Universidad de Valladolid

MÁSTER EN PROFESOR DE EDUCACIÓN SECUNDARIA OBLIGATORIA Y BACHILLERATO, FORMACIÓN PROFESIONAL Y ENSEÑANZAS DE IDIOMAS

ESPECIALIDAD DE FÍSICA Y QUÍMICA

TRABAJO FIN DE MÁSTER

OBSERVACIÓN DEL TRABAJO Y CIRCUSTANCIAS SOCIALES DE LOS ALUMNOS Y ANÁLISIS MEDIANTE IA

Estudiante: D Alberto Francisco Cepeda Nieto

Tutor: Dr. D. Manuel Ángel González Delgado

Valladolid, 2025

Resumen/Abstract.

Este trabajo explora la aplicación de la inteligencia artificial, concretamente modelos de regresión y aprendizaje automático, para predecir el rendimiento académico en Física y Química de alumnos de Educación Secundaria Obligatoria (ESO) en el Colegio San Agustín de Valladolid. A través de Python y bibliotecas especializadas de este, como Scikit-learn, se ha desarrollado un conjunto de modelos que analizan factores personales, familiares y académicos, y son capaces, a partir de dichos factores, de predecir el rendimiento académico del alumno. El estudio muestra cómo el modelo Random Forest es el que mejor precisión logra en la predicción del rendimiento y en especial en el curso de 2ºESO. Este trabajo es una muy buena primera aproximación a los modelos de predicción y accesible para quien quiera introducirse en el análisis predictivo aplicado a la educación mediante inteligencia artificial.

This paper explores the application of artificial intelligence, specifically regression and machine learning models, to predict the academic performance in Physics and Chemistry of students in Compulsory Secondary Education (ESO) at the San Agustín School in Valladolid. Using Python and specialised Python libraries, such as Scikit-learn, a set of models have been developed that analyse personal, family and academic factors, and are capable, based on these factors, of predicting the student's academic performance. The study shows that the Random Forest model is the one that achieves the best accuracy in predicting performance, especially in the second year of secondary school. This work is a very good first approach to prediction models and is accessible to anyone who wants to enter into predictive analysis applied to education using artificial intelligence.

ÍNDICE

ÍNDICE DE TABLAS, GRÁFICAS Y FIGURAS.

1. Introducción.

Comprender y mejorar el rendimiento académico de los estudiantes ha sido siempre una de las principales preocupaciones en el ámbito educativo, especialmente en etapas clave como la Educación Secundaria Obligatoria (ESO). Factores personales, familiares, escolares y sociales se unen de forma compleja para influir en los resultados académicos, lo que ha llevado a los educadores a buscar herramientas más eficaces para identificar y poder anticipar así las situaciones de bajo rendimiento.

En este contexto, los avances en la tecnología han dado lugar a nuevos métodos de análisis educativo. En particular, la inteligencia artificial (IA) y el aprendizaje automático (machine learning) tienen un gran potencial en el tratamiento de este tipo de datos, permitiendo no solo detectar patrones ocultos, sino también predecir resultados académicos con un alto grado de precisión. Estas herramientas permiten dejar atrás un enfoque que solo actúa cuando el problema ya ha aparecido, y avanzar hacia uno más preventivo, donde sea posible anticiparse y adaptar las intervenciones antes de que la situación se convierta en un problema educativo visible o difícil de revertir.

El trabajo se ha realizado con alumnos de la ESO del Colegio San Agustín de Valladolid, y tiene como objetivo principal desarrollar un modelo predictivo del rendimiento académico para dichos alumnos. Este modelo se ha construido a partir de datos recogidos mediante una encuesta que aborda factores personales (hábitos de estudio, motivación, uso de tecnología), familiares (apoyo en casa, ambiente de estudio), escolares (relación con el profesorado, atención en clase) y sociales (relaciones entre iguales, actividades extraescolares).

Todo esto se ha realizado mediante el lenguaje de programación Python y librerías especializadas en machine learning como Scikit-learn, que permitirán entrenar modelos como los árboles de decisión y Random Forest.

Por tanto, este trabajo pretende aportar una primera aproximación práctica al uso de técnicas de inteligencia artificial mediante los modelos de predicción en la ESO, investigando cómo los datos pueden convertirse en un recurso muy útil para mejorar la planificación y las estrategias pedagógicas del centro.

2. Objetivos.

2.1. Objetivo general.

Ofrecer una primera aproximación al uso de modelos predictivos aplicados al rendimiento académico de estudiantes de la ESO del Colegio San Agustín de Valladolid en la asignatura de Física y Química, a partir del análisis de diversos factores personales, familiares, escolares y sociales obtenidos mediante una encuesta.

2.2. Objetivos específicos.

Diseñar y realizar una encuesta a los estudiantes de ESO del Colegio San Agustín para recoger información sobre factores personales, familiares, escolares y sociales relacionados con su rendimiento académico.

Procesar y preparar los datos obtenidos para su análisis, incluyendo la transformación de variables y la preparación de los datos para su tratamiento computacional.

Entrenar los modelos de predicción del rendimiento académico, regresión lineal, árboles de decisión y Random Forest, utilizando Python y sus librerías especializadas, Panda y Scikit-learn.

Evaluar la capacidad de los modelos poniendo a prueba su capacidad predictiva sobre nuevos datos.

Interpretar los resultados obtenidos, identificando los factores con mayor peso predictivo, su futura utilidad y que líneas de mejora seguir para investigaciones futuras.

3. Fundamento Teórico.

El rendimiento académico es un concepto fundamental en el ámbito educativo, ya que permite evaluar el grado de aprendizaje y desarrollo de los estudiantes a lo largo de su formación. Se mide a través de distintos indicadores, como calificaciones, participación en clase, habilidades adquiridas y cumplimiento de objetivos curriculares (Arribas, 2012). Sin embargo, este desempeño no depende únicamente del esfuerzo individual del estudiante, sino que está influenciado por una variedad de factores, como el entorno familiar, los métodos de enseñanza, la motivación, el acceso a recursos educativos y el bienestar emocional (Beneyto Sánchez, 2013).

En el contexto de la LOMLOE (Ley Orgánica 3/2020, de 29 de diciembre) el rendimiento académico se entiende como el resultado del proceso educativo que no solo se limita a las calificaciones, sino también al desarrollo integral del estudiante. La ley pone un énfasis especial en la personalización del aprendizaje, la equidad en la educación y la evaluación continua, lo que impacta directamente en la forma en que se mide el rendimiento académico.

Además, la LOMLOE refuerza la necesidad de un enfoque inclusivo y personalizado en la educación, lo cual afecta directamente al rendimiento académico de los estudiantes. Esto se refleja en el artículo 78, que subraya la importancia de adaptar la enseñanza a las necesidades y características individuales del alumnado, buscando reducir las desigualdades y promoviendo la igualdad de oportunidades. Por lo tanto, el rendimiento académico debe ser evaluado teniendo en cuenta las diversas condiciones de los estudiantes, reconociendo que cada uno tiene un ritmo y estilo de aprendizaje diferente.

En base a todo lo anterior, el éxito académico se define como el resultado de diversos factores, entre los cuales se incluyen el rendimiento académico, el cumplimiento de los objetivos de aprendizaje, el desarrollo de habilidades y competencias, así como la satisfacción personal y la persistencia (York et al., 2015).

Figura 1. *Modelo del éxito académico (York et al., 2015).*

Por todo esto, a lo largo de los años, numerosas investigaciones han buscado comprender los determinantes del rendimiento académico con el fin de diseñar estrategias que favorezcan el aprendizaje y reduzcan las brechas educativas. Desde esta perspectiva, se ha señalado que aspectos como los hábitos de estudio, el apoyo docente y los recursos institucionales juegan un papel clave en el desempeño estudiantil (Miguéis et al., 2018). En este contexto, resulta esencial analizar las condiciones que pueden potenciar o dificultar el éxito estudiantil, así como las formas en que el sistema educativo puede adaptarse para garantizar un desarrollo óptimo en los alumnos.

En los últimos años, la Inteligencia Artificial (IA) ha surgido como una de las tecnologías más disruptivas en distintos sectores, incluyendo la educación. Su aplicación en entornos académicos ha permitido innovaciones significativas en la manera en que se diseñan, imparten y evalúan los procesos de enseñanza y aprendizaje. De acuerdo con la UNESCO (2023), la IA tiene el potencial de transformar el sistema educativo al proporcionar herramientas avanzadas para personalizar la instrucción, automatizar tareas administrativas y predecir el rendimiento académico de los estudiantes. Esta tecnología no solo optimiza la gestión educativa, sino que también facilita la detección temprana de estudiantes en riesgo, permitiendo la implementación de estrategias de apoyo específicas que pueden mejorar sus resultados académicos.

Como se ha dicho anteriormente, el rendimiento académico es un aspecto central en la educación y está determinado por una variedad de factores personales, familiares, institucionales y sociales. La capacidad de predecir este rendimiento a través de herramientas tecnológicas es un avance significativo, ya que permite la identificación de

patrones que pueden orientar la toma de decisiones tanto a nivel institucional como individual. La predicción del desempeño estudiantil mediante IA es un campo de investigación en crecimiento, con aplicaciones que van desde el análisis de datos educativos hasta la optimización de estrategias pedagógicas (Zawacki-Richter et al., 2019).

En este contexto, las instituciones educativas han comenzado a implementar modelos basados en IA para analizar grandes volúmenes de datos y generar predicciones sobre el rendimiento de los estudiantes. Estas predicciones pueden basarse en variables como el historial académico, la asistencia a clases, la participación en actividades extracurriculares y el entorno socioeconómico del estudiante. Estudios recientes han demostrado que la combinación de estas variables con algoritmos de aprendizaje automático permite obtener modelos predictivos altamente precisos, lo que facilita la toma de decisiones informadas por parte de docentes y administradores educativos (Oppong, 2023; Romero & Ventura, 2020).

Los enfoques de inteligencia artificial para la predicción del rendimiento escolar incluyen modelos de clasificación y regresión. Entre los algoritmos más utilizados destacan los árboles de decisión, las máquinas de soporte vectorial (SVMs), las redes neuronales artificiales (ANNs) y los Random Forest, cada uno con diferentes capacidades para modelar relaciones jerárquicas y no lineales en grandes volúmenes de datos (Rastrollo et al., 2020).

Sin embargo, la incorporación de IA en la educación también plantea una serie de desafíos. Entre ellos, se encuentran cuestiones éticas relacionadas con la privacidad de los datos, la interpretabilidad de los modelos utilizados y el riesgo de sesgos algorítmicos que puedan afectar la equidad en la educación (Mishara, 2024). Además, a pesar de los avances en este campo, aún existe una brecha significativa en la aplicación de estas tecnologías en la educación superior, donde se requiere un análisis más detallado de los factores que influyen en el rendimiento académico de los estudiantes (Zawacki-Richter et al., 2019).

3.1.Impacto de la IA en la Educación.

