUSPENDIDOS	-0.647767	-0.699725	0.442470	-0.522662	0.413842	0.303760	0.095270	-0.857056
CRE_SUSPENDIDOSxAÑO	-0.627742	-0.690571	0.441524	-0.521980	0.290036	0.189090	-0.014396	-0.932948

	ING_NOTA	ANY_ING	UPV_DESDE	EDAD	CRE_MATxAÑO	CRE_SUPxAÑO	CRE_SUSPENDIDOS	CRE_SUSPENDIDOSxAÑO
MEDIA_SUS	0.400088	0.323178	-0.156485	0.200685	-0.098859	0.098009	-0.647767	-0.627742
MEDIAS_SUS_NP	0.409432	0.313018	-0.112035	0.152729	-0.033965	0.177600	-0.699725	-0.690571
ORDEN_PROMOCION	-0.236705	-0.207864	0.037203	-0.069209	0.021854	-0.113422	0.442470	0.441524
MEDIA_OFICIAL	0.374938	0.212024	-0.082445	0.142708	-0.024775	0.135080	-0.522662	-0.521980
DUR_REAL	-0.131957	-0.906173	0.534236	-0.513349	0.320702	0.211331	0.413842	0.290036
CRE_MAT	-0.014975	-0.847183	0.635041	-0.622063	0.675547	0.573013	0.303760	0.189090
CRE_SUP	0.069695	-0.812463	0.656613	-0.639535	0.712045	0.668738	0.095270	-0.014396
T_EFI	0.311575	0.138830	0.114572	-0.089495	0.246043	0.512319	-0.857056	-0.932948
ING_NOTA	1.000000	0.138770	0.091182	-0.164616	0.153735	0.248632	-0.384234	-0.347080
ANY_ING	0.138770	1.000000	-0.476772	0.488936	-0.348746	-0.259768	-0.327284	-0.216550
UPV_DESDE	0.091182	-0.476772	1.000000	-0.761884	0.550697	0.520600	0.031024	-0.022343
EDAD	-0.164616	0.488936	-0.761884	1.000000	-0.544604	-0.508191	-0.046870	0.000156
CRE_MATxAÑO	0.153735	-0.348746	0.550697	-0.544604	1.000000	0.953181	-0.028002	-0.066418
CRE_SUPxAÑO	0.248632	-0.259768	0.520600	-0.508191	0.953181	1.000000	-0.315354	-0.365040
CRE_SUSPENDIDOS	-0.384234	-0.327284	0.031024	-0.046870	-0.028002	-0.315354	1.000000	0.954326
CRE_SUSPENDIDOSxAÑO	-0.347080	-0.216550	-0.022343	0.000156	-0.066418	-0.365040	0.954326	1.000000

Figura 12: Gráfico de correlación

Hay que destacar que la nota media de los alumnos se relaciona de manera positiva con la tasa de eficiencia y la nota de ingreso a la universidad. Por otro lado, la nota media se relaciona de manera negativa con los créditos suspendidos y la duración real. Se refuerza por tanto la idea de que la nota de ingreso y la tasa de eficiencia están fuertemente relacionados con la media oficial.

Por último, se analiza la relación entre la nota media obtenida por los estudiantes durante su paso por la universidad y la nota de acceso a la misma. Para ello, se ha realizado un gráfico de dispersión. Cada punto en el gráfico representa un estudiante, y se ha utilizado una línea de regresión para ilustrar la tendencia general de la relación entre las dos variables.

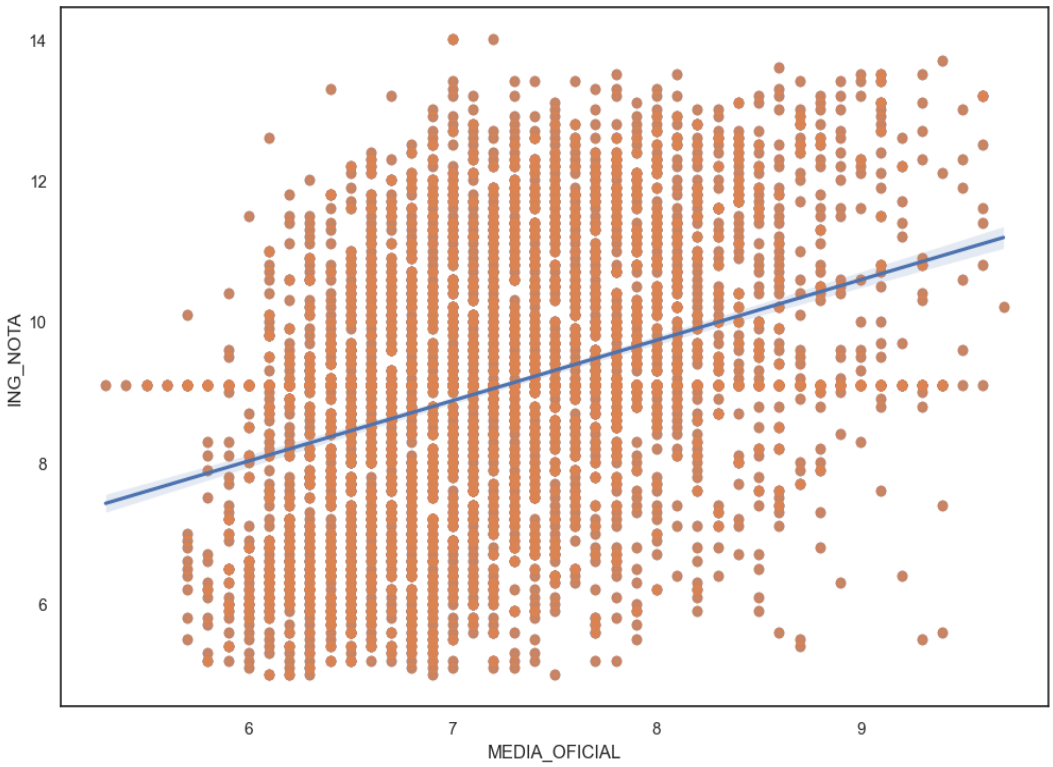

Figura 13: Gráfico de dispersión

El gráfico de dispersión muestra una relación positiva entre la nota media y la nota de acceso a la universidad. Se puede observar que la mayoría de los puntos se encuentran en una nube central, lo que sugiere una fuerte concentración de estudiantes con un rango de notas medias y de acceso similares. Además, se puede apreciar que la línea de regresión muestra una tendencia ascendente hacia la derecha, lo que indica que a medida que la nota de acceso aumenta, también lo hace la nota media.

Esto sugiere que la nota de acceso a la universidad puede ser un buen indicador del rendimiento académico del estudiante, ya que parece haber una correlación positiva entre ambas variables. Sin embargo, es importante tener en cuenta que la correlación no implica necesariamente causalidad, por lo que se necesitarían más estudios para determinar si existe una relación causal entre estas variables.

En general, los resultados de este gráfico de dispersión indican que los estudiantes que han obtenido una buena nota de acceso a la universidad también tienen más probabilidades de tener un buen rendimiento académico.

4.2 Modelos predictivos de rendimiento académico

En el análisis exploratorio de datos hemos identificado las variables explicativas que podrían tener una relación con la nota media de los estudiantes al acabar el grado en la universidad. Ahora, el siguiente paso es especificar y estimar modelos predictivos que puedan ayudarnos a entender mejor las relaciones entre estas variables y la nota media, y así predecir la nota media de los estudiantes y además poder identificar cuáles son los factores que más influyen en el rendimiento académico de los estudiantes. Para ello, nos basaremos en la revisión científica y las relaciones observadas entre las variables a nivel descriptivo. Las variables explicativas incluyen información del alumno, como los trabajos, nivel de estudios de los padres, sus notas anteriores, entre otros. El objetivo es encontrar un modelo que tenga una buena capacidad de predicción y que pueda ayudar a entender mejor los factores que influyen en el rendimiento académico de los estudiantes.

