INFORME

Este informe esta dividido en las siguientes partes:

- [PARTE 1] Introducción y objetivos del proyecto
- [PARTE 2] Origen y descripción de las bases de datos
- [PARTE 3] Proceso de limpieza y transformación de datos en Python
- [PARTE 4] Integración de datos en Power BI y diseño del modelo
- [PARTE 5] Análisis de datos y principales insights
- [PARTE 6] Conclusiones y posibles líneas futuras de investigación

PARTE 1: Introducción y objetivos del proyecto

Este informe detalla el desarrollo de un proyecto de análisis de datos realizado con información oficial del sistema educativo ecuatoriano, específicamente relacionada con las matrículas estudiantiles desde el año 2009 hasta el año 2023. El objetivo principal de este proyecto fue identificar patrones, tendencias, problemáticas y puntos críticos en la evolución de la matrícula estudiantil en el Ecuador, así como comprender la dinámica del abandono escolar, la influencia de la nacionalidad de los estudiantes y otros factores clave del sistema.

Este proyecto se desarrolló de forma completamente autónoma, como parte de una iniciativa personal para aplicar y fortalecer habilidades en análisis de datos, limpieza de información, visualización y generación de insights de valor. Se utilizaron herramientas como Python (para la preparación y limpieza de los datos), Power BI (para el análisis visual e interpretación de los resultados), y Jupyter Notebooks dentro de Visual Studio Code como entorno de desarrollo.

Entre los objetivos específicos del proyecto destacan:

- Consolidar y limpiar dos fuentes de datos con más de 500,000 registros para construir una base confiable.
- Detectar y corregir inconsistencias en nombres de instituciones, estructuras de columnas y valores atípicos.
- Analizar el comportamiento de las matrículas iniciales y finales a lo largo de los años.
- Evaluar indicadores clave como la tasa de deserción escolar, el crecimiento de la población extranjera estudiantil y la distribución geográfica del abandono escolar.
- Presentar estos hallazgos mediante visualizaciones interactivas que permitan una comprensión clara por parte de cualquier usuario o tomador de decisiones.

Este informe busca no solo documentar el proceso técnico llevado a cabo, sino también resaltar la importancia del análisis de datos en la comprensión de fenómenos sociales complejos, como la educación, desde una perspectiva basada en evidencia.

PARTE 2: Origen y descripción de las bases de datos

El proyecto se fundamenta en dos bases de datos públicas extraídas del portal oficial de datos abiertos del gobierno del Ecuador: datosabiertos.gob.ec. Estas bases de datos, correspondientes a los registros de inicio y fin de matrícula escolar, abarcan un extenso período entre los años 2009 y 2023,

proporcionando una visión longitudinal sobre el comportamiento del sistema educativo ecuatoriano durante más de una década.

1. Base de datos de Inicio de Matrícula (2009–2023)

Esta primera base contiene registros de la cantidad de estudiantes matriculados al inicio de cada período lectivo, organizados por provincia, cantón e institución educativa. Los principales campos que contiene esta base son:

- Provincia
- Cantón
- Institución Educativa
- Código Único de la Institución
- Total de Estudiantes Matriculados
- Año Lectivo
- Regímenes y características adicionales (como acceso a infraestructura)
- Códigos administrativos (provincia, cantón, parroquia)

Esta base sirvió como punto de partida para entender la distribución geográfica y temporal de los estudiantes en el sistema educativo ecuatoriano al comienzo de cada año escolar.

2. Base de datos de Fin de Matrícula (2009–2023)

La segunda base de datos se enfocó en los resultados al final del mismo periodo escolar. Esta base incluía los mismos campos básicos que la anterior, pero también incorporaba información mucho más detallada y valiosa para análisis educativos, tales como:

- Número de estudiantes que finalizaron el año escolar
- Estudiantes promovidos (aprobados)
- Estudiantes reprobados
- Estudiantes que abandonaron sus estudios
- Distribución por nacionalidad (ecuatorianos, venezolanos, colombianos, peruanos, otros extranjeros)
- Total de docentes por institución
- Tipo de institución (Fiscal, Particular, Municipal, Fiscomisional)
- Régimen (Costa o Sierra)

Gracias a esta base, fue posible analizar la efectividad del sistema educativo, los niveles de retención o abandono, y cómo varían estos según factores geográficos, institucionales y sociales.