El impacto de la inteligencia artificial (IA) en la educación es cada vez más significativo y abarca diversos aspectos que transforman tanto la manera en que los estudiantes aprenden como la forma en que los docentes enseñan. Uno de los avances más destacados

es la personalización del aprendizaje. Gracias a los algoritmos de IA, es posible ajustar los contenidos educativos al ritmo, nivel y necesidades concretas de cada estudiante, permitiendo un enfoque mucho más individualizado que el de los modelos tradicionales. Esta capacidad de personalización se traduce en una mejora notable en la retención de conocimiento, ya que cada alumno puede avanzar según su propio ritmo, profundizando en los temas que el alumno necesita reforzar y de mayor interés. Además, esta personalización contribuye al desarrollo de habilidades específicas, adaptadas no solo a las capacidades del estudiante, sino también a sus intereses y potencialidades (Zawacki-Richter et al., 2019).

Otro avance relevante dentro del impacto de la IA son los sistemas de tutoría inteligente, que funcionan como asistentes virtuales diseñados para proporcionar apoyo al estudiante en tempo real. Estos sistemas analizan tanto las respuestas dadas por los alumnos como sus patrones de interacción, lo que les permite detectar cuándo un estudiante presenta dificultades o errores conceptuales. A partir de este análisis, los tutores inteligentes no solo ofrecen explicaciones adicionales o más claras, sino que también son capaces de sugerir ejercicios personalizados o adaptar las rutas de aprendizaje para reforzar los conocimientos. Este tipo de herramientas no solo mejora la experiencia de aprendizaje, sino que también fomenta la autonomía del estudiante, al brindarle recursos inmediatos para resolver dudas y avanzar de manera más independiente (Liu et al., 2025).

Además, la evaluación automatizada representa uno de los avances más prácticos y extendidos derivados de la integración de la IA en la educación. Gracias a técnicas como el procesamiento de lenguaje natural y el análisis de patrones, los sistemas automatizados son capaces de corregir exámenes, tareas y trabajos de forma rápida y precisa. Esto no solo permite al estudiante recibir retroalimentación casi inmediata, sino que también reduce de forma considerable la carga administrativa para los docentes. Al liberar tiempo que antes dedicaban a tareas repetitivas de corrección, los profesores pueden centrarse en aspectos mucho más cualitativos, como planificación pedagógica, el diseño de materiales o el acompañamiento emocional y motivacional de sus alumnos (Chen et al., 2020).

Además de las aplicaciones centradas en la personalización del aprendizaje, la tutoría inteligente y la evaluación automatizada, la IA también tiene un impacto significativo en otros aspectos clave del ámbito educativo. Uno de ellos es el análisis predictivo del rendimiento académico. A partir del estudio de datos históricos y en tiempo real, los

algoritmos de IA pueden prever el rendimiento futuro de los estudiantes, identificando patrones que podrían pasar desapercibidos para los docentes. Esta capacidad predictiva permite detectar de manera temprana a aquellos alumnos que podrían estar en riesgo de bajo rendimiento o abandono escolar, lo que a su vez facilita la implementación de estrategias de intervención personalizadas para mejorar sus resultados académicos (Miguéis et al., 2018).

Otro ámbito donde la IA muestra un impacto importante es en la optimización de la gestión educativa. Las instituciones académicas pueden aprovechar las herramientas basadas en inteligencia artificial para gestionar de manera más eficiente los recursos disponibles, desde la planificación de horarios y la asignación de docentes, hasta la administración de materiales didácticos. Este tipo de aplicaciones no solo mejora la organización interna de los centros educativos, sino que también permite liberar tiempo y esfuerzos que puedan ser redirigidos hacia la mejora de la calidad educativa (Sposato, 2025).

Por último, resulta fundamental destacar el papel de la IA en la creación de entornos de realidad virtual y aprendizaje inmersivo. La combinación de inteligencia artificial con tecnologías como la realidad aumentada y la realidad virtual permite generar experiencias educativas mucho más interactivas y atractivas para los estudiantes. Estas herramientas favorecen la compresión de conceptos complejos mediante la experimentación, la simulación y el aprendizaje activo, abriendo nuevas posibilidades para el diseño de actividades innovadoras que despierten el interés y la motivación del alumnado (Liu et al., 2017).

En conjunto, todas las aplicaciones muestran como la IA no solo transforma el aprendizaje individual, sino que también impacta en la organización, la gestión y las metodologías educativas, posicionándose como una herramienta clave en la educación.

3.2.Beneficios y desafíos de la IA en la Educación.

El uso de inteligencia artificial en el ámbito educativo puede traer una serie de beneficios que tienen un impacto directo tanto en la calidad del aprendizaje como en la eficiencia general del sistema educativo. Uno de los beneficios más relevantes es la posibilidad de avanzar hacia una mayor equidad en el acceso a la educación. Las herramientas basadas en IA pueden ser de gran ayuda para reducir barreras que hasta ahora limitaban en gran medida a los estudiantes. Por ejemplo, los sistemas de aprendizaje adaptativo pueden

ajustarse a alumnos con distintas capacidades, permitiendo que personas con discapacidades, dificultades específicas de aprendizaje o incluso con problemas de acceso físico a las aulas puedan beneficiarse de materiales adaptados a sus necesidades. Esto no solo amplía las oportunidades de aprendizaje, sino que también contribuye a generar entornos educativos más inclusivos y justos (Pagliara et al., 2024).

Otro beneficio destacado es la mejora en el seguimiento y la retroalimentación que reciben tanto los estudiantes como los docentes. Gracias a los sistemas de IA, es posible supervisar de forma continua el progreso de los alumnos, generando datos en tiempo real sobre su desempeño. Esto permite detectar rápidamente si un estudiante está teniendo dificultades en un área concreta y adaptar las estrategias pedagógicas antes de que el problema empeore. Además, al contar con información precisa y actualizada, los docentes pueden tomar decisiones para ajustar sus clases, priorizar contenidos o diseñar actividades acordes con las necesidades reales de su alumnado (Zawacki-Richter et al., 2019).

Por último, un beneficio que no se puede pasar por alto es la optimización del tiempo del docente. Las tareas repetitivas y administrativas suelen consumir gran parte del tiempo del profesorado. La automatización de estas tareas mediante herramientas de IA permite liberar tiempo que puede ser redirigido hacia aspectos más valiosos del proceso educativo (Chen et al., 2020).

En conjunto, todos los beneficios muestran que la IA no es solo una herramienta tecnológica, sino una oportunidad de transformar la educación, haciéndola más personalizada y eficiente.

Aunque los beneficios de la IA en la educación son numerosos y prometedores, también es necesario reconocer que su implementación presenta una serie de desafíos importantes para tener en cuenta.

Uno de los desafíos más relevantes es el relacionado con la ética y la privacidad de los datos. Para que las herramientas de IA funcionen correctamente, es necesario recopilar y procesar grandes cantidades de datos personales sobre los estudiantes: sus respuestas, sus progresos, sus patrones de aprendizaje, entre otros. Esto plantea preguntas sobre quién tiene acceso a esos datos, cómo se almacenan y con qué fines se utilizan. Es necesario que las instituciones educativas cuenten con regulaciones claras y transparentes que protejan la privacidad de los alumnos y garanticen el uso responsable y ético de la información recogida (Hakimi et al., 2021).

Otro desafío importante es la resistencia al cambio, algo que no sorprende si se piensa en la magnitud de cambio que suponen dichas tecnologías. Tanto docentes como estudiantes pueden llegar a sentirse inseguros o reticentes a modificar sus métodos tradicionales de enseñanza y aprendizaje. Implementar herramientas de IA, no es solo introducir tecnología, sino que supone repensar dinámicas, roles y formas de relacionarse en el aula, esto implica un proceso de adaptación que requiere tiempo, formación y apoyo institucional (Liu et al., 2025).

Por último, se ha de tener en cuenta la brecha tecnológica. No todas las instituciones educativas cuentan con los recursos necesarios para incorporar las herramientas IA en sus prácticas diarias. Esto puede generar desigualdades en el acceso a las ventajas que ofrece esta tecnología, ampliando la diferencia entre aquellos centros que pueden permitirse estas innovaciones y aquellos que no (Liu et al., 2017).

Además de la brecha tecnológica que existe entre instituciones, también se puede observar desigualdad significativa entre alumnos. Los estudiantes que asisten al mismo centro pueden tener experiencias educativas muy distintas dependiendo de su disponibilidad de Internet o los recursos materiales disponibles por cada uno. Esta diferencia condiciona significativamente el grado en que los estudiantes pueden sacar provecho de las herramientas digitales o recursos IA que el centro educativo les ofrece (Pierce & Cleary, 2024).

En definitiva, aunque la inteligencia artificial abre numerosas puertas en el ámbito educativo, también se considera necesario reflexionar cuidadosamente sobre cómo implementarla de manera justa y ética, de forma que sus beneficios puedan llegar realmente a todos los estudiantes, evitando que se convierta en un nuevo factor de desigualdad.

3.3.Factores determinantes en la predicción del Rendimiento Académico.

Uno de los usos más destacados de la inteligencia artificial en educación, como se ha mencionado anteriormente, es su capacidad de realizar análisis predictivos sobre el rendimiento académico de los estudiantes. Sin embargo, para que estas predicciones sean acertadas, es fundamental entender primero qué factores influyen en el rendimiento académico del alumnado. Solo conociendo bien estos elementos es posible construir buenos modelos que permitan personalizar la enseñanza, anticiparse a las dificultades de los estudiantes y diseñas intervenciones pedagógicas efectivas.

El rendimiento académico de los estudiantes viene determinado por una combinación de factores personales, familiares, escolares y sociales, los cuales interactúan entre sí influyendo de manera directa o indirecta en el rendimiento académico de los alumnos. Entre los factores personales se incluyen elementos como los hábitos de estudio, la motivación individual, la disciplina, el uso del tiempo libre y el acceso a recursos educativos, todos ellos fundamentales para establecer rutinas efectivas de aprendizaje (Yao et al., 2019). Por otro lado, los factores familiares abarcan aspectos como los ingresos económicos del hogar, el nivel educativo de los padres, el ambiente emocional y físico en el que vive el estudiante, así como el grado de apoyo y acompañamiento que recibe para poder realizar las tareas y actividades académicas (González-Pienda & Núñez, 2004). En cuanto a los factores escolares o institucionales, destacan la calidad y preparación de los docentes, las técnicas pedagógicas aplicadas en el aula, la disponibilidad de recursos tecnológicos y materiales didácticos, y los métodos de enseñanza utilizados por la institución educativa (Etxeberria et al., 2017). Finalmente, los factores sociales hacen referencia al estrato socioeconómico del entorno del estudiante, al acceso a una infraestructura educativa adecuada y al contexto social general en el que se desenvuelve el estudiante (Fernandes et al., 2019).

El uso de herramientas de IA permite analizar todos estos factores, identificando correlaciones relevantes entre ellos y estableciendo modelos predictivos que faciliten la personalización de la enseñanza según las necesidades individuales de cada alumno (Sajja et al., 2023).

3.4.Algoritmos de IA para la predicción del Rendimiento Académico.

Una vez se han identificado los principales factores que influyen en el rendimiento académico, se puede explorar cómo la inteligencia artificial puede utilizar esta información para construir modelos predictivos eficaces. En este sentido, diversos algoritmos y técnicas de IA se han desarrollado específicamente para analizar datos educativos y anticipar el desempeño de los estudiantes.

