**Análisis y predicción de la demanda de educación superior en Colombia con Machine Learning.**

**Integrantes**

*Jefferson Alejandro Ruiz Rojas*

*Juan Camilo Muriel Pérez*

**1. Definición del Problema**

* **Descripción del problema:**  
  El sistema de educación superior en Colombia enfrenta desafíos en la planificación estratégica, la asignación de recursos y la toma de decisiones. Actualmente, la falta de información predictiva sobre la demanda futura de estudiantes en programas académicos y universidades específicas limita la capacidad de las instituciones para optimizar su oferta.

Esto genera problemas como:

Subutilización o saturación de la infraestructura (aulas, laboratorios).

Desequilibrios en la oferta académica, con programas saturados y otros con baja demanda.

Dificultades en la asignación de presupuestos y personal docente.

La ausencia de un modelo predictivo robusto basado en datos históricos resulta en una planificación reactiva en lugar de proactiva, lo que afecta la calidad y la eficiencia del servicio educativo. Este proyecto busca resolver este problema.

* **Objetivos del proyecto:**

**Objetivo general**

Desarrollar un modelo predictivo de Machine Learning que, utilizando el historial de matrícula de los últimos 10 años, pueda estimar con precisión el número de estudiantes matriculados en los próximos años en diversas universidades y programas académicos, tomando como punto de partida dos de las universidades con una importancia relevante en Colombia.

**Objetivos específicos**

**Análisis exploratorio de datos (EDA):** Examinar los datos históricos para identificar tendencias, patrones de estacionalidad y relaciones con variables externas, como indicadores económicos.

**Selección y preprocesamiento de los datos:** Limpiar, transformar y estructurar el conjunto de datos para su uso en los modelos de Machine Learning.

**Desarrollo y entrenamiento de modelos predictivos:** Entrenar y evaluar diferentes algoritmos de Machine Learning, destacando el uso de Random Forest Regressor y Gradient Boosting Machines por su eficacia en el manejo de datos estructurados, con el fin de comparar su rendimiento y seleccionar el modelo que brinde un equilibrio óptimo entre precisión, interpretabilidad y eficiencia computacional en la predicción de la demanda en educación superior.

**Evaluación y validación del modelo:** Medir la precisión de las predicciones utilizando métricas como el Error Cuadrático Medio (RMSE) y el Error Absoluto Medio (MAE).

**Implementación y visualización:** Crear una herramienta o *dashboard* simple que permita a los usuarios visualizar las predicciones de matrícula de manera intuitiva.

* **Justificación:**

Este proyecto es fundamental para mejorar la eficiencia y la planificación en el sector educativo superior.

**Optimización de recursos:** Permite a las universidades **planificar de manera anticipada** el número de cupos, la infraestructura necesaria (aulas, laboratorios) y la contratación de personal docente, lo que reduce costos y mejora la calidad.

**Toma de decisiones estratégica:** Las instituciones pueden usar las predicciones para **ajustar su oferta académica**, identificar programas con alta demanda potencial e invertir de manera más inteligente sus recursos.

**Beneficio social:** Al mejorar la eficiencia y la oferta académica, se puede **garantizar una mejor experiencia educativa** para los futuros estudiantes y una mejor distribución de la oferta a nivel nacional. Esto puede ayudar a reducir la deserción y mejorar el rendimiento de los programas.

En resumen, este proyecto convierte los datos históricos en información valiosa y procesable, lo que contribuye a un **sistema de educación superior más resiliente y eficiente**.

**2. Recolección de Datos**

* **Fuente de los datos:**  
  SNIES (Sistema Nacional de Información de la Educación Superior)  
  <https://snies.mineducacion.gov.co/>
* **Descripción del dataset:**

El dataset es una colección histórica que abarca la matrícula de estudiantes en las universidades de Colombia durante los últimos 10 años. Este conjunto de datos está estructurado para permitir un análisis detallado y multidimensional del sector de educación superior. Las variables clave incluidas son:

* + **Nombre de la institución:** Identificador de la universidad (ej. Universidad Nacional, Universidad de los Andes).
  + **Sector:** Clasificación de la institución (p. ej., Pública, Privada).
  + **Nombre del programa:** El programa académico en el que el estudiante está matriculado (ej. Ingeniería de Sistemas, Medicina).
  + **Categoría:** Tipo de programa (ej. Pregrado, Posgrado, Técnico).
  + **Sexo:** Género del estudiante (Masculino, Femenino).
  + **Semestre:** El periodo académico en el que se realizó la matrícula (ej. Primer Semestre, Segundo Semestre).
  + **Número de matriculados:** La variable objetivo, que representa la cantidad de estudiantes matriculados por cada combinación de las variables anteriores.

