FUNDACIÓN UADE
UNA GRAN UNIVERSIDAD

TRABAJO PRÁCTICO OBLIGATORIO

CIENCIA DE DATOS

DOCENTE: MARIANO DANIEL FRANCISCO

Grupo 1:
Valentino Cattaneo Luna,
Mateo Maestromey,
Diego Eduardo Cibeira

ANÁLISIS DE DESFAVORABILIDAD EDUCATIVA EN LA PROVINCIA DE BUENOS AIRES

1. Introducción

Contexto y problema: La equidad educativa en la Provincia de Buenos Aires se ve influenciada por múltiples factores socioeconómicos y geográficos. Uno de los principales indicadores para comprender estas desigualdades es el índice de desfavorabilidad educativa, que clasifica los establecimientos escolares según su contexto social, económico y territorial.

Este índice, definido oficialmente por la Dirección General de Cultura y Educación (DGCyE), toma valores entre 0 y 5, donde 0 representa contextos muy favorables y 5 contextos altamente desfavorables.

Cada nivel implica, además, un porcentaje adicional en el salario docente, como compensación por las condiciones del entorno educativo:

Nivel	Porcentaje adicional
0	0%
1	30%
2	60%
3	90%
4	100%
5	120%

El objetivo del presente trabajo es analizar los factores asociados a la desfavorabilidad educativa y predecir el nivel de desfavorabilidad de los establecimientos, integrando datos poblacionales y educativos de la Provincia de Buenos Aires.

Objetivos y preguntas:

Objetivo general:
Analizar la relación entre las características institucionales, geográficas y poblacionales de los municipios bonaerenses y el nivel de desfavorabilidad educativa de sus escuelas, para construir un modelo predictivo que permita estimar la desfavorabilidad de nuevos establecimientos.

Objetivos específicos:

- Unir y depurar datasets de establecimientos educativos y población por municipio.
- Analizar los patrones de distribución de la desfavorabilidad a nivel municipal.
- Evaluar la influencia de variables como subvención, ámbito, área, y población total.
- Entrenar y evaluar modelos de *Machine Learning* supervisados (Random Forest Classifier) para predecir el nivel de desfavorabilidad.

Pregunta guía:

> ¿Qué variables socioeducativas y geográficas influyen en la desfavorabilidad de los establecimientos educativos en la Provincia de Buenos Aires, y cómo pueden predecirse sus niveles?

Antecedentes:

Históricamente, la DGCyE utilizó el Índice de Vulnerabilidad Socioeducativa (IVS), basado en datos censales, para determinar el grado de desfavorabilidad. Sin embargo, este indicador quedó obsoleto, ya que no contempla las condiciones físicas de las escuelas ni la matrícula actual.
La presente investigación busca actualizar este enfoque mediante un análisis de datos recientes (2022–2025), integrando información del censo de población, el NBI (Necesidades Básicas Insatisfechas) y los establecimientos educativos públicos y privados.

Fuentes oficiales:

- Índice de vulnerabilidad socioeducativa del establecimiento — DGCyE
- Metadatos de establecimientos educativos — Gobierno de la Provincia de Buenos Aires
- Datos NBI por municipio — Dirección Provincial de Estadística

2. Metodología

Fuentes y recolección:

Se utilizaron dos fuentes principales de datos abiertos:

1. Establecimientos Educativos (2025) — Dataset con más de 21.000 registros de escuelas bonaerenses, conteniendo información sobre ámbito, nivel, subvención, dependencia, matrícula, y nivel de desfavorabilidad.
   Fuente: Datos Abiertos PBA.

2. Población por Municipio (Censo 2022) — Dataset del INDEC con la población total, nativa y extranjera por cada municipio bonaerense.

3. La variable objetivo (target) del modelo fue **'desfavorabilidad'**, considerada como una variable categórica con valores entre 0 y 5.

Los datos fueron recolectados en formato CSV y leídos mediante `pandas` en Google Colab.

Procesamiento de datos

Se aplicaron las siguientes etapas de preparación:

- **Estandarización de texto:** conversión a minúsculas, eliminación de espacios y caracteres especiales.
- **Integración de datasets:** fusión (`merge`) entre el dataset de establecimientos (`df1`) y el de población (`df2`) utilizando la clave `municipio_nombre`.
- **Eliminación de columnas irrelevantes:** se conservaron únicamente las variables de interés para el análisis.

