TRABAJO DE GRADO
Opción Seminario-Diplomado.

Transformando el éxito educativo con una aproximación predictiva de machine learning

Corporación Universitaria Remington.
Ingeniería de Sistemas.
Machine Learning.

Nicolas Alfonso Charry Diaz – Nicolas Zuluaga Zuluaga.
Juan Pablo Vélez Uribe
Seminario de Machine Learning
2023

Tabla de Contenidos

Resumen

Este trabajo de grado se adentra en la convergencia entre la educación y la inteligencia artificial, específicamente explorando el desarrollo y la implementación de algoritmos de machine learning para predecir el rendimiento académico de los estudiantes. El enfoque se centra en un proyecto especifico que utiliza el historial académico como base para determinar si un estudiante aprobara o reprobara un examen.

A lo largo de este espacio y basados en los cursos de crehana, hemos examinado los fundamentos del machine learning y como estos pueden ser aplicados de una manera responsable y efectiva en el contexto predictivo educativo. Se tiene como fin, la importancia de comprender el comportamiento de los estudiantes a lo largo de su trayectoria académica para lograr las predicciones mas precisas. Además, se exploran las implicaciones éticas y practicas en utilizar los modelos vistos y aprendidos.

Dando como continuación, con este proyecto buscamos abordar la necesidad de anticipar el rendimiento estudiantil y ofrecer una solución practica para mejorar la toma de decisiones en el ámbito educativo. Los resultados preliminares muestran prometedores niveles de precisión en las predicciones del éxito académico, abriendo así las puertas para futuras aplicaciones y refinamientos en este campo emergente con todo el avance de la analítica educativa y el potencial de la transformación de la inteligencia artificial en la mejora de proceso, junto con las posibilidades que ofrece la aplicabilidad del machine learning.

Palabras clave

Machine Learning, Inteligencia Artificial (IA), Historial Académico, Innovación, Modelo Predictivo.

Marco conceptual y contextual

Machine Learning (ML) es una rama de la Inteligencia Artificial (IA) que se centra en el desarrollo de algoritmos y modelos que permiten a las máquinas aprender patrones y tomar decisiones sin intervención humana explícita. En el contexto de la Inteligencia Artificial, Machine Learning implica la capacidad de las máquinas para aprender de los datos, adaptarse a nuevas situaciones y mejorar su rendimiento a lo largo del tiempo sin una programación específica para cada tarea.

El proceso de Machine Learning implica la exposición de un sistema informático a datos relevantes, permitiéndole aprender patrones y estructuras subyacentes. Estos modelos pueden realizar tareas específicas, como clasificación, regresión, clustering, reconocimiento de patrones, entre otras, en función de la naturaleza del problema.

La esencia del Machine Learning radica en la capacidad de generalización, lo que significa que el modelo puede aplicar lo que ha aprendido a nuevas situaciones o datos no vistos anteriormente. Los enfoques comunes en Machine Learning incluyen aprendizaje supervisado, no supervisado y por refuerzo.

En resumen, Machine Learning en el contexto de la Inteligencia Artificial representa la capacidad de las máquinas para aprender de los datos y mejorar su rendimiento en tareas específicas, lo que tiene aplicaciones en una amplia variedad de campos, desde reconocimiento de voz y visión por computadora hasta toma de decisiones empresariales y recomendaciones personalizadas.

Fundamentos aplicados de Machine Learning

El Machine Learning (ML) aborda situaciones de manera autónoma mediante un análisis de datos, y su desempeño mejora proporcionalmente a la cantidad de datos que se emplean (Juárez, G., 2017). Para llevar a cabo dicho análisis, se aplican algoritmos que se adaptan a las necesidades específicas (Juárez, G., 2017). ML ejecuta un algoritmo a través de los datos de entrada, generando así información adicional para resolver el problema (Bishop, C., 2007). El propósito de generar más datos se fundamenta en diversas técnicas, entre las cuales se incluyen la regresión lineal y polinómica, árboles de decisión, redes neuronales, redes bayesianas y cadenas de Markov. Estas técnicas capacitan a ML para reconocer patrones, extraer conocimientos, descubrir información y realizar predicciones.

En términos de aprendizaje, cada individuo adopta métodos específicos basados en los sentidos, la experiencia y las habilidades cognitivas, como tomar notas, la resolución de ejercicios y la lectura. En el ámbito informático, se busca que las computadoras alcancen autonomía y aprendan automáticamente habilidades definidas por algoritmos para el aprendizaje y la gestión de datos (Jordan, M. I., & Mitchell, T. M., 2015).

Es importante destacar que el ML no implica auto programación, sino más bien un autoaprendizaje continuo a partir de datos y experiencia para identificar patrones y abordar

nuevas tareas. Este proceso de aprendizaje se configura mediante la combinación de diversas técnicas, datos, conceptualización del análisis de datos y algoritmos, con el fin de generar nuevos patrones o modelos predictivos.

Para su aplicación, existen diversas herramientas tales como las bibliotecas Numpy, Pandas y Scikit learn. NumPy es una biblioteca en Python diseñada para realizar operaciones numéricas eficientes. Su función principal es proporcionar estructuras de datos como arreglos multidimensionales (ndarrays), junto con funciones matemáticas para operar en estos arreglos. NumPy es esencial en ML para la manipulación eficiente de datos numéricos, realizando operaciones matriciales y algebraicas fundamentales. A su vez, está Pandas que se utiliza para la manipulación y análisis de datos. Proporciona estructuras de datos flexibles, como los DataFrames, que permiten organizar y analizar datos de manera tabular. Pandas es especialmente útil para cargar datos desde diversas fuentes, limpiar datos, realizar agregaciones y manipulaciones eficientes en conjuntos de datos. Por último, está Scikit-learn, biblioteca de aprendizaje automático en Python que proporciona herramientas simples y eficientes para la minería y el análisis de datos. Incluye implementaciones de una amplia variedad de algoritmos de aprendizaje supervisado y no supervisado, así como herramientas para evaluar el rendimiento de los modelos, realizar selección de características y ajustar parámetros. Scikit-learn facilita la implementación y experimentación con diferentes algoritmos de Machine Learning.

Por otro lado, está Jupyter Notebooks, entornos interactivos basados en web que permiten la creación y el intercambio de documentos que contienen código, visualizaciones y texto narrativo. Son ampliamente utilizados en el ámbito de la ciencia de datos y Machine Learning para desarrollar y documentar código de manera iterativa. Los Notebooks Jupyter permiten ejecutar código por bloques, lo que facilita la experimentación y la visualización de resultados en tiempo real, lo que es valioso para la exploración de datos y la creación de modelos en ML.

Otro fundamento a tener en cuenta bajo un contexto de ML es la estadística descriptiva, la cual se refiere a la aplicación de técnicas estadísticas para describir y resumir características fundamentales de un conjunto de datos (Rendon, 2016). Este enfoque estadístico es crucial en varias etapas del proceso de Machine Learning, desde la exploración inicial de datos hasta la evaluación de modelos. Aquí hay algunas formas en que la estadística descriptiva se utiliza en el ámbito de Machine Learning:

Exploración de Datos:
Identificación de valores atípicos: La estadística descriptiva puede ayudar a identificar valores atípicos o anómalos en los datos que podrían afectar la calidad del modelo.
Resumen de variables: Proporciona resúmenes estadísticos como la media, la mediana y la desviación estándar para comprender la distribución y la variabilidad de las variables.
Preprocesamiento de Datos: Manejo de valores perdidos: Permite calcular estadísticas descriptivas para determinar cómo manejar los valores perdidos en el conjunto de datos.

Normalización y escalamiento: La comprensión de las estadísticas descriptivas de las variables ayuda en la toma de decisiones sobre la normalización o escalamiento de características.

Selección de Características:
Correlación entre variables: La estadística descriptiva, como la matriz de correlación, ayuda a identificar relaciones entre variables y puede ser útil en la selección de características.

Evaluación de Modelos:
Evaluación de métricas: Se utilizan estadísticas descriptivas, como la media y la desviación estándar, para evaluar el rendimiento del modelo en términos de métricas como el error medio cuadrático o la precisión.

Visualización de Datos:
Histogramas y gráficos de dispersión: Utilizados para visualizar la distribución y las relaciones entre variables, lo que puede influir en la elección y la evaluación del modelo.