El desarrollo de modelos de IA en la educación se basa principalmente en dos tipos de aprendizaje automático, el aprendizaje supervisado y el aprendizaje no supervisado. El aprendizaje supervisado implica entrenar el modelo utilizando conjuntos de datos previamente etiquetados, es decir, datos donde se conoce de antemano la respuesta correcta o la categoría deseada. En el contexto educativo, esto significa que se introducen

al modelo ejemplos históricos (por ejemplo, datos de las calificaciones finales de los estudiantes), y el algoritmo aprende a reconocer patrones que le permitan predecir resultados similares en casos nuevos. Al contrario, el aprendizaje no supervisado trabaja con datos no etiquetados, tratando de identificar por sí mismo estructuras dentro de la información (Almuqati et al., 2024).

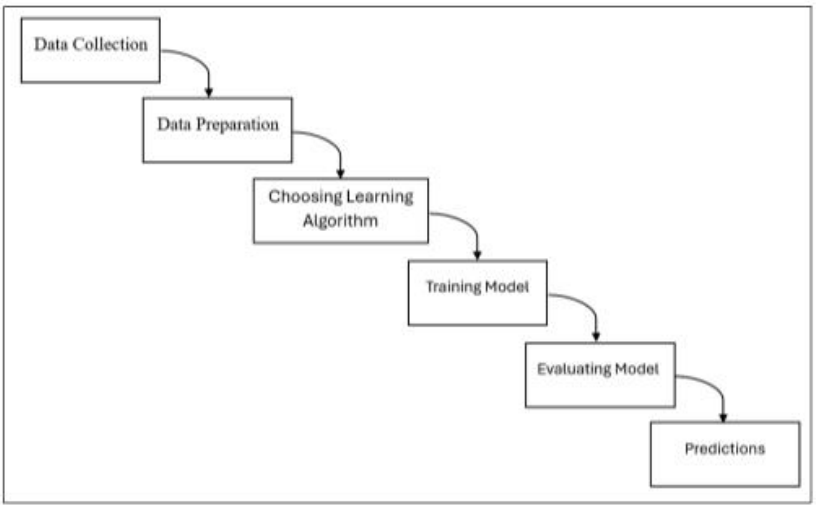


Figura 2. *Flujo de trabajo para el Machine Learning (Almuqati et al., 2024).*

Dentro de los modelos o algoritmos más utilizado para la predicción del rendimiento académico destacan varias aproximaciones que han demostrado su efectividad en distintos contextos educativos. Uno de ellos es la regresión logística, que permite predecir la probabilidad de que un estudiante alcance un determinado nivel de rendimiento en función de variables como la asistencia, los hábitos de estudio o la interacción con plataformas educativas (Miguéis et al., 2018). Este tipo de modelo es especialmente útil cuando se trabaja con resultados binarios o categóricos, como aprobar o suspender una materia.

Otro grupo de algoritmos ampliamente utilizados son los árboles de decisión y los Random Forest, los cuales ayudan a identificar patrones en el comportamiento de los estudiantes y a clasificarlos en diferentes categorías de rendimiento. Estos modelos permiten no solo hacer predicciones, sino también visualizar de manera clara cuáles son las variables que más peso tienen en las decisiones del modelo (Huynh-Cam et al., 2021).

Los árboles de decisión funcionan dividiendo reiterativamente un conjunto de datos en subconjuntos más pequeños basándose en pruebas de atributos, creando una estructura similar a un árbol, donde cada nodo representa una pregunta o condición sobre una

variable, y cada rama representa la respuesta a esa pregunta. El modelo va "bajando" por el árbol, tomando decisiones en cada paso, hasta llegar a un nodo hoja final que da el resultado. (por ejemplo, una clasificación de aprobado/suspendido o una predicción de nota). Este modelo permite ver de forma clara y gráfica como se llega a una conclusión, mostrando que factores pesan más en cada predicción (Quinlan, 1986).

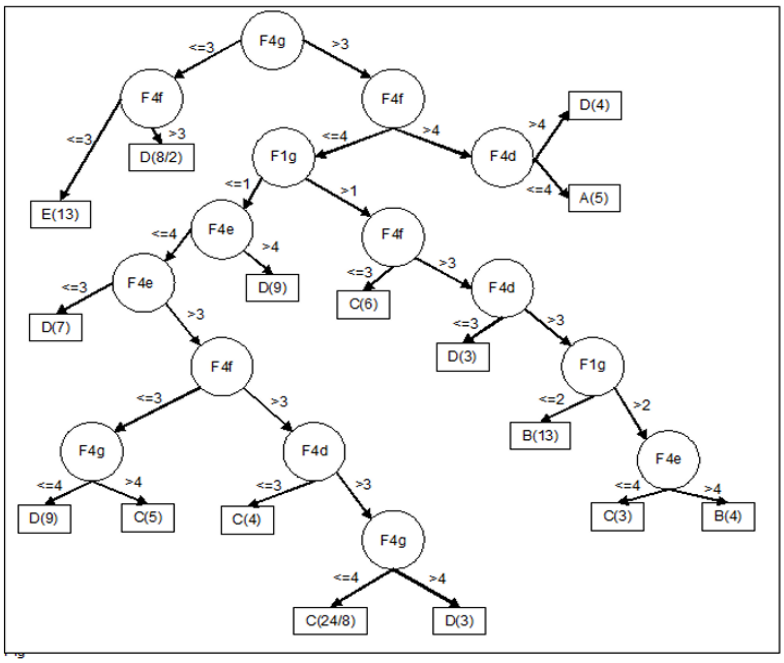


Figura 3. *Árbol de decisión, (Castrillón et al., 2020).*

Por otro lado, el algoritmo Random Forest es una técnica de ensamblado que construye múltiples arboles de decisión durante el entrenamiento y produce la media de las predicciones (regresión) de los árboles individuales. Por lo tanto, no depende de un solo árbol, sino que construye múltiples arboles de decisión, cada uno usando una parte diferente de los datos y seleccionando variables al azar. Después, combina las respuestas de todos esos árboles para dar lugar a una predicción final. Esta técnica mejora la precisión del modelo y reduce errores como el sobreajuste, que ocurre cuando un árbol de decisión se adapta demasiado a los datos de entrenamiento y falla con datos nuevos (Breiman, 2001).

Otro modelo de predicción son las redes neuronales, las cuales representan modelos avanzados, inspirados en el funcionamiento del cerebro humano, que son capaces de aprender patrones complejos y realizar predicciones más precisas al identificar interacciones no lineales entre múltiples factores. Cada neurona recibe información, la procesa y transmite el resultado a las neuronas de la siguiente capa. Lo interesante de este

modelo es que es capaz de aprender a partir de ejemplos. Durante el proceso de entrenamiento, la red neuronal ajusta sus conexiones internas para mejorar su capacidad de predecir correctamente. Cuantos más datos recibe, más aprende a reconocer patrones complejos que pueden ser difíciles de detectar a simple vista. La principal limitación de este modelo es que su interpretación a menudo es muy compleja, es decir, aunque ofrece resultados precisos, no es fácil entender como ha llegado a ellos (Gil-Vera et al., 2021).

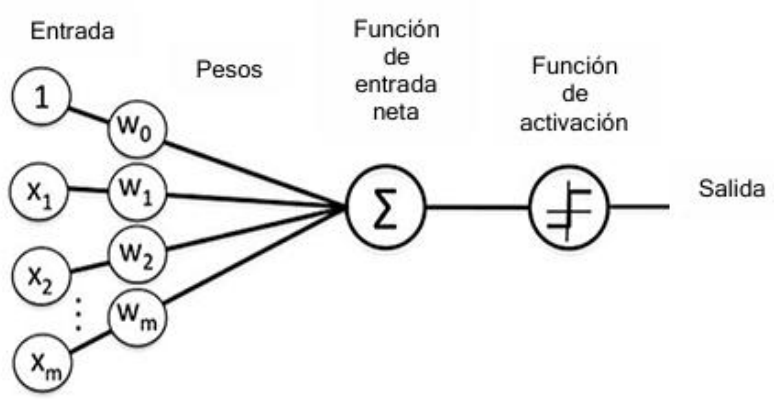


Figura 4. *Capas red neuronal. Adaptado de (Nicholson 2019).*

Por último, las máquinas de soporte vectorial (SVM) son utilizadas para clasificar a los estudiantes en función de su desempeño, basándose en datos históricos que permiten encontrar los límites óptimos entre diferentes grupos de rendimiento. Estos algoritmos de aprendizaje supervisado son utilizados en tareas de clasificación, es decir, su objetivo principal es encontrar la mejor manera de separar distintos grupos de datos. En el contexto educativo, se utilizaría para clasificar a los estudiantes en diferentes categorías de rendimiento a partir de datos históricos como calificaciones, asistencia o hábitos de estudio (Yağcı, 2022).

Diversos estudios han evidenciado la eficacia de los algoritmos de inteligencia artificial en el análisis y predicción del rendimiento académico de los estudiantes. Por ejemplo, Castrillón et al. (2020) aplicaron técnicas de clasificación, incluyendo redes neuronales artificiales y árboles de decisión, para predecir el rendimiento académico de estudiantes universitarios en Colombia, logrando una precisión del 91,7%. Otro estudio es el de Maulana et al. (2023), donde utilizaron diferentes modelos con un porcentaje de precisión mayor del 85% en todos ellos, el que mayor precisión tuvo fue el de Random Forest.

Estos estudios enseñan cómo los algoritmos de inteligencia artificial pueden proporcionar información valiosa para anticipar resultados académicos, permitiendo diseñar

intervenciones educativas más ajustadas a las necesidades de los estudiantes y, en última instancia, contribuir a la mejora del sistema educativo. No obstante, para poder aprovechar al máximo estos modelos, es fundamental comprender que cada uno de ellos presenta ventajas y limitaciones, por lo que su elección depende del tipo de datos disponibles, el tamaño del conjunto de datos y los objetivos específicos de la predicción.

3.5.Aplicación de los factores en los Modelos de Inteligencia Artificial.

El análisis del rendimiento académico mediante técnicas de inteligencia artificial requiere, en primer lugar, la transformación de los factores influyentes en variables estructuradas que puedan ser interpretadas por los algoritmos. Lo factores anteriormente mencionados, son aplicados como entradas numéricas en los modelos predictivos (Etxeberria et al., 2017).

Dicha transformación es fundamental para permitir una aplicación efectiva de las diferentes técnicas de análisis de datos sobre el aprendizaje. Las diferentes variables tratadas anteriormente son recogidas mediante encuestas o cuestionarios y se codifican en las bases de datos (Romero & Ventura, 2014).

Posteriormente, estos atributos son introducidos en los algoritmos de aprendizaje, como por ejemplo los árboles de decisión, con el fin de predecir el rendimiento académico futuro. En el caso de los árboles de decisión, los factores son utilizados para crear diferentes reglas de clasificación que separan a los estudiantes en diferentes categorías de desempeño (Miguéis et al., 2018; Fernandes et al., 2019).

Cabe destacar la importancia que tiene la selección adecuada de las diferentes variables predictivas, para poder obtener así un modelo de predicción acertado, ya que esto es más determinante que la complejidad del modelo de IA utilizado (Tempelaar et al., 2015).

En resumen, la predicción del rendimiento académico a través de la IA parte de una correcta selección de los factores y variables para tener en cuenta para el modelo, permitiendo así crear modelos precisos que puedan facilitar la atención temprana de estudiantes de riesgo y el diseño de estrategias o intervenciones educativas más efectivas (Zawacki-Richter et al., 2019).