En este proceso, se utilizaron técnicas estadísticas avanzadas, como PLS, K vecinos más cercanos, árboles de decisión, Random Forest y Support Vector Regression (SVR), basándonos en la revisión científica y las relaciones observadas entre las variables a nivel descriptivo. El objetivo de esta sección es presentar los resultados de estos modelos y evaluar su capacidad predictiva en términos de precisión y capacidad de generalización. De esta manera, se busca determinar cuál de estos modelos es el más adecuado para predecir la nota media de los estudiantes.

4.2.1 Árboles de decisión

En esta sección se presentarán los resultados de dos modelos de aprendizaje automático aplicados para predecir la nota media de los estudiantes: el modelo MART y el modelo CART. Además, se incluirán los resultados de la evaluación utilizando validación cruzada y el conjunto de prueba para cada modelo.

El modelo MART (Multiple Additive Regression Trees) es una técnica de aprendizaje automático que utiliza múltiples árboles de regresión para modelar la relación entre las variables de entrada y la variable de salida. Por otro lado, el modelo CART (Classification and Regression Trees) es un método de aprendizaje automático que utiliza árboles de decisión para realizar tareas de clasificación y regresión.

Para evaluar el rendimiento de los modelos, se utilizó la validación cruzada, una técnica de validación que permite evaluar la capacidad de generalización del modelo. También se evaluó el modelo con el conjunto de prueba para tener una idea del rendimiento del modelo en datos nunca vistos.

En resumen, se presentarán los resultados de dos modelos de aprendizaje automático (MART y CART) aplicados para predecir la nota media de los estudiantes, y se evaluarán utilizando validación cruzada y el conjunto de prueba.

4.2.1.1 Resultados del modelo MART

Los parámetros son una parte esencial de cualquier modelo de aprendizaje automático. Estos parámetros controlan cómo el modelo toma decisiones y cómo se ajusta a los datos de entrenamiento.

La columna "Parámetro" contiene el nombre de cada parámetro que se utiliza en el modelo. La "Descripción" es una breve descripción de lo que hace el parámetro y cómo afecta al modelo.

"Valores estudiados" son todos los posibles valores que se han probado para cada parámetro durante la búsqueda de parámetros. La búsqueda de parámetros es un proceso en el que se evalúan diferentes combinaciones de parámetros para encontrar la mejor combinación que maximiza el rendimiento del modelo.

Por último, "Mejores parámetros" es el valor de cada parámetro que se ha encontrado como el óptimo en la búsqueda de parámetros. Este valor se usa para entrenar el modelo y obtener los mejores resultados en términos de precisión o desempeño en el conjunto de datos de prueba.

Tabla 6: Parámetros del modelo MART

Parámetros	Descripción	Valores estudiados	Mejores parámetros
n_estimators	Es el número de árboles de decisión que se utilizan para construir el modelo	[100,500,1000]	500
learning_rate	Es la tasa de aprendizaje del modelo. Esta tasa controla cuánto se ajustan los pesos del modelo en cada iteración.	[0.001,0.01, 0.1]	0.01
max_depth	Es la profundidad máxima del árbol de decisión. Controla la cantidad máxima de ramificaciones que puede tener el árbol.	[3,5,7]	3
min_sample_split	Es el número mínimo de observaciones necesarias para que un nodo del árbol pueda dividirse en subnodos adicionales.	[2,5,10]	5
min_samples_leaf	Es el número mínimo de observaciones que deben estar en una hoja (el nodo final de una rama) para que se considere válida.	[1,2,4]	2

Evaluación con Validación Cruzada:

Tabla 7: Resultados I Modelo MART

R^2	RMSE	MSE	MAE
0.40	0.62	0.36	0.46

Evaluación con el conjunto de prueba:

Tabla 8: Resultados II modelo MART

R^2	RMSE	MSE	MAE
0.45	0.61	0.34	0.44

Los resultados para el modelo MART entrenado y evaluado utilizando validación cruzada indican que el modelo tiene una capacidad moderada para predecir la nota media de los estudiantes. El RMSE indica que, en promedio, las predicciones del modelo se desvían de las notas reales en aproximadamente 0.62 puntos, mientras que el MAE indica una desviación promedio de 0.46 puntos. El valor de R^2 indica que el modelo explica el 40% de la variabilidad en las notas de los estudiantes.

Por otro lado, los resultados para el modelo MART evaluado utilizando el conjunto de prueba independiente sugieren que el modelo tiene un rendimiento ligeramente mejor en el conjunto de prueba en comparación con la evaluación de validación cruzada. El valor de R^2 indica que el modelo explica el 45% de la variabilidad en las notas de los estudiantes en el conjunto de prueba. El RMSE, MSE y el MAE son similares a los obtenidos en la evaluación de validación cruzada, lo que sugiere que el modelo generaliza bien en datos no vistos.

En resumen, ambos métodos han indicado que el modelo es capaz de generaliza y predecir de manera precisa la variable objetivo en datos nuevos.

4.2.1.2 Resultados del modelo CART

Tabla 9: Parámetros modelo CART

Parámetros	Descripción	Valores estudiados	Mejores parámetros
max_depth (máxima profundidad)	Es la profundidad máxima del árbol de decisión. Controla la cantidad máxima de ramificaciones que puede tener el árbol.	[3, 5, 7, 9, 11, 13]	3
max_features (máximas características)	Es el número máximo de características que se consideran al buscar la mejor división en cada nodo.	[4, 6, 8]	4
max_leaf_nodes (max nodo hoja)	Es el número máximo de hojas que se permiten en el árbol.	[5,10,15,20,25,30,35]	20
min_samples_leaf (min observaciones	Es el número mínimo de observaciones que deben estar en una hoja (el nodo final de una rama) para que se considere válida.	[1, 2, 3]	3
min_samples_split (min observaciones	Es el número mínimo de observaciones necesarias para que un nodo del árbol pueda dividirse en subnodos adicionales.	[2, 4, 6, 8]	8

Evaluación con Validación Cruzada:

Tabla 10: Resultados modelo CART con Validación Cruzada

R^2	RMSE	MSE	MAE
0.39	0.62	0.38	0.46

Evaluación con el conjunto de prueba:

Tabla 11: Resultados modelos CART con el conjunto de prueba

R^2	RMSE	MSE	MAE
0.37	0.62	0.37	0.46

Basándonos en los resultados del modelo CART, su capacidad para predecir la nota media de los estudiantes es moderada. La evaluación con validación cruzada nos muestra un RMSE de 0.62, lo que indica que el modelo tiene un margen de error de aproximadamente 0.62 puntos al predecir la nota media de los estudiantes. El R^2 obtenido de 0.39 indica que el modelo es capaz de explicar solo el 39% de la variabilidad de los datos de la variable objetivo, la nota media. Además, el MAE obtenido de 0.46 y el MSE de 0.38 indican que el modelo puede estar cometiendo errores en la predicción de la nota media.

La evaluación con el conjunto de prueba muestra un R^2 de 0.37, lo que indica que el modelo explica el 37 % de la variabilidad de los datos. Además, el RMSE y MSE son iguales a los obtenidos con la evaluación mediante validación cruzada, lo que indica que el modelo no ha mejorado su rendimiento al generalizar en datos nuevos.

En general, podemos decir que el modelo CART presenta un buen rendimiento moderado en cuanto a su capacidad de predicción y generalización para la variable objetivo de la nota media de los estudiantes.

Ambos modelos tienen resultados similares en cuanto a la precisión de la predicción de la nota media de los estudiantes. Sin embargo, el modelo MART parece generalizar mejor a través de la evaluación con el conjunto de prueba, lo que indica que el modelo es más efectivo al predecir datos que nunca ha visto antes.

En resumen, el modelo MART parece ser más efectivo tanto en términos de precisión como de generalización, por lo que podría considerarse como el modelo preferido para predecir la nota media de los estudiantes.

