**PREDICCIÓN DE MUNICIPIOS Y PRODUCTOS CON POTENCIAL DE EXPORTACIÓN**

Aporte Diego Hatun

1. **INTRODUCCIÓN**

El comercio exterior es un motor fundamental para el desarrollo económico de las regiones.

Identificar municipios y productos con potencial de exportación permite optimizar políticas públicas, estrategias comerciales y recursos. En este trabajo se desarrolla un modelo de inteligencia artificial (IA) que clasifica municipios y productos con alto potencial de exportación basado en datos de producción y crecimiento económico.

1. **ANÁLISIS DE DATOS**

El conjunto de datos (42 columnas y 18,697 filas) combina múltiples fuentes relacionadas con municipios, características socioeconómicas y volúmenes de producción por producto, suministradas por Procolombia, para garantizar la calidad del modelo se aplicaron las siguientes etapas:

* 1. LIMPIEZA DE DATOS

Se seleccionaron solo las columnas necesarias para el modelo, excluyendo aquellas irrelevantes o redundantes, con el objetivo de reducir la dimensionalidad del conjunto de datos y evitar ruido en el modelo.

* Se corrigieron problemas de formato, como:
  + Eliminación de separadores de miles (como comas en números: 3,000.5).
  + Conversión de todas las columnas numéricas a tipo float.

*Esto asegura que todas las variables estuvieran en un formato numérico para evitar errores durante el modelado.*

* Se imputaron filas/columnas con valores faltantes en las variables seleccionadas, para prevenir que los valores nulos generen errores o sesguen el modelo.
* Aunque no se aplicó escalamiento explícito (debido a la naturaleza del Random Forest, que no lo requiere), se verificó que los datos estuvieran en rangos esperados, esto nos permite mantener consistencia en los valores y asegurar que los datos no estuvieran desbalanceados.
  1. ANÁLISIS EXPLORATORIO DE DATOS (EDA)

Se generaron estadísticas descriptivas para comprender la distribución de las variables numéricas clave en el conjunto de datos. Estas estadísticas están anexas en un documento e incluyen valores como promedio, mediana, desviación estándar, valores mínimos y máximos.

**Principales Métricas:**

* Producción (t): Resumen de los volúmenes de producción en toneladas.
* Crecimiento 2022: Tasa de crecimiento económico anual.
* Valor agregado 2022: Indicador económico de relevancia municipal.
* Distribución sectorial: Porcentaje de actividades primarias, secundarias y terciarias.

**Identificación de Productos con Mayor Volumen de Producción:**

Se realizó un análisis para identificar los productos con mayor volumen de producción en toneladas por municipio. Este análisis permitió resaltar aquellos productos líderes en cada región y sus respectivos niveles de producción.

* 1. ESTADÍSTICAS DESCRIPTIVAS

Se calcularon estadísticas generales para las variables numéricas.

Destacan los siguientes puntos:

* **Producción (t)**:
  + Media: 12,153 toneladas.
  + Mediana: 425 toneladas.
  + Producción máxima: 14,268,340 toneladas.
* **Valor agregado 2022**:
  + Media: 1,162.
  + Máximo: 316,317.

