Plan Overview

A Data Management Plan created using DMPTool

Title: Predicción de Resultados en Pruebas Saber Pro a partir de Saber 11 utilizando Modelos de Machine Learning

Creator: Mateo Grisales

Affiliation: Universidad de Los Andes (uniandes.edu.co)

Funder: Universidad de Los Andes (uniandes.edu.co)

Template: Digital Curation Centre

Project abstract:

El proyecto tiene como objetivo predecir el rendimiento de los estudiantes en las pruebas Saber Pro a partir de los datos obtenidos en las pruebas Saber 11, utilizando modelos de aprendizaje supervisado. El análisis se centrará en la identificación de un conjunto de características relevantes (socioeconómicas y académicas) que permitan determinar patrones en el rendimiento de los estudiantes en sus evaluaciones de secundaria y prever su éxito en la educación superior. Para lograr esto, se emplearán técnicas de reducción de dimensionalidad como PCA, así como diferentes algoritmos de machine learning para generar un modelo predictivo que sirva de base para la intervención educativa temprana, buscando mejorar áreas críticas como matemáticas y comprensión de lectura.

Start date: 09-30-2024

End date: 11-30-2024

Last modified: 10-20-2024

Predicción de Resultados en Pruebas Saber Pro a partir de Saber 11 utilizando Modelos de Machine Learning

Data Collection

What data will you collect or create?

Los datos de las pruebas realizadas por el Instituto Colombiano para la Evaluación de la Educación (ICFES), tanto en las categorías Saber 11 como Saber Pro, están disponibles en los datos abiertos del gobierno. Para la prueba Saber 11, existe un conjunto de datos que abarca el periodo de 2010 a 2022, mientras que para la prueba Saber Pro se cuenta con un conjunto de datos anonimizados desde 2018 hasta 2022.

En el caso de Saber 11, el conjunto de datos contiene 7,11 millones de registros, donde cada fila representa a un individuo que presentó el examen. Las 51 columnas proporcionan información sobre la ubicación geográfica del evaluado, la institución educativa a la que pertenece, su estrato socioeconómico, condiciones de vivienda, y los resultados obtenidos en la prueba.

Por su parte, los datos de Saber Pro incluyen 1,2 millones de registros que corresponden a los estudiantes que realizaron la prueba. Las 57 columnas ofrecen información similar, abarcando aspectos como la ubicación geográfica, la institución educativa, el financiamiento de la carrera, el estrato socioeconómico, las condiciones de vivienda y los resultados del examen.

Ambas bases de datos también contienen un identificador único para cada persona que presentó la prueba, lo que permite, en algunos casos, analizar el desempeño de un ciudadano tanto en la prueba Saber 11 como en Saber Pro.

How will the data be collected or created?

Los datos serán descargados en formato CSV desde la página de datos abiertos del gobierno. Posteriormente, estos archivos se subirán a un repositorio público en GitHub, con el fin de facilitar su acceso y manipulación para el desarrollo de los análisis correspondientes.

Documentation and Metadata

What documentation and metadata will accompany the data?

La documentación de las bases de datos está previamente disponible en las páginas de datos abiertos del gobierno. Adicionalmente, se anexarán documentos en el repositorio de GitHub con las definiciones de cada columna, de manera que cualquier persona pueda entender la naturaleza de los datos y las variables presentes en las bases de datos.

Ethics and Legal Compliance

How will you manage any ethical issues?

Los resultados no incluirán ningún documento que permita identificar a las personas que presentaron la prueba, garantizando así la protección de su identidad en el análisis propuesto. El objetivo es identificar y caracterizar los perfiles que obtienen mejores resultados en la prueba Saber Pro, basándose en sus condiciones socioeconómicas y en los resultados obtenidos previamente en la prueba Saber 11.

How will you manage copyright and Intellectual Property Rights (IP/IPR) issues?

Los datos publicados en la página de datos abiertos del gobierno están bajo la Licencia de Datos Abiertos (LDA), la cual promueve su uso como información pública, siempre que se les otorgue la referencia adecuada y no se modifiquen de manera indebida.

Storage and Backup

How will the data be stored and backed up during the research?

Por cuestiones de recursos, GitHub es la mejor herramienta para almacenar las versiones disponibles. Aunque AWS es óptimo, conlleva un costo asociado. Por lo tanto, en el repositorio se creará una carpeta destinada a guardar las versiones de avance para cada entrega.

How will you manage access and security?

Todos los miembros del equipo tendrán acceso a los datos para el desarrollo del proyecto.

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved?