Dentro de la estadística descriptiva se utilizan dos tipos fundamentales de variables: variables categóricas y variables objetivas (o variables de respuesta). Una variable categórica es una variable que puede tomar un conjunto finito y discreto de valores. Estos valores representan categorías o grupos, y no tienen un orden inherente entre ellos. Las variables categóricas pueden ser nominales u ordinales. Mientras que por su lado la variable objetiva, también llamada variable de respuesta o variable dependiente, es aquella que se intenta predecir en un modelo de Machine Learning. En un problema de clasificación, la variable objetivo suele representar las clases que el modelo debe predecir; en un problema de regresión, la variable objetivo es una cantidad continua que se desea estimar (Rendon, 2016).

Continuando con los fundamentos, están las métricas de evaluación en Machine Learning las cuales según Gomez (2023), son herramientas que permiten medir el rendimiento y la eficacia de un modelo predictivo. Estas métricas son esenciales para comprender cómo se comporta un modelo en términos de precisión, generalización y capacidad para hacer predicciones correctas sobre datos no vistos. Unas de las más comunes son:

Precisión (Accuracy): la cual mide la proporción de predicciones correctas en relación con el total de predicciones. Es útil cuando las clases están equilibradas en el conjunto de datos.

Recall (Sensibilidad o Tasa Positiva Verdadera): mide la capacidad del modelo para identificar todos los casos positivos. Es importante en situaciones donde la omisión de casos positivos es crítica.

Precisión (Precisión): mide la precisión de las predicciones positivas. Es relevante cuando el costo de los falsos positivos es alto.

F1-Score: es una métrica que combina precisión y recall en un solo número. Es útil cuando hay un desequilibrio entre las clases o cuando tanto falsos positivos como falsos negativos son críticos.

Área bajo la Curva ROC (AUC-ROC): mide la capacidad del modelo para discriminar entre clases. Cuanto mayor sea el AUC-ROC, mejor será el modelo en distinguir entre clases.

Matriz de Confusión: proporciona una visión detallada del rendimiento del modelo, mostrando el número de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos.

Mean Squared Error (MSE): es utilizado en problemas de regresión, mide el promedio de los errores al cuadrado entre las predicciones y los valores reales.

Log Loss: especificado para problemas de clasificación, mide la eficiencia del modelo en términos de la probabilidad asignada a las clases correctas.

Cohen's Kappa: evalúa la concordancia entre las predicciones del modelo y las observaciones reales, ajustando la precisión por la probabilidad de concordancia aleatoria.

De acuerdo con el tipo de problema se selecciona la métrica, es común utilizar varias métricas en conjunto para obtener una comprensión más completa del rendimiento del modelo en diferentes aspectos (Gomez, 2023).

Finalmente, todo proyecto de ML tiene un ciclo de aprendizaje automático que es continuo, y la elección del algoritmo de aprendizaje automático adecuado es solo uno de los pasos. Según Hurwitz y Kirsch (2018), las etapas en el ciclo de aprendizaje automático son las siguientes:

Identificar los datos: Reconocer las fuentes de datos pertinentes es el primer paso en el ciclo. Además, al desarrollar tu algoritmo de aprendizaje automático, considera la posibilidad de ampliar los datos objetivo para mejorar el sistema.

Preparar los datos: Asegurarse de que los datos estén limpios, seguros y gestionados. Si creas una aplicación de aprendizaje automático basada en datos inexactos, la aplicación fallará.

Seleccionar el algoritmo de aprendizaje automático: Puedes tener varios algoritmos de aprendizaje automático aplicables a tus datos y al desafío comercial.

Entrenar: Necesitas entrenar el algoritmo para crear el modelo. Dependiendo del tipo de datos y del algoritmo, el proceso de entrenamiento puede ser supervisado, no supervisado o de aprendizaje por refuerzo.

Evaluar: Evaluar tus modelos para encontrar el algoritmo que tenga mejor rendimiento.

Implementar: Los algoritmos de aprendizaje automático crean modelos que pueden implementarse tanto en aplicaciones en la nube como en locales.

Predecir: Después de la implementación, comienza a hacer predicciones basadas en nuevos datos entrantes.

Evaluar predicciones: Evaluar la validez de tus predicciones. La información obtenida al analizar la validez de las predicciones se retroalimenta en el ciclo de aprendizaje automático para ayudar a mejorar la precisión.

Machine Learning: Análisis contrafactual

El Análisis Contrafáctico, una rama reciente del Aprendizaje Automático Explicable se enfoca en desarrollar técnicas que generen explicaciones comprensibles para humanos sobre los resultados de modelos predictivos. Dada la rápida expansión de los algoritmos de Aprendizaje Automático en sistemas de toma de decisiones automáticas con impacto directo en vidas humanas, la investigación en este campo se vuelve esencial.

El término contrafáctico proviene de los campos de la filosofía, psicología y otras ciencias sociales, donde se refiere a una afirmación condicional cuyo antecedente es falso. Es una evaluación de las supuestas consecuencias de una situación o acción que nunca ocurrió, de ahí el término "contrafáctico": "contrario a los hechos". Un ejemplo en este sentido podría ser: Si no hubiera habido la Segunda Guerra Mundial, entonces la ONU no existiría hoy en día (Agüera, 2022).

El análisis contrafactual debe presentar las siguientes propiedades:

Validez: Decimos que una instancia contrafáctica es válida si su clase predicha es diferente a la del original. Si el número de clases posibles es mayor que dos, a menudo designaremos una clase deseada. La validez es fundamental para la definición de la instancia contrafáctica.
Proximidad a la instancia original: Una instancia contrafáctica debe mantenerse cerca de la instancia original para ser significativa para el usuario correspondiente.
Viabilidad: Llamamos a una instancia contrafáctica viable si representa una combinación realista de características. Esto es esencial para que el análisis tenga algún significado. La viabilidad se garantiza prohibiendo todas las combinaciones no realistas y prestando atención a las relaciones causales entre las características. También se puede mejorar teniendo en cuenta la cercanía al manifiesto de datos.

Acción: Una característica de los datos es ejecutable si representa un atributo que puede cambiarse razonablemente. Debe ser mutable y estar bajo el control de la persona a la que se aplica. Un ejemplo sería "consumo diario de proteínas" en oposición a "etnia". Una instancia contrafáctica también se denomina ejecutable si difiere solo en características ejecutables. La acción es necesaria para que las instancias contrafácticas sean útiles para los usuarios.
Dispersión: Una instancia contrafáctica se denomina dispersión si solo difiere del original en algunas características. Se ha encontrado 13 que las personas comprenden más fácilmente explicaciones más cortas, y estas proporcionan una guía más clara hacia el cambio de etiqueta. La priorización de características también puede ofrecer una visión más clara sobre el funcionamiento y los sistemas del clasificador (Agüera, 2022).

En este contexto es importante tener en cuenta que pueden presentarse sesgos, como por el ejemplo, sesgo por variables omitidas, el cual ocurre cuando no se tienen en cuenta ciertas variables relevantes al realizar el análisis contrafactual. Si variables importantes no se incluyen en el modelo o análisis, la estimación de los resultados puede estar sesgada o incompleta. Puede conducir a conclusiones incorrectas o incompletas sobre las relaciones causales, ya que no se están considerando todos los factores que podrían influir en los resultados.
También se tiene el sesgo por selección, el cual surge cuando la selección de observaciones o datos no se realiza de manera aleatoria o representativa, sino que está sesgada hacia ciertos grupos o condiciones. En el análisis contrafactual, puede ocurrir al elegir ciertos casos de estudio de manera no aleatoria, lo que puede distorsionar la validez de las conclusiones contrafácticas, ya que los resultados pueden no ser generalizables o aplicables a toda la población si la selección no es representativa.

Ambos sesgos son importantes considerar en el análisis contrafactual para asegurar que las conclusiones sean sólidas y aplicables a un rango más amplio de situaciones (Agüera, 2022). Por otro lado, se tiene el análisis causal que hace referencia al estudio de las relaciones de causa y efecto entre variables. La teoría causal busca comprender los mecanismos subyacentes que generan ciertos resultados observados. Las relaciones causales implican que un cambio en una variable tiene un impacto directo en otra.

Los métodos estadísticos causales buscan identificar y cuantificar efectos causales en datos observacionales. Esto implica abordar el problema de confusión, donde factores adicionales pueden influir en las variables de interés.

