Proyecto Big Data

JUAN JOSE ALVAREZ
JUAN PABLO ARIAS
KEVIN CALDERON
PAULA ANDREA ROMERO

Contenido

introducción y problema de negocio

- El rendimiento en las pruebas Saber 11 en Colombia refleja desigualdades sociales y educativas.
- Objetivo del proyecto: Identificar factores asociados al puntaje global del ICFES y segmentar a los estudiantes por perfiles socioeducativos.
- Enfoque: Análisis predictivo y de segmentación usando técnicas de Machine Learning aplicadas en PySpark (MLlib).
- Alcance: Datos de Armenia, Bogotá, Medellín y Neiva (ICFES), combinados en un solo dataset.

Filtros y Transformación

- Unión de 4 datasets (Armenia, Bogotá, Medellín y Neiva).
- Selección de 15 columnas relevantes (puntajes, contexto socioeconómico).
- Tratamiento de valores faltantes:
- Imputación de puntajes por media.
- Eliminación de registros con nulos persistentes.
- Cálculo del puntaje global:

$$\text{Índice Global} = \frac{3(\text{Lectura Critica}) + 3(\text{Matemáticas}) + 3(\text{Ciencias}) + 3(\text{Sociales}) + 1(\text{Inglés})}{13}$$

- Transformaciones adicionales:
- Indexación de variables categóricas.
- Vectorización con VectorAssembler.
- Escalado de variables para clustering.
-
- Eliminación de registros con puntajes fuera del rango oficial (0–100 por área, 0–500 total).
- Eliminación de columnas con más del 35% de valores nulos o inconsistencias.
- Eliminación de columnas categóricas originales luego de ser indexadas para evitar redundancia.

Respuesta preguntas del negocio

1. ¿Existe una correlación significativa entre el índice de pobreza y los puntajes globales del ICFES en los municipios analizados?

2. ¿Qué relación hay entre la penetración de internet en los hogares y los resultados Infraestructura y Desarrollo Municipal?

3. ¿Cómo afecta el nivel de inversión municipal en educación a los puntajes de ICFES en secundaria en los municipios con mayor y menor inversión?

4. ¿Cuál es la evolución de los puntajes promedio del ICFES entre 2016 y 2022 en los municipios seleccionados, y qué factores explican las principales variaciones?

Respuesta preguntas del negocio

1. ¿Qué características socioeconómicas comunes presentan los municipios con los peores desempeños en las pruebas ICFES?

2. ¿Qué tan predictivo es el índice de matrícula escolar sobre los puntajes obtenidos en las pruebas Saber 11?

3. ¿Cuáles son los factores que más influyen en el puntaje global del ICFES según un modelo predictivo multivariable?

4. ¿Qué diferencias existen entre municipios capitales (como Bogotá, Medellín, Neiva y Armenia) en cuanto al impacto de las variables macroeconómicas en el desempeño educativo?

Selección de técnicas de ML

MODELOS SUPERVISADOS (PREDDICIÓN DEL PUNTAJE)

- Regresión Lineal
- Árbol de Decisión
- Random Forest
- Gradient Boosting (GBT)

MODELOS NO SUPERVISADOS (SEGMENTACIÓN):

- K-Means
- Gaussian Mixture Models (GMM)

Técnicas seleccionadas por su escalabilidad, interpretabilidad y capacidad de capturar relaciones no lineales.

Preparación para el modelado

- Se eliminaron variables altamente correlacionadas con PUNT_GLOBAL.
- Se normalizaron variables con StandardScaler.
- Se seleccionaron variables relevantes según criterios de negocio: conectividad, nivel socioeconómico, jornada, educación de padres.

Mlib en colab

Configuración del entorno:

- Instalación y configuración de PySpark en Colab con entorno virtual.
- Procesamiento distribuido en Colab:
- Aunque Colab no es un clúster real, Spark ejecuta tareas en paralelo sobre el driver local.
- Se trabajó con datasets de miles de registros usando transformaciones distribuidas (e.g. withColumn, select, filter).