**Identificación de Columnas Clave**

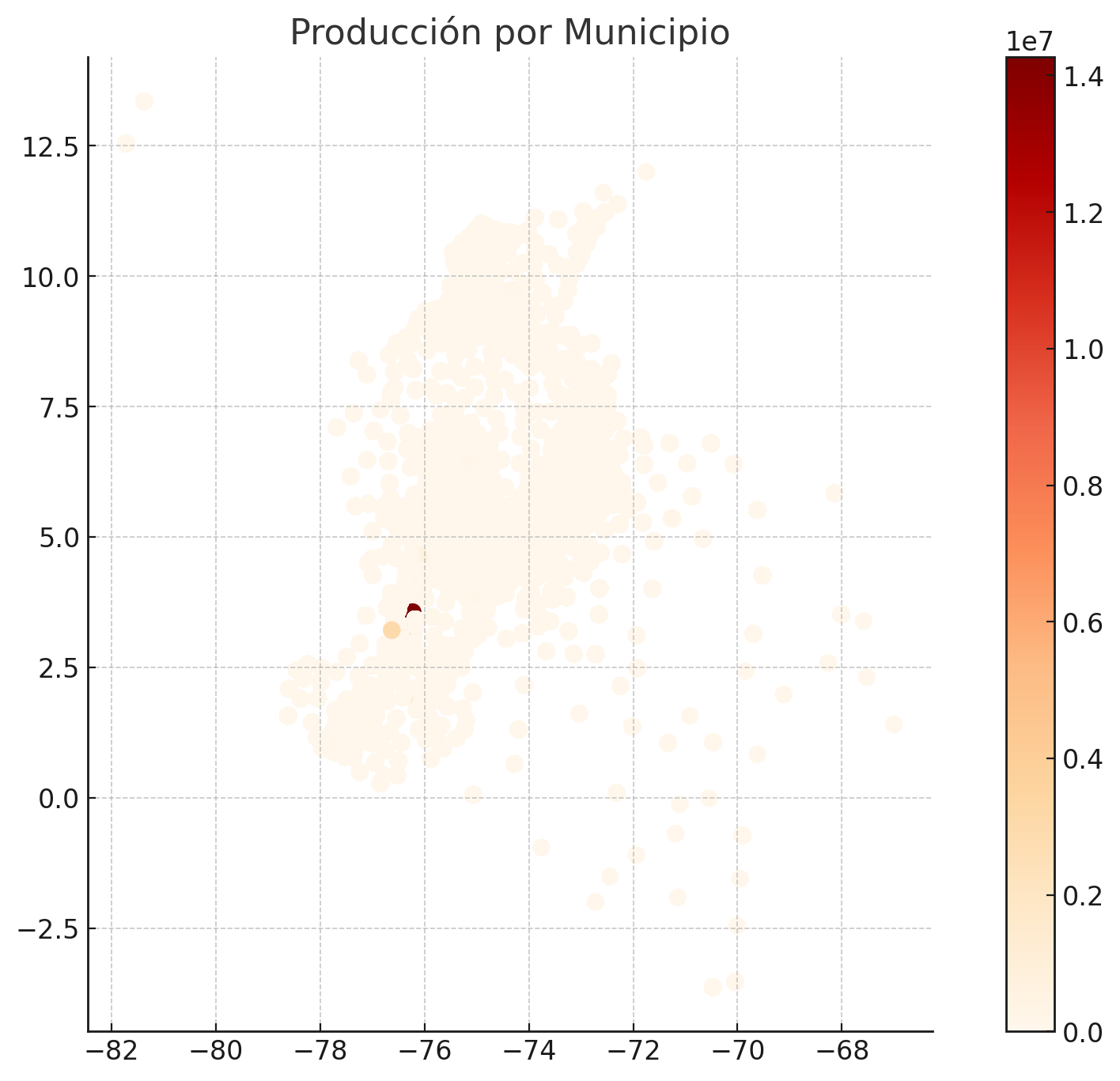
* **Municipios**: codigo\_municipio, Nombre Municipio.
* **Productos con Potencial de Exportación**: Desagregación cultivo, Producción (t).

**Resumen de los Productos con Mayor Producción**

Los municipios con la mayor producción total incluyen:

* **Palmira**: 14,348,700 toneladas.
* **Candelaria**: 8,822,511 toneladas.
* **El Cerrito**: 6,160,745 toneladas.

**VISUALIZACIÓN**



Los puntos representan la producción, destacando áreas con alta densidad de producción.

* 1. VARIABLES EXPLORADAS

Estas variables deben estar relacionadas con factores económicos, geográficos, demográficos y productivos.

Aquí está la lista de variables sugeridas:

**Variables relacionadas con los municipios:**

* **LATITUD** y **LONGITUD**: Ubicación geográfica del municipio.
* **% Act. primarias municipio**: Proporción de actividades primarias (agricultura, minería).
* **% Act. secundarias municipio**: Proporción de actividades secundarias (manufactura).
* **% Act. terciarias municipio**: Proporción de actividades terciarias (servicios).
* **Valor agregado 2022**: Contribución económica total del municipio.
* **Peso relativo municipal en el valor agregado departamental (%)**: Relevancia del municipio en el contexto departamental.
* **% pobl. con pregrado municipio**, **% pobl. con posgrado municipio**: Nivel educativo promedio de la población.
* **Crecimiento 2022**: Crecimiento económico anual del municipio.

**Variables relacionadas con los productos:**

* **Desagregación cultivo**: Nombre del producto o cultivo.
* **Producción (t)**: Producción en toneladas.
* **Crecimiento 2022**: Crecimiento asociado al cultivo/producto en el municipio.
* **% Act. primarias municipio**: Indirectamente relacionada con el sector agrícola y productos primarios.

**Variable objetivo (target):**

* **Potencial de exportación**: Variable binaria (1 si el municipio y producto tienen potencial de exportación, 0 si no).

