RIQUEZA BAJO LA TIERRA

23 de noviembre de 2025

Link de repositorio de github: LansFacha/Miner-a-ciencia-de-datos-

Alumno: Ian de Jesús Bonilla Diaz

Profesor: Jaime Romero Sierra

BENEMERITA UNIVERSIDAD AUTONOMA DE PUEBLA

Ciencia de datos

Objetivo del Proyecto.

El objetivo de este esfuerzo es estudiar los patrones de minería y distribución en México, utilizando ciencia de datos, para la determinación del patrón de producción, el aumento y disminución, la evolución de los recursos de extracción y los factores que influyen en el sistema minero.

1. También se creó un modelo de pronóstico para predecir el crecimiento de la producción en el período subsiguiente para apoyar la toma de decisiones estratégicas basadas en el análisis de costo-beneficio, el uso óptimo de los recursos y para una operación más efectiva y económicamente sostenible con menor inversión.

Justificación.

La minería en México es una de las industrias que tiene especial importancia en términos de la significancia de sus consecuencias económicas, sociales y ecológicas. Sin embargo, no incluye una consideración metódica de los detalles relevantes del volumen de producción, tiempo y lugar, o la relación entre tipos de minerales. La falta de un análisis significa que también falta el análisis de datos, y es por eso que las empresas, corporaciones, sindicatos y organizaciones públicas tienen que tomar decisiones sin ellos.

Este proyecto es significativo porque:

1. Permite la detección de tendencias y patrones. Se analizan volúmenes, productos, estados y municipios para encontrar tendencias y patrones estacionales que permitan al planificador estratégico planificar con anticipación.

2. Mejora la toma de decisiones. Ofrece claridad sobre:

regiones más productivas,

minerales más extraídos,

fluctuaciones temporales.

3. Aumenta la transparencia. El proyecto destila datos en bruto en datos visuales claros a través de gráficos, métricas y análisis interpretables.

4. Minimiza los riesgos operativos y financieros. Identifica anomalías, cambios masivos y valores atípicos pertinentes.

5. Impulsa la innovación. Fomenta el uso de Python, análisis exploratorio, visualización avanzada y modelos predictivos, en línea con las tendencias de digitalización internacional.

Fuentes de Datos. Se utilizó una base de datos oficial de producción minera en México por año, mes, estado, municipio, clasificación de minerales y volumen para analizar la minería.

Origen. Registros oficiales de producción minera a nivel estatal y municipal.

Tamaño del conjunto de datos. 89,591 registros. 10 columnas.

Características. Variables numéricas o categóricas, con explicación de:

Año y mes.

Estado y municipio.

Producto y grupo de productos.

Unidad de medida.

Cobertura y estado.

Volumen de producción.

4. Proceso de Limpieza de Datos. Se aseguró la calidad, consistencia y utilidad analítica de los datos mediante la limpieza.

4.1 Valores faltantes. No encontramos valores numéricos nulos. En columnas categóricas, se imputó la moda cuando fue apropiado.

4.2 Tipos de datos. AÑO y VOLUMEN → numérico. ESTADO, PRODUCTO, GRUPO_PRODUCTO, COBERTURA, ESTADO → categórico. MES → entero (1-12).

4.3 Eliminación de duplicados. Se identificaron y eliminaron registros duplicados.

4.4 Valores atípicos. Se descubrieron más de 14,000 valores atípicos, principalmente en minerales de alto valor. No se eliminaron ya que no son valores atípicos y son representativos de picos reales de producción.

5. Análisis Exploratorio de Datos (EDA).

5.1 Descripción general. Conjunto de datos: 89,591 registros, 10 variables. Variables numéricas: AÑO, VOLUMEN. MES, ESTADO, PRODUCTO, GRUPO_PRODUCTO, COBERTURA, ESTADO – variables categóricas. El resumen estadístico confirma una fuerte asimetría en el volumen.

5.2 Distribución de variables.

5.2.1 Numérica (VOLUMEN). Histograma: fuerte asimetría a la derecha. Diagrama de caja: abundantes valores atípicos típicos de picos de producción.

5.2.2 Categórica. Estado: Zacatecas, Chihuahua, Sonora y Durango producen la mayor parte de la producción. PRODUCTO: Los más comunes son Plata, Oro, Plomo y Cobre.

5.3 Correlaciones. La correlación entre AÑO y VOLUMEN es en casi todos los casos nula (0.03). Las relaciones entre grupo de productos y volumen solo se aprecian cuando las variables están codificadas. Se utilizaron mapas de calor y diagramas de dispersión.

5.4 Valores atípicos. Detectados con el método IQR. Reflejan producción real → retenerlos.

5.5 Valores faltantes. Tenga en cuenta que son muy mínimos y solo en categóricos → imputados con la moda.

5.6 Relación categórica - numérica. Los diagramas de caja revelan extensas fluctuaciones voluminosas por estado y producto. Los productos preciosos tienden a tener mayor dispersión.

5.7 Hallazgos clave. Buen conjunto de datos grande sin valores faltantes significativos. Los valores atípicos se explican por picos de producción. Dominio de algunos minerales y estados. Grandes diferencias por temporada (mes). La variable objetivo para ML: ¿El volumen va a aumentar, en relación con el período anterior?

Modelo de Aprendizaje Automático.

6.1 Modelo utilizado. Clasificador de Bosque Aleatorio. Tipo: clasificación supervisada. Objetivo: pronosticar si el volumen aumentará en el próximo registro (0/1).

6.2 Justificación. Fuerte contra el ruido y los valores atípicos. Maneja grandes cantidades de datos con OneHotEncoding; puede trabajar con variables categóricas. Buena precisión sin normalización compleja.

6.3 Entrenamiento. División 80/20 usando train_test_split. OneHotEncoding para las columnas categóricas. Entrenamiento de más de 60 variables codificadas en un modelo.

6.4 Resultados. Precisión: 0.76. Precisión y Recall: equilibrados. La matriz de confusión: clasificación correcta en general. Factores más importantes: PRODUCTO, GRUPO_PRODUCTO, ESTADO, MES.

7. Tablero. Resumo visualmente los hallazgos clave en un tablero. Incluye:

Mapa de producción por estado.

Línea de tiempo de volumen anual.

Distribución mensual.

Productos principales.

Distribución por Grupo de Productos.

KPIs de producción total, promedio y máxima.

Utilidad:

Facilita la toma de decisiones.

Permite visualizar tendencias y anomalías.

Expone datos complejos de manera simple.

8. Conclusiones. La minería en México muestra grandes picos relacionados con minerales preciosos. Zacatecas, Chihuahua y Sonora están a la cabeza en producción. El volumen varía ampliamente y rara vez hay relación temporal entre ellos. La precisión es del 76% cuando el modelo predictivo es Bosque Aleatorio. El tablero ayuda a interpretar el sector de un vistazo.

Líneas de Trabajo Futuro. Usar modelos avanzados, como XGBoost o Prophet. ¡Hacer un análisis municipal en profundidad! Crear un tablero interactivo con Streamlit. Automatizar actualizaciones de modelos. Incorporar precios internacionales de metales como herramienta económica para el análisis.

Referencias. (De Estadística y Geografía, I. N. (s. f.-b). *Minería*. https://www.inegi.org.mx/temas/mineria/

.)