- **Codificación de variables categóricas:**
    - *One-Hot Encoding* para variables con menos de 5 categorías (`ambito`, `categoria`, `dependencia`).
    - *Binary Encoding* para variables con más de 5 categorías (`subvencion`, `region_educativa`, `depend_func`, `tipo_organizacion`).
- **Creación de variables binarias:**
    - `sector_bin` → 1 = privado, 0 = estatal
    - `area_bin` → 1 = conurbano, 0 = interior

Métodos de análisis

<u>El análisis se dividió en dos partes:</u>

1. Análisis Exploratorio de Datos (AED):
    - Distribución de la desfavorabilidad por municipio.
    - Relación entre subvención, área y desfavorabilidad.
    - Correlación entre la población municipal y el índice de desfavorabilidad.
2. Modelado de Machine Learning:

Para la etapa de modelado se utilizó un algoritmo de Random Forest Classifier, un método de *machine learning* supervisado basado en un conjunto de árboles de decisión entrenados sobre diferentes subconjuntos del dataset.
Este modelo fue elegido por su robustez ante ruido, su capacidad de manejar variables categóricas y numéricas, y su bajo riesgo de sobreajuste gracias al promedio de predicciones entre múltiples árboles.

El conjunto de datos fue dividido en un 80% para entrenamiento y un 20% para prueba, utilizando una semilla aleatoria (42) para asegurar la reproducibilidad de los resultados.

Para evaluar el rendimiento del modelo y garantizar que no dependiera del azar en la división de los datos, se aplicó una validación cruzada con K-Fold (k = 5).
En este método, el conjunto de entrenamiento se divide en cinco partes: en cada iteración, cuatro se utilizan para entrenar y una para validar, rotando sucesivamente. Esto permite evaluar la estabilidad del modelo en distintas particiones.

El modelo se configuró con los siguientes hiperparámetros:

- `n_estimators = 300` → número de árboles del bosque.
- `max_depth = 15` → profundidad máxima de cada árbol, seleccionada para lograr un equilibrio entre *underfitting* (modelo simple) y *overfitting* (modelo sobreajustado).
- `random_state = 42` → garantiza la reproducibilidad.

Se calcularon las siguientes métricas de rendimiento mediante validación cruzada:

- Accuracy (exactitud): mide el porcentaje de aciertos totales.
- Precision ponderada: mide cuántas de las predicciones positivas fueron correctas, considerando el peso de cada clase.

Los resultados mostraron valores de precisión y exactitud consistentes entre los pliegues, con un promedio estable, lo que indica que el modelo generaliza bien y no depende del azar de la partición de los datos.

3. Resultados

Presentación visual:

Se generaron diversos gráficos utilizando *Matplotlib* y *Seaborn* para el análisis exploratorio de datos (AED) y la interpretación de los resultados del modelo. Entre ellos:

- Promedio de desfavorabilidad por municipio.
  Permitió identificar diferencias geográficas en la distribución del índice.

- Boxplots de matrícula por sector (privado/estatal).
  Mostraron la presencia de *outliers* y mayor dispersión en el sector estatal.

- Gráficos de dispersión entre población y desfavorabilidad promedio.
  Indicaron que la cantidad de habitantes del municipio no se asocia directamente con la desfavorabilidad educativa.

- Gráfico de importancia de variables (feature importance).
  El modelo Random Forest determinó que las variables con mayor peso en la predicción del índice de desfavorabilidad fueron:
  latitud, longitud, ámbito urbano/rural, población total, matrícula total y dependencia oficial.

Hallazgos clave:

- Los municipios con mayor población no necesariamente presentan mayor desfavorabilidad.

- Las escuelas rurales y de gestión estatal tienden a concentrar los valores más altos del índice.

- Existe una correlación moderada entre la subvención y la desfavorabilidad, donde las instituciones sin aporte estatal presentan índices más bajos.

- Factores geográficos (latitud y longitud) y demográficos (población, matrícula) tienen una influencia significativa en el modelo predictivo.

Desempeño del modelo predictivo

Se utilizó un modelo de clasificación Random Forest, ajustado con `max_depth=15` y validado con una validación cruzada de 5 particiones (K-Fold).
El conjunto de datos se dividió en 80% entrenamiento y 20% prueba.

Los resultados promedio de las métricas fueron:

Métrica	Promedio
Accuracy	0.809
Precision	0.788
Recall	0.809
F1-Score	0.792

Estos valores reflejan un rendimiento sólido y equilibrado, donde el modelo logra identificar correctamente la mayoría de los casos, con una buena relación entre precisión (exactitud de las predicciones positivas) y *recall* (capacidad de detectar correctamente las categorías verdaderas).