3. Formato y volumen de los datos

Ambas bases fueron descargadas en formato CSV y juntas sumaban aproximadamente 500,000 registros. Esto implicó la necesidad de aplicar técnicas eficientes de limpieza y transformación de datos para garantizar su manejabilidad y calidad. La magnitud y la variedad de campos permitieron múltiples posibilidades de análisis, desde tendencias temporales hasta segmentaciones por regiones o tipos de instituciones.

PARTE 3: Limpieza y transformación de datos en Python

Una vez descargadas las dos bases de datos en formato CSV, se procedió a su tratamiento y limpieza utilizando Python, a través de un entorno Jupyter Notebook alojado en Visual Studio Code. El enfoque se centró en preparar los datos para que sean consistentes, comparables y analíticamente útiles, minimizando errores comunes como duplicados, formatos inconsistentes y valores atípicos.

1. Carga de datos y bibliotecas utilizadas

Se utilizaron las siguientes bibliotecas principales para trabajar con los datos:

- pandas: para manipulación y análisis de datos.
- datetime: para trabajar con formatos de fechas.
- matplotlib: aunque fue importada, no se llegó a utilizar en esta etapa.

Ambos archivos CSV fueron cargados en DataFrames separados y se hizo una primera inspección para identificar columnas relevantes, inconsistencias, valores nulos y errores comunes.

2. Limpieza de columnas irrelevantes

En ambas bases se eliminaron columnas no necesarias para el análisis, tales como:

Códigos de provincia, cantón y parroquia (ya que los nombres eran más representativos).

Información de infraestructura como “acceso a edificios” u otros metadatos administrativos que no aportaban al análisis de matrícula.

Esto ayudó a reducir el tamaño de los DataFrames y facilitar su manejo.

3. Estandarización del campo “Año Lectivo”

En ambas bases de datos, el campo Año Lectivo venía en formato “2009-2010”. Para facilitar su análisis por separado, se creó una transformación que dividía ese string en dos columnas nuevas:

- Inicio_Periodo: usando funciones como split() o slice() se extrajo el primer año del rango.
- Fin_Periodo: se extrajo el segundo año.

Estas dos nuevas columnas facilitaron las comparaciones año a año y la alineación con otras variables.

4. Estandarización de nombres de instituciones

Uno de los desafíos más importantes fue la inconsistencia en los nombres de las instituciones. Muchas instituciones compartían el mismo código único, pero aparecían con nombres ligeramente distintos, probablemente por errores humanos, diferencias de escritura o falta de normalización.

Para resolver este problema:

- Se aplicó un conteo de combinaciones entre Código Único de Institución y Nombre de Institución utilizando value_counts().
- Se generó un nuevo DataFrame llamado MAPS_Frecuente, que retenía solo el nombre más repetido por cada código único (asumiendo que este es el nombre correcto).

- Posteriormente, se hizo un left join entre la base original y MAPS_Frecuente, para reemplazar los nombres inconsistentes por los estandarizados.

Este proceso redujo drásticamente la cantidad de combinaciones únicas entre código y nombre de institución, pasando de aproximadamente 53,000 a 29,000, mejorando así la calidad del análisis posterior.

5. Eliminación de valores atípicos

Para evitar que instituciones con números exageradamente altos de estudiantes distorsionaran los análisis, se implementó una función personalizada que:

- Recibía un DataFrame y una columna objetivo.
- Calculaba el percentil 0.98 (98%) de esa columna.
- Filtraba los registros cuyo valor excedía ese límite.