Algunas técnicas comunes incluyen el uso de experimentos controlados para establecer causalidad y métodos como el matching, la regresión discontinua y la instrumentalización en datos observacionales para aproximar relaciones causales. A su vez, pueden aplicarse métodos como la experimentación aleatoria, donde los participantes o unidades de estudio se asignan aleatoriamente a diferentes grupos de tratamiento. Enfoque que ayuda a controlar los sesgos potenciales y a garantizar que los grupos sean comparables, lo que fortalece la validez interna de un experimento. También está la asignación estratificada, un

enfoque en el que se divide la población en grupos homogéneos llamados estratos, y luego se asigna aleatoriamente a cada grupo de tratamiento dentro de cada estrato. Esto se hace para asegurarse de que cada estrato esté representado en cada condición experimental, lo que ayuda a controlar las variables de confusión y a mejorar la validez externa del experimento.

Ahora bien, para abordar problemas de causalidad, mejorar la estimación de parámetros y mitigar problemas de sesgo, confusión, mejorar la validez interna y externa, y permitir inferencias más robustas sobre relaciones causales en entornos no experimentales, se tienen métodos cuasiexperimentales y métodos de regularización como double lasso, propensity score, y double debiased machine learning. Los métodos cuasiexperimentales se utilizan cuando no es posible realizar un experimento aleatorio puro, pero aún se busca establecer relaciones causales. Técnicas como la regresión discontinua y la diferencia en diferencias son ejemplos de métodos cuasiexperimentales que buscan replicar la aleatorización en situaciones no experimentales.

Ahora bien, en cuanto a métodos de regularización Propensity Score busca equilibrar las características observadas entre grupos tratados y no tratados en estudios observacionales. Calcula la probabilidad de recibir tratamiento condicional a las covariables observadas. Puede utilizarse en combinación con métodos cuasiexperimentales para mejorar la equidad entre grupos; Double Lasso se utiliza en modelos de regresión para seleccionar variables relevantes y reducir el riesgo de sobreajuste. Puede mejorar la precisión de las estimaciones y controlar la dimensionalidad de las covariables. Finalmente, Double Debiased Machine Learning (DML) busca corregir sesgos en las estimaciones causales al desvincular la estimación del efecto causal de la estimación de las características de confusión. Puede ser valioso en situaciones donde los modelos de regresión tradicionales podrían estar sesgados.

Evaluación de modelos de Machine Learning

Validación Cruzada (Cross-Validation):
Esta Técnica utilizada para evaluar el rendimiento de un modelo al dividir el conjunto de datos en subconjuntos de entrenamiento y prueba de manera iterativa. Que permite una evaluación más robusta del modelo al utilizar múltiples divisiones, mitigando el riesgo de sobreajuste y proporcionando una medida más precisa de la capacidad de generalización.

Matriz de Confusión:
La podemos definir como una tabla que muestra la relación entre las predicciones de un modelo y las clases reales, destacando los verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. Proporcionando información detallada sobre el rendimiento del modelo, especialmente útil en problemas de clasificación para evaluar la precisión y el impacto de errores específicos.

Curva ROC (Receiver Operating Characteristic):
Este tiene como punto una representación gráfica de la sensibilidad frente a la tasa de falsos positivos, proporcionando una medida de la capacidad de discriminación del modelo.

Generando así una mayor utilidad para evaluar y comparar modelos en términos de rendimiento y tomar decisiones informadas sobre el umbral de clasificación.

AUC-ROC (Área bajo la Curva ROC):
Esta métrica cuantifica la integral del área bajo la curva ROC, brindando una medida resumida del rendimiento del modelo que facilita la comparación de modelos en un solo número, donde un valor más alto indica un mejor rendimiento de clasificación.

Métricas de Regresión: MAE, MSE, RMSE:
Esta métrica evalúa la precisión de modelos de regresión midiendo la diferencia entre las predicciones y los valores reales, generando información sobre el rendimiento del modelo en la predicción de valores numéricos, permitiendo una evaluación cuantitativa de la precisión.

Curvas de Aprendizaje (Learning Curves):
Esta es una representación gráfica que muestra cómo el rendimiento del modelo varía con el tamaño del conjunto de datos de entrenamiento y ayuda a identificar problemas de sobreajuste o subajuste, permitiendo ajustes en la complejidad del modelo.

Optimización de Hiperparámetros: Grid Search y Random Search:
Este método es útil para encontrar la combinación óptima de hiperparámetros que maximice el rendimiento del modelo y a la vez mejore la capacidad predictiva del modelo al ajustar los parámetros que no se aprenden durante el entrenamiento, como la tasa de aprendizaje o la profundidad del árbol.

Análisis de Residuos:
Este análisis consta de una evaluación de la diferencia entre los valores predichos y los valores reales, permitiendo identificar patrones sistemáticos de error. Para si, generar modelos útiles de regresión para validar supuestos y mejorar la interpretación del rendimiento del modelo.

Importancia de Características:
Este análisis consta de una evaluación de la contribución de cada característica en la toma de decisiones del modelo el cual facilita la identificación de características clave que afectan significativamente el rendimiento del modelo.

Machine Learning: Aprendizaje supervisado

Machine Learning Supervisado: Según lo indicado por (Russo, 2016), a través del aprendizaje supervisado en machine learning, se emplean algoritmos que se basan en un conjunto de datos de entrenamiento previamente etiquetados, es decir, con clasificaciones conocidas. Estos algoritmos procesan este conjunto para realizar predicciones, ajustándolas y corrigiéndolas en caso de errores. El proceso de entrenamiento persiste hasta que el modelo logra alcanzar el nivel de precisión deseado.

Machine Learning No Supervisado: A través del aprendizaje automático no supervisado, se aborda un escenario en el cual el conjunto de datos carece de etiquetas o resultados predefinidos. En este contexto, los algoritmos se enfrentan al desafío de deducir las estructuras inherentes a los datos de entrada sin disponer de información conocida previa. Este enfoque implica la aplicación de procesos matemáticos que buscan sistemáticamente reducir la redundancia en los datos o estructurarlos según similitudes intrínsecas. Esta perspectiva, Según lo indicado por (Russo, 2016), la capacidad de los algoritmos de aprendizaje automático no supervisado para descubrir patrones y relaciones inherentes en conjuntos de datos no etiquetados, contribuyendo así a la comprensión y organización de la información sin la necesidad de guía o supervisión externa.

Machine Learning Semi Supervisado: En el enfoque de aprendizaje automático semi supervisado, se fusionan conjuntos de datos que contienen tanto instancias etiquetadas como no etiquetadas con el propósito de construir una función específica o clasificador. Estos modelos se enfrentan al desafío de aprender las estructuras subyacentes en los datos, lo que implica la organización y comprensión de patrones presentes en las instancias tanto etiquetadas como no etiquetadas. La tarea principal de estos modelos es, por lo tanto, la generación de funciones que permitan organizar y clasificar los datos, así como realizar predicciones en instancias no etiquetadas.
Según lo indicado por (Russo, 2016), el aprendizaje semi supervisado se erige como un enfoque que aprovecha la información disponible en instancias etiquetadas mientras explora y extrae conocimiento de datos no etiquetados, contribuyendo así a la construcción de modelos más robustos y precisos. Este método se posiciona como una estrategia efectiva para abordar problemas en los cuales se dispone de conjuntos de datos mixtos, facilitando la generalización del modelo y mejorando su capacidad predictiva.

La variedad de datos utilizados para el modelo de aprendizaje automático o Machine Learning puede clasificarse en cuatro categorías principales, cada una con sus propias características distintivas:

Datos Numéricos: También conocidos como datos cuantitativos, constituyen información donde los puntos de datos son valores numéricos precisos. Esta categoría se subdivide en datos continuos, que pueden asumir infinitos valores en un rango específico, y datos discretos, que poseen un número finito de valores posibles dentro de un conjunto determinado. (Palop Alcaide,2022)

Datos Categóricos: Representan características que, aunque pueden expresarse numéricamente, carecen de un significado matemático específico. En situaciones de súper-clasificación, estos datos corresponden a las etiquetas de clase. Dentro de esta categoría se encuentran los datos ordinales, que combinan elementos numéricos y categóricos al estar clasificados en categorías con un orden particular. (Palop Alcaide,2022)

Datos de Series Temporales: Consisten en una secuencia de números recopilados en intervalos regulares a lo largo de un período de tiempo. Estos datos poseen un valor

temporal que facilita su ordenación cronológica, brindando información valiosa sobre la evolución a lo largo del tiempo. (Palop Alcaide, 2022)

Datos de Texto: Se componen principalmente de palabras. Para que estos datos sean útiles en el aprendizaje automático, es esencial procesarlos mediante técnicas como las bolsas de palabras, que les otorgan una estructura significativa y permiten su análisis en contextos algorítmicos. (Palop Alcaide, 2022)

Estos tipos de datos son esenciales para la construcción de modelos efectivos en machine learning, cada uno con sus propias consideraciones y desafíos particulares.
En relación con las métricas de desempeño y otros conceptos como "Data set", "train", "test", y "evaluation", estos son términos fundamentales en el proceso de entrenamiento y evaluación de modelos de aprendizaje automático los abordaremos a continuación para seguir entendiendo sobre la contribución y la optimización y validación de los resultados obtenidos.