Precisión: Las métricas se validaron mediante cross-validation (K-Fold), mostrando una desviación estándar menor a 0.02 entre los pliegues, lo cual indica una alta estabilidad y capacidad de generalización del modelo.

Esto sugiere que el Random Forest logra un equilibrio adecuado, evitando tanto el *overfitting* como el *underfitting*.

4. Discusión

Interpretación

Los resultados obtenidos muestran que la desfavorabilidad educativa no está determinada únicamente por el tamaño poblacional del municipio, sino que responde a un conjunto de factores institucionales, geográficos y sociales.
Las variables con mayor peso predictivo —latitud, longitud, ámbito urbano/rural, población total y dependencia estatal— revelan que la ubicación geográfica y el contexto territorial influyen directamente en las condiciones de acceso y recursos educativos.

Las escuelas ubicadas en zonas rurales dispersas o del interior presentan una mayor probabilidad de alcanzar índices de desfavorabilidad elevados (niveles 4 o 5). Asimismo, la dependencia estatal y los menores niveles de subvención privada se asocian con contextos más vulnerables, lo cual sugiere que las políticas de financiación y localización escolar impactan significativamente en la equidad educativa.

El modelo Random Forest confirmó que la georreferenciación (latitud y longitud) tiene una fuerte relación con la desfavorabilidad, lo que refuerza la idea de que la distribución territorial de los establecimientos educativos continúa siendo desigual, especialmente fuera del Conurbano bonaerense.

Comparación con antecedentes

El patrón identificado coincide parcialmente con los resultados históricos del Índice de Vulnerabilidad Socioeducativa (IVS) utilizado por la DGCyE, que también vinculaba la desfavorabilidad con las condiciones socioeconómicas y de infraestructura del entorno.
Sin embargo, el presente análisis aporta una actualización metodológica y temporal, integrando datos recientes (2022–2025) y variables adicionales como la población total, la matrícula y el tipo de subvención, lo que mejora la granularidad y precisión del diagnóstico.

En comparación con los estudios previos, este enfoque predictivo basado en *machine learning* permite anticipar zonas o establecimientos con alta probabilidad de vulnerabilidad educativa, ofreciendo un instrumento útil para la toma de decisiones en políticas públicas, asignación de recursos y planificación escolar.

5. Conclusión

Resumen de hallazgos

El análisis realizado permitió identificar los principales factores que inciden en la desfavorabilidad educativa de los establecimientos de la provincia de Buenos Aires, integrando información institucional, demográfica y geográfica.
A través del modelo predictivo de *machine learning* basado en Random Forest, se comprobó que variables como latitud, longitud, ámbito urbano/rural, población total, matrícula y dependencia estatal poseen la mayor influencia en la predicción del índice de desfavorabilidad.

El modelo alcanzó un rendimiento promedio del 80% de exactitud (accuracy) y valores consistentes de precisión, recall y F1-score cercanos al 0.79–0.81, demostrando estabilidad y capacidad de generalización mediante validación cruzada (*K-Fold*).

Estos resultados evidencian que el modelo puede predecir con solidez el nivel de vulnerabilidad educativa de una escuela a partir de características básicas, sin necesidad de información económica directa.

Recomendaciones

- Implementar este tipo de modelos predictivos para priorizar la asignación de recursos educativos, especialmente en zonas rurales y de baja subvención estatal.

- Integrar indicadores complementarios, como el porcentaje de hogares con Necesidades Básicas Insatisfechas (NBI) o la infraestructura escolar, para mejorar la precisión del modelo y reflejar de forma más completa la realidad socioeducativa.

- Utilizar la visualización geográfica de la desfavorabilidad (a partir de latitud y longitud) como herramienta de apoyo para planificación territorial y gestión educativa.

Limitaciones y futuro

Entre las limitaciones, se reconoce que la desfavorabilidad actual se mide con criterios parcialmente obsoletos (basados en el antiguo IVS) y que no considera aspectos cualitativos como el estado edilicio, conectividad digital o acceso a servicios básicos.

Como línea futura, se propone actualizar los criterios de clasificación incorporando variables más recientes y representativas, además de entrenar el modelo con nuevos censos educativos y socioeconómicos.

De esta forma, el presente trabajo sienta las bases para el desarrollo de sistemas predictivos que permitan anticipar escenarios de desigualdad educativa, optimizando la toma de decisiones a nivel provincial y promoviendo una educación más equitativa.

En síntesis, este trabajo demuestra que las técnicas de *machine learning* pueden aplicarse eficazmente al análisis de desigualdades sociales, proporcionando herramientas objetivas para la gestión educativa y la reducción de brechas territoriales.