3.6.Python como herramienta de análisis de datos.

Python se ha consolidado como uno de los lenguajes de programación más utilizados en el ámbito del análisis de datos y la inteligencia artificial, esto es debido a la gran cantidad

de librerías especializadas que presenta. Una librería es un conjunto de funciones, módulos y herramientas ya desarrolladas que permiten realizar tareas específicas sin tener que programarlas desde cero (Lubanovic, 2015). En el ámbito educativo, puede resultar una herramienta especialmente útil para el tratamiento de grandes volúmenes de datos académicos.

Entre las librerías más destacadas para el análisis de datos se encuentran Pandas (para la manipulación de estructura de datos), NumPy (para el cálculo numérico), Matplotlib y Seaborn (para la visualización) y Scikit-learn, esta última es una de las herramientas más potentes para el desarrollo de modelos de machine learning (Lubanovic, 2015; Chollet, 2018).

Scikit-learn integra una amplia variedad de algoritmos supervisados y no supervisados que facilitan tareas como clasificación, regresión y agrupamiento. Destacan los árboles de decisión y los Random Forest, ya que estos modelos no solo son capaces de realizar predicciones precisas, sino que también ofrecen interpretabilidad, al mostrar que variables influyen más en las decisiones del modelo (Chollet, 2018).

Por lo tanto, trabajar con Python es rápido, cómodo y eficiente, ya que muchas de las tareas más complejas ya están resueltas gracias a las diferentes librerías. Estas se pueden importar fácilmente y permiten aplicar modelos o funciones sin necesidad de programar desde cero, lo que ahorra tiempo y reduce errores.

La IA se ha presentado, por tanto, como una herramienta con muchísimo potencial para este tipo de trabajos, ya que es capaz de identificar diferentes patrones complejos a través de grandes volúmenes de datos. Diferentes modelos como arboles de decisión, redes neuronales, etc. Estos, han demostrado ser muy útiles en la predicción del rendimiento académico, eso sí, siempre teniendo en cuenta que no es solo importante elegir un buen modelo, si no también, contar con datos de calidad y bien seleccionados.

4. Metodología.

4.1.Diseño del estudio.

El trabajo realizado se enmarca en un enfoque aplicado, exploratorio y cuantitativo. Aplicado, porque se utiliza una herramienta de análisis predictivo en un contexto educativo real; exploratorio, porque se trata de una primera aproximación al uso de modelos de predicción sin tratar de construir el modelo predictivo definitivo; y cuantitativo, porque se trabaja con datos numéricos recogidos mediante una encuesta y analizados por técnicas estadísticas y algoritmos de aprendizaje automático.

El estudio se ha llevado a cabo en los cursos de Educación Secundaria Obligatoria (ESO) del Colegio San Agustín de Valladolid, en la asignatura de Física y Química. La muestra con la que se ha trabajado está formada por una clase de 2ºESO, una clase de 3ºESO y dos clases de 4ºESO. El hecho de elegir estos cursos, y no cursos de Bachillerato es debido a diferentes motivos. Por una parte, se buscaba un grupo de alumnos que tuvieran una pequeña base académica, pero que todavía se encuentren en una etapa educativa en la cual sea posible actuar con margen de mejora. Por otra parte, se ha podido ver en el estudio Mercader Rubio et al. (2022) que la ESO representa una fase crítica para el desarrollo de hábitos de estudio, motivación y rendimiento, factores que son clave en el proceso educativo del alumnado, ya que detectar dificultades en esta etapa, especialmente en 2ºESO, puede ayudar a evitar que los estudiantes desarrollen una percepción negativa de la asignatura o que normalicen el bajo rendimiento académico en la asignatura, esto ha sido algo que se ha podido ver durante el periodo de prácticas. Además, trabajar con cursos de ESO permite explorar cómo influyen otros factores en el rendimiento académico en una etapa en la que todavía no existe una presión tan marcada como en Bachillerato, ya que en este último hay que tener en cuenta elementos como la nota media necesaria de cada uno para acceder a estudios superiores y la mayor presión de resultados para quien los necesite.

Otro factor para tener en cuenta es que, en la ESO, hay un mayor número de clases, lo que aumenta el número estudiantes con los que llevar a cabo este tipo de estudios. El contar con un mayor número de alumnos hace que se pueda trabajar con un número de muestras más amplias y, por tanto, puede dar lugar a resultados estadísticamente más fiables.

El número de clases que se han seleccionado es debido a una cuestión de accesibilidad, ya que estas se corresponden a grupos que impartía el tutor de prácticas, lo que ha facilitado el contacto con el alumnado y la recogida de datos. Esto ha permitido contar con una muestra manejable, sin perder de vista el aspecto representativo. Se trata, por tanto, de un diseño sencillo y realista, pero que puede ofrecer una primera aproximación muy útil para comprobar cómo funcionan los modelos de predicción en un entorno educativo real.

4.2.Recogida de datos.

La recogida de datos se ha llevado a cabo a través de una encuesta diseñada específicamente para el alumnado de 2º, 3º y 4º de ESO del Colegio San Agustín de Valladolid. Como ya se ha comentado anteriormente, la elección de estos cursos no solo responde al enfoque del trabajo, sino también a cuestiones de accesibilidad, ya que se trata de los grupos donde impartía clase el tutor de prácticas.

La encuesta se elaboró utilizando Microsoft Forms, ya que es una herramienta integrada en el entorno digital del cetro y accesible para todo el alumnado, además, es muy cómoda para poder posteriormente tratar los datos, ya que permite exportarlos a Microsoft Excel. Dado que, como se comentó en la introducción, todos los estudiantes disponen de dispositivos digitales y acceso a la plataforma Microsoft 365, esta opción es la más cómoda y eficaz para poder compartir el cuestionario y recoger las encuestas. En caso de que algún alumno hubiera podido tener dificultades para realizar la encuesta en ese formato, se tuvo en cuenta la alternativa de facilitarle la encuesta en formato papel, aunque no fue necesario recurrir a ello.

Previo a compartir la encuesta con el alumnado, el cuestionario fue revisado por el tutor, quien dio su aprobación y se encargó de subirlo a los grupos de Microsoft Teams, desde donde los estudiantes podían acceder fácilmente. La encuesta se escribió con un tono cercano y sencillo, ya que va dirigida a estudiantes de entre 13 y 16 años. Las preguntas eran de elección, en ningún momento tenían que redactar nada, solo en la variable *extraescolares* donde tenían que escribir cual era la extraescolar. El objetivo era que resultara fácil de entender y que pudieran responder con comodidad, sin que el lenguaje utilizado pudiera suponer una dificultad.

En cuanto a la protección de datos, se optó por algo sencillo. En lugar de pedir el nombre, cada estudiante indicó el número de clase, lo que garantizaba el anonimato a la hora de

tratar los datos. Posteriormente, el tutor de prácticas facilitó la nota media de cada alumno de la signatura de Física y Química asociada a ese número, lo que permitió relacionar los datos de la encuesta sin comprometer la identidad de los estudiantes en ningún momento.

Lo factores incluidos en la encuesta se han basado en diferentes estudios, el factor "padres" del estudio González-Pienda & Núñez, 2004, este factor trata sobre la insistencia y ayuda de los padres en los estudios del alumno, del estudio Etxeberria et al., 2017, y del artículo se han sacado factores relacionados con el entorno del alumno, y de Castrillón et al., 2020, se han sacado varios factores como la lectura, el uso de videojuegos o actividades extraescolares, que aunque en el estudio se habla de alumnos de universidad, pueden ser factores representativos también para alumnos de ESO. Además, se incluyeron algunas variables propuestas de forma personal. Por ejemplo, no se incluyeron preguntas relacionadas con situaciones de exclusión social o dificultades económicas grandes, ya que el alumnado del Colegio San Agustín de Valladolid, en general, no pertenece a un entorno socioeconómico desfavorecido.

estudio	**videojuego**	**lectura**	**autoestima**	**previsor**	**deberes**	**ansiedad**
Tiempo de estudio en casa	Uso de la tecnología para el ocio	Tiempo de lectura en casa	Predisposición de lograr algo que cueste	Entrega o no al límite las tareas habitualmente	Entrega o no los deberes habitualmente	Siente presión o ansiedad por los estudios
padres	**ambiente estudio**	**atención**	**matemáticas**	**extraescolares**	**profesores**	**amigos**
Ayuda e implicación de los padres en los estudios	Ambiente adecuado en el hogar	Atención en clase	Gusto por las matemáticas	Realización de actividades extraescolares	Relación con el profesorado	Relaciones sociales en el colegio

Tabla 1. *Descripción de factores recogidos en la encuesta.*

4.3.Tratamiento y preparación de datos.

Una vez recogidas todas las respuestas, se exportaron todos los datos desde Microsoft Forms a Excel, con el fin de revisarlos y prepararlos para posteriormente utilizarlos en los modelos de predicción. Al final, esta parte es fundamental para contar con datos ordenados, fiables y con un formato adecuado para su posterior uso.

En primer lugar, se lleva a cabo una "limpieza" de los datos, eliminando respuestas que no estuvieran bien completadas o respuestas en las cuales el número del alumno en la clase coincidiera por confusión. A continuación, se procede a organizar los datos por curso, separando los resultados de 2º, 3º y 4º de ESO, ya que también se quiere observar cómo predice el modelo en los diferentes cursos por separado. Dentro de cada curso, se ordenaron los datos por su rendimiento en la asignatura de Física y Química, así estaría los datos ordenados de forma muy visual.

Posteriormente se transformaron las respuestas a un formato numérico, utilizando una escala del 1 al 5, escala Likert (Likert, 1932), en función de la opción elegida en cada pregunta (por ejemplo, "Nunca"=1, "Siempre"=5). De la misma forma, para el rendimiento, se utilizó una escala del 1-10, que refleja directamente la calificación académica del alumno en la materia.

estudio	videojuego	lectura	autoestima	previsor	deberes	ansiedad
4	3	4	4	2	4	1
4	3	3	5	2	4	2
5	4	3	3	2	4	3
4	2	3	5	3	5	3
2	1	1	5	3	3	2
3	4	1	4	3	4	4
padres	**ambiente estudio**	**atención**	**matemáticas**	**extraescolares**	**profesores**	**amigos**
4	5	4	4	1	4	4
4	5	4	4	1	5	5
4	3	4	3	1	4	5
5	3	5	4	2	4	5
4	4	3	4	1	4	5
3	5	4	1	1	4	5

Tabla 2. *Pequeño ejemplo de escala Likert*.

Con todo esto los datos quedan completamente estructurados. Para las diferentes pruebas que se realizaran posteriormente se seguirá el mismo esquema, se crea un archivo Excel que contengan los datos que se utilizaran para entrenar los modelos, y otro que contenga

los datos de prueba, donde se elimina la columna rendimiento, con los que se evaluará la capacidad predictiva de los modelos.

Una vez los datos están listos y organizados en Excel, se exportan al formato .CSV, es un tipo de formato plano, utilizando la coma como separador, ya que es el estándar que reconoce Python al leer los archivos. En este paso hay que tener cuidado durante la exportación de no tener columnas vacías o espacios sin contenido entre las comas, mucho cuidado también con caracteres no ASCII, ya que todo esto puede producir errores al intentar cargar los datos en el entorno de programación y dificultar la escritura del código. Por ello, se revisan los archivos antes de continuar, asegurando que no haya filas incompletas, columnas sobrantes o encabezados con tildes y a mayores, en el propio código se incluyeron algunas instrucciones específicas para prevenir estos posibles errores.