**Cómo calcular los valores objetivo dinámicamente**

En lugar de establecer valores fijos, puedes calcular los umbrales directamente desde tus datos. Como se muestra a continuación:

python

Copiar código

# Calcular dinámicamente los cuartiles superiores

umbral\_produccion = data['Producción (t)'].quantile(0.75) # Cuartil superior para producción

umbral\_crecimiento = data['Crecimiento 2022'].quantile(0.75) # Cuartil superior para crecimiento

print(f"Umbral de Producción: {umbral\_produccion}")

print(f"Umbral de Crecimiento: {umbral\_crecimiento}")

**Ventajas del cálculo dinámico:**

1. **Adaptabilidad:**
   * Los umbrales se ajustan automáticamente a los datos reales.
2. **Relevancia:**
   * Reflejan mejor las condiciones específicas del conjunto de datos proporcionado.

**Casos Especiales**

* **Datos con muchos valores extremos:** Si el cuartil superior no representa bien la realidad, considera usar percentiles personalizados (por ejemplo, 90th en lugar de 75th).
* **Cambio en los criterios:** Puedes ajustar los umbrales según las necesidades del negocio, por ejemplo, basándote en valores históricos o indicadores de la industria.

**Código fuente utilizado para el análisis:**

import pandas as pd

import matplotlib.pyplot as plt

import geopandas as gpd

# Estadísticas descriptivas

data.describe()

# Identificación de productos con mayor volumen de producción

produccion\_municipio\_producto = data.groupby(['Nombre Municipio', 'Desagregación

cultivo'])['Producción (t)'].sum().reset\_index()

produccion\_municipio\_producto = produccion\_municipio\_producto.sort\_values(by='Producción (t)',

ascending=False)

# Visualizaciones avanzadas

plt.figure(figsize=(10, 6))

plt.hist(data['Producción (t)'].dropna(), bins=30, edgecolor='k')

plt.title('Distribución de la Producción (t)')

plt.xlabel('Producción (t)')

plt.ylabel('Frecuencia')

plt.show()

plt.figure(figsize=(10, 6))

plt.scatter(data['Crecimiento 2022'], data['Producción (t)'], alpha=0.6)

plt.title('Relación entre Crecimiento (2022) y Producción (t)')

plt.xlabel('Crecimiento 2022 (%)')

plt.ylabel('Producción (t)')

plt.show()

if 'LATITUD' in data.columns and 'LONGITUD' in data.columns:

gdf = gpd.GeoDataFrame(data, geometry=gpd.points\_from\_xy(data['LONGITUD'],

data['LATITUD']))

gdf.plot(column='Producción (t)', cmap='OrRd', legend=True, figsize=(12, 8), markersize=50)

plt.title('Producción por Municipio')

plt.show()

1. DESARROLLO DEL MODELO

Se utilizó un Random Forest Classifier, debido a su robustez frente a datos desbalanceados y su capacidad para manejar variables no lineales.

**Definición de la Variable Objetivo**

Se clasificó un municipio o producto como de potencial de exportación si cumplía ambos criterios:

1. Producción (t): Igual o superior al cuartil superior (2,237 toneladas).

2. Crecimiento 2022: Igual o superior al cuartil superior (33.64%).

***Potencial de exportación: Valor binario (1 si el municipio/producto tiene alta producción y crecimiento, 0 en caso contrario).***

**Selección de Variables para el Modelo**

Se seleccionaron variables predictoras clave para el modelo:

- `Nombre del municipio`,

- `LATITUD`, `LONGITUD`

- `Producción (t)`

- `Crecimiento 2022`

- `% Act. primarias municipio`, `% Act. secundarias municipio`, `% Act. terciarias municipio`

- `% pobl. con pregrado municipio`

- `Valor agregado 2022`

- `Peso relativo municipal en el valor agregado departamental (%)`

- `Producto`

* 1. **ENTRENAMIENTO Y EVALUACIÓN**

Conjunto de Datos: Dividido en entrenamiento (80%) y prueba (20%).

Resultados: Precisión promedio del modelo: 99.95% (validación cruzada).

Reporte de clasificación: Excelente desempeño en ambas clases (con y sin potencial de

exportación).

**Consideraciones para Clasificar un Municipio o Producto como Potencial de Exportación**

La clasificación de municipios y productos como potenciales de exportación se basa en criterios cuantitativos definidos para capturar los factores más relevantes que influyen en la capacidad de un municipio o producto para competir en mercados internacionales. Estas consideraciones son fundamentales para la toma de decisiones estratégicas en desarrollo económico y comercio exterior.

**Criterios de Clasificación**

Producción (t)

* Definición: El volumen de producción en toneladas de un producto en un municipio.
* Umbral: Debe estar en el cuartil superior de la distribución general de producción. Esto asegura que solo los municipios con niveles de producción significativamente altos sean considerados.
* Justificación: Un alto volumen de producción es necesario para garantizar la capacidad de abastecimiento y competitividad en mercados internacionales.