Data Set (Conjunto de Datos): Un conjunto de datos es simplemente un conjunto de instancias o ejemplos que se utilizan para entrenar, validar o probar un modelo de machine learning. Este conjunto puede incluir datos numéricos, categóricos, de series temporales o de texto, dependiendo de la naturaleza del problema y del tipo de modelo que se esté construyendo.

Entrenamiento (Train): En el contexto de machine learning, el conjunto de entrenamiento (train set) es una porción del conjunto de datos que se utiliza para enseñar al modelo. Durante el entrenamiento, el modelo ajusta sus parámetros para aprender patrones y relaciones presentes en los datos. Este proceso es crucial para que el modelo pueda realizar predicciones precisas en situaciones futuras.

Prueba (Test): El conjunto de prueba (test set) es otra porción del conjunto de datos que no se utiliza durante el entrenamiento. Una vez que el modelo ha sido entrenado, se evalúa en el conjunto de prueba para medir su capacidad de generalización. Es decir, se comprueba cómo se comporta el modelo al enfrentarse a datos que no ha visto previamente. Esto ayuda a estimar el rendimiento del modelo en situaciones del mundo real.

Evaluación (Evaluation): La evaluación implica medir el rendimiento del modelo en función de métricas específicas. Estas métricas pueden incluir precisión, recall, F1-score, entre otras, dependiendo del tipo de problema y de las metas del modelo. La evaluación proporciona información sobre la capacidad del modelo para realizar predicciones precisas y generalizar a datos nuevos y no vistos.

Dando como continuidad al aprendizaje supervisado, abordaremos las opciones para estimar un modelo de regresión, donde según (Peláez, I. M, 2016). existen dos donde se destacan por su facilidad de aplicación e interpretación, el modelo de regresión lineal y el modelo de regresión logística. Teniendo en cuenta el tipo de variable que deseemos estimar

(variable dependiente o respuesta) aplicaremos un modelo de regresión u otro. Simplificando, cuando la variable dependiente es una variable continua, el modelo de regresión más frecuentemente utilizado es la regresión lineal, mientras que cuando la variable de interés es dicotómica (es decir, toma dos valores como sí/no, hombre/mujer) se utiliza la regresión logística. Otro tipo de modelos de regresión utilizados, aunque no tan frecuentemente, son la regresión no lineal o la regresión ordinal que estiman una serie de modelos matemáticos que pueden ajustarse mejor que un modelo lineal.

El enfoque de los mínimos cuadrados constituye un método fundamental en el análisis de regresión, que implica la determinación de la recta de mejor ajuste a través de la minimización de la suma de las distancias al cuadrado entre los puntos reales y los predichos por la recta estimada. Este proceso se basa en la introducción de variables en el modelo de regresión lineal, y la elección del mejor modelo se realiza al comparar las estadísticas F parciales obtenidas en cada uno de los modelos construidos.
La selección de variables, mediante técnicas previas, implica el cálculo repetido de este coeficiente al eliminar o introducir variables, ya que este procedimiento es esencialmente la estimación de nuevos modelos de regresión. Este proceso es gestionado automáticamente por paquetes estadísticos, a menos que se opte por la técnica de forzar la inclusión de todas las variables, donde la estimación manual de todos los modelos posibles se convierte en una responsabilidad del analista.

Otra estrategia para validar un modelo implica la evaluación de los residuos de la regresión, que representan las discrepancias entre los valores estimados por el modelo y los valores observados. La distribución de estos residuos, bajo la suposición de que el modelo es adecuado, debería seguir una ley normal con media cero y varianza constante. Este supuesto puede ser visualmente confirmado mediante la representación gráfica de una nube de puntos que ilustre la distribución de los residuos, permitiendo así diagnosticar posibles problemas como la falta de linealidad o la heterocedasticidad.
Se destaca la importancia de considerar cuidadosamente los valores extremos al ajustar el modelo de regresión, ya que, aunque representen observaciones legítimas, pueden distorsionar significativamente los resultados, especialmente al utilizar el método de los mínimos cuadrados. Por lo tanto, se recomienda estimar dos modelos distintos: uno que incluya estos valores y otro que los excluya, evaluando finalmente cuál de ellos se adapta mejor a los objetivos de la investigación.
En el contexto de la regresión logística, la identificación del mejor modelo se lleva a cabo mediante la comparación de modelos utilizando el cociente de verosimilitud, que cuantifica la probabilidad relativa de un modelo en comparación con otro. La diferencia entre los cocientes de verosimilitud se evalúa mediante la distribución de la Ji-cuadrado, y si no hay evidencia convincente de que un modelo sea superior al otro, se prefiere el modelo más simple, como propone (Peláez, 2016).

Árbol de Decisión: Según (Suguiura,F.O.R.,2022). Un árbol de decisión representa de manera gráfica y analítica todos los posibles eventos que pueden derivarse de una decisión tomada en un momento específico. Su función principal es asistir en la toma de decisiones

al proporcionar una visión probabilística de las opciones disponibles. Estos árboles ofrecen una representación visual que permite examinar los resultados y comprender cómo se desarrolla el modelo en distintas situaciones. Su utilidad radica en la capacidad de visualizar patrones, buscar subgrupos específicos y revelar relaciones que podrían no ser evidentes mediante métodos estadísticos más convencionales.
Desde una perspectiva técnica, los árboles de decisión son una poderosa técnica estadística que se utiliza para diversas tareas, como la segmentación, estratificación, predicción, reducción y filtrado de datos, identificación de interacciones, fusión de categorías y discretización de variables continuas. La función de árboles de decisión en herramientas como SPSS facilita la creación de árboles de clasificación y decisión. Estos árboles se emplean para identificar grupos, explorar relaciones entre ellos y prever eventos futuros. Variados tipos de árboles, como CHAID, CHAID exhaustivo, CRT y QUEST, están disponibles para adaptarse de manera óptima a la naturaleza específica de los datos que se están analizando.

Algoritmo de Random Forest: De acuerdo con (Espinosa-Zúñiga, J. J.,2020). Los algoritmos de Bosques Aleatorios, como el Random Forest, son técnicas de aprendizaje automático que se destacan por su capacidad para mejorar la precisión y la robustez de los modelos predictivos. Este enfoque, ampliamente utilizado en diversas disciplinas, se basa en la construcción de múltiples árboles de decisión independientes durante el proceso de entrenamiento. Cada árbol se desarrolla utilizando diferentes subconjuntos de datos y características, introduciendo aleatoriedad en el proceso.
La Fuerza de los Bosques Aleatorios radica en su capacidad para mitigar el sobreajuste (overfitting) inherente a los árboles de decisión individuales, al promediar las predicciones de múltiples árboles. Este enfoque no solo mejora la precisión de las predicciones, sino que también proporciona una estimación más confiable de la importancia de cada característica en el modelo.
Además de su eficacia en problemas de clasificación y regresión, los Bosques Aleatorios destacan por su resistencia al ruido y su capacidad para manejar conjuntos de datos grandes y complejos. Estas características los convierten en una opción popular en la comunidad científica y empresarial para una variedad de aplicaciones, desde el análisis de datos hasta la toma de decisiones en tiempo real.
En resumen, los algoritmos de Bosques Aleatorios representan una herramienta valiosa en el campo del aprendizaje automático, proporcionando modelos predictivos más robustos y precisos gracias a su enfoque innovador y su capacidad para abordar desafíos comunes en el análisis de datos.

Redes Neuronales: En el marco del aprendizaje supervisado, demuestran su versatilidad al manejar tanto datos continuos como categóricos. Este enfoque innovador en el procesamiento de información permite a las redes neuronales adaptarse eficazmente a una variedad de tipos de datos, abordando así una amplia gama de problemas en diferentes dominios.
Esta fuente confiable destaca la capacidad de las Redes Neuronales para aprender patrones complejos y relaciones no lineales en datos continuos. Asimismo, señala que estas redes

son capaces de manejar de manera efectiva variables categóricas, a través de técnicas como la codificación one-hot, que transforma las variables categóricas en representaciones numéricas comprensibles para el modelo.