Para la construcción de los modelos se utilizó Python, tal como se ha ido justificando a lo largo del trabajo, ya que presenta los modelos ya integrados en una de sus bibliotecas. Esta accesibilidad es muy importante para este trabajo, ya que, de esta forma, no es necesario crear el algoritmo desde cero. La biblioteca más destacada usada en este trabajo ha sido Scikit-learn, la cual contenía todos los modelos de predicción que se han utilizado.

Otras bibliotecas utilizadas destacadas Pandas, necesaria para gestionar y manipular los datos en formato tabla, así como NumPy, útil para operaciones numéricas. También se ha utilizado Matplotlib y Seaborn, empleadas para visualizar tanto la distribución de los datos como los resultados de las predicciones. Además, para generar automáticamente un archivo Excel con los datos de predicción procesados, se utilizó la biblioteca openpyxl, que facilita la creación y edición de hojas de cálculo directamente desde Python, esto permitió automatizar parte del flujo de trabajo.

Todo el código se escribió y guardó en el Bloc de notas, en archivos con extensión .py, lo cual permite su ejecución desde la línea de comandos o cualquier entorno compatible con Python, esto hace que sea sencillo y accesible. Esta forma de trabajo ha permitido entender cómo funcionan los diferentes modelos y como tratan los datos.

4.4.Aplicación de los modelos predictivos.

Una vez tratados los datos, y con los archivos .CSV generados y revisados, se procede a escribir el código necesario para aplicar los modelos de predicción. Como ya se ha

comentado en el párrafo anterior, el código se escribe en el Bloc de notas y se guarda como extensión .py para ejecutarlo fácilmente.

Los modelos que se han utilizado en este trabajo han sido el árbol de decisión (Quinlan, 1986), Random Forest (Huynh-Cam et al., 2021) y regresión lineal (Miguéis et al., 2018). La elección se basa en los resultados revisados en el fundamento teórico, donde estos modelos demuestran ofrecer un buen equilibrio entre rendimiento y facilidad de interpretación, especialmente Random Forest, ya que es más difícil de que de problemas como el sobreajuste debido a que está formado por muchos árboles de decisión. Por esto mismo, no se han utilizado modelos como las redes neuronales, ya que su complejidad dificulta el interpretar como han llegado a un resultado determinado.

En primer lugar, y antes de realizar pruebas más amplias, se partió de algo más sencillo, entrenar los modelos con una muestra muy reducida de datos, con el único objetivo de comprobar cómo se comportaban. Es una fase inicial que sirve para familiarizarse con el funcionamiento interno de los algoritmos y visualizar así elementos como el árbol de decisión generado o la estructura de los modelos Random Forest. En esta fase solo se busca entender la respuesta de los modelos a la información y cómo organizan las decisiones.

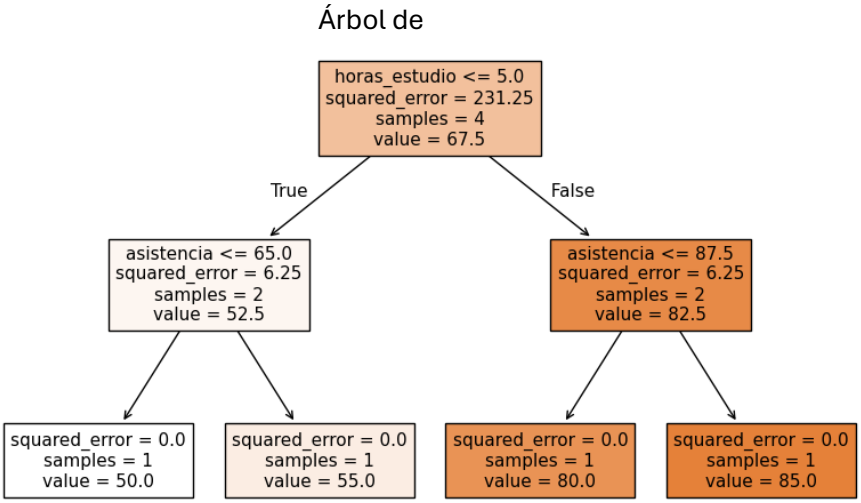


Figura 5. *Ejemplo visual del árbol de decisión.*

Una vez se realizó esta primera toma de contacto, se diseñó el procedimiento para las pruebas. En todas ellas se decidió utilizar entre un 80% de las muestras para el entrenamiento del modelo, ya que se realizaron varias pruebas con porcentajes menores y los resultados obtenidos fueron peores, mientras que el resto se utilizó para comprobar si los modelos eran capaces de predecir correctamente los resultados de los datos que no habían formado parte del entrenamiento. Esta proporción se mantuvo como referencia a lo largo de todo el trabajo.

Se realizan tres pruebas diferentes, en las que se cambian las muestras empleadas tanto para entrenar como para predecir. De este modo, se puede observar si el comportamiento del modelo se mantenía estable o si variaba mucho en función de los datos utilizados. Antes de ejecutar cada una de estas pruebas, se ejecuta un código que permite entrenar a los modelos y mostrar qué variables tienen más peso en sus decisiones, es decir, a la hora de predecir el rendimiento. Esto permite comparar el comportamiento de los diferentes modelos y comprobar si coinciden o no en que factores consideran importantes.

Después, se ejecuta un segundo código en el que se integran ambas partes: entrenamiento y predicción final. Dicho código permite entrenar y aplicar los tres modelos, y generar automáticamente un archivo Excel con los rendimientos académicos predichos por el modelo para los datos seleccionados.

Adicionalmente, se diseña una prueba final donde se utilizan solo un grupo reducido de variables, aquellas que los modelos de predicción habían identificado como más relevantes. La intención de esta prueba es comprobar si la precisión de los modelos mejora centrándose en las variables relevantes y eliminando las variables que sean más irrelevantes para la predicción del rendimiento.

4.5.Evaluación de los modelos predictivos.

Para valorar cómo se han comportado los modelos de predicción utilizados, se aplican dos formas de evaluación, una centrada en la precisión de las diferentes predicciones y otra más orientada a medir la fiabilidad del modelo frente a diferentes conjuntos de datos.

En cuanto a la precisión, se ha calculado como el porcentaje de aciertos sobre el total de muestras, considerando como acierto aquellas predicciones que coinciden con la nota real del alumno o difieren en un punto. Es decir, si uno de los datos tiene un 8 en rendimiento, se considera correcto si el modelo predice 7, 8 o 9. Esto resulta realista en un contexto educativo, donde no se busca exactitud milimétrica, sino comprobar si el modelo se

aproxima de manera razonable al rendimiento real. En el artículo Castrillón et al., 2020, por ejemplo, valoran el rendimiento de la A-E. Dicho cálculo se ha aplicado a cada uno de los modelos, obteniendo así una precisión media general para cada uno, y también hemos podido comprobar la precisión media para cada curso.

Por otro lado, también se ha de valorar la fiabilidad del modelo, para ello, se comparan los resultados obtenidos en las tres pruebas que se han planteado en el trabajo. La idea es que, si un modelo da resultados similares, aunque se cambie las muestras de entrenamiento y predicción, se podría decir que el modelo es fiable y que no tiene una gran dependencia de los datos que se usen. En cambio, si las diferencias entre pruebas son notorias, podría indicar que el modelo depende mucho de los datos que se utilicen como entrenamiento y predicción y, por tanto, sería un modelo menos fiable.

Estas dos formas de evaluar permiten hacerse una idea bastante completa del funcionamiento de los modelos en este contexto educativo concreto. Además, esta evaluación permite orientar “el rumbo” del trabajo en el futuro.

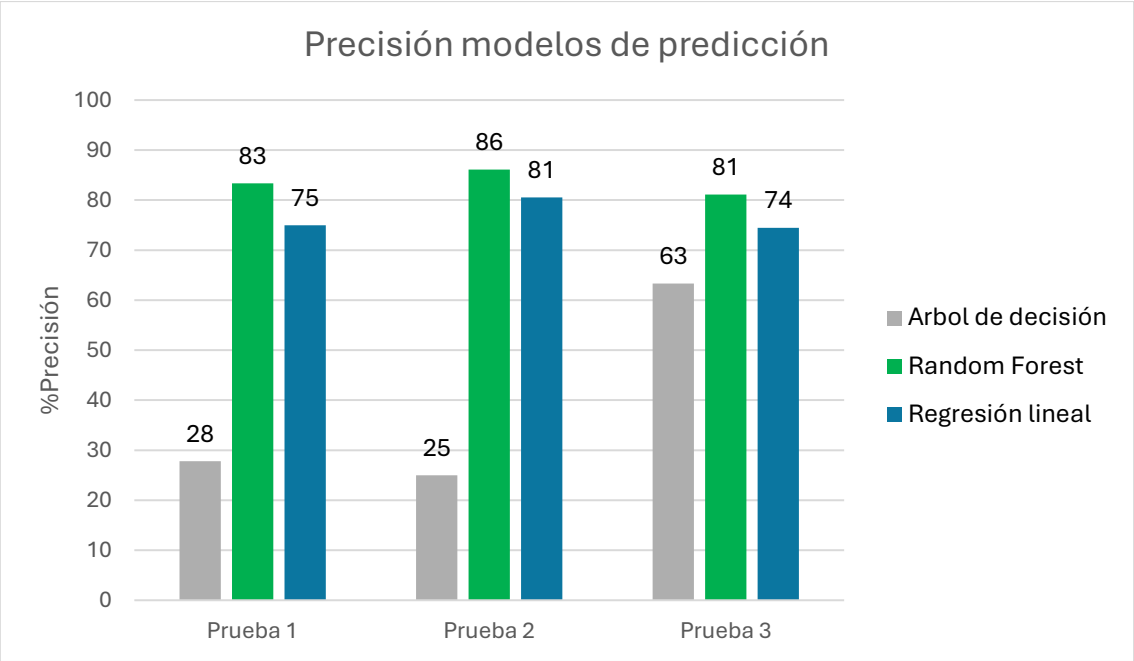
5. Análisis de resultados y discusiones.

A lo largo de este apartado se presentan los resultados obtenidos con los distintos modelos de predicción utilizados en el trabajo. Además de mostrar los porcentajes de acierto y comparar el rendimiento entre modelos, se tratará de entender los resultados desde un doble punto de vista, por un lado, un enfoque más técnico o numérico, y por otro, desde una visión más educativa o pedagógica. La idea es no quedarse solo con si el modelo acierta con mayor o menor precisión, sino también discutir sobre por qué ocurre esto, qué factores parecen influir en mayor medida y cómo se puede interpretar todo ello dentro del contexto escolar.

5.1. Análisis general de los modelos.

Tras el tratamiento de datos, se contó finalmente con un total de 76 muestras de estudiantes de 2º, 3º y 4º de ESO. Algunas respuestas tuvieron que descartarse, ya que varios alumnos confundieron su número de clase o introdujeron alguna respuesta con valores que no tenían sentido dentro del cuestionario. Para las pruebas de predicción se han utilizado 60 muestras para el entrenamiento de los modelos y 16 para comprobar su capacidad predictiva, cambiando en cada prueba qué alumnos se usaban para una función u otra. Esto permitió observar cómo se comportaban los modelos ante distintas combinaciones de datos, manteniendo siempre una estructura de prueba coherente.