Esta función fue aplicada principalmente sobre la columna Total de Estudiantes Matriculados, permitiendo mantener un análisis más realista sin que valores extremos afectaran las visualizaciones o las métricas.

6. Preparación para Power BI

Una vez finalizado el proceso de limpieza:

- Se generó un nuevo DataFrame consolidado con todos los datos limpios.
- Este DataFrame fue cargado directamente en Power BI a través de un script de Python, utilizando la opción de conectarse a una fuente externa por script.
- Esta estrategia alivió la carga del Power Query en Power BI, ya que gran parte de la transformación ya había sido ejecutada en Python.

Cualquier transformación adicional sería realizada posteriormente dentro de Power BI, solo en caso de ser necesaria.

PARTE 4: Integración y modelado de datos en Power BI

Una vez completado el proceso de limpieza y transformación en Python, se procedió a cargar los datos a Power BI, empleando tanto scripts de Python como funciones propias del entorno de Power Query para terminar de preparar los datos para el análisis visual e interactivo.

1. Carga de datos en Power BI desde Python

Power BI permite conectarse directamente a scripts de Python. Aprovechando esta funcionalidad:

- Se copió el mismo script utilizado en Jupyter Notebook y se pegó dentro del entorno de Power BI.
- Esto permitió cargar el DataFrame ya limpio, lo cual redujo considerablemente la carga de trabajo en Power Query.

Este enfoque fue eficiente porque delegó las tareas pesadas de limpieza a Python, haciendo que Power BI se enfocara principalmente en la visualización y en cálculos DAX.

2. Transformaciones adicionales en Power Query

Ya dentro de Power Query, se realizaron algunas transformaciones complementarias necesarias para el modelado correcto:

- Formateo de campos de año: Se aseguraron que las columnas Inicio_Periodo y Fin_Periodo tuvieran un formato numérico o de fecha según el caso, lo que facilitaría su análisis en línea de tiempo.
- Unificación de columnas clave: Se creó una nueva columna en ambas bases (inicio y fin de matrícula) llamada Periodo_Institucion, combinando el año lectivo y el código único de la institución (ejemplo: “2019-2023_1032847”).

Esta columna sirvió como clave de unión entre ambas bases de datos, para garantizar que cada institución en cada año escolar tuviera un único registro.

3. Modelado relacional y combinación de bases

Debido a que se contaban con dos bases de datos distintas pero relacionadas (Inicio de Matrícula y Fin de Matrícula), fue fundamental unirlas correctamente para poder realizar comparaciones año a año entre matrícula, abandono, promoción, entre otros.

- Se utilizó un Inner Join entre las dos bases, usando como columna principal la recientemente creada Periodo_Institucion.
- Esto permitió mantener solo aquellas instituciones que tenían datos tanto de inicio como de fin de matrícula para un mismo año lectivo.

Esta decisión garantizó la consistencia de las métricas y que no se generaran análisis parciales o desalineados.

4. Creación de tabla de medidas (DAX)

Para organizar las métricas de manera clara y accesible, se creó una tabla específica llamada "Medidas" que agrupó todos los cálculos realizados con DAX. Algunas de las medidas clave desarrolladas fueron:

Totales generales:

- Total de estudiantes matriculados
- Total de estudiantes que abandonaron
- Total de estudiantes promovidos
- Total de docentes
- Total de estudiantes extranjeros

Indicadores de rendimiento y abandono:

- Tasa de abandono escolar por año
- Tasa de promoción nacional
- Ratio estudiantes/docentes
- Porcentaje de crecimiento en matrícula

- Porcentaje de crecimiento de estudiantes promovidos
- Porcentaje de crecimiento de abandono
- Porcentaje de crecimiento de estudiantes extranjeros

Insights específicos:

- Año con mayor tasa de abandono
- Provincia con mayor abandono
- Institución con mayor tasa de deserción escolar

Estas medidas permitieron evaluar de forma dinámica los datos a lo largo de los años y entre diferentes regiones e instituciones.