Esta granularidad en los datos permitirá a los modelos de machine learning no solo predecir la matrícula total, sino también identificar tendencias específicas por programa, sector o género, lo cual es crucial para una planificación más precisa y estratégica.

* **Herramientas utilizadas:**  
  Python, Google Colab, VS Code

**3. Preparación de los Datos**

* **Limpieza de datos:**

**Integración de datos:** Se unifican los documentos históricos de cada año en un solo conjunto de datos. Este proceso, realizado a través de un *script* de Python, garantiza que toda la información esté consolidada y lista para el análisis, facilitando el manejo y la consistencia del dataset.

**Agregación de variables:** La variable de **género** no es relevante para el objetivo principal de este proyecto, que es predecir la matrícula total de estudiantes. Por ello, se agrupan los datos y se **suma el número de matriculados** por programa, institución, año, etc., lo que permite tener una sola fila para cada combinación sin distinción de género.

Este proceso de limpieza y agregación nos permite crear una base de datos más concisa y adecuada para el modelado, lo que mejora la eficiencia y la precisión de las predicciones.

* **Transformación de variables:**  
  Se seleccionan las siguientes variables del conjunto de datos unificado:
  + **Institución Educativa:** Se mantiene como una variable categórica para capturar la matrícula específica de cada universidad.
  + **Programa Académico:** Similar a la institución, esta variable permite que el modelo identifique patrones de matrícula por tipo de programa.
  + **Año:** Una variable numérica fundamental para el análisis de series de tiempo. Representa el factor temporal y la progresión de la matrícula a lo largo de los años.
  + **Semestre:** Esta variable categórica es esencial para capturar la **estacionalidad** en los datos (p. ej., un aumento en la matrícula en el segundo semestre del año).
  + **Cantidad de Matriculados:** Esta es nuestra **variable objetivo**, el valor que el modelo deberá predecir.

Este proceso de filtrado asegura que el modelo solo se entrene con los datos más significativos, lo que reduce la complejidad y mejora su rendimiento. Las variables no seleccionadas, como "categoría" o "sector," pueden ser exploradas más adelante en caso de que el modelo base necesite una mayor granularidad.

* **Visualización y análisis exploratorio:**

*Grafica 1.*

Gráfico, Gráfico de dispersión

El contenido generado por IA puede ser incorrecto.

**Descripción de la gráfica:**La gráfica de dispersión muestra la relación entre los valores reales de matrícula (eje X) y los valores predichos por el modelo (eje Y) en el conjunto de prueba. Cada punto representa una observación, mientras que la línea diagonal negra indica la referencia ideal en la que las predicciones coinciden exactamente con los valores reales. Se observa que la mayoría de los puntos se concentran cerca de la diagonal, lo cual refleja un buen ajuste del modelo y un nivel de error relativamente bajo. No obstante, algunos puntos alejados de la línea sugieren la existencia de casos con mayor variabilidad o posibles valores atípicos en los datos.

*Grafica 2.*

Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

**Descripción de la gráfica:**  
La gráfica muestra la evolución histórica del número de matriculados en el programa de Ingeniería Informática en la Corporación Universitaria Lasallista, diferenciados por año y semestre. Se observa una alta variabilidad en la demanda a lo largo del tiempo, con picos en algunos periodos y caídas en otros. El punto rojo marca la predicción realizada por el modelo para el primer semestre de 2026, estimando aproximadamente 35 estudiantes matriculados, lo que refleja una proyección acorde con las tendencias recientes del programa. (Programa de Ingeniería Informática en la Corporación Universitaria Lasallista, fue la información seleccionada para ilustración, que funcionará con total eficiencia para otra Institución o carrera que el usuario desee predecir).

**4. Selección y Entrenamiento del Modelo**

* **Modelo(s) seleccionado(s):**  
  El enfoque inicial para la predicción de la matrícula estudiantil se basará en el uso de **modelos de regresión lineal**. Esta elección metodológica se justifica por varias razones clave que se alinean con los objetivos del proyecto.
* **Motivo de la selección:**  
  Esta decisión se justifica por su **simplicidad, eficiencia y capacidad de interpretación**. Un modelo de regresión lineal nos permitirá establecer de manera rápida y clara una relación base entre las variables históricas (como el año) y la matrícula de estudiantes. Esta aproximación no solo es computacionalmente eficiente, sino que también facilita la comprensión de los resultados para cualquier persona, ya que podemos identificar fácilmente cómo cada factor contribuye a la predicción.
* **Entrenamiento del modelo:**  
  La variable dependiente tomada es la cantidad de estudiantes matriculados.  
  Los datos son divididos en dos grupos principales: Entrenamiento (80%) y prueba (20%)