Durante el desarrollo del proyecto, se deberán almacenar los datos utilizados para el modelamiento y evaluación de cada perfil, según la técnica de machine learning empleada para el pronóstico de resultados. Esta información debe ser resguardada en el repositorio para facilitar su replicación en futuras iteraciones del modelo.

What is the long-term preservation plan for the dataset?

Las versiones del conjunto de datos se guardarán en GitHub. Al utilizar una base de datos pública, podremos preservar los resultados del proyecto, almacenándolos en el repositorio para facilitar ajustes o proyectos futuros con la información disponible. En este caso, conservaremos los archivos CSV de los datos iniciales, así como las versiones de avance, incluyendo los resultados finales del modelo, junto con sus datos de entrenamiento y evaluación necesarios para la posible replicación del ejercicio.

Data Sharing

How will you share the data?

El repositorio de GitHub será público, lo que permitirá a cualquier usuario interesado en utilizar la información presente en este proyecto crear una rama y acceder a los datos. El acceso será libre.

Are any restrictions on data sharing required?

Los datos no cuentan con ninguna restricción y van a poder ser accedidos por todos los miembros del equipo.

Responsibilities and Resources

Who will be responsible for data management?

Todo el equipo de trabajo estará involucrado en el manejo de la información para el proyecto. Sin embargo, Sebastián Barrera será el encargado de organizar las versiones del repositorio y asegurarse de que todo esté disponible para el público en general.

What resources will you require to deliver your plan?

Necesitamos acceso a GitHub, Python a través de Notebooks, datos abiertos mediante la página del gobierno colombiano y, finalmente, una herramienta de visualización como Looker Studio para la posible elaboración de interfaces. Estas herramientas son gratuitas, por lo que no se requerirán gastos adicionales para el desarrollo del proyecto.

Planned Research Outputs

Dataset - "Conjunto de datos procesados de Saber 11 y Saber Pro"

Este conjunto de datos incluye las variables relevantes extraídas de los resultados de las pruebas Saber 11 (2010-2022) y los resultados correspondientes en las pruebas Saber Pro (2018-2022). Los datos han sido limpiados y transformados, aplicando técnicas de reducción de dimensionalidad como PCA para seleccionar las características más influyentes. El dataset contiene información socioeconómica y académica de los estudiantes, y está destinado a ser utilizado para el entrenamiento y validación de un modelo predictivo que estima los resultados de Saber Pro.

Model representation - "Modelo predictivo de desempeño en Saber Pro"

Modelo predictivo supervisado basado en técnicas de machine learning, como regresión lineal, árboles de decisión y redes neuronales, que utiliza los datos de las pruebas Saber 11 para predecir el desempeño de los estudiantes en Saber Pro. El modelo ha sido entrenado y validado utilizando datos históricos, con el objetivo de identificar patrones de rendimiento relacionados con características socioeconómicas y académicas. La representación del modelo incluye su arquitectura, parámetros y métricas de rendimiento, lo que facilita su evaluación y replicación.

Text - "Informe final de resultados y recomendaciones"

El informe final documenta los resultados del análisis de datos y la implementación del modelo predictivo. Incluye una descripción detallada del problema de investigación, la metodología de análisis de datos, los hallazgos principales, y visualizaciones descriptivas que muestran la relación entre los perfiles de Saber 11 y los resultados de Saber Pro. Además, ofrece recomendaciones para intervenciones educativas que podrían ser implementadas para mejorar el desempeño de los estudiantes en las áreas más débiles antes de su ingreso a la educación superior.

Interactive resource - "Prototipo interactivo para predicción de desempeño en Saber Pro"

Este prototipo interactivo permite ingresar el perfil académico y socioeconómico de un estudiante basado en los resultados de Saber 11, y genera una predicción personalizada de su desempeño en Saber Pro. El prototipo muestra gráficamente las áreas de mejora sugeridas, como matemáticas o lectura crítica, y proporciona recomendaciones para intervenciones específicas. Esta herramienta está diseñada para apoyar a las instituciones educativas en la identificación de estudiantes que podrían beneficiarse de programas de fortalecimiento antes de su ingreso a la educación superior.

Planned research output details

Title	Type	Anticipated release date	Initial access level	Intended repository(ies)	Anticipated file size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
Conjunto de datos procesados de Saber 11 y Saber P ...	Dataset	Unspecified	Open	None specified		None specified	None specified	No	No
Modelo predictivo de desempeño en Saber Pro	Model representation	Unspecified	Open	None specified		None specified	None specified	No	No
Informe final de resultados y recomendaciones	Text	Unspecified	Open	None specified		None specified	None specified	No	No
Prototipo interactivo para predicción de desempeño ...	Interactive resource	Unspecified	Open	None specified		None specified	None specified	No	No