4.2.2 Random Forest

En esta sección se presentarán los resultados obtenidos del modelo de Random Forest, Random Forest es una técnica de aprendizaje automático que utiliza un conjunto de árboles de decisión para realizar la predicción. En este caso, se ha aplicado un modelo de Random Forest para predecir la nota media de los estudiantes.

Para evaluar el rendimiento del modelo, se utilizó la validación cruzada, una técnica de validación que permite evaluar la capacidad de generalización del modelo. También se evaluó el modelo con el conjunto de prueba para tener una idea del rendimiento del modelo en datos nunca vistos.

Por tanto, se presentarán los resultados del modelo de aprendizaje automático (Random Forest) aplicado para predecir la nota media de los estudiantes, y se evaluarán utilizando validación cruzada y el conjunto de prueba.

Tabla 12: Parámetros para el modelo Random Forest

Parámetros	Descripción	Valores estudiados	Mejores parámetros
n_estimator (núm de estimadores)	Es el número de árboles de decisión que se utilizan para construir el modelo	[50,100,200,500]	200
max_depth (max profundidad)	Es la profundidad máxima del árbol de decisión. Controla la cantidad máxima de ramificaciones que puede tener el árbol.	[5,10,20, None]	10
min_samples_split (min observaciones para división)	Es el número mínimo de observaciones necesarias para que un nodo del árbol pueda dividirse en subnodos adicionales.	[2,5,10]	2
min_samples_leaf (min observaciones para hojas)	Es el número mínimo de observaciones que deben estar en una hoja (el nodo final de una rama) para que se considere válida.	[1,2,4]	6
max_features (max características)	Es el número máximo de características que se consideran al buscar la mejor división en cada nodo.	[auto, sqrt, log2]	Log2
bootstrap	Si se debe o no realizar el muestreo con reemplazo al construir cada árbol en el bosque.	[True, False]	True

Evaluación con Validación Cruzada:

Tabla 13: Resultados modelo Random Forest con validación cruzada

R^2	RMSE	MSE	MAE
0.38	0.62	0.38	0.46

Evaluación con el conjunto de prueba:

Tabla 14: Resultados modelo Random Forest con el conjunto de prueba

R^2	RMSE	MSE	MAE
0.42	0.61	0.35	0.44

Los resultados obtenidos indican que el modelo de Random Forest presenta una capacidad moderada para explicar la variabilidad en los datos, como lo demuestra el valor del coeficiente de determinación (R^2) de 0.38 en la evaluación con validación cruzada y 0.42 en la evaluación con el conjunto de prueba. Además, el error cuadrático medio (MSE) y la raíz del error cuadrático medio (RMSE) también presentan valores moderados, siendo de 0.35 y 0.62 respectivamente en la evaluación con validación cruzada y de 0.34 y 0.61 respectivamente en la evaluación con el conjunto de prueba. Por otro lado, el error absoluto medio (MAE) presenta un valor bajo en ambas evaluaciones, lo que indica que el modelo presenta una buena capacidad para predecir la nota media de los estudiantes en términos generales. En conclusión, estos resultados sugieren que el modelo puede ser útil para predecir la nota media de los estudiantes.

En resumen, basándonos en los resultados del conjunto de prueba que son los que mejor desempeño han dado, el modelo de Random Forest tiene un buen desempeño para generalizar a datos nuevos. El R^2 de 0.42 indica que el modelo es capaz de explicar el 42% de la variabilidad de la variable objetivo en el conjunto de prueba. Además, los valores de RMSE, MSE y MAE son relativamente bajos, lo que sugiere que el modelo tiene un bajo error de predicción en el conjunto de prueba.

Es importante entender cómo funciona y cómo influyen las diferentes variables en las predicciones en los otros modelos probados. En el caso del modelo de Random Forest, se puede obtener información sobre la importancia de las variables utilizadas en la construcción del modelo.

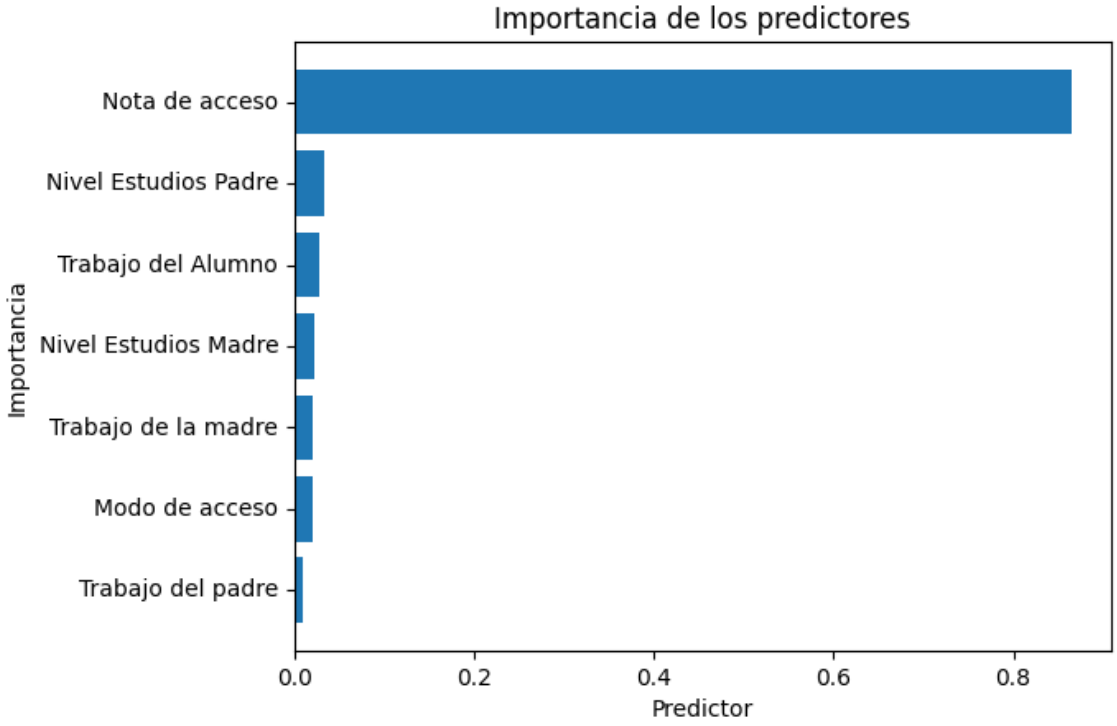

Figura 14: Gráfico de barras de la importancia de los predictores en modelo Random Forest

Al igual que ocurre en el modelo MART, en el modelo de Radom Forest la nota de acceso tiene un peso en el modelo que sobrepasa el 80 %, seguidos de los niveles de estudios del padre y de la madre y el trabajo del estudiante que rondan un peso de entre 3 y 6 % cada una.

4.2.3 Partial Least Squares Path Modeling (PLS)

En esta sección, se aplicó un modelo de Regresión PLS para predecir la nota media de los estudiantes a partir de un conjunto de variables explicativas. Para evaluar el rendimiento del modelo, se utilizó una técnica de validación cruzada para optimizar los parámetros del modelo y evaluar su precisión, así como una evaluación del conjunto de pruebas para comprobar la generalización del modelo.

Tabla 15: Parámetros modelo PLS

Parámetros	Descripción	Valores estudiados	Mejores parámetros
n_components	Indica el número de componentes que se deben retener en la proyección de los datos de entrada en el espacio de componentes latentes.	[2,3,4,5,6,7,8]	3
scale	Indica si los predictores y la variable objetivo deben escalarse antes de ajustar el modelo.	[true, false]	True

Evaluación con Validación Cruzada:

Tabla 16: Resultados del modelo PLS con validación cruzada

R^2	RMSE	MSE	MAE
0.26	0.67	0.45	0.53

Evaluación con el conjunto de prueba:

Tabla 17: Resultados del modelo PLS con el conjunto de prueba

R^2	RMSE	MSE	MAE
0.30	0.65	0.43	0.51

Basándonos en los resultados obtenidos por el modelo de PLS con el objetivo de predecir la nota media de los estudiantes. Se observa que para la evaluación mediante validación cruzada el valor de RMSE de 0.67 indica que en promedio el modelo se desvía un 0.67 de la nota media real, que junto con un valor de MSE de 0.45 y un valor de MAE de 0.53 indican que el modelo tiene una precisión moderada de la nota media de los estudiantes. Además, el valor de R^2 indica que el modelo explica aproximadamente el 26 % de la variabilidad de los datos.

