Lead University

Curso: 2025- I BCD7214 Administración de Datos (Sección 1)

Docente: Alejandro Zamora

Estudiantes: Carolina Salas, Kristhel Porras

Trabajo Grupal 1

# **Análisis de resultados en el Contexto del Proyecto de Predicción de Exportación de Fertilizantes**

El conjunto de imágenes representa un análisis exploratorio de datos (EDA) enmarcado dentro del pipeline de datos descrito en el documento técnico del sistema de predicción de exportación de fertilizantes. Este análisis es fundamental en la preparación de datos para el modelo de XGBoost, permitiendo identificar patrones, correlaciones, distribuciones y posibles sesgos en las variables.

1. **Mapa de Calor de Correlación**

**Análisis Técnico**

El mapa de calor de correlación muestra la relación entre variables en el dataset utilizando el coeficiente de correlación de Pearson. Las correlaciones se interpretan de la siguiente manera:

* Valores cercanos a 1 indican una fuerte correlación positiva.
* Valores cercanos a -1 representan una fuerte correlación negativa.
* Valores cercanos a 0 sugieren ausencia de correlación.

**Hallazgos Clave**

* "Peso" y "Valor" presentan una alta correlación positiva (~0.91), lo que indica que el peso de los envíos influye directamente en su valor. Dado este alto nivel de redundancia, podría no ser necesario incluir ambas variables en el modelo.
* "Nombre Comercial" y "País de Origen" (-0.46) tienen una correlación negativa significativa, lo que sugiere que ciertos nombres comerciales están vinculados a países específicos. Esta relación puede ser relevante en la predicción del destino de la exportación.
* "Número de Registro" y "País de Destino" (~0.05) muestran una correlación muy baja, lo que indica que el número de registro no tiene un impacto significativo en la determinación del destino de exportación.

**Implicaciones para el Modelo**

* Variables con alta correlación pueden ser eliminadas o combinadas mediante técnicas de reducción de dimensionalidad, como PCA o selección de atributos, para evitar problemas de multicolinealidad en el modelo de XGBoost.
* Variables con baja correlación con la variable objetivo podrían ser descartadas o transformadas para capturar relaciones no lineales.

1. **Distribución de Número de Registro**

**Análisis Técnico**

El histograma representa la distribución de la variable Número de Registro, acompañado de una estimación de densidad mediante KDE (Kernel Density Estimation). Se observan picos en ciertos rangos específicos, lo que sugiere que algunos valores de registro son más comunes que otros.

**Hallazgos Clave**

* La distribución presenta una estructura multimodal, con varios picos significativos.
* Esta estructura sugiere que los datos pueden estar organizados en lotes o categorías dentro del sistema de registro.

**Implicaciones para el Modelo**

* Un patrón multimodal podría indicar sesgos en los datos, donde ciertos registros tienen mayor representación.
* Dependiendo de la relación con la variable objetivo, podría ser necesario aplicar normalización o discretización.
* Si el número de registro no aporta información significativa, podría eliminarse del dataset para evitar ruido en el modelo.

1. **Distribución de País de Origen**

**Análisis Técnico**

Este gráfico muestra la frecuencia de los países de origen en las exportaciones de fertilizantes. Se observa que un país tiene un número desproporcionadamente alto de registros.

**Hallazgos Clave**

* La distribución está fuertemente **desequilibrada**, lo que indica que un país domina los registros.
* Este sesgo puede afectar la capacidad del modelo de XGBoost para generalizar correctamente, ya que podría sesgarse hacia el país más representado.

**Implicaciones para el Modelo**

* Se debe aplicar balanceo de clases, mediante técnicas como undersampling, oversampling o la asignación de pesos en la función de pérdida.
* Es fundamental evaluar si la representación de ciertos países está influyendo en la precisión del modelo.

1. **Distribución de País de Destino**

**Análisis Técnico**

Este gráfico muestra la frecuencia de los países de destino en las exportaciones. Se observa un pico extremadamente alto en un solo país.

**Hallazgos Clave**

* La presencia de un solo país con la mayor parte de las exportaciones sugiere un fuerte **desbalance de clases**.
* Si el modelo solo predice correctamente este país, pero falla en los demás, se podría obtener una alta precisión global, pero una baja precisión en las clases minoritarias.

**Implicaciones para el Modelo**

* La precisión global no es suficiente si la predicción de clases minoritarias es deficiente. Se recomienda evaluar métricas como recall, F1-score y balance de clases en el entrenamiento.
* Se pueden aplicar técnicas de re-weighting en la función de pérdida o muestreo estratificado en el entrenamiento.

1. **Distribución de