Introducción a la Inteligencia Artificial

Aprendizaje Automático (Machine Learning) en IA:
Este aprendizaje es un subcampo de la IA que se centra en el desarrollo de algoritmos que permiten a las máquinas aprender patrones y tomar decisiones sin ser programadas explícitamente. Generando un gran valor en sistemas de recomendación, reconocimiento de voz, procesamiento de imágenes y más.

Redes Neuronales Artificiales en IA:
Las redes neuronales son unas estructuras inspiradas en la arquitectura del cerebro humano, utilizadas para modelar y resolver problemas complejos. Generando aplicaciones en reconocimiento de patrones, procesamiento de lenguaje natural y visión por computadora.

Procesamiento del Lenguaje Natural (PLN):
Este campo de estudio que se enfoca en la interacción entre las computadoras y el lenguaje humano. Generando así la traducción automática, análisis de sentimientos, chatbots y comprensión del lenguaje.

Ética en Inteligencia Artificial:
Esta ética muestra consideraciones de cuestiones éticas relacionadas con la creación y el uso de sistemas de inteligencia artificial. Garantizando así la toma de decisiones justa, transparente y sin sesgos, así como abordar la privacidad y la seguridad.

Introducción a la ética en la Inteligencia Artificial

Transparencia y Explicabilidad:
Este modelo ético exige que los sistemas de IA sean capaces de explicar sus decisiones y procesos de manera comprensible para los usuarios y afectados. Con el fin de garantizar la confianza, además, facilita la rendición de cuentas y permite la detección de posibles sesgos.

Equidad y Sesgo en los Algoritmos:
Esta etapa ética aborda la necesidad de evitar la discriminación y el sesgo injusto al diseñar y entrenar modelos de IA. Con el fin de asegurar que los sistemas no perpetúen o amplifiquen desigualdades existentes y traten a todos los usuarios de manera justa.

Privacidad y Protección de Datos:
En este proceso se considera la importancia de preservar la privacidad y la seguridad de la información personal en el contexto de la IA. Para si, garantizar el manejo ético de datos, evitar el uso indebido y proteger la intimidad de los individuos.

Innovación tecnológica con inteligencia artificial

Web scraping: De acuerdo con (Parikh et al, 2018), Web scraping con machine learning es un enfoque que combina la extracción de datos web (web scraping) con técnicas de aprendizaje automático. En este proceso, se utilizan herramientas de scraping para recopilar información de páginas web, y luego se aplican algoritmos de machine learning para analizar y extraer conocimientos significativos de esos datos. La sinergia entre ambas disciplinas permite la automatización de la adquisición de datos en línea y la aplicación de modelos de machine learning para tareas más avanzadas, como clasificación, predicción o agrupación. A continuación, abordaremos alguno de los funcionamientos de esta:

Extracción de Datos con Web Scraping: Se utilizan técnicas de web scraping para extraer datos estructurados o no estructurados de páginas web. Esto puede incluir la descarga de texto, imágenes, tablas u otros elementos relevantes.

Preprocesamiento de Datos: Los datos recopilados a menudo requieren limpieza y preprocesamiento para convertirlos en un formato utilizable para el machine learning. Esto puede incluir la normalización de texto, la conversión de imágenes a matrices numéricas, entre otros.
Entrenamiento del Modelo de Machine Learning: Se selecciona y entrena un modelo de machine learning según la tarea específica que se desea realizar. Por ejemplo, si se busca clasificar noticias, se podría entrenar un modelo de clasificación.
Aplicación del Modelo: Una vez entrenado, el modelo se aplica a los datos extraídos mediante web scraping. Puede realizar tareas como etiquetar automáticamente datos, predecir tendencias, agrupar información o realizar cualquier tarea para la cual haya sido diseñado.
Iteración y Mejora: El proceso es iterativo, y se pueden realizar ajustes en las técnicas de scraping o en los modelos de machine learning para mejorar la calidad de los resultados con el tiempo.

Matplotlib: De acuerdo con **(**Richert, 2013), Matplotlib en el contexto de Machine Learning es una biblioteca gráfica en Python que facilita la creación de visualizaciones informativas y expresivas. Este recurso es esencial para representar de manera efectiva datos complejos generados por algoritmos de aprendizaje automático. Matplotlib proporciona una interfaz flexible y fácil de usar que permite a los científicos de datos y profesionales de machine learning crear gráficos detallados, como gráficos de dispersión, histogramas y mapas de calor, contribuyendo así a una comprensión más profunda de los patrones y tendencias presentes en los conjuntos de datos. Su versatilidad y capacidad para adaptarse a diversas necesidades hacen de Matplotlib una herramienta fundamental en la visualización de resultados y la toma de decisiones informada en el ámbito del machine learning.

Permutation Importance: De acuerdo con Petch, & Nelson, (2022). la explicabilidad mediante Permutation Importance, en el marco de la innovación tecnológica con inteligencia artificial, se característica afecta el rendimiento del modelo. Se mide la pérdida

de rendimiento después de permutar los valores y se utiliza como indicador de la importancia de la característica.

Interpretación de Importancia Relativa: La importancia relativa de cada característica se determina observando cuánto disminuye la precisión del modelo cuando se permutan sus valores. Características con una disminución significativa en el rendimiento se consideran más importantes.

Visualización y Explicación: Para facilitar la comprensión en el ámbito de la innovación tecnológica, los resultados de Permutation Importance se pueden visualizar mediante gráficos o representaciones visuales, ofreciendo una explicación intuitiva de la contribución de cada característica.
refiere a una técnica que busca proporcionar transparencia y comprensión sobre cómo un modelo de machine learning toma decisiones, centrándose en la importancia relativa de las características. Esta metodología implica la evaluación de la contribución de cada característica al modelo al permutar sus valores, revelando así la influencia específica de cada variable en la precisión del modelo con cuyo objetivo funcional es el siguiente.

Shap Values: De acuerdo con Mahecha, & Barreto, La explicabilidad de Shap Values dentro de la innovación tecnológica con inteligencia artificial se refiere a un enfoque que busca ofrecer interpretaciones claras sobre cómo un modelo de machine learning toma decisiones, centrándose en la contribución de cada característica a las predicciones. Los Shap Values, una técnica específica en este contexto, se utilizan para cuantificar la influencia de cada característica en las decisiones del modelo, permitiendo una comprensión más profunda de los procesos de toma de decisiones. En ellas, a continuación, algunos de sus funcionamientos principales.

Cálculo de Shap Values: Dentro de la innovación tecnológica, los Shap Values se calculan al considerar todas las combinaciones posibles de características, evaluando cómo las predicciones varían con la inclusión o exclusión de características específicas.

Interpretación Individual: Cada Shap Value representa la contribución única de una característica a una predicción, proporcionando información sobre si esa característica afecta positiva o negativamente la salida del modelo.

Suma de Shap Values: La suma de los Shap Values, junto con el valor base del modelo, resulta en la predicción final del modelo para una instancia dada.

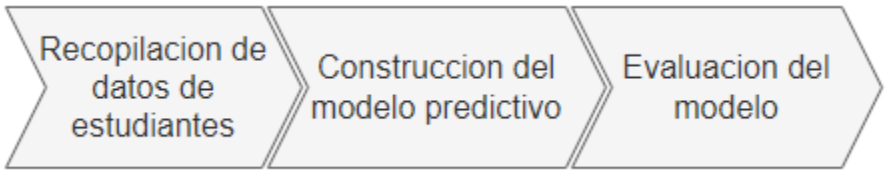
Visualización y Explicación: Dentro del contexto de innovación tecnológica, los Shap Values se pueden visualizar mediante gráficos y representaciones visuales para ofrecer explicaciones intuitivas sobre cómo cada característica influye en las decisiones del modelo.

Desarrollo e implementación del aprendizaje

De acuerdo con (Páez, & Ramírez, 2022). En la revisión de todo este entorno, hemos podido identificar la falta de técnicas para el aprendizaje y la creación de modelos predictivos ante el rendimiento académico de los estudiantes. Con este potencial, hemos visto la oportunidad de analizar e implementar estrategias de prevención para la reprobación de exámenes y/o cursos. la predicción anticipada del rendimiento académico brinda a los profesores e instituciones educativas la capacidad de intervenir de una manera efectiva y temprana para evitar la reprobación del estudiante.

