README - SHEDATA

**📌 Descripción del Reto**

Este proyecto corresponde a una propuesta de análisis de datos educativos en el contexto colombiano. Utiliza un enfoque de ciencia de datos para explorar si existen diferencias en la proporción de escuelas con atención a estudiantes con discapacidad según la ubicación geográfica de los establecimientos educativos (zonas rurales vs. urbanas).

**🔍 Hipótesis**

**Problema de investigación**

¿Las escuelas ubicadas en zonas rurales de la Región Caribe colombiana presentan una menor proporción de atención a estudiantes con discapacidad en comparación con aquellas ubicadas en zonas urbanas o mixtas?

**Hipótesis de investigación**

En la Región Caribe colombiana, se presentó una baja proporción de escuelas ubicadas en zonas rurales que atienden a estudiantes con discapacidad en comparación con aquellas ubicadas en zonas urbanas o mixtas.

**Hipótesis del modelo**

**Hipótesis nula (H₀):** La proporción de escuelas con atención a estudiantes con discapacidad es igual en zonas rurales y urbanas.

**Hipótesis alternativa (H₁):** La proporción de escuelas con atención a estudiantes con discapacidad es menor en zonas rurales que en urbanas.

**🎯 Enfoque**

Para abordar el problema planteado, se utilizó un enfoque de ciencia de datos dividido en tres etapas principales:

1. **Limpieza de datos:** Se trabajó con un conjunto de datos de establecimientos educativos de Colombia. Se hizo una revisión de valores nulos, los cuales fueron reemplazados por ceros o limpiados según su naturaleza (por ejemplo, valores como “No Aplica” fueron homologados). También se crearon nuevas variables binarias para representar la presencia o ausencia de atención a discapacidad, niveles educativos y tipos de discapacidad.
2. **Análisis exploratorio (EDA):** Se analizaron las frecuencias y proporciones de atención a estudiantes con discapacidad en función de la ubicación geográfica (zona y departamento). Se utilizaron visualizaciones básicas como gráficos de barras para ilustrar comparaciones entre zonas rurales y urbanas, y entre departamentos.
3. **Modelado predictivo:** Se implementaron modelos de clasificación como el Árbol de Decisión para evaluar si la inclusión en las escuelas de los diferentes tipos de discapacidades puede predecir si esta está ubicada en una zona rural o urbana. También se aplicaron métricas de desempeño como accuracy y matriz de confusión para evaluar los modelos.

Este enfoque permitió identificar patrones clave y contrastar empíricamente la hipótesis planteada, brindando insumos relevantes tanto para análisis descriptivo como predictivo.

**👥 Roles del Equipo**

El trabajo fue realizado de manera colaborativa, con una distribución flexible de tareas. A continuación, se describen las principales responsabilidades asumidas por cada integrante:

* **Ivana Machuca.** Representante del equipo. Encargada de las reuniones de seguimiento con el profesor, elaboración de dashboards en Power BI, diseño y redacción del discurso final, diseño de las diapositivas y presentación del proyecto.
* **Carolina Molina.** Encargada del análisis exploratorio de datos y de la elaboración de los gráficos exploratorios.
* **Elsa Marín.** Encargada del diseño visual del equipo (colores, logo, backgrounds), del diseño de las diapositivas y del discurso final.
* **Helen Cervera.** Encargada de la limpieza de datos, análisis exploratorio, actualización de los datos en Power BI y de la creación del README.
* **Yurle Araujo.** Encargada de la limpieza de datos, del desarrollo del modelo predictivo y de la generación de gráficas predictivas.

**📝 Pasos Realizados**

1. **Diseño de identidad visual del equipo:** se definió el nombre del grupo, paleta de colores, logo y estilo visual para mantener coherencia en las piezas gráficas y presentaciones.
2. **Organización de tareas:** se definieron las diferentes tareas a realizar en un tablero de Trello, al que todas las integrantes tenían acceso. En este, se iba actualizando diariamente el estado de las tareas a medida que se iba trabajando. Enlace: <https://trello.com/invite/b/683f9f5dfca3a519c60413cd/ATTIca0f20506930c28eab8970aa91a36db19318A35F/shedata-datathon-trabajo>
3. **Definición del problema:** se formuló la pregunta de investigación sobre la diferencia en la proporción de escuelas que atienden a estudiantes con discapacidad según su ubicación geográfica (zona rural, urbana o mixta) en la Región Caribe colombiana.
4. **Recolección de datos:** se trabajó con la base de datos oficial del Ministerio de Educación Nacional de Colombia, correspondiente al año 2022.
5. **Limpieza de datos:** se trataron los valores nulos, se recodificaron y unificaron categorías de variables clave como discapacidad, zona y niveles, y se organizaron los datos para el análisis.
6. **Análisis exploratorio (EDA):** se aplicaron conteos, proporciones y visualizaciones para identificar patrones por zona, nivel educativo y departamento.
7. **Modelado predictivo:** se implementaron modelos de clasificación como el Árbol de Decisión para evaluar si la atención a estudiantes con diferentes tipos de discapacidad permiten predecir si una institución está ubicada en zona rural o urbana. Se evaluaron los modelos con métricas como accuracy y matriz de confusión.
8. **Visualización de resultados:** se crearon gráficos en Python y dashboards interactivos en Power BI para facilitar la comprensión de los resultados y su comunicación.
9. **Documentación y presentación:** se diseñaron las diapositivas de presentación en Google Slides, se redactó el discurso de exposición y se organizó la entrega final del proyecto, incluyendo la documentación en el archivo README.

**🚀 Resultados obtenidos**

Durante el desarrollo del proyecto se obtuvieron los siguientes productos:

* **Dataset limpio y transformado:** base de datos procesada con nuevas variables y lista para análisis, generada a partir del conjunto original del Ministerio de Educación.
* Análisis exploratorio con visualizaciones: gráficos básicos y cruzados (por zona, departamento y atención a discapacidad) que facilitaron la comprensión del problema.
* **Modelo predictivo (DecisionTree):** se entrenó un modelo de clasificación para predecir la ubicación geográfica de las escuelas (zona rural o urbana) a partir de variables relacionadas con la atención a estudiantes con discapacidad.
* **Dashboard interactivo en Power BI:** herramienta visual para explorar la atención a estudiantes con discapacidad por zona, departamento, nivel educativo, y otros filtros.
* **Presentación final en Google Slides:** diapositivas que resumen el planteamiento, el proceso y los hallazgos más importantes del proyecto, usadas para la sustentación final.
* **README Documentado:** archivo que sistematiza el trabajo realizado, incluyendo la descripción del reto, hipótesis, metodología, pasos ejecutados y productos entregables.

**📂 Estructura del Repositorio**

* DATATHON-SHEDATA.pbix # Dashboard interactivo en Power BI
* Datathon-SheData.ipynb # Jupyter Notebook con la limpieza, análisis descriptivo, modelo predictivo y gráficos asociados.
* ESTABLECIMIENTOS\_EDUCATIVOS\_COLOMBIA\_20250603 # Base de datos cruda
* README.md # Este archivo
* SheData\_Logo # Archivo del logo del grupo
* SheData\_TedTalk # Archivo con presentación final

**🛠 Herramientas Utilizadas**

* Python (Pandas, Numpy, Sklearn, Matplotlib, Seaborn)
* Trello
* Jupyter Notebook
* GitHub
* Power BI

*Otras:*

* Google Slides
* PhotoShop
* ChatGPT
* Gemini