Por otro lado, los resultados de la evaluación del conjunto de prueba indican un valor de RMSE de 0.65, en promedio el modelo se desvía un 0.65 de la nota media real, un valor de MAE de 0.51, un valor de MSE de 0.43 y el valor de R^2 indica que el modelo explica aproximadamente el 30% en la evaluación del conjunto de prueba.

Aunque el modelo tenga valores de R^2 es bajo, esto no significa necesariamente que el modelo sea malo. Se puede decir que el modelo de PLS es razonablemente preciso en la predicción de la nota media de los estudiantes.

En resumen, como el rendimiento del modelo en el conjunto de prueba es mejor que el rendimiento de la validación cruzada esto puede indicar que el modelo está generalizando bien y es más adecuado para su uso en nuevos datos.

Variable Importance in Projection (VIP) es una medida que se utiliza para evaluar la importancia relativa de los predictores en un modelo de Proyección en Componentes Latentes (PLS). El VIP se basa en el análisis de varianza (ANOVA) de cada predictor, y mide la contribución de cada uno de ellos a la variabilidad explicada por el modelo. Ayuda a identificar los predictores más importantes en un modelo de PLS. Esta tabla VIP permite identificar los predictores más importantes en el modelo de PLS y su relevancia en la predicción de la variable de respuesta.

Tabla 18: Análisis VIP PLS

Predictor	**VIP**
Nota de acceso	0.55
Modo de acceso	0.15
Trabajo Alumno	0.08
Nivel estudios de la madre	0.01
Trabajo madre	0.01
Nivel estudios del padre	0.02
Trabajo del padre	0.00

Los resultados del análisis indican que la variable de mayor importancia es la nota de acceso, con un valor VIP de 0.55, lo que sugiere que es el factor que más influye en la nota media de los estudiantes. Otros factores relevantes son el modo de acceso con un valor VIP de 0.15 y el trabajo del alumno con un valor VIP de 0.08. Los demás factores como el nivel de estudios de la madre, trabajo de la madre, nivel de estudios del padre y trabajo del padre, tienen valores VIP menores a 0.02, lo que sugiere que tienen una importancia marginal en la predicción de la nota media. En general, estos resultados proporcionan una buena guía para comprender la importancia relativa de cada variable y para identificar los factores clave que influyen en el rendimiento académico de los estudiantes universitarios.

El análisis de la importancia de las variables es una técnica fundamental en el proceso de modelado predictivo. Esta técnica nos permite determinar qué variables tienen el mayor impacto en el rendimiento de un modelo de regresión y, por lo tanto, nos ayuda a entender mejor el problema que estamos tratando de resolver.

4.2.4 k vecinos más próximos (KNN)

En esta sección se abordan los resultados del algoritmo de vecinos más cercanos (KNN) en un modelo para predecir la nota media de los estudiantes.

Además de las métricas de regresión, también se utilizaron la técnica de validación cruzada para evaluar el rendimiento del modelo. La validación cruzada nos permitió evaluar cómo se generaliza el modelo a nuevos datos, y se proporcionó una idea de la robustez y la estabilidad del modelo.

Además, se utilizó un conjunto de prueba para evaluar el rendimiento del modelo en datos que no habían sido utilizados ni en el entrenamiento ni en la validación cruzada.

Tabla 19: Parámetros del modelo KNN

Parámetros	Descripción	Valores	Mejores
n_neighbors	El número de vecinos más cercanos que se utilizarán para predecir la clase o el valor de un punto de datos de prueba.	range [1, 101]	29
weights	Define cómo se ponderan las contribuciones de los vecinos más cercanos en la predicción.	[uniform, distance]	Uniform
metric	Define la métrica de distancia utilizada para medir la similitud entre los puntos de datos en el espacio de características.	[euclidean, manhattan]	manhattan

Evaluación con Validación Cruzada:

Tabla 20: Resultados del modelo KNN con validación cruzada

R^2	RMSE	MSE	MAE
0.22	0.69	0.47	0.56

Evaluación con el conjunto de prueba:

Tabla 21: Resultados del modelo KNN con el conjunto de prueba

R^2	RMSE	MSE	MAE
0.25	0.67	0.44	0.53

En primer lugar, se evaluó el modelo mediante validación cruzada, obteniendo un coeficiente de determinación R^2 de 0.22, lo que indica que el modelo puede explicar el 22% de la variabilidad en la variable objetivo. El error cuadrático medio (MSE) fue de 0.47, lo que significa que el modelo tuvo un error promedio de 0.47 en la predicción de la variable objetivo. El error cuadrático medio de raíz (RMSE) fue de 0.69, lo que indica que el modelo tuvo un error promedio de 0.69 en la predicción de la variable objetivo. El error absoluto medio (MAE) fue de 0.56, lo que significa que el modelo tuvo un error promedio de 0.56 en la predicción de la variable objetivo.

Posteriormente, se evaluó el modelo con el conjunto de prueba, obteniendo un coeficiente de determinación R^2 de 0.25, lo que indica que el modelo puede explicar el 25% de la variabilidad en la variable objetivo. El error cuadrático medio (MSE) fue de 0.44, lo que significa que el modelo tuvo un error promedio de 0.44 en la predicción de la variable objetivo. El error cuadrático medio de raíz (RMSE) fue de 0.67, lo que indica que el modelo tuvo un error promedio de 0.67 en la predicción de la variable objetivo. El error absoluto medio (MAE) fue de 0.53, lo que significa que el modelo tuvo un error promedio de 0.53 en la predicción de la variable objetivo.

En resumen, los resultados indican que el modelo KNN tiene una capacidad moderada para explicar la variabilidad en la variable objetivo, con un R^2 de 0.22 en validación cruzada y 0.25 en el conjunto de prueba. Además, el modelo tiene un error promedio moderado en la predicción de la nota media, con un MSE de 0.47 en validación cruzada y 0.44 en el conjunto de prueba, un RMSE de 0.69 en validación cruzada y 0.67 en el conjunto de prueba, y un MAE de 0.56 en validación cruzada y 0.53 en el conjunto de prueba. Estos resultados pueden ser útiles para comprender la eficacia del modelo KNN en la predicción de la nota media y para identificar posibles mejoras en el modelo.

En general, estos resultados sugieren que el modelo de KNN puede ser útil para hacer predicciones sobre el rendimiento académico de los estudiantes universitarios, pero que puede haber margen de mejora para mejorar su capacidad de generalización y reducir la discrepancia entre las predicciones y los valores reales de la nota media.

En el contexto del modelo de KNN Regresor, una técnica útil para calcular la importancia de las variables es el VIP (Variable Importance in Projection). La técnica VIP mide cómo afecta la

eliminación de cada variable al rendimiento del modelo, lo que nos permite identificar las variables que tienen el mayor impacto en la precisión de las predicciones.

Utilizamos la técnica VIP para calcular la importancia de las variables en nuestro modelo de KNN Regresor. Los resultados muestran que la nota de acceso es la más importante para el modelo, ya que su eliminación tendría el mayor efecto en el rendimiento del modelo. Las demás variables, como trabajo del alumno, trabajo de la madre, nivel de estudios de la madre, modo de acceso, nivel de estudios del padre y trabajo del padre, tienen una importancia mucho menor en el modelo.

En resumen, el análisis de la importancia de las variables utilizando la técnica VIP es una herramienta valiosa para entender mejor los modelos de KNN Regresor. Nos permite identificar las variables más importantes para el modelo, lo que nos ayuda a enfocar nuestros esfuerzos en las áreas más importantes para mejorar la precisión de las predicciones.