A continuación, se muestran los porcentajes de precisión obtenidos en las tres pruebas realizadas, en las cuales se aplicaron los modelos de árbol de decisión, Random Forest y regresión lineal. Para cada prueba se calculó la precisión general del modelo, entendida como el porcentaje de aciertos obtenidos al comparar la predicción con la nota real, considerando como correctas aquellas predicciones ±1 punto con respecto al valor real. Esta forma de evaluar, ya explicada en la metodología, permite comprobar si los modelos se aproximan de forma razonable al rendimiento académico del alumnado.



Gráfica 1. *Resultados modelos de predicción.*

Random Forest ha sido el modelo de predicción que mejores resultados ha dado de forma más constante. En las tres pruebas realizadas, ha mantenido porcentajes de precisión superiores al 80%, mayores que los otros dos modelos.

Una de las razones por las que Random Forest ha funcionado mejor que el resto, es porque no se basa en una única predicción, sino que combina hasta 100 árboles de decisión para obtener una respuesta final más fiable. Gracias a esto, se reduce el riesgo de errores grotescos que se dan cuando depende de un único árbol, lo que explica su estabilidad incluso cuando se cambian los datos. Todo esto hace que Random Forest no solo sea el modelo más adecuado para la realización de este estudio, sino también una herramienta de gran potencial para futuras investigaciones que busquen predecir el rendimiento académico del alumnado a partir de múltiples factores.

En cuanto al modelo de regresión lineal, su comportamiento ha sido menos positivo, manteniéndose en porcentajes entre 74%-80% de precisión en las tres pruebas. Aunque en menor medida que el modelo Random Forest, si ha demostrado cierta estabilidad a lo largo de los diferentes conjuntos de datos utilizados.

Se trata de un modelo muy simple, esto tiene sus ventajas y sus desventajas. Su ventaja es que es muy fácil de interpretar, la regresión lineal permite ver claramente cómo afecta cada variable al resultado final, tanto positiva como negativamente. Esto es especialmente

útil en el contexto educativo, ya que no solo nos interesa saber si acierta, sino que interesa también que factores están influyendo más y en qué dirección. Pero, esta simpleza también provoca que pueda tener sus limitaciones a la hora de captar relaciones complejas o no lineales entre variables. Por ello, la regresión lineal es una buena alternativa para estudios que busquen una primera aproximación en la predicción del rendimiento académico y para detectar patrones generales en los datos.

Por último, en el caso del árbol de decisión, los resultados han sido claramente más irregulares. En dos de las tres pruebas su precisión estuvo por debajo del 30%, y solo en la tercera alcanzó un resultado del 63%, inferior a cualquier resultado de los otros dos modelos. Esta falta de consistencia sugiere que el modelo es muy sensible al conjunto de datos que se le proporciona, y que es necesario un número mayor de muestras para poder utilizarlo con mayor fiabilidad.

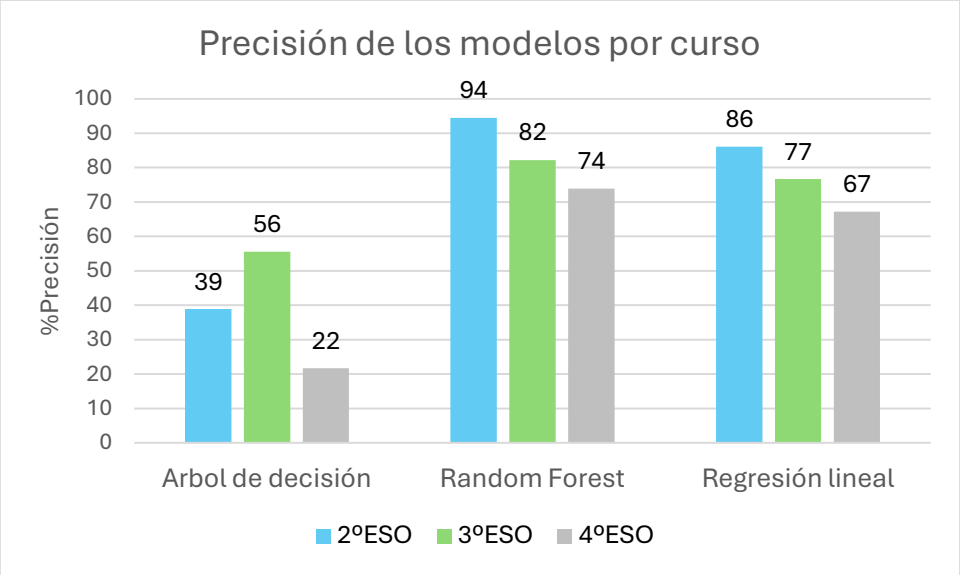
Aunque su principal ventaja es la interpretación visual, ya que permite ver claramente cómo se van tomando las decisiones, en este estudio no ha logrado ofrecer predicciones fiables de manera continua. El tamaño reducido de la muestra ha afectado a este modelo, ya que tiende a sobre ajustarse si no tiene suficientes datos. Por tanto, no ha demostrado ser una opción válida a la hora de realizar una primera aproximación, aunque puede resultar interesante, por lo visto en el marco teórico, trabajar con el modelo en investigaciones con un mayor número de muestras.

En cualquier caso, es importante tener en cuenta que las precisiones obtenidas están condicionadas por el tamaño limitado de la muestra. En este estudio se ha trabajado con 76 alumnos, mientras que en los estudios revisados en la bibliografía emplean muestras mucho más amplias, por encima de las 300. Por tanto, aunque los resultados no reflejen niveles de precisión extremadamente altos, sí permite obtener una primera aproximación realista y funcional sobre cómo se comportan estos modelos en el contexto educativo.

5.2.Análisis por cursos de los modelos.

Además del análisis general de cada modelo, también resulta interesante observar las predicciones de estos modelos según el curso del alumnado. Dado que se ha trabajado con estudiantes de 2º, 3º y 4º ESO, se han calculado los porcentajes de acierto de cada modelo para cada uno de estos niveles, en las tres pruebas realizadas. Esto permite analizar cómo se comportan los modelos en función del curso, es decir de cómo han respondido los alumnos de cada curso, lo que puede estar influido por su madurez y por

su conocimiento de la física, y conocer en qué nivel educativo ofrecen, en este estudio, mejores resultados.



Gráfica 2. *Resultados modelos de predicción por curso.*

Antes de comenzar con el análisis por curso, conviene señalar que, debido a su comportamiento irregular, el modelo de árbol de decisión no será tenido en cuenta en este apartado.

Los resultados de 2ºESO para los modelos de Random Forest y regresión lineal son especialmente llamativos, en especial en Random Forest. En ambos casos, este curso ha sido el que ha obtenido las mejores medias de acierto en las tres pruebas, superando a 3º y 4º tanto en precisión como en fiabilidad. Esto quiere decir que los modelos han captado mejor los patrones de rendimiento en este nivel, dando lugar a predicciones más ajustadas.

Como hemos dicho, el Random Forest ha mostrado un porcentaje de precisión especialmente alto en este curso, con un porcentaje medio de 94%, teniendo un 100% en dos de las pruebas. El modelo de regresión lineal también ha obtenido resultados positivos, aunque en menor medida, teniendo un 100% de precisión en una de las pruebas. Este elevado acierto en la predicción puede deberse a que, para el alumnado de 2º, existe una correlación más clara entre los factores elegidos para la encuesta y el rendimiento en Física y Química en este nivel académico.

Los resultados de 3º y 4ºESO muestran una predicción más irregular por parte de los modelos, especialmente en 4º y si lo comparamos con 2º. En la etapa de 3ºESO el modelo Random Forest ha dado predicciones positivas, superiores al 80% e inferiores y menos

fiables para 4ºESO. En el caso del modelo de regresión lineal, ha dado resultados peores, por debajo del 80% en 3ºESO, aunque con cierta estabilidad en las pruebas 1 y 3, y resultados bajos y poco fiables en cuanto a la predicción en 4ºESO.

Esta diferencia entre cursos puede deberse a que, a medida que el alumno avanza de curso, entran en juego un mayor número de factores que afectan de forma significativa al rendimiento, lo que dificulta la precisión de los modelos, ya que los patrones son más complicados de encontrar, y, por tanto, será necesario un mayor número de muestras, esto puede venir dado debido a la maduración personal y escolar del alumnado. Por lo que, sería necesario ajustar las preguntas o factores descritos en la encuesta, ya que, si pueden estar bien ajustados para 2ºESO, pero pueden necesitar un pequeño ajuste dependiendo del curso.

5.3.Variables con mayor influencia.

Una de las ventajas más interesantes de los modelos utilizados en este trabajo es que permiten identificar qué variables han tenido mayor peso a la hora de realizar las predicciones. Se ha podido comprobar cómo en las tres pruebas hay determinados factores que se repiten como relevantes, mientras que otros apenas han tenido influencia, independientemente del modelo.

El modelo del árbol de decisión ha mostrado resultados algo irregulares. Aunque destaca la importancia de la lectura, de la ayuda de los padres, de las extraescolares o del ambiente de estudio, en general se aprecia poca consistencia entre las tres pruebas, en algunos casos, incluso hay variables con peso cero, como la atención en clase, esto refuerza la idea de que este modelo ha sido menos robusto y más sensible a los cambios de muestra.

En el caso del modelo Random Forest, la variable más influyente ha sido de forma constante el gusto por las Matemáticas, seguida de cerca por la lectura, el ambiente de estudio, y en menor medida la autoestima y la ayuda de los padres. Estos resultados coinciden en gran medida con lo planteado en la bibliografía comentada en el fundamento teórico.

En cuanto a la regresión lineal, el modelo permite no solo conocer el peso de la variable, sino también el signo de cada variable, es decir, si tiene un efecto positivo o negativo sobre le rendimiento predicho. Las variables con mayor peso positivo han sido, en las tres pruebas, el gusto por las Matemáticas, la lectura, el ambiente de estudio y la relación con el profesorado, este último factor puede ser debido a que cuando una relación profesor-

alumno es mala, es fácil que el alumno vea al profesor como un obstáculo y no como un apoyo, lo que refuerza una actitud negativa hacia la asignatura y puede acabar afectando directamente al rendimiento. Las variables con peso negativo ha sido de forma reiterada los amigos que tienen en el colegio, aunque en principio tener buenos amigos podría asociarse a un ambiente positivo y de apoyo, es posible que en ciertos casos una alta sociabilización en el aula provoque distracciones, falta de concentración y una menor implicación académica. Puede que aquellos alumnos que den más importancia a lo social dentro del aula estén menos centrados en los estudios, lo que podría reflejarse en un rendimiento menor.

Uno de los factores que más peso ha tenido en todos los modelos ha sido el gusto por las Matemáticas, lo cual tiene mucho sentido teniendo en cuenta que la asignatura que estamos evaluando en Física y Química. Ambas comparten una base numérica y lógica, por lo que es normal que quien se sienta cómodo en Matemáticas, también tenga una mayor seguridad en Física y Química. En este sentido, lo que marca la diferencia es la actitud positiva hacia la asignatura y la predisposición a enfrentarse a retos que exijan razonamiento lógico.