**5. Evaluación del Modelo**

* **Métricas utilizadas:**

Para evaluar el desempeño del modelo se emplearon métricas de regresión, dado que la variable objetivo corresponde al número de estudiantes matriculados. En particular, se calcularon:

* Error Absoluto Medio (MAE): mide la diferencia promedio en valores absolutos entre las predicciones y los valores reales.
* Error Cuadrático Medio (MSE): penaliza con mayor peso los errores grandes, lo que permite identificar desviaciones significativas.
* Raíz del Error Cuadrático Medio (RMSE): facilita la interpretación del MSE al expresarlo en las mismas unidades que la variable objetivo (estudiantes).
* Coeficiente de Determinación (R²): mide la proporción de la varianza explicada por el modelo, donde valores cercanos a 1 indican un mayor ajuste.
* **Resultados obtenidos:**  
  El modelo entrenado con Random Forest Regressor arrojó los siguientes resultados al evaluarse sobre el conjunto de prueba:

|  |  |
| --- | --- |
| **MÉTRICA** | **VALOR OBTENIDO** |
| Error Absoluto Medio (MAE) | 58.38 |
| Error Cuadrático Medio (MSE) | 23550.012 |
| Raíz del Error Cuadrático Medio (RMSE) | 153.46 |
| Coeficiente de Determinación () | 0.9436 |

* **Comparación entre modelos:**  
  Aunque el enfoque inicial incluyó la posibilidad de probar múltiples algoritmos, en esta fase se implementó principalmente el Random Forest Regressor, el cual mostró un buen rendimiento en términos de precisión y robustez frente a datos con alta dimensionalidad. Este algoritmo resultó adecuado debido a su capacidad de reducir el sobreajuste mediante la combinación de múltiples árboles de decisión y su habilidad para capturar relaciones no lineales entre las variables.

**6. Despliegue y Comunicación de Resultados**

* **Aplicación práctica:**  
  El modelo predictivo desarrollado con Random Forest Regressor fue implementado de manera experimental en un entorno de programación con Python. A través del código, se generaron predicciones específicas para programas y universidades seleccionadas, proyectando el número estimado de estudiantes en periodos futuros. Adicionalmente, se diseñaron visualizaciones que integran el historial de matrículas con los valores predichos, lo cual facilita la interpretación de los resultados. Como paso complementario, se plantea la creación de un dashboard interactivo, que permita a directivos de instituciones educativas consultar las proyecciones de matrícula de forma intuitiva y tomar decisiones estratégicas sobre cupos, contratación de docentes e inversión en infraestructura.
* **Conclusiones del proyecto:**  
  El proyecto logró implementar un modelo de Machine Learning capaz de predecir la demanda de matrícula en programas de educación superior en Colombia. A través del análisis exploratorio se identificaron patrones relevantes, como la estacionalidad en los semestres y el crecimiento sostenido en áreas tecnológicas y de ingeniería. El modelo alcanzó un buen desempeño (con un R² cercano a 0.94 en pruebas iniciales), lo que indica un alto nivel de confiabilidad para la estimación de futuros escenarios de matrícula.  
  Entre los principales aprendizajes se destaca la importancia de un preprocesamiento riguroso de los datos (codificación de variables categóricas, integración de fuentes históricas y eliminación de redundancias), así como la utilidad de técnicas de visualización para comprender mejor los patrones en los datos.
* **Trabajo futuro:**  
  Para fortalecer el proyecto y ampliar su alcance, se sugieren las siguientes líneas de mejora:
* Comparación de algoritmos: Implementar modelos adicionales, como Gradient Boosting Machines y redes neuronales, con el fin de evaluar si aportan mayor precisión que el Random Forest.
* Incorporación de variables externas: Integrar indicadores económicos y demográficos (PIB, tasa de empleo, población en edad universitaria) para enriquecer el modelo y mejorar la capacidad predictiva.
* Automatización y despliegue en la nube: Crear un sistema automatizado que actualice periódicamente las predicciones con nuevos datos, desplegado en plataformas como AWS, GCP o Azure.
* Dashboard interactivo: Desarrollar una interfaz web accesible para instituciones educativas, que les permita explorar predicciones personalizadas por institución, programa y semestre.

En síntesis, el proyecto representa un avance significativo hacia la utilización de técnicas de Machine Learning en la planificación estratégica de la educación superior en Colombia, abriendo la puerta a futuras aplicaciones de análisis predictivo en políticas públicas y gestión institucional.

**Anexos**

* Ministerio de Educación Nacional de Colombia. (s. f.). SNIES – Sistema Nacional de Información de la Educación Superior. Recuperado de <https://snies.mineducacion.gov.co/>
* Géron, A. (2022). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O’Reilly Media.
* Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
* Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.