Tabla 22: Análisis VIP KNN

Predictor	**VIP**
Nota de acceso	0.45
Modo de acceso	0.02
Trabajo Alumno	0.02
Nivel estudios de la madre	0.02
Trabajo madre	0.00
Nivel estudios del padre	0.00
Trabajo del padre	0.00

Los resultados obtenidos muestran que la variable ING_NOTA es la más importante para nuestro modelo de KNN Regresor. Esta variable representa la nota de acceso y, según los resultados obtenidos, es la variable que tiene el mayor impacto en el rendimiento del modelo. De hecho, la eliminación de esta variable tendría el mayor efecto en la precisión de las predicciones.

Por otro lado, las variables trabajo del alumno, trabajo de la madre, nivel de estudios de la madre, modo de acceso, nivel de estudios del padre y trabajo del padre tienen una importancia mucho menor en el modelo. Estas variables, aunque tienen cierto impacto en el rendimiento del modelo, tienen una importancia mucho menor en comparación con la nota de acceso.

En resumen, el análisis de la importancia de las variables utilizando la técnica VIP es una herramienta valiosa para entender mejor los modelos de KNN Regresor. Nos permite identificar las variables más importantes para el modelo, lo que nos ayuda a entender que factores son los que más influyen en el rendimiento académico.

4.2.5 Supper Vector Regression (SVR)

En esta sección se presentarán los resultados de un modelo de aprendizaje automático aplicado para predecir la nota media de los estudiantes: el del modelo de Regresión de Vectores de Soporte (SVR). Además, se incluirán los resultados de la evaluación utilizando validación cruzada y el conjunto de prueba para cada modelo.

El modelo de Regresión de Vectores de Soporte (SVR) es una técnica de aprendizaje supervisado utilizada para la predicción de variables continuas. El SVR se basa en la idea de encontrar un hiperplano o una función no lineal que se ajuste a los datos de entrenamiento de manera óptima.

Tabla 23: Parámetros para el modelo SVR

Parámetros	Descripción	Valores estudiados	Mejores parámetros
C	Parámetro de regularización que controla el trade-off entre el ajuste del modelo y la penalización por error.	[0.1,1, 10,100]	0.1
Kernel	Función kernel utilizada para mapear los datos de entrada en un espacio de características de mayor dimensión.	[linear, poly, rbf]	linear
degree (Grado)	Grado del polinomio en el kernel polinómico	[2,3,4]	3
Gamma	Parámetro del kernel utilizado para controlar la forma de la función kernel.	[0.1,0.01, 0.001,10,	10
shrinking	Parámetro booleano que indica si se utiliza la técnica de reducción de tamaño de los vectores de soporte.	[True, False]	*False*

Evaluación con Validación Cruzada:

Tabla 24: Resultados modelo SVR con validación cruzada

R^2	RMSE	MSE	MAE
0.32	0.68	0.53	0.58

Evaluación con el conjunto de prueba:

Tabla 25: Resultados del modelo SVR con el conjunto de prueba

R^2	RMSE	MSE	MAE
0.38	0.61	0.50	0.54

La evaluación del modelo se realizó utilizando validación cruzada y un conjunto de prueba. La evaluación por validación cruzada dio como resultado un coeficiente de determinación (R^2) de 0.32, un error cuadrático medio (MSE) de 0.53, un error absoluto medio (MAE) de 0.58 y una raíz del error cuadrático medio (RMSE) de 0.68. La evaluación con el conjunto de prueba dió como resultado un R^2 de 0.38, un MSE de 0.5, un MAE de 0.54 y un RMSE de 0.61.

A partir de los resultados obtenidos, se puede observar que el modelo SVR tiene un rendimiento moderado en la predicción de los valores de la nota media. El R^2 obtenido en ambas evaluaciones indica que el modelo explica sólo una proporción moderada de la variación de los datos. Además, los valores de MSE, MAE y RMSE son relativamente altos, lo que sugiere que el modelo tiene dificultades para capturar la variabilidad de los datos.

La importancia de las características es una forma de entender qué variables están contribuyendo más a la predicción de un modelo de aprendizaje automático. En particular, Permutation Feature Importance es un método de importancia de características que consiste en evaluar cómo cambia el rendimiento del modelo cuando se permutan aleatoriamente los valores de cada característica.

En el contexto de un modelo de regresión de Vector de Soporte (SVR), la importancia de las características nos permite entender qué variables están contribuyendo más a la predicción de los valores continuos de la variable objetivo.

En este caso, hemos utilizado el método de Permutation Feature Importance para analizar la importancia de las características en un modelo SVR. Hemos realizado 10 repeticiones del análisis para obtener una medida más robusta de la importancia de cada característica.

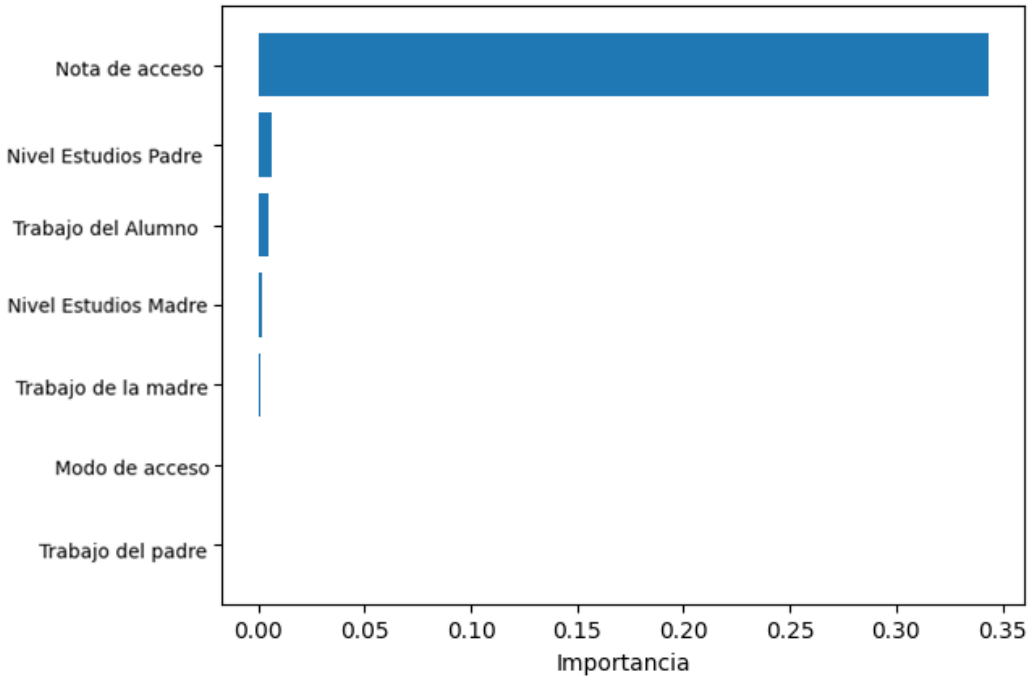

Figura 15: Gráfico de barras de la importancia de los predictores en el modelo SVR

Después de realizar el análisis de Permutation Feature Importance en un modelo de regresión SVR, podemos concluir que la nota de acceso es la característica más importante en la predicción de los valores continuos de la variable objetivo. Con un peso de 0.35, se puede ver que esta característica tiene un gran impacto en la predicción del modelo.

Por otro lado, las características relacionadas con el nivel de estudios del padre, de la madre y el trabajo del alumno tienen un peso mucho más bajo, no llegando a 0.02. Esto indica que estas características tienen un impacto mucho menor en la predicción del modelo y, por lo tanto, podrían no ser tan relevantes en la toma de decisiones.