El trabajo en desarrollo propone una metodología que se fundamenta en la recopilación de métricas o información sobre los estudiantes en algunos puntos como el inicio o fin de los cursos. Tal como se ilustra en la figura 1. Posterior a ello, se desarrollarán modelos predictivos que posibiliten la anticipación del rendimiento académico que los estudiantes podrán alcanzar. Ya en la fase final, se llevará a cabo la evaluación del modelo, comparándolo con base a métricas definidas y representativas.

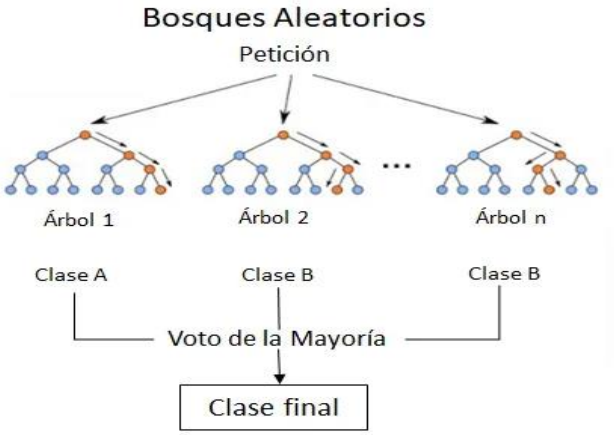
Figura 1
Metodología propuesta



Fuente: Páez & Ramírez (2022).

Bosque aleatorio
En el presente trabajo hemos tratado por optar el modelo del algoritmo de bosques aleatorios tal como se muestra en la figura 2, ya que este presenta una mayor precisión en la predicción del rendimiento académico debido a que son aprendizajes conjuntos en base a diferentes arboles de decisiones que se construyen a partir de una selección aleatoria de variables independientes del conjunto de datos. Cada árbol participe del conjunto analiza una muestra aleatoria extraída del conjunto mediante el proceso de muestreo, también conocido como bootstrapping. Posteriormente, se toman en consideración los resultados de todos los arboles clasificadores y se les aplica el principio de la sabiduría de las masas y finalmente se adoptan las clasificaciones más frecuentes como solución definitiva (Breiman, 2001); (Breiman, 2004).

Figura 2
Bosques aleatorios

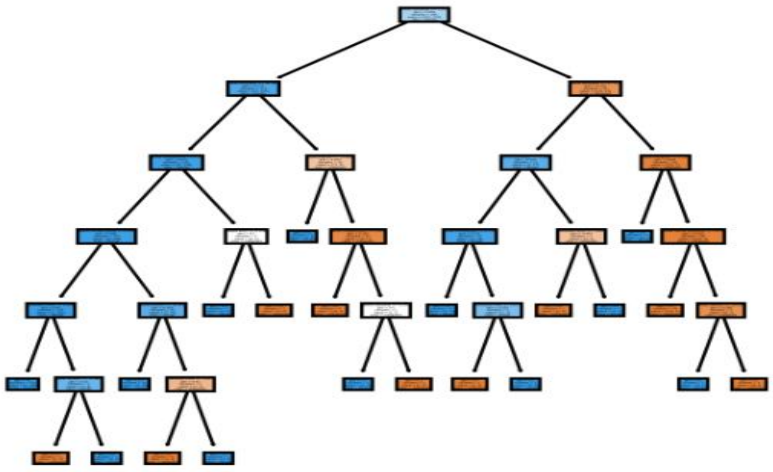


Fuente: Hebert, P., H., (2020).

Árbol de decisión

Las técnicas de aprendizaje automático conocidas como Árboles de Decisión operan dividiendo de manera sucesiva un conjunto de datos en segmentos más específicos, siguiendo una estructura jerárquica que se asemeja a la configuración de un árbol invertido. Inicialmente, el "nodo raíz" representa la totalidad del conjunto de datos. A medida que se desarrolla el árbol, se produce una subdivisión recursiva de los nodos, seleccionando atributos que facilitan la creación de segmentos homogéneos dentro del conjunto de datos, conduciendo eventualmente a los nodos "terminales". Estos últimos contienen los valores correspondientes a la clasificación deseada o la variable objetivo (Charbuty & Abdulazeez, 2021). Este enfoque estructurado permite una representación visual y lógica del proceso de toma de decisiones, desglosando la complejidad del conjunto de datos en segmentos manejables para facilitar la interpretación y análisis.

Figura 3
Árbol de decisiones



Los factores empleados para determinar la división de los nodos, tales como Gini, Entropía, Chi-cuadrado, entre otros, ejercen una influencia significativa en la exactitud de un árbol, y la elección entre ellos depende de si se utilizan los árboles para tareas de clasificación o regresión. La selección del algoritmo para construir el árbol de decisión también se fundamenta en el tipo de variable objetivo. Por ejemplo, se opta por algoritmos como C4.5 y CART cuando la variable a predecir es categórica o continua, mientras que CHAID y QUEST son preferidos cuando las variables son exclusivamente categóricas (Maimon & Rokach, 2014). Esta adaptabilidad permite al usuario seleccionar el método más adecuado según la naturaleza de los datos que está analizando.

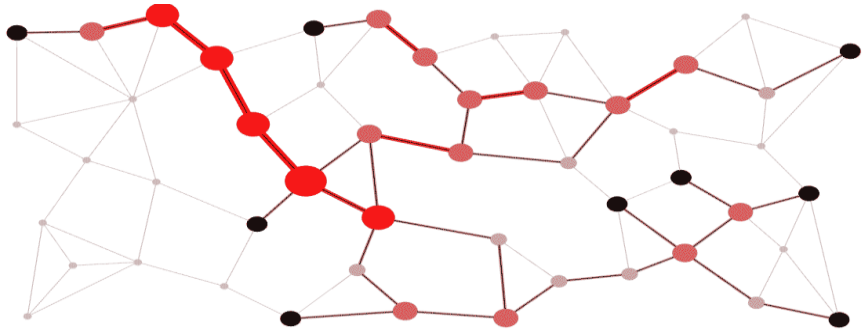
No obstante, a pesar de la valiosa contribución de los Árboles de Decisión, se observa que su eficiencia se ve comprometida cuando el conjunto de datos presenta una elevada cantidad de atributos, lo que conlleva a una profundidad considerable en la estructura del árbol. Esto puede resultar en una predicción precisa de la variable objetivo con un 100% de exactitud en los datos de entrenamiento, pero evidenciar una baja precisión al enfrentarse a datos nuevos. Para abordar este desafío, una estrategia viable implica gestionar la profundidad del árbol o recurrir a enfoques como la implementación de Bosques Aleatorios (Hartshorn, 2016). Esta última alternativa, al emplear múltiples árboles y promediar sus resultados, ayuda a mitigar los riesgos de sobreajuste y mejora la capacidad de generalización del modelo, elevando así su robustez y utilidad en escenarios con complejidades inherentes.

Redes neuronales

El algoritmo conocido como Redes Neuronales, también denominado redes neurales artificiales (Hosseini, Hosseini, & Ahi, 2021), ha encontrado aplicaciones significativas en diversos estudios médicos, demostrando ser altamente confiable en la predicción de diversas enfermedades (Site, Nurmi, & Lohan, 2021). En su esencia, las redes neuronales buscan emular las funciones del cerebro humano en el proceso de toma de decisiones. Este paradigma se asemeja a cómo nosotros, al tomar decisiones, evaluamos múltiples alternativas, identificamos criterios relevantes y les asignamos pesos ponderados para llegar a una solución. De manera similar, las redes neuronales adoptan este enfoque al enfrentar diversas alternativas o entradas de datos, asignando pesos a diferentes criterios o nodos con el propósito de evaluar las opciones y, finalmente, proporcionar la salida correspondiente a la decisión deseada.

En la representación visual proporcionada en la figura 4, se observa la estructura fundamental de las redes neuronales. Esta estructura comprende una capa de entrada, una capa de salida y capas intermedias conocidas como capas ocultas, en las cuales se llevan a cabo los cálculos asociados a los pesos para alcanzar una solución. Cada nodo en estas capas ocultas contribuye de manera única al proceso de toma de decisiones, destacando la complejidad y la capacidad adaptativa de las redes neuronales para modelar patrones complejos en los datos y ofrecer predicciones precisas. Esta flexibilidad inherente las convierte en una herramienta valiosa en diversas disciplinas, particularmente en el ámbito médico, donde la fiabilidad y la precisión son imperativos para la toma de decisiones clínicas fundamentales.

