Deep Learning

Primera Entrega

Profesor

RAUL RAMOS POLLAN

Responsables

CAMILO VÉLEZ PALACIO

JUAN PABLO YARCE ESCOBAR

Ingeniería de Sistemas

Facultad de Ingeniería

Universidad de Antioquia

Semestre 2023-2

2023

Contexto de aplicación: Identificar los factores socioeconómicos y circunstanciales que más influyen en el desempeño en el ICFES de un año específico, en esté caso el 2019.
Objetivo de machine learning: Analizar la importancia relativa de cada característica en relación con los resultados del examen.
Dataset: Datos de aproximadamente 500,000 estudiantes con resultados en el ICFES y características socioeconómicas.
Al contar con una cantidad considerablemente grande de filas, se planea limitar el estudio a solamente los estudiantes de Antioquia, igualmente también se planea hacer una delimitación de las columnas para poder centrarnos en sólo unas cuantas que nos sea de interés
Tamaño del dataset: 546.000 filas y 82 columnas
Distribución de las clases: No aplicable, ya que se trata de un análisis de importancia de características.
Métricas de desempeño: Las métricas planeadas de momento serían la de importancia de características(Feature importancia), gráficos de influencia, análisis de varianza y análisis de correlación
Referencias: El dataset fue sacado de la siguiente fuente
https://www.datos.gov.co/Educaci-n/Saber-11-2019-2/ynam-yc42
De resultados previos se pueden observar distintos análisis sobre los ICFES pero en años anteriores y siguientes:

1. https://www.kaggle.com/code/sorelyss/icfes-eda
    - En este apartado de Kaggle, un usuario de nombre SORELYS utiliza un dataset similar al escogido por nosotros solo que en su caso el dataset incluye los datos desde el año 2018 a 2021.
    - Con este recurso, el usuario muestra varias métricas encontradas en los datos como por ejemplo: puntajes de cada materia por género, puntajes de cada materia por edad y puntajes de cada materia basándose en el puntaje alimenticio del estudiante.
    - Estas métricas nos pueden dar una idea general de cómo se ven algunos apartados de los datos que usaremos nosotros.
2. https://manglar.uninorte.edu.co/bitstream/handle/10584/9877/Documento%20principal.pdf?sequence=6&isAllowed=y

- Este es un artículo que se encuentra en el repositorio de la Universidad del Norte de Colombia, sus autores Luis K. Avila, Emanuel Ospino y Arlinton J. Paez, usaron una data de los ICFES desde el año 2017-1 a 2021-1 para analizar cómo diferentes variables socioeconómicas se relacionan con el puntaje obtenido.
- Según los autores, usaron algoritmos de aprendizaje supervisado y árboles de decisión para generar unas reglas de clasificación y definir lo que ellos llaman “por encima de la media” y “por debajo de la media” y ver el impacto de estas variables socioeconómicas sobre el puntaje global.
- Este trabajo puede sernos de mucha utilidad pues buscaba algo muy similar a nuestra idea de trabajar con este dataset.

3. https://repository.ucatolica.edu.co/server/api/core/bitstreams/7463099f-9b5b-466f-a19d-e5ec7bc2ff6c/content
    - Este trabajo de grado, elaborado por Andres G. Neita Sanchez y Jose D. Mora Giraldo desde la Universidad Católica de Colombia, usa un dataset de las pruebas icfes desde el año 2010 a 2021 y buscaba especificarlos a los estudiantes de la institución “San Ricardo Pampuri” para predecir los resultados de los estudiantes de esta misma institución en el año 2023.
    - Según los autores, usaron dos modelos predictivos supervisados, regresión lineal múltiple y árboles de regresión para poder llegar a predecir los resultados del año 2023.
    - Si bien este artículo busca algo un poco diferente a lo que nuestro trabajo apunta, puede servirnos para guiarnos en base al dataset y qué esperar del comportamiento de algunos datos.