4.3 Propuesta / Selección del modelo predictivo final

En este apartado se presentará el proceso de selección del modelo predictivo final. Después de haber explorado diferentes modelos, como árboles de decisión, Random Forest, máquinas de vectores de soporte, k vecinos más cercanos y regresión de mínimos cuadrados parciales, se llevará a cabo una evaluación comparativa para seleccionar el mejor modelo predictivo. Para ello, se analizarán los diversos resultados obtenidos por los distintos modelos con el fin de identificar el modelo que mejor se ajuste a las necesidades del proyecto.

En resumen, la idea de este apartado era la de escoger el modelo predictivo final que mejores resultados haya obtenido y que cumpla con los objetivos establecidos en este proyecto.

Por lo tanto, tras haber realizado diversas pruebas con diferentes modelos predictivos, se ha seleccionado el modelo de Árbol de Decisión MART como el más adecuado para nuestra tarea de predicción. Este modelo ha demostrado una precisión competente y capacidad de generalización en la predicción de la nota media de los estudiantes de la universidad. Además, su simplicidad y facilidad de interpretación lo convierten en una herramienta muy útil para entender cómo influyen las variables explicativas en el rendimiento académico.

Saber qué variables son las más importantes para un modelo puede proporcionar información valiosa para entender el fenómeno que se está modelando y para mejorar la precisión y la generalización del modelo.

La identificación de las variables más importantes para el rendimiento de los estudiantes nos ayudará a determinar cuáles son los factores que influyen en el rendimiento de los estudiantes. Este gráfico de barras de la importancia de los predictores del modelo MART permite visualizar el papel que juega cada uno de los factores para predecir la nota media.

Los predictores que tienen un mayor impacto en la variable objetivo son seleccionados automáticamente por el modelo mediante el análisis de la contribución de cada predictor en la reducción del error.

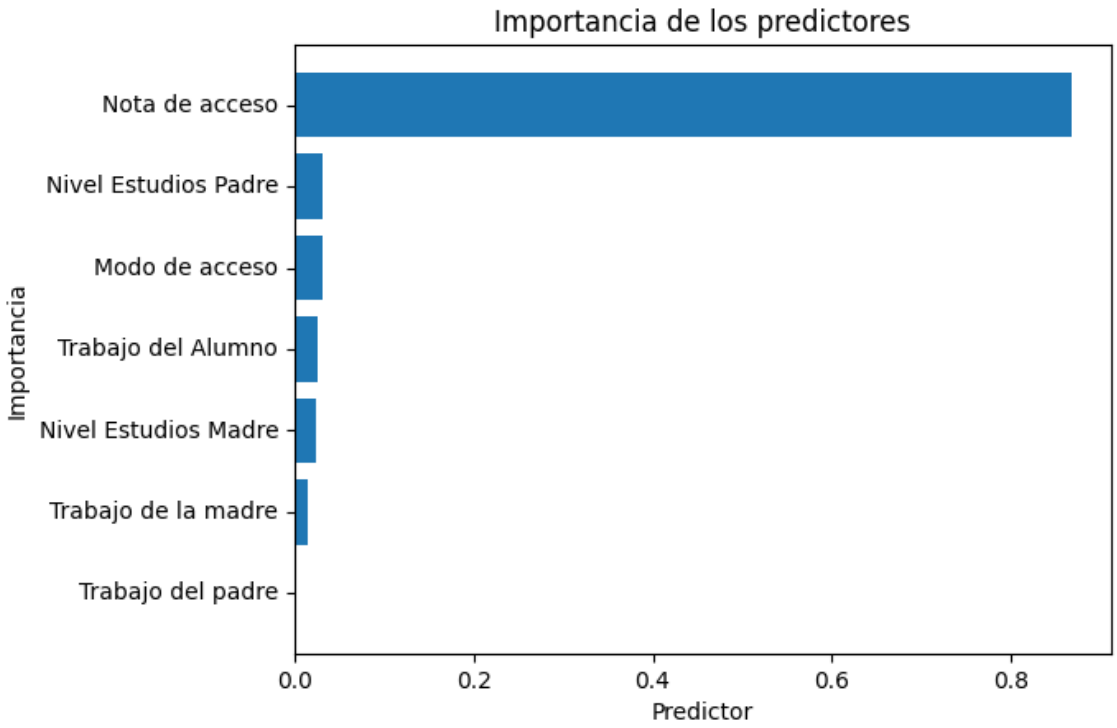

Figura 16: Gráfico de barras de la importancia de los predictores en modelo MART

Este gráfico de barras muestra la importancia de los predictores para el modelo MART muestra que la nota de acceso, que es la nota de acceso a la universidad, tiene una importancia de alrededor del 80%, lo que sugiere que esta variable es un fuerte predictor del rendimiento académico de los estudiantes. En segundo lugar, se encuentra el nivel de estudios del padre, con una importancia de alrededor del 6%, seguido en menor medida por el trabajo del estudiante, el nivel de estudios de la madre y el modo de acceso. Estos resultados indican que la nota de acceso y el nivel de estudios del padre son variables clave a tener en cuenta para mejorar el rendimiento académico de los estudiantes, y que el trabajo del estudiante y el nivel de estudios, así como el modo de acceso, también pueden tener un impacto en menor medida.

5. Conclusiones

La realización de este trabajo final de grado puede ser una oportunidad para profundizar en el conocimiento de la realidad educativa y contribuir a la mejora de la calidad de la educación. La aplicación de técnicas de estadística multivariante y aprendizaje automático permite obtener resultados más precisos y significativos, lo que añade valor a la investigación y aumenta su impacto en la comunidad académica.

El rendimiento académico de los estudiantes universitarios es un factor importante en la evaluación de la calidad de la educación superior. Como se ha visto a lo largo del trabajo, diferentes estudios han identificado varios grupos de factores que influyen en el rendimiento académico de los estudiantes, incluyendo factores personales, sociales e institucionales. Según estos estudios los factores personales incluyen competencia cognitiva, motivación, bienestar psicológico y satisfacción con los estudios. Los factores sociales incluyen el entorno familiar y el nivel socioeconómico y los factores institucionales incluyen normas y condiciones de la institución educativa. Sin embargo, la calidad del entorno de estudio y la planificación a largo plazo son predictores positivos del rendimiento académico, mientras que el apoyo social percibido, especialmente por parte de la familia y las amistades, también se ha demostrado que influye en el rendimiento académico.

Este proyecto se ha centrado en explorar las relaciones existentes entre diferentes variables y el rendimiento académico de los estudiantes universitarios. Para ello, se ha utilizado una variedad de técnicas de estadística multivariante y de aprendizaje automático, con el fin de modelar y predecir el rendimiento académico en función de diversas variables explicativas.

Por lo tanto, se presentan los resultados obtenidos a través del análisis exploratorio y la modelización de los datos. Además, se discuten las implicaciones prácticas y teóricas de los resultados, así como las posibles limitaciones y futuras líneas de investigación. En definitiva, esta investigación busca contribuir a una mejor comprensión de los factores que influyen en el rendimiento académico de los estudiantes universitarios y a la mejora de las políticas y estrategias educativas.

Como ya se explicó en otros estudios cuyo objetivo principal era predecir el rendimiento académico de los estudiantes mediante el uso de técnicas de aprendizaje automático y minería de datos han utilizado diferentes algoritmos como regresión lineal múltiple, SVM, Random Forest, árboles de decisión, y K vecinos cercanos. Los resultados varían en términos de precisión y efectividad, pero en general, se ha concluido que los algoritmos basados en árboles de decisión y Random Forest son los que tienen una mayor precisión. Pero siempre dejando la puerta a la posibilidad de estudiar más a fondo y ser considerados en futuras investigaciones que factores socioeconómicos y psicológicos y su posible impacto en el rendimiento académico.

En concordancia con las conclusiones obtenidas por Hamoud (2018) en "Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis" y Segura (2018) en

"Using Decision Trees For Predicting Academic Performance Based On Socio-Economic Factors", donde se buscó determinar si los factores socioeconómicos de los estudiantes afectan el rendimiento académico, se considera que los factores socioeconómicos de los estudiantes no tienen una relación ni influencia en las calificaciones de los estudiantes.