Figura 4 Redes neuronales



Fuente: CIIIA. (2021). Redes neuronales de grafos: Recuperado de https://www.ciiia.mx/noticiasciiia/redes-neuronales-de-grafos-qu-son

Recopilación de Datos

La información que se analizara fue obtenida de algunos estudiantes de la corporación universitaria Remington, esta recolección de información abarco elementos claves como la asistencia a clase, las materias aprobadas, las horas diarias de estudio, etc. Este conjunto de características especificas fueron seleccionadas deliberadamente debido a la practicidad del proceso de recolección y a su comprobada relevancia en estudios previos centrados en la predicción del rendimiento académico, tal como se ha venido evidenciando en trabajo base como (Shahiri et al., 2015). La participación en este estudio abarco un total de 100 estudiantes, para la cual se capturaron detalladamente los atributos mencionados en la (tabla 1). Este enfoque metodológico proporcionara una base robusta para el análisis y la posterior interpretación de los resultados en el contexto de rendimiento académico de los estudiantes de la Remington.

Tabla 1.

Atributos con posibles valores.

Atributos	*Valor*
Asistencia a clase	Si/No
Participación de Actividades	Si/No
Acceso a recursos educativos	Si/No
Nivel de motivación	Alto/Bajo/Medio
Hábitos de estudio	Frecuente/Poco Frecuente
Factores de estrés	Alto/Medio/Bajo
Uso de la tecnología	Si/No
Horas de estudio diarias	1/2/3 horas
Resultado de Exámenes Previos	Bueno/Malo/Medio
Materias reprobadas	>5/>7/>8
Materias aprobadas	<5/<3/<2
Promedio actual	>3.5/<3.0
Preferencia de estudio	Diurna/Nocturna

Procesamiento de la información

Bosque aleatorio

De acuerdo con los resultados basándonos en el bosque aleatorio, se realizarán los siguientes pasos en la aplicación de algoritmos Random Forest:

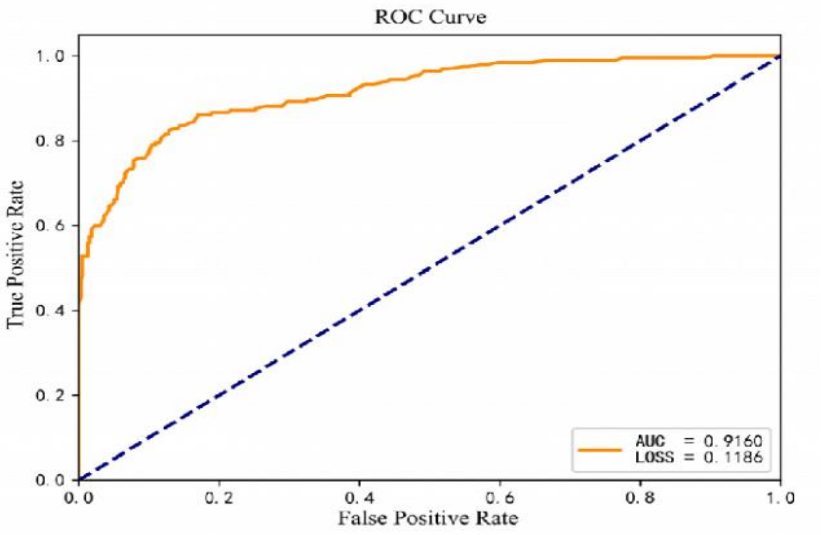
1. Subir las librerías.
2. Definir la función rf_bch.
3. Subir la data
4. Calcular si el alumno aprueba la asignatura.
5. Tenemos presente y muy detallado los filtros de consulta.
6. Convertir los datos cualitativos en datos numéricos (0-1).
7. Definir la función rf_arq
8. usaremos la función rf_arq dentro de un ciclo de 10 a 101con el fin de ver diferentes arquitecturas del random forest y mostrar el resultado del mayor nivel de precisión.

9. Luego de usar el resultado de mayor precisión, llamamos nuevamente la función rf_bch y mostramos el resultado, la matriz de confusión y la curva ROC
10. Ajustar a un árbol de decisión.
11. Calcular y mostrar el Accuracy, así como mostrar la calificación promedio.
12. Calcular y mostrar para profundidades.
13. Mostrar o imprimir gráfico.
14. Calcular y mostrar la matriz de confusión, con la máxima profundidad (2) que tiene el nivel de precisión mayor.
15. Calcular y mostrar la curva ROC

Luego de ejecutar este primer modelo de Random Forest con ROC (Figura 5), se puede constatar la precisión alcanzada en la predicción del rendimiento académico de los estudiantes de la Corporación Universitaria Remington ver la figura. La elección de Random Forest como modelo inicial se justifica por varias razones fundamentales como:

1. Lidiar con la complejidad
2. Reducción de sobreajuste
3. Manejo de variables
4. Versatilidad y escalabilidad

Figura 5 Curva ROC



Fuente: Riyanti, I., E. (2021)

Árbol de decisiones

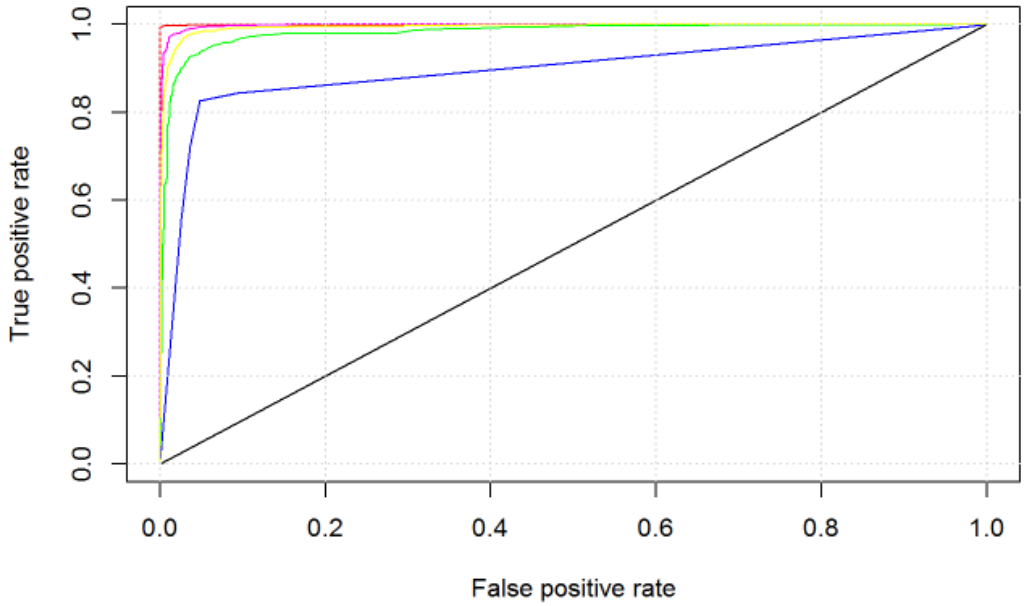
De acuerdo con los resultados basándonos en el árbol de decisiones, se realizarán los siguientes pasos:

1. Subir las librerías.

2. Subir la data
3. Calcular si el alumno aprueba la asignatura.
4. Tenemos presente y muy detallado los filtros de consulta.
5. Convertir los datos cualitativos en datos numéricos (0-1).
6. Barajar las filas
7. División de la data
8. Numero de estudiantes que pasan por el dataset
9. Ajustar a un árbol de decisión.
10. Calcular y mostrar el Accuracy, así como mostrar la calificación promedio.
11. Calcular y mostrar para profundidades.
12. Mostrar o imprimir gráfico.
13. Calcular y mostrar la matriz de confusión, con la máxima profundidad (2) que tiene el nivel de precisión mayor.
14. Calcular y mostrar la curva ROC

Basándonos en los resultados obtenidos mediante el modelo de Árbol de Decisiones y la curva ROC (Figura 6), se debe tener presente las etapas plasmadas anteriormente para seguir una serie de pasos clave que garanticen una interpretación precisa y efectiva.

Figura 6 Curva ROC árbol de decisiones



Fuente: Herrera, O. (2015).

Redes neuronales

De acuerdo con los resultados basándonos en las redes neuronales, se realizarán los siguientes pasos aplicando los debidos algoritmos.