Por otro lado, la lectura por placer también ha aparecido como una variable muy influyente. Leer con frecuencia mejora la comprensión lectora, entrena la atención, la capacidad de concentración y el manejo del lenguaje (Cremin & Moss, 2018), habilidades clave a la hora de estudiar o entender diferentes conceptos. Por tanto, la lectura regular puede actuar como entrenamiento, el cual no es directo en la asignatura de Física y Química, pero sí refuerza habilidades para rendir en ella.

Por último, llama especialmente la atención que la variable horas de estudio, que en principio parecía fundamental, no haya tenido un peso mayor en los modelos de predicción. Este resultado puede verse de dos formas diferentes, por un lado, puede que los alumnos no respondieran a esta pregunta de manera objetiva, y, por otro lado, puede indicar que lo que más influye en el rendimiento no es solo cuántas horas se estudia, sino cómo se estudia y en qué condiciones, ya que la variable ambiente de estudio sí ha tenido un gran peso en los modelos de predicción. Este tipo de resultados dan lugar a la idea de que no siempre los factores más evidentes son los más determinantes, y que conviene mirar con profundidad detrás de cada variable.

5.4.Análisis de casos concretos.

Con todo lo analizado hasta este punto, resulta interesante dar un paso más y examinar algunos casos concretos donde los modelos han funcionado especialmente bien, así como otros en los que han fallado de manera grotesca. Este tipo de análisis permite entender qué tipo de perfiles o respuestas pueden favorecer o dificultar la precisión en la predicción.

Alumno	Árbol de Decisión	Random Forest	Regresión Lineal	rendimiento
Alumno 1	8	8	8	9
Alumno 2	5	7	6	8
Alumno 3	9	7	7	7
Alumno 4	5	5	5	6
Alumno 5	9	8	8	9
Alumno 6	8	8	8	8
Alumno 7	4	6	6	7
Alumno 8	9	7	6	6
Alumno 9	7	6	6	5
Alumno 10	9	7	6	4
Alumno 11	9	7	7	7
Alumno 12	4	7	9	6
Alumno 13	9	6	6	5
Alumno 14	7	8	10	10
Alumno 15	6	7	7	8
Alumno 16	9	7	8	4

Tabla 3. *Resultados Prueba de predicción 1.*

Alumno	Árbol de Decisión	Random Forest	Regresión Lineal	rendimiento
Alumno 1	5	8	9	9
Alumno 2	5	7	7	8
Alumno 3	5	7	7	8
Alumno 4	9	8	8	7
Alumno 5	8	6	5	6
Alumno 6	6	6	6	5
Alumno 7	4	7	7	9
Alumno 8	8	8	8	8
Alumno 9	5	6	5	7
Alumno 10	8	6	6	6
Alumno 11	6	6	4	5
Alumno 12	8	5	5	4
Alumno 13	7	7	7	8
Alumno 14	7	6	5	5
Alumno 15	6	6	8	4
Alumno 16	7	9	10	10

Tabla 4. *Resultados Prueba de predicción 2.*

Alumno	Árbol de Decisión	Random Forest	Regresión Lineal	rendimiento
Alumno 1	8	8	8	9
Alumno 2	8	7	7	8
Alumno 3	4	6	6	7
Alumno 4	5	6	7	7
Alumno 5	6	6	7	6
Alumno 6	7	7	7	4
Alumno 7	8	7	8	9
Alumno 8	9	7	8	8
Alumno 9	6	5	7	6
Alumno 10	5	7	8	6
Alumno 11	6	6	6	5
Alumno 12	5	7	9	7
Alumno 13	9	6	6	6
Alumno 14	6	6	6	5
Alumno 15	5	7	7	10
Alumno 16	8	8	9	8

Tabla 5. *Resultados Prueba de predicción 3.*

Puede apreciarse de un vistazo que los errores más llamativos se dan cuando el rendimiento real del alumno es muy bajo o alto. Es decir, los modelos tienden a fallar más en los extremos. Esto puede ser debido a la escasez de ejemplos de ese tipo dentro del conjunto de datos de entrenamiento. Al haber menos alumnos con notas especialmente altas o bajas, los modelos tienen menos referencias con las que aprender a predecir correctamente estos perfiles, por tanto, sus estimaciones se alejarán en mayor medida de la realidad.

Uno de estos casos es el Alumno 16 en la Prueba 1, cuya nota fue mal predicha por parte de los modelos de predicción. Lo que llama la atención, es que la predicción realizada por los modelos de Random Forest y regresión lineal es muy similar a la del Alumno 11 de esa misma prueba, la cual si es acertada.

Al comparar sus variables clave, se entiende por qué los modelos los han tratado de forma tan parecida, ambos alumnos tienen un "ambiente de estudio" de 5, un rendimiento bajo en "lectura" (1 y 2 respectivamente) y un 3 en "Matemáticas". Estas tres variables han sido, como se ha visto en los apartados anteriores, las que más peso han tenido en los modelos, especialmente Random Forest. Por tanto, es lógico que, al presentar perfiles tan similares en los factores más influyentes, el modelo haya asignado un rendimiento intermedio en ambos casos.

Este tipo de situación da lugar a una de las limitaciones de los modelos, esto es cuando los factores clave coinciden, el modelo puede generalizar, aunque haya otros factores de menor influencia que puedan marcar la diferencia en el rendimiento real.

Otro ejemplo que se ha analizado es el Alumno 16 de la Prueba 2 y el Alumno 15 de la Prueba 3, los cuales tienen un rendimiento real de 10. En el primer caso, la predicción fue acertada por los modelos Random Forest y regresión lineal, pero fue incorrectamente predicho para el segundo caso.

Al revisar sus respuestas, se aprecia que hay diferencias significativas en las variables más influyentes. Por ejemplo, aunque ambos tienen un 5 en "ambiente de estudio", y una diferencia mínima en Matemáticas (4 en el fallido y 5 en el acertado), en "lectura" la diferencia es mucho más marcada, el Alumno 15 tiene un 4, mientras que el Alumno 16 tiene un 1. Teniendo en cuenta el peso que ha tenido la variable lectura en los modelos, especialmente en Random Forest, es comprensible que esta diferencia haya influido notablemente en el resultado.

Además, aunque la variable "padres" no ha sido una de las más determinantes, sí ha tenido algo de peso en la regresión lineal y su impacto ha sido negativo (mayor puntuación en esa variable se asocia a menor rendimiento, según los coeficientes del modelo). En este caso, el Alumno 16 tiene un 5 en "padres", mientras que el Alumno 15 tiene un 1, lo que podría haber contribuido a inclinar la predicción hacia una nota más baja, al menos desde el punto de vista del modelo lineal.

5.5.Discusión general.

Teniendo en cuenta los resultados obtenidos, el modelo Random Forest es el más adecuado para este tipo de investigaciones en el que se aplica la inteligencia artificial a la predicción del rendimiento académico. No solo demuestra ser el más preciso y fiable en la predicción, sino que su funcionamiento es perfecto tanto para primeras aproximaciones exploratorias como para estudios con conjuntos de datos más amplios.

Este modelo se ve bien acompañado por el modelo de regresión lineal, que, aunque está algo más limitado cuando los patrones son complejos, resulta muy útil en fases iniciales para detectar las variables de mayor peso y si influyen de manera positiva o negativa. Su sencillez en la interpretación es ideal para comprender qué hay detrás de los resultados. En cuanto al árbol de decisión, el hecho de que haya sido tan irregular en este trabajo significa que no funciona bien con pocas muestras, y que podría funcionar mucho mejor con un mayor conjunto de datos.

Finalmente, este estudio ha dejado algunas apreciaciones curiosas, como que factores como la lectura por placer o el ambiente de estudio hayan tenido una mayor relevancia que las propias horas de estudio, también, que trabajar con alumnado exige claridad, sencillez y un enfoque bien adaptado.

6. Conclusiones y trabajo futuro.

El objetivo principal de este trabajo puede considerarse cumplido, ya que, se han analizado tres modelos predictivos diferentes utilizando datos reales, es decir aplicado a una situación real, del Colegio San Agustín de Valladolid, lo que permite tener una primera visión de cómo puede funcionar este tipo de trabajo en contextos educativos reales. Otra opción, sería poder realizar el trabajo en un centro educativo con mayor heterogeneidad de alumnos o con otro tipo de alumnado, lo que daría pie a pensar nuevos factores y cambios en los factores clave para el rendimiento. Este estudio no deja de ser un trabajo realizado para un tipo de alumnado referido al Colegio San Agustín de Valladolid.

Los resultados obtenidos han permitido extraer algunas conclusiones claras. Por un lado, el modelo de Random Forest ha destacado por su precisión y fiabilidad con respecto al resto, siendo el que mejor ha funcionado de forma general y también por cursos. La regresión lineal, aunque algo más limitada con los patrones complejos, ha resultado útil por su fácil interpretación. El árbol de decisión, en cambio, ha mostrado un comportamiento mucho más irregular, lo que puede sugerir que necesita un mayor conjunto de datos para tener mayor fiabilidad. Este trabajo también ha servido para ver que ciertos factores como la lectura por placer, el ambiente de estudio o el gusto por las matemáticas tienen una influencia en la predicción del rendimiento, y como un el número de horas de estudio no resulta tan clave como se pensaba en estos cursos.

Desde el punto de vista de la aplicación de este trabajo, puede tener una utilidad real en centros educativos si se continúa desarrollando. Para ello habría que revisar y adaptar la encuesta, sobre todo en los cursos de 3º y 4ºESO, donde puede ser la causa de haber obtenido peores resultados. Sería interesante complementar la información de la encuesta con factores objetivos recogidos por el profesorado en las primeras semanas de clase, o tomar datos dados por las familias, es decir trabajar en colaboración con estos, aunque seguramente esto último resulte bastante más complicado. Otro aspecto clave en la mejora del trabajo es ampliar el número de muestras, esto hará que los modelos de predicción tengan más datos y sean capaces de predecir con una mayor precisión. Además, ya en un futuro más lejano, se deberá ir reduciendo el número de datos de entrenamiento e ir aumentando los de prueba, ya que los modelos estarán cada vez más entrenados.

Ahora bien, es importante dejar claro que este estudio no deja de ser una primera aproximación. Aún queda mucho camino por recorrer para poder construir un modelo de predicción sólido y aplicable de forma regular en centros educativos. El tamaño de muestra ha sido limitado y la encuesta, aunque funcional, debe irse mejorando a partir de los resultados obtenidos para futuros estudios, especialmente en cursos como 3º y 4ºESO, donde pueden aparecer otro tipo de factores diferentes a los cursos inferiores. Además, este trabajo puede resultar muy útil para docentes o investigadores que quieran iniciarse en el uso de Python con fines educativos, ya que muestra paso a paso como diseñar una encuesta, tratar los datos, construir los modelos y analizar sus resultados. También puede servir como base para futuros estudios más ambiciosos, con muestras más amplias, nuevas variables, datos más objetivos e incluso una colaboración directa con las familias o equipos docentes. En definitiva, se trata de un estudio sencillo, pero con una proyección muy interesante para seguir explorando cómo la inteligencia artificial puede convertirse en una herramienta útil en el ámbito educativo, tanto para el análisis como para la toma de decisiones pedagógicas más informadas.