Así, después de explorar diferentes modelos, se decidió utilizar el modelo de Árbol de Decisión MART debido a su precisión y capacidad de generalización en la predicción. Además, su simplicidad y facilidad de interpretación lo hacen útil para entender las variables explicativas que influyen en el rendimiento académico de los estudiantes.

El análisis de la importancia de los predictores mostró que la nota de acceso a la universidad es un fuerte predictor del rendimiento académico, seguido por el nivel de estudios de los padres, pero en menor medida. También se puede decir que otros factores, como el trabajo del alumno y de la madre, así como el modo de acceso, pueden tener un ligero impacto.

En conclusión, se ha identificado el modelo de Árbol de Decisión MART como el más adecuado para predecir la nota media de los estudiantes y se ha destacado la importancia de la nota de acceso como predictor clave del rendimiento académico. Y por tanto se ha demostrado que la nota de acceso al grado universitario es el factor más determinante, con diferencia, en la predicción de la nota media de los estudiantes universitarios, y por ello del rendimiento académico global en los años en los que el estudiante cursa sus estudios universitarios.

Para futuras investigaciones sería interesante estudiar la diferencia entre el rendimiento el primer año de carrera, diferenciándolo del resto. Como se ha revisado en otras investigaciones científicas acerca del rendimiento académico, el primer año del grado suele ser un año de adaptación y de cambios que se ven influidos por otros factores aparte de la nota de acceso a la universidad, sin embargo, una vez acabado ese año de adaptación, parece ser que ya no hay otros factores externos al estudiante que influyan en su rendimiento si no que es la propia capacidad y predisposición del estudiante lo que se considera clave para el éxito académico durante los estudios universitarios. Aquellos alumnos que ya han demostrado tener un buen rendimiento académico anterior a su paso por la Universidad son aquellos que obtienen las mejores calificaciones y tienen más probabilidades se seguir demostrando ese buen rendimiento a lo largo de su carrera universitaria.

6. Referencias

Burman, I. & Som, S. (2019, febrery). Predicting Students Academic Performance Using Support Vector Machine. In 2019 *Amity International Conference on Artificial Intelligence* (AICAI), (pp. 756-759). IEEE.

Canagareddy, D., Subarayadu, K., & Hurbungs, V. (2019). A machine learning model to predict the performance of university students. In *Smart and Sustainable Engineering for Next Generation Applications: Proceeding of the Second International Conference on Emerging Trends in Electrical, Electronic and Communications Engineering (ELECOM 2018), November 28–30, 2018, Mauritius 2* (pp. 313-322). Springer International Publishing.

Canagareddy, D., Subarayadu, K., Hurbungs, V. (2019). A Machine Learning Model to Predict the Performance of University Students. In: *Fleming, P., Lacquet, B., Sanei, S., Deb, K., Jakobsson, A. (eds)* Smart and Sustainable Engineering for Next Generation Applications. ELECOM 2018. Lecture Notes in Electrical Engineering, vol 561. Springer, Cham.

Chauhan, N., Shah, K., Karn, D., & Dalal, J. (2019, April). Prediction of student's performance using machine learning. In *2nd International Conference on Advances in Science & Technology (ICAST)*.

Chiheb, F., Boumahdi, F., Bouarfa, H. y Boukraa, D., 2017. Predicting students' performance using decision trees: Case of an Algerian University. Proceedings of the 2017 International Conference on Mathematics and Information Technology, ICMIT 2017, vol. 2018-January, pp. 113-121.

Carrión Pérez, E. (2002). Validación de características al ingreso como predictores del rendimiento académico en la carrera de medicina. *Revista Cubana de Educación Médica Superior*, 16(1), 1-2.

Díez Villoria, E., Alonso, A., Verdugo Alonso, M. Á., Campo Blanco, I., Sancho, I., Sánchez, S. & Moral Cabrero, E. (2011). *Espacio Europeo de Educación Superior: estándares e indicadores de buenas prácticas para la atención a estudiantes universitarios con discapacidad*. Instituto Universitario de Integración en la Comunidad (Salamanca, España).

Esparza-Paz, F., Sánchez-Chávez, R., Esparza-Zapata, S., Esparza-Zapata, E. & Villacrés-Lara, A. (2020). Factores de rendimiento académico en estudiantes universitarios, componentes de calidad de la educación superior. Estudio de caso Facultad de Administración de Empresas, Escuela Superior Politécnica de Chimborazo. *Revista de Innovaciones Educativas* 22(33), 46-61.

Fernández, Y. O. (2011). Variables académicas que influyen en el rendimiento académico de los estudiantes universitarios. *Investigación Educativa*, 15, 27, 165-180.

Garbanzo, G. (2007). Factores asociados al rendimiento académico en estudiantes universitarios, una reflexión desde la calidad de la educación superior pública. *Revista Educación* 31(1), 43-63.

Hamoud, A., Hashim, A.S. & Awadh, W.A. (2018). Predicting student performance in higher education institutions using decision tree analysis. *International Journal of Interactive Multimedia and Artificial Intelligence, 5, 26-31.*

Lasarte, O.F., Díaz, E.R. & Sáez I.A. (2019). Rendimiento académico, apoyo social percibido e inteligencia emocional en la universidad. *European Journal of Investigation in Health, Psychology and Education*, 9(1), 39-49.

Martinez Heras, J. (18 septiembre 2020). Random Forest (Bosque Aleatorio): combinando árboles. www.IArtificial.net

Martinez Heras, J. (19 septiembre 2020). Las 7 fases del proceso de Machine Learning. www.IArtificial.net

de Miguel Díaz, M., Urquijo, P.A., Blanco, J.M.A., Escorza, T.E., Espinar, S.R.& García, J.V. (2002). Evaluación del Rendimiento Académico en la Enseñanza Superior. Comparación de resultados entre alumnos procedentes de la LOGSE y del COU. *Revista de Investigación Educativa*, 20(2), 357-383.

Plant, E. A., Ericsson K.A., Hill, L. & Asberg, K. (2005). Why study time does not predict grade point average across college students: Implications of deliberate practice for academic performance. *Contemporary Educational Psychology*, 30, (1), 96-116.

Real Decreto 1393, de 29 de octubre de 2007. Ordenación de las enseñanzas universitarias oficiales. Boletín Oficial del Estado de España, Madrid, n. 260, p. 44037-44048, 30 oct. 2007. Aprobado en Consejo de Ministros de 26 de octubre de 2007.

Real Decreto 43, de 2 de febrero de 2015. Modificación del Real Decreto 1393/2007 por el que se establece la ordenación de las enseñanzas universitarias oficiales, y del Real Decreto 99/2011 por el que se regulan las enseñanzas oficiales de doctorado. Boletín Oficial del Estado de España, Madrid, n. 29, p. 8088-8091, 03 febr. 2015. Aprobado en Consejo de Ministros el 30 enero 2015.

Real Decreto 822/2021, de 28 de septiembre de 2021, por el que se establece la organización de las enseñanzas universitarias y del procedimiento de aseguramiento de su calidad**.** Boletín Oficial del Estado de España, Madrid, n. 233, p. 119.537-119.578, 29 sep. 2021. Aprobado en Consejo de Ministros el 28 septiembre 2021.

Romero, S. (2015). Uso de Técnicas de Machine Learning para predecir el rendimiento académico de los estudiantes de la carrera de ingeniería civil en informática de la Universidad del Bio Bio, Chillán.

Segura-Morales, M.& Loza Aguirre, E. (2017, december). “Using decision trees for predicting academic performance based on socio-economic factors. In 17 *International Conference on Computational Science and Computational Intelligence (*CSCI), (

Simeone, O. (2018). A Very Brief Introduction to Machine Learning With Applications to Communication Systems. IEEE *Transactions on Cognitive Communications and Networking*, 4, (4), 648-664.