Crecimiento Económico (Crecimiento 2022)

* Definición: La tasa de crecimiento económico anual del municipio en el año más reciente.
* Umbral: Debe estar en el cuartil superior de la distribución general de crecimiento económico.
* Justificación: Municipios con crecimiento económico alto tienden a tener mejores condiciones para el desarrollo de actividades económicas adicionales, incluyendo exportaciones.

**Cómo se Clasifica un Municipio o Producto**

Un municipio o producto será clasificado como de potencial de exportación si cumple ambos criterios:

* Su producción (t) es igual o superior al umbral definido para la producción.
* Su crecimiento económico es igual o superior al umbral definido para el crecimiento.

**Ejemplo de Aplicación:**

* Umbral Producción: 2,237 toneladas (cuartil superior).
* Umbral Crecimiento Económico: 33.64% (cuartil superior).
* Municipio Evaluado: Palmira
* Producción: 367,915.46 toneladas (cumple con el criterio de producción).
* Crecimiento: 21.91% (no cumple con el criterio de crecimiento).
* Conclusión: Palmira no es clasificado como potencial de exportación.

**3.2 IMPLEMENTACIÓN**

El modelo fue guardado como `modelo\_exportacion.pkl` y puede integrarse fácilmente en

aplicaciones mediante un script en Python.

Algunos de los ejemplos que se le puede dar al modelo, se describen a continuación:

**Script de ejemplo:**

*import pandas as pd*

*import joblib*

*from sklearn.ensemble import RandomForestClassifier*

*# Cargar el modelo entrenado*

*model = joblib.load('/mnt/data/modelo\_exportacion.pkl')*

*# Función para predecir potencial de exportación*

*def predecir\_potencial(datos\_nuevos):*

*# Convertir a DataFrame si los datos no lo son*

*if not isinstance(datos\_nuevos, pd.DataFrame):*

*datos\_nuevos = pd.DataFrame(datos\_nuevos)*

*# Asegurar que los datos tienen las columnas necesarias*

*required\_columns = [*

*'LATITUD', 'LONGITUD', 'Producción (t)', 'Crecimiento 2022',*

*'% Act. primarias municipio', '% Act. secundarias municipio',*

*'% Act. terciarias municipio', '% pobl. con pregrado municipio',*

*'Valor agregado 2022', 'Peso relativo municipal en el valor agregado departamental (%)'*

*]*

*if not all(col in datos\_nuevos.columns for col in required\_columns):*

*raise ValueError(f"Los datos deben contener las siguientes columnas: {required\_columns}")*

*# Predecir*

*predicciones = model.predict(datos\_nuevos[required\_columns])*

*return predicciones*

*# Ejemplo de uso*

*if \_\_name\_\_ == "\_\_main\_\_":*

*# Datos de ejemplo*

*datos\_ejemplo = {*

*'LATITUD': [6.25],*

*'LONGITUD': [-75.56],*

*'Producción (t)': [5000],*

*'Crecimiento 2022': [10],*

*'% Act. primarias municipio': [30],*

*'% Act. secundarias municipio': [20],*

*'% Act. terciarias municipio': [50],*

*'% pobl. con pregrado municipio': [25],*

*'Valor agregado 2022': [10000],*

*'Peso relativo municipal en el valor agregado departamental (%)': [15]*

*}*

*datos\_df = pd.DataFrame(datos\_ejemplo)*

*print(predecir\_potencial(datos\_df))*

**Script de ejemplo csv:**

*import pandas as pd*

*# Cargar archivo CSV*

*datos\_nuevos = pd.read\_csv('ruta\_al\_archivo.csv')*

*# Predecir usando el modelo from prediccion\_exportacion*

*import predecir\_potencial*

*predicciones = predecir\_potencial(datos\_nuevos)*

*print(predicciones)*

Script de MySQL:

*import pandas as pd*

*import mysql.connector*

*# Configuración de la conexión*

*conexion = mysql.connector.connect( host='tu\_host', user='tu\_usuario', password='tu\_contraseña', database='tu\_base\_de\_datos' )*

*# Ejecutar una consulta SQL para extraer datos*

*consulta = 'SELECT \* FROM tu\_tabla'*

*datos\_nuevos = pd.read\_sql(consulta, conexion)*

*# Cerrar la conexión*

*conexion.close()*

*# Predecir usando el modelo from prediccion\_exportacion*

*import predecir\_potencial*

*predicciones = predecir\_potencial(datos\_nuevos)*

*print(predicciones)*