PARTE 5: Visualización y descubrimiento de insights (Dashboards e Interpretación)

Con el modelo de datos ya limpio y estructurado en Power BI, se procedió al desarrollo de dashboards interactivos que permitieran explorar los patrones de matrícula, deserción escolar y otros indicadores educativos clave en Ecuador desde el año 2009 hasta 2023.

Diseño y estructura del dashboard

El tablero fue organizado de manera clara y funcional, con secciones que permitieran visualizar tanto información general como análisis específicos por provincia, cantón, tipo de institución y régimen escolar.

Principales elementos visuales utilizados:

- Gráficos de líneas y áreas para mostrar la evolución de estudiantes, abandono y otros indicadores a lo largo del tiempo.
- Gráficos de barras para comparaciones entre provincias, cantones o instituciones.
- Tarjetas y KPIs para destacar métricas importantes como la tasa de deserción, número total de estudiantes extranjeros, docentes, etc.
- Segmentadores (slicers) por año, provincia, cantón, tipo de institución, régimen y nacionalidad, para personalizar el análisis dinámicamente.

Principales insights obtenidos del análisis

Durante el análisis exploratorio y la interacción con el dashboard, se identificaron hallazgos clave con implicaciones sociales, educativas y políticas. A continuación se presentan los más relevantes:

1. Tendencia decreciente en la matrícula estudiantil

- La cantidad de estudiantes matriculados en el sistema educativo ecuatoriano ha venido disminuyendo progresivamente desde 2009.
- El periodo 2022-2023 fue particularmente crítico, con un colapso drástico en la matrícula: solo 16.000 estudiantes registrados, lo que representa una reducción del 99.04% respecto al periodo anterior.
- Este comportamiento atípico puede reflejar fallos en la recolección de datos, efectos posteriores a la pandemia o cambios estructurales en el sistema.

2. Comportamiento de la población estudiantil extranjera

- A lo largo de los años se ha observado un aumento constante en la presencia de estudiantes extranjeros, con un pico en 2020, antes del impacto de la pandemia.
- Quito es la ciudad que más concentra estudiantes extranjeros, superando ampliamente al resto del país.

3. Distribución del sistema educativo por tipo de institución

- Los colegios fiscales representan más del 50% del sistema educativo nacional, siendo el pilar principal de la educación pública.
- En contraste, las instituciones particulares demostraron tener una menor tasa de deserción escolar (apenas 2.35%) comparada con:
- Instituciones municipales y fiscomisionales, ambas cercanas al 4%, es decir, casi el doble.

4. Concentración de estudiantes por región

- Guayaquil y Quito son consistentemente las ciudades con mayor número de estudiantes a lo largo de todos los años analizados.
- Estas ciudades también concentran buena parte de los recursos docentes y presentan desafíos particulares en cuanto a retención escolar.

5. Provincias con mayores niveles de abandono escolar

Las provincias con mayor volumen de abandono fueron:

- Guayas, Manabí y Pichincha.

A nivel de cantones, los cinco con mayores cifras de abandono fueron:

- Guayaquil, Quito, Santo Domingo de los Tsáchilas, Cuenca y Portoviejo.

6. Evolución de la tasa de deserción escolar

En general, la tasa de deserción escolar se mantuvo por debajo del 5% entre 2009 y 2023.

Excepciones destacadas:

En 2011, se registró el pico más alto con una tasa de 5.15%.

Incluso el crítico año 2022, con la caída abrupta en matrícula, presentó una tasa de deserción relativamente más baja (3.63%), lo cual podría deberse a la baja cantidad total de estudiantes registrados ese año.

7. Relación entre número de docentes y tasa de abandono

Aunque no se identificó una correlación directa fuerte entre el número de docentes por institución y la tasa de abandono...

Se observó que en cantones con menos de 800 docentes, la tasa de deserción se mantiene cercana al 2% y con poca variación, lo cual podría sugerir una posible relación entre menor cantidad de docentes y estabilidad en retención estudiantil.