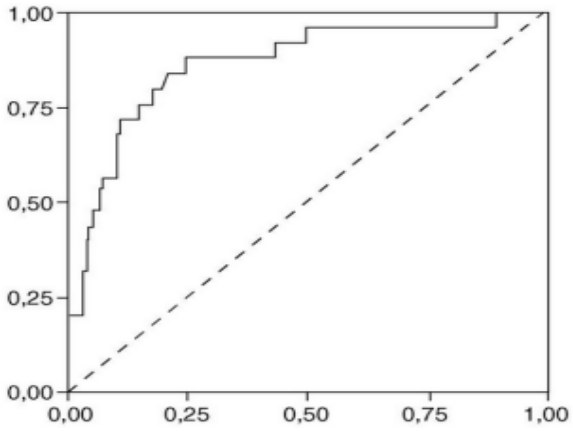
1. Subir las librerías.

2. Definir la función rn_bch para
3. Subir la data
4. Calcular si el alumno aprueba la asignatura.
5. Tenemos presente y muy detallado los filtros de consulta.
6. Convertir los datos cualitativos en datos numéricos (0-1).
7. Definir la función rn_arq para
8. Llamamos a la función rn_arq dentro de un ciclo, identificando las diferentes arquitecturas para una red neuronal y con este podemos retornar los resultados del mayor nivel de precisión.
9. Usando la arquitectura que nos dio en mejor nivel de precisión llamamos a la función rn_bch y mostramos el resultado del nivel de precisión, curva de ROC

Basándonos en los resultados obtenidos mediante el modelo de redes neuronales y la curva ROC (Figura 7), se debe tener presente las etapas plasmadas anteriormente para seguir una serie de pasos clave que garanticen una interpretación precisa y efectiva.

Figura 7 Curva ROC red neuronal

Fuente: Pazos Mandiá, J. C. (2015).



Conclusiones

En trabajo de grado, exploramos y pusimos en práctica tres modelos como el Árbol de Decisiones, Las Redes Neuronales y Bosque Aleatorio, se destaca una valiosa comprensión sobre la capacidad de estos enfoques para abordar problemas complejos, específicamente en el contexto de la predicción del rendimiento académico. Los resultados obtenidos brindan una visión integral de las fortalezas y limitaciones de cada modelo, permitiendo extraer conclusiones significativas.

Las Redes Neuronales, por otro lado, demostraron su poder en la captura de relaciones no lineales y complejas entre las variables, permitiendo una modelación más adaptativa y refinada. La flexibilidad inherente de las redes neuronales es especialmente valiosa cuando se trata de lidiar con conjuntos de datos ricos en información y multidimensionales.

La elección del Bosque Aleatorio como modelo inicial destacó la importancia de la diversidad y la agregación en la mejora de la precisión predictiva. La combinación de múltiples árboles de decisión mitigó los riesgos de sobreajuste y mejoró la generalización del modelo, proporcionando así resultados más robustos y fiables.

Con el modelo de Árbol de Decisiones, se evidenció la capacidad de este enfoque para desentrañar patrones complejos en los datos y proporcionar reglas lógicas discernibles. La interpretación transparente de las decisiones tomadas por el modelo facilita una comprensión clara de los factores determinantes del rendimiento académico.

En conjunto, estos modelos han proporcionado una comprensión profunda de los determinantes del rendimiento académico, permitiendo la identificación de factores clave y el diseño de estrategias de intervención educativa más personalizadas.

Referencias

Agüera, M. (2022). Explainable Machine Learning: Mathematical Optimization in Counterfactual Analysis (Tesis de maestría). Universidad de Sevilla https://idus.us.es/bitstream/handle/11441/142703/DGFM%20MARTIN%20AGUERA%2c%20AGUSTIN.pdf?sequence=1&isAllowed=y.

Bishop, C. (2007). Pattern recognition and machine learning. New York, NY: Springer

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Breiman, L. (2004). Consistency for a simple model of random forests. Statistics Department. . Berkeley: University of California at Berkeley.

Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends, 2(1), 20-28

Espinosa-Zúñiga, J. J. (2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. *Ingeniería, investigación y tecnología*, *21*(3).

Gomez, J. (23 de septiembre de 2023) *Métricas De Evaluación De Modelos En El Aprendizaje Automático*. DataSource. https://www.datasource.ai/es/data-science-articles/metricas-de-evaluacion-de-modelos-en-el-aprendizaje-automatico

Gutiérrez Villaverde, H., Linares Barbero, M., Agüero Correa, A. y Pérez Nuñez, J. (2022). Predicción de rendimiento académico de alumnos usando machine learning. Universidad de Lima, Facultad de Ciencias Empresariales y Económicas, Carrera de Negocios Internacionales.

Hartshorn, S. (2016). Machine learning with random forests and decision trees: A Visual guide for beginners.

Herrera, O. (2015). Tarea3_OscarHerrera. Recuperado de https://rpubs.com/Oskarh2/84197

Hosseini, M.-P., Hosseini, A., & Ahi, K. (2021). A Review on Machine Learning for EEG Signal Processing in Bioengineering. IEEE Reviews in Biomedical Engineering , 14, 204-218. doi:10.1109/RBME.2020.2969915

Hurwitz, J. & Kirsh, D. (2018). *Machine learning for dummies*. , IBM Limited Edition. https://www.ibm.com/downloads/cas/GB8ZMQZ3

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects, Retrieved from: http:// science.sciencemag.org/content/349/6245/255?casa_ token=PngLtzsuefoAAAAA:w2Nq8oJ899bxwJ5nyLMEc01l9nJx6_ O30cobyne8JJfw28Q2k_FHMT6DHc0FvjsIa21Hn5Fa_mylOeEW

Juárez, G. (2017). ¿Cómo funciona el aprendizaje automático (Machine Learning) Retrieved from: http://www.nexolution.com/como-funciona-el-aprendizajeautomatico-machine-learning/

LORENZO, M., & María, J. (2007). *Estadística descriptiva*. ALFA CENTAURO.

Mahecha, C. C. Z., & Barreto, O. S. F. IA Explicable en administración de riesgo de crédito.

Maimon, O. Z., & Rokach, L. (2014). Data mining with decision trees: theory and applications. World scientific, 81.

Páez, A. R., & Ramírez, N. D. G. (2022). Modelos predictivos del rendimiento académico a partir de características de estudiantes de ingeniería. *IE Revista de Investigación Educativa de la REDIECH*, *13*, 1-18.

Palop Alcaide, F. (2022). Predicción de dominios maliciosos utilizando técnicas de Machine Learning.

Parikh, K., Singh, D., Yadav, D., & Rathod, M. (2018). Detection of web scraping using machine learning. Open access international journal of Science and Engineering, 3, 114-118.

Pazos Mandiá, J. C. (2015). TFG_Juan_Carlos_Pazos_Mandia.pdf. Recuperado de https://earchivo.uc3m.es/bitstream/handle/10016/23211/TFG_Juan_Carlos_Pazos_Mandia.pdf

Peláez, I. M. (2016). Modelos de regresión: lineal simple y regresión logística. *Revista Seden*, *14*, 195-214.

Petch, J., Di, S., & Nelson, W. (2022). Opening the black box: the promise and limitations of explainable machine learning in cardiology. Canadian Journal of Cardiology, 38(2), 204-213.

Rendón-Macías, M. E., Villasís-Keever, M. Ángel, & Miranda-Novales, M. G. (2016). Estadística descriptiva. *Revista Alergia México*, *63*(4), 397–407. https://doi.org/10.29262/ram.v63i4.230

Richert, W. (2013). Building machine learning systems with Python. Packt Publishing Ltd.

Russo, C., Ramón, H., Alonso, N., Cicerchia, B., Esnaola, L., & Tessore, J. P. (2016). Tratamiento masivo de datos utilizando técnicas de Machine Learning.

Shahiri, A. M., Husain, W., y Rashid, N. (2015). A review on predicting student's performance using data mining techniques. Procedia Computer Science, 72, 414-422. https://doi.org/10.1016/j.procs.2015.12.157

Site, A., Nurmi, J., & Lohan, E. S. (2021). Systematic Review on Machine-Learning Algorithms Used in Wearable-Based eHealth Data Analysis. IEEE Access, 9, 112221-112235. doi:10.1109/ACCESS.2021.3103268

Suguiura, F. O. R. (2022). Árbol de Decisión en Aprendizaje Automático. REVISTA VARIANZA, 39-46.