7. Bibliografía

Almuqati, M. T., Sidi, F., Mohd Rum, S. N., Zolkepli, M., & Ishak, I. (2024). Challenges in Supervised and Unsupervised Learning: A Comprehensive Overview. *International Journal of Advanced Science, Engineering and Information Technology, 14*(4), 1449–1455. https://doi.org/10.18517/ijaseit.14.4.20191

Arribas, J. M. (2012). El rendimiento académico en función del sistema de evaluación empleado. *RELIEVE. Revista Electrónica de Investigación y Evaluación Educativa*, 18(1), art. 3. https://www.redalyc.org/pdf/916/91624440003.pdf

Beneyto Sánchez, S. (2013). *Entorno familiar y rendimiento académico*. Universidad de La Rioja. https://dialnet.unirioja.es/descarga/libro/657731.pdf

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Castrillón, Omar D., Sarache, William, & Ruiz-Herrera, Santiago. (2020). Prediction of academic performance using artificial intelligence techniques. *Formación universitaria*, *13*(1), 93-102. https://dx.doi.org/10.4067/S0718-50062020000100093

Chen, X., Xie, H., Hwang, G.-J., & Wang, F. L. (2020). A multi-perspective study on artificial intelligence in education: Grants, conferences, journals, software tools, institutions, and researchers. Computers & Education: Artificial Intelligence, 1, 100005. https://doi.org/10.1016/j.caeai.2020.100005

Chollet, F. (2018). *Deep learning with Python*. Manning Publications.

Cremin, T., & Moss, G. (2018). *Reading for pleasure: Scrutinising the evidence base.* Language and Education. https://doi.org/10.1080/09500782.2024.2324948

Etxeberria, F., Garay, J., & Azpillaga, V. (2017). Impacto de los factores institucionales en el rendimiento académico. Revista de Psicología Educativa, 23(2), 135–150. https://doi.org/10.1016/j.pse.2017.05.003

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Van Erven, G. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. Journal of Business Research, 94, 335–343. https://doi.org/10.1016/j.jbusres.2018.02.012

Gil-Vera, V. D., & Quintero-López, C. (2021). Predicción del rendimiento académico estudiantil con redes neuronales artificiales. *Información Tecnológica, 32*(6), 221–230. https://doi.org/10.4067/S0718-07642021000600221

Hakimi, L., Eynon, R., & Murphy, V. A. (2021). The Ethics of Using Digital Trace Data in Education: A Thematic Review of the Research Landscape. *Review of Educational Research*, *91*(5), 671-717. https://doi.org/10.3102/00346543211020116

Huynh-Cam, T.-T., Chen, L.-S., & Le, H. (2021). Using Decision Trees and Random Forest Algorithms to Predict and Determine Factors Contributing to First-Year University Students' Learning Performance. *Algorithms, 14*(11), 318. https://doi.org/10.3390/a14110318

Ley Orgánica 3/2020, de 29 de diciembre, por la que se modifica la Ley Orgánica 2/2006, de 3 de mayo, de Educación (LOMLOE). Boletín Oficial del Estado, núm. 340, de 30 de diciembre de 2020. https://www.boe.es/buscar/act.php?id=BOE-A-2020-17264

Likert, R. (1932). A technique for the measurement of attitudes. Archives of Psychology, 22 140, 55.

Liu, D., Dede, C., Huang, R., & Richards, J. (Eds.). (2017). *Virtual, augmented, and mixed realities in education*. Springer. https://doi.org/10.1007/978-981-10-5490-7

Liu, V., Latif, E., & Zhai, X. (2025). Advancing Education through Tutoring Systems: A Systematic Literature Review. *arXiv preprint arXiv:2503.09748*. https://arxiv.org/abs/2503.09748

Lubanovic, B. (2015). *Introducing Python: Modern computing in simple packages*. O'Reilly Media.

González-Pienda, J. A., & Núñez, J. C. (2004). La implicación de los padres y su incidencia en el rendimiento de los hijos. *Revista de Psicología y Educación*, 1(1), 115–134. https://www.revistadepsicologiayeducacion.es/pdf/9.pdf

Maulana, A., Idroes, G. M., Kemala, P., Maulydia, N. B., Sasmita, N. R., Tallei, T. E., ... & Rusyana, A. (2023). Leveraging Artificial Intelligence to Predict Student

Performance: A Comparative Machine Learning Approach. *Journal of Educational Management and Learning*, *1*(2), 64-70.

Mercader Rubio, I.; Oropesa Ruiz, N.F.; Ángel, N.G.; Fernández Martínez, M.M. Motivational Profile, Future Expectations, and Attitudes toward Study of Secondary School Students in Spain: Results of the PISA Report 2018. Int. J. Environ. Res. Public Health 2022, 19, 3864. https://doi.org/10.3390/ijerph19073864

Miguéis, V. L., Freitas, A., García, P. J. V., & Silva, A. (2018). Early segmentation of students according to their academic performance: A predictive modeling approach. Decision Support Systems, 115, 36–51. https://doi.org/10.1016/j.dss.2018.09.001

Mishara, P. (2024). The Ethical Implications of AI in Education: Privacy, Bias, and Accountability. *J. Inform. Educ. Res*, *4*, 3550. https://doi.org/10.52783/jier.v4i2.1827

Nicholson, C. (2019). A beginner's guide to neural networks and deep learning. *Retrieved January*, *30*, 2020.

Oppong, S. O. (2023). Predicting students' performance using machine learning algorithms: A review. *Asian Journal of Research in Computer Science, 16*(3), 128–148. https://doi.org/10.9734/AJRCOS/2023/v16i3351

Pagliara, S. M., Bonavolontà, G., Pia, M., Falchi, S., Zurru, A. L., Fenu, G., & Mura, A. (2024). The integration of artificial intelligence in inclusive education: A scoping review. *Information, 15*(12), 774. https://doi.org/10.3390/info15120774

Pierce, G. L., & Cleary, P. F. (2024). *The persistent educational digital divide and its impact on societal inequality*. PLOS ONE, 19(4), e0286795. https://doi.org/10.1371/journal.pone.0286795

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81–106. https://doi.org/10.1023/A:1022643204877

Rastrollo, J., López, V., García, A., & Pérez, R. (2020). Analyzing and predicting students' performance. Applied Sciences, 10(3), 1042. https://doi.org/10.3390/app10031042

Romero, C., & Ventura, S. (2014). A survey on pre-processing educational data. En A. Peña-Ayala (Ed.), *Educational data mining: Applications and trends* (pp. 29–64). Springer. https://doi.org/10.1007/978-3-319-02738-8_2

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), e1355. https://doi.org/10.1002/widm.1355

Sajja, R., Sermet, Y., Cwiertny, D., & Demir, I. (2023). Integrating AI and Learning Analytics for Data-Driven Pedagogical Decisions and Personalized Interventions in Education. *arXiv preprint arXiv:2312.09548*. https://arxiv.org/abs/2312.09548

Sposato, M. (2025). Artificial intelligence in educational leadership: A comprehensive taxonomy and future directions. *International Journal of Educational Technology in Higher Education, 22*(20). https://doi.org/10.1186/s41239-025-00517-1

Tempelaar, D. T., Rienties, B., & Giesbers, B. (2015). In search for the most informative data for feedback generation: Learning analytics in a data-rich context. Computers in Human Behavior, 47, 157–167. https://doi.org/10.1016/j.chb.2014.05.038

UNESCO. (2023). *Guidance for generative AI in education and research*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000386693

Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments, 9*(1), 11. https://doi.org/10.1186/s40561-022-00192-z

Yao, H., Lian, D., Cao, Y., Wu, Y., & Zhou, T. (2019). Predicting academic performance for college students: A campus behavior perspective. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *10*(3), 1-21.

York, T. T., Gibson, C., & Rankin, S. (2015). Defining and measuring academic success. Practical Assessment, Research, and Evaluation, 20(5). https://doi.org/10.7275/hz5x-tx03

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – Where are the educators? International Journal of Educational Technology in Higher Education, 16(1), 39. https://doi.org/10.1186/s41239-019-0171-0

8. Anexos.

Anexo I. Enlace encuesta.

Encuesta realizada al alumnado para la toma de datos:

Categoría	Factor específico (para preguntar)	Respuesta (valores entre 1 y 5)
Hábitos personales	¿Dedicas tiempo a estudiar en casa? (1 antes del exámen, 5 todos los días)	1-5
	¿Usas la tablet para ocio y juegos? (1 muy poco, 5 todos los días)	1-5
	¿Te gusta leer en casa? (1 muy poco, 5 todos los días)	1-5
	¿Crees que, aunque algo te cueste, puedes aprenderlo si lo trabajas? (1 no lo creo, 5 lo creo firmemente)	1-5
	¿Dejas los deberes para el último día? (1 nunca, 5 siempre)	1-5
	¿Sientes ansiedad, estrés o presión respecto a los estudios? (1 nada, 5 mucha)	1-5
Familiares	¿Te preguntan tus padres/tutores por los estudios? ¿Te ayudan con tareas si lo necesitas? (1 nunca, 5 siempre)	1-5
	¿Tienes espacio tranquilo en casa para estudiar? (1 no tengo espacio de estudio, 5 tengo un espacio de estudio perfecto)	1-5
Escolares / institucionales	¿Crees que tus profesores te ayudan cuando lo necesitas? (1 nunca, 5 siempre)	1-5
	¿Atienedes en clase? (1 nunca, 5 siempre)	1-5
	¿Haces los deberes? (1 nunca, 5 siempre)	1-5
	¿Tienes buena relación con los profesores? (1 muy mala, 5 muy buena)	1-5
Sociales / comunitarios	¿Participas en actividades extracurriculares (deporte, música, clubes, etc.)? ¿En que actividad?	
	¿Tienes buenos amigos en clase? (1 no tengo amigos, 5 tengo muchos amigos)	1-5
	¿Te gustan las matemáticas? (1 no me gustan nada, 5 es de mis asignaturas favoritas)	1-5

ID	¿Cuál es tu clase? ¿Cuál es tu número de clase?

Anexo II. Enlace tabla datos recogidos.

https://drive.google.com/file/d/1wxTgADVCCXH0qhT7F_fxiFkf0MNtkdFL/view?usp=sharing

Anexo III. Códigos para comprender el funcionamiento.

Códigos de árbol de decisión, Random Forest y regresión lineal para comprender como funcionan:

Árbol de decisión:

https://drive.google.com/file/d/1f_dKM6i4qpmOv1mI8ofws0U1pZ7g57Ql/view?usp=sharing

Random Forest:

https://drive.google.com/file/d/1EPyFUomLSG8D9mTGqv6c2y5oMyRlddLN/view?usp=sharing

Regresión lineal:

https://drive.google.com/file/d/1SJcwQ6Ob2hkvGS_XfG88fKuWHPEQKKmt/view?usp=sharing

Anexo IV. Códigos para realizar las pruebas.

Entrenar modelos:

https://drive.google.com/file/d/1PfvA0zz0PjMZnv0WamGlsqgKEqoNv3nV/view?usp=sharing

Predicción con modelos:

https://drive.google.com/file/d/17uqFaVmAdQw3vAZOEhFbjsSzIOLVxiv3/view?usp=sharing