Mlib en colab

Modelado con MLlib:

- Uso de VectorAssembler para consolidar variables predictoras.
- Transformación de variables categóricas con StringIndexer.
- Escalamiento con StandardScaler para clustering.
- Entrenamiento de modelos supervisados (LinearRegression, RandomForestRegressor, GBTRegressor) y no supervisados (KMeans, GaussianMixture).
- Evaluación con RegressionEvaluator y ClusteringEvaluator.

Ventajas del enfoque:

- Permite reproducir análisis de Big Data sin recursos costosos.
- Fácil de integrar con visualizaciones de matplotlib y seaborn para análisis complementario.

mlib cluster de alto rendimiento

Se aplicaron modelos supervisados y no supervisados utilizando MLlib sobre un clúster de alto rendimiento con Apache Spark. Random Forest y Gradient Boosting ofrecieron el mejor desempeño predictivo, mientras que K-Means y GMM permitieron identificar clústeres de alto rendimiento académico asociados a mejores condiciones socioeconómicas y mayor conectividad.

Aplicacion de metricas

Modelos Supervisados:

- RMSE, MAE, R^2
- Comparación:

Modelo	RMSE	MAE	R^2
Regresión Lineal	45.62	36.72	0.1809
Árbol de Decisión	40.89	32.91	0.3419
Random Forest	40.67	32.75	0.3488
Gradient Boosting	39.79	31.84	0.3767

- Objetivo: Predecir el puntaje global del ICFES.
- Métricas aplicadas: RMSE, MAE, R^2.
- Hallazgos:
- Regresión Lineal tuvo bajo poder explicativo.
- Árbol de Decisión y Random Forest mejoraron notablemente el ajuste.
- Gradient Boosting fue el modelo con mejor desempeño, con el menor error y mayor R^2.
- Conclusión: Es adecuado para sistemas de predicción del desempeño académico.

Aplicacion de metricas

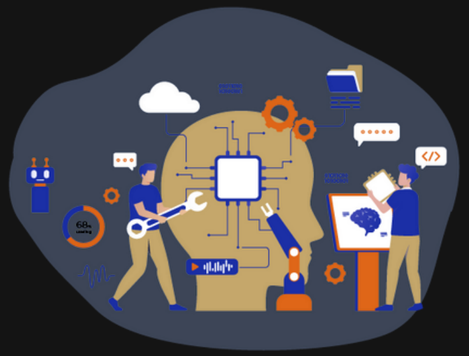
Silhouette Score, distancia entre clústeres, log-likelihood.

Modelo	Silhouette	Distancia	Log-Likelihood
K-Means	0.21	N/A	N/A
GMM	0.40	5.79	-5,637,004.94

- Objetivo: Agrupar estudiantes según condiciones y rendimiento.
- Métricas aplicadas: Silhouette Score, distancia entre clústeres, log-likelihood.
- Hallazgos:
- GMM superó a K-Means en cohesión y separación de grupos.
- Silhouette de 0.40 indica buena segmentación.
- Log-likelihood constante demuestra estabilidad del modelo.
- Conclusión: GMM permite identificar perfiles útiles para políticas educativas focalizadas.

Conclusiones

- El puntaje ICFES está significativamente influenciado por factores como el estrato, la escolaridad de los padres y el acceso a internet.
- El modelo Gradient Boosting logró predecir el puntaje con alta precisión relativa.
- La segmentación con GMM identificó perfiles estudiantiles útiles para diseñar políticas educativas focalizadas.
- Se recomienda:
- Priorizar intervenciones a estudiantes en clústeres de alto riesgo.
- Incorporar estos modelos en herramientas de monitoreo educativo.
- Ampliar el análisis con más años y regiones.
- Incluir nuevas variables (infraestructura, apoyo escolar).



GRACIAS POR LA ATENCIÓN