8. Comportamiento post pandemia

A pesar del colapso en el número de estudiantes en 2022, la tasa de deserción no se disparó, indicando que posiblemente el problema se concentró en la inscripción o registro, más que en la permanencia.

Para 2023, la tendencia parece haberse normalizado, retomando valores de crecimiento y retención.

9. Régimen escolar: Costa vs Sierra

No se encontraron diferencias significativas en términos de deserción escolar entre los dos regímenes principales del país (Costa y Sierra).

Esto sugiere que los factores que afectan el abandono escolar podrían estar más relacionados con aspectos estructurales, económicos o institucionales que con el régimen académico.

10. Casos extremos

El periodo 2011-2012 fue el que presentó la mayor tasa de abandono escolar, destacando la provincia de Morona Santiago con un 9.1%, donde cerca de 4.000 estudiantes abandonaron sus estudios.

La institución con la mayor tasa de deserción fue José María Velaz, con un 17%, el doble que la segunda institución más alta, a pesar de tener una base de estudiantes relativamente baja (5.000 alumnos).

PARTE 6: Conclusiones, recomendaciones y aprendizajes personales

1. Conclusiones generales del análisis

Tras realizar un análisis exhaustivo de los datos del sistema educativo ecuatoriano entre 2009 y 2023, se concluye lo siguiente:

- Desaceleración del sistema educativo: Existe una clara disminución en la cantidad de estudiantes matriculados en los últimos años, culminando en un desplome inédito en el periodo 2022-2023, lo cual podría tener múltiples causas: errores en la recolección de datos, efectos pospandemia, falta de políticas efectivas de retención escolar o abandono del sistema por migración o desmotivación estudiantil.
- Persistencia de la deserción escolar en ciertas regiones: Aunque la tasa nacional se ha mantenido relativamente baja, ciertas provincias y cantones concentran niveles más altos de abandono. Esto apunta a problemas regionales estructurales que requieren atención focalizada.
- Diferencias marcadas por tipo de institución: Las instituciones particulares ofrecen mayor retención estudiantil, posiblemente debido a factores como calidad de enseñanza, infraestructura o acompañamiento familiar. Por otro lado, las instituciones municipales y fiscomisionales presentan tasas de deserción significativamente mayores.
- Aumento en la diversidad estudiantil: El incremento en el número de estudiantes extranjeros, particularmente en años previos a la pandemia, refleja cambios demográficos y desafíos nuevos para el sistema educativo en cuanto a inclusión y adaptación cultural.

- Desigualdad en recursos docentes: La distribución desigual de docentes entre cantones podría estar afectando indirectamente la retención escolar. Aunque no se estableció una correlación fuerte, sí se identificaron tendencias que sugieren esta posibilidad.

2. Recomendaciones

A partir del análisis, se proponen las siguientes recomendaciones orientadas a mejorar la calidad y cobertura educativa en Ecuador:

- Fortalecer los mecanismos de recopilación y validación de datos: El desplome abrupto en la matrícula del periodo 2022-2023 sugiere posibles fallos en el sistema de reporte o recopilación de datos. Es vital implementar auditorías de datos anuales y sistemas automáticos de control de calidad.
- Desarrollar políticas específicas de intervención regional: Provincias como Guayas, Manabí y Pichincha requieren programas personalizados que atiendan las causas del abandono escolar, desde problemas económicos hasta falencias institucionales.
- Invertir en instituciones fiscales y fiscomisionales: Dado que estas instituciones concentran la mayoría del alumnado y presentan mayores tasas de deserción, deben ser prioridad en cuanto a infraestructura, formación docente y recursos pedagógicos.
- Atender la inclusión de estudiantes extranjeros: Se necesitan políticas de integración cultural y apoyo psicosocial para garantizar que los estudiantes migrantes permanezcan y se desarrollen en el sistema educativo.
- Equilibrar la distribución de docentes: Incentivar a los docentes a trasladarse a cantones con baja cobertura y garantizar condiciones laborales atractivas podría mejorar la calidad educativa y reducir la deserción.