Sistema de Análisis Territorial y Predicción de Riesgo de Abandono Escolar

Paola Michelle Figueroa

Benítez Lowenski Paredes

Rosario Carlos Damián Rodríguez Uitzil

Leonard Jose Cuenca Roa

Contenido

Resumen Ejecutivo. 2
Introducción y objetivos. 2
Objetivos 3
Descripción de los datos 4
Identificación empresa 5
Metodología. 5

Resumen Ejecutivo

Este proyecto aborda la deserción escolar como un fenómeno multifactorial mediante el desarrollo de un Sistema de Análisis Territorial y Predicción de Riesgo. Actualmente, la fragmentación de datos entre instituciones educativas y organismos socioeconómicos impide una visión integral del problema.

El objetivo principal es integrar fuentes de datos abiertas (SEP, INEGI, CONEVAL) para construir un modelo de Machine Learning capaz de clasificar zonas y estudiantes según su riesgo de abandono y visualizarlo en un dashboard conciso y claro que pueda ayudar en demostrar lo grabe de la situación. La metodología incluye la ingesta y limpieza de datos heterogéneos, un análisis exploratorio para identificar correlaciones entre pobreza y eficiencia terminal y el entrenamiento de algoritmos de clasificación supervisada. Como resultado final, se entregará un dashboard interactivo con capacidades geoespaciales que permitirá a los tomadores de decisiones visualizar "**zonas calientes**" y anticipar el riesgo educativo, transformando el enfoque de política pública de uno reactivo a uno preventivo basado en evidencia.

Palabras clave: Deserción escolar, Machine Learning, Análisis geoespacial, Datos abiertos, Políticas públicas.

Introducción y objetivos

La deserción escolar representa uno de los desafíos más críticos para el desarrollo social y económico de la región. Si bien es un problema ampliamente documentado, su análisis tradicionalmente se ha realizado mediante silos de información, donde las métricas educativas se evalúan desconectadas de la realidad territorial y socioeconómica de los estudiantes. Esta falta de integración limita la capacidad de las instituciones gubernamentales para actuar antes de que el estudiante abandone el sistema.

El presente proyecto propone una solución tecnológica basada en la Ciencia de Datos para unificar estas variables dispersas. Al correlacionar indicadores de desempeño académico con índices de marginación y demografía, es posible identificar patrones latentes que preceden al abandono. La propuesta trasciende el entendimiento del 'por qué' de la deserción, buscando anticipar **'dónde'** y **'cuándo'** es más probable que ocurra. El objetivo central es dotar a los organismos públicos, medios sociales, organismos humanitarios, de una herramienta estratégica para la toma de decisiones basada en evidencia. Asimismo, se busca visibilizar esta problemática mediante datos, contribuyendo a la mejora de las políticas educativas y ofreciendo un recurso tangible para la acción social.

Objetivos

Objetivo General

Desarrollar un sistema de análisis de datos y predicción recopilando la información en un dashboard logrando mostrar los riesgos de abandono escolar, utilizando técnicas de Machine Learning y visualización geoespacial, permita identificar zonas prioritarias de intervención basándose en factores socioeconómicos y educativos.

Objetivos Específicos

1. **Integración de Datos (ETL):** Consolidar y normalizar datasets heterogéneos provenientes de fuentes oficiales (SEP, INEGI, CONEVAL) para crear un almacén de datos unificado apto para el análisis.
2. **Análisis Exploratorio (EDA):** Identificar correlaciones estadísticas significativas entre las variables de desigualdad social (pobreza, rezago) y los indicadores de desempeño educativo (tasas de abandono, eficiencia terminal).
3. **Modelado Predictivo:** Entrenar y validar un modelo de clasificación supervisada Random Forest, Regresión Logística que permita calcular la probabilidad de riesgo académico a nivel territorial.
4. **Visualización para la Toma de Decisiones:** Diseñar e implementar un dashboard interactivo con mapas de calor y filtros dinámicos que facilite la

interpretación de los hallazgos para usuarios no técnicos y gestores de políticas públicas.

Descripción de los datos

El proyecto se fundamenta en la explotación de **Fuentes de Datos Abiertas (Open Data)** gubernamentales, garantizando la transparencia y replicabilidad del estudio. Se estructurarán tres dimensiones principales de información:

1. **Datos Educativos (Fuente: SEP):**
   - **Variables:** Tasa de abandono intra-curricular, eficiencia terminal, tasa de reprobación y matrícula total.
   - **Granularidad:** Se trabajará a nivel municipal y estatal, dependiendo de la disponibilidad de los microdatos.
   - **Uso:** Estas variables funcionarán como la "variable objetivo" (target) y métricas de desempeño a predecir o explicar.
2. **Datos Demográficos y de Vivienda (Fuente: INEGI):**
   - **Variables:** Población total por rangos de edad, nivel de escolaridad promedio de los padres, acceso a servicios básicos (internet, electricidad) y densidad poblacional.
   - **Uso:** Variables explicativas (features) que aportan el contexto del entorno inmediato del estudiante.
3. **Datos de Pobreza y Rezago Social (Fuente: CONEVAL):**
   - **Variables:** Índice de Rezago Social (IRS), porcentaje de población en situación de pobreza y pobreza extrema, carencias sociales.
   - **Uso:** Variables explicativas críticas para ponderar el impacto de la desigualdad económica en el fenómeno educativo.

Identificación empresa

- **Tipo de Proyecto**: Académico / Investigación Aplicada al Sector Público.
- **Ámbito de Aplicación**: Instituciones gubernamentales encargadas de la educación y el desarrollo social.
- **Naturaleza**: Desarrollo de producto de datos para el bien social.

Metodología

Para el desarrollo del proyecto se utilizará una adaptación del marco de trabajo **CRISP-DM** (Cross-Industry Standard Process for Data Mining) o un ciclo de vida estándar de Ciencia de Datos, estructurado en las siguientes fases:

1. **Entendimiento de los Datos y Adquisición:**
    - Recolección de datasets desde los portales oficiales.
    - Diccionario de datos y validación de la calidad inicial (detección de nulos, formatos inconsistentes).
2. **Preprocesamiento y Limpieza (Data Wrangling):**
    - Normalización de claves geoestadísticas para el cruce de tablas.
    - Imputación de valores faltantes y tratamiento de outliers.
    - Ingeniería de características (Feature Engineering): Creación de nuevas variables (ej. ratio alumnos/población).
3. **Análisis Exploratorio de Datos (EDA):**
    - Análisis univariante para entender la distribución de cada variable.
    - Análisis bivariante/multivariante (Matrices de correlación) para validar hipótesis sobre la relación entre pobreza y deserción.
4. **Modelado (Machine Learning):**
    - División del dataset en conjuntos de entrenamiento y prueba (Train/Test split).

- Selección de algoritmos: Se probarán modelos de clasificación (como Decision Trees o Random Forest) para categorizar el nivel de riesgo (Alto/Medio/Bajo).
- Ajuste de hiperparámetros y validación cruzada.

5. **Evaluación:**
   - Medición del desempeño del modelo utilizando métricas como *Accuracy, Precision, Recall* y *F1-Score*, priorizando la capacidad del modelo para detectar correctamente los casos de alto riesgo.
6. **Visualización y Comunicación:**
   - Desarrollo del dashboard en PowerBI o Tableau.
   - Implementación de mapas coropléticos utilizando GeoPandas para la representación espacial del riesgo