Subdirección General de Actividad Universitaria Investigadora de la Secretaría General de Universidades. (2022). Datos y cifras del sistema universitario español: curso 2021-2022. Secretaría General Técnica del Ministerio de Universidades

Anexos

Anexo 1: Código

Carga de los datos

```
import pandas as pd
path = "/Users/josep/OneDrive/TFG_ordenado/datos_limpios_finales1.xlsx"
data = pd.read_excel(path)
pd.options.display.max_columns = None
data.head()
```

Hiper parámetros KNN

```
# Definir los posibles valores de k y las métricas de distancia
param_grid = {
    'n_neighbors': list(range(1, 101)),
    'weights': ['uniform', 'distance'],
    'metric': ['euclidean', 'manhattan']
}
```

Hiper parámetros SVR

```
# Definir los parámetros a ajustar en la búsqueda de grid
param_grid = {'C': [0.1, 1, 10, 100],
              'kernel': ['linear', 'rbf'],
              'degree': [2, 3, 4, 6, 8, 10],
              'gamma': [0.1, 0.01, 10, 'scale', 'auto'],
              'shrinking': [True, False]}
```

Hiper parámetros Random Forest

```
param_grid = {'n_estimators': [50, 100, 200, 500],
              'max_depth': [5, 10, 20, None],
              'min_samples_split': [2, 5, 10],
              'min_samples_leaf': [1, 2, 4, 6, 8],
              'max_features': [1.0, 'sqrt', 'log2'],
              'bootstrap': [True, False]}
```

Hiper parámetros árboles de decisión

```
param_grid = {'max_depth': [ 3,5, 7, 9, 11, 13],
              'min_samples_split': [2, 4, 6, 8],
              'max_features': [ 4, 6, 8],
              'min_samples_leaf': [1, 2, 3],
              'max_leaf_nodes': [10, 15, 20, 25, 30, 35]}
```

Evaluación el modelo con Validación Cruzada

```
# Evaluar el modelo con validación cruzada
scores = cross_val_score(best_model, X_train, y_train, cv=5, scoring='neg_mean_squared_error')
rmse_scores = np.sqrt(-scores)
mae_scores = -cross_val_score(best_model, X_train, y_train, cv=5, scoring='neg_mean_absolute_error')
mse_scores = -scores
r2_scores = cross_val_score(best_model, X_train, y_train, cv=5, scoring='r2')


# Mostrar los resultados de la evaluación con validación cruzada
print("Evaluación con validación cruzada:")
print("RMSE: ", rmse_scores.mean())
print("MAE: ", mae_scores.mean())
print("MSE: ", mse_scores.mean())
print("R2: ", r2_scores.mean())
```

Evaluación del modelo conjunto de prueba

```
# Evaluar el modelo con el conjunto de prueba
y_pred = grid_search.best_estimator_.predict(X_test)
print("RMSE en el conjunto de prueba:")
print(mean_squared_error(y_test, y_pred, squared=False))
print("MSE en el conjunto de prueba:")
print(mean_squared_error(y_test, y_pred))
print("MAE en el conjunto de prueba:")
print(mean_absolute_error(y_test, y_pred))
print("R2 en el conjunto de prueba:")
print(r2_score(y_test, y_pred))
```

Anexo 2: ODS

OBJETIVOS DE DESARROLLO SOSTENIBLE

Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS).

Objetivos de Desarrollo Sostenibles	Alto	Medio	Bajo	No Procede
ODS 1. **Fin de la pobreza.**				X
ODS 2. **Hambre cero.**				X
ODS 3. **Salud y bienestar.**				X
ODS 4. **Educación de calidad.**	X			
ODS 5. **Igualdad de género.**				
ODS 6. **Agua limpia y saneamiento.**				X
ODS 7. **Energía asequible y no contaminante.**				X
ODS 8. **Trabajo decente y crecimiento económico.**		X		
ODS 9. **Industria, innovación e infraestructuras.**				X
ODS 10. **Reducción de las desigualdades**.				X
ODS 11. **Ciudades y comunidades sostenibles**.				X
ODS 12. **Producción y consumo responsables.**				X
ODS 13. **Acción por el clima.**				X
ODS 14. **Vida submarina.**				X
ODS 15. **Vida de ecosistemas terrestres.**				X
ODS 16. **Paz, justicia e instituciones sólidas.**				X
ODS 17. **Alianzas para lograr objetivos.**				X

Reflexión sobre la relación del TFG/TFM con los ODS y con el/los ODS más relacionados.

Los **Objetivos de Desarrollo Sostenible** pretenden ampliar los Objetivos de Desarrollo del Milenio (ODM) y alcanzar aquellos objetivos que no se cumplieron. Todos los países, con independencia de su nivel de desarrollo o riqueza, se comprometen a promover la prosperidad y a proteger el medioambiente. Los Objetivos de Desarrollo Sostenible no son obligatorios pero cada país asume la responsabilidad de trabajar por su cumplimiento.

En este TFG de los 17 Objetivo de Desarrollo Sostenible (ODS) podemos hablar de relación directa en mayor o menor grado con 2 de ellos.

La educación es en sí mismo uno de los pilares de la sociedad y por tanto se convierte en un ODS determinante. De esta forma el Objetivo de Desarrollo Sostenible (ODS) número 4 de la Agenda 2030 de las Naciones Unidas se refiere a la educación de calidad, y tiene como objetivo garantizar una educación inclusiva, equitativa y de calidad para todos. Ya que en el presente trabajo se ha pretendido determinar el cómo algunos factores pueden afectar o no al rendimiento académico de los estudiantes universitarios, podemos decir que existen varias formas en que el rendimiento académico de los estudiantes universitarios puede estar relacionado con este ODS:

1. Acceso a la educación: El ODS 4 busca asegurar que todos los niños y jóvenes tengan acceso a una educación de calidad. Si los estudiantes universitarios no tienen acceso a una educación de calidad, es probable que su rendimiento académico se vea afectado. Por lo tanto, para mejorar el rendimiento académico de los estudiantes universitarios, es necesario garantizar que tengan acceso a una educación de calidad.

2. Calidad de la educación: El ODS 4 también busca mejorar la calidad de la educación. Si la calidad de la educación no es buena, es probable que los estudiantes universitarios no aprendan lo suficiente y su rendimiento académico se vea afectado. Por lo tanto, para mejorar el rendimiento académico de los estudiantes universitarios, es necesario mejorar la calidad de la educación.

3. Equidad en la educación: El ODS 4 busca garantizar que la educación sea equitativa para todos, independientemente de su género, origen socioeconómico o discapacidad. Si los estudiantes universitarios enfrentan barreras para acceder a una educación de calidad debido a su género, origen socioeconómico o discapacidad, es probable que su rendimiento académico se vea afectado. Por lo tanto, para mejorar el rendimiento académico de los estudiantes universitarios, es necesario abordar las desigualdades en el acceso a la educación y garantizar que todos tengan acceso a una educación de calidad.

En resumen, el rendimiento académico de los estudiantes universitarios está estrechamente relacionado con el ODS número 4 de la Agenda 2030 de las Naciones Unidas. Para mejorar el rendimiento académico de los estudiantes universitarios, es necesario garantizar que tengan acceso a una educación de calidad, mejorar la calidad de la educación y abordar las desigualdades en el acceso a la educación.

En relación con el objetivo de Desarrollo Sostenible (ODS) número 8 de la Agenda 2030 de las Naciones Unidas, se refiere al trabajo decente y crecimiento económico, y tiene como objetivo

promover el crecimiento económico sostenible, inclusivo y sostenido, y el empleo pleno y productivo. Podemos decir que esto se relaciona con el rendimiento académico de los estudiantes en la medida en que el tener un buen rendimiento académico supondrá mayor habilidad para conseguir empleos mejores, para Desarrollo de habilidades y competencias que les permita ser más productivos y apoyar el crecimiento y también un buen rendimiento académico también pueden ser más propensos a ser innovadores y emprendedores, lo que puede contribuir al crecimiento económico sostenible y al desarrollo de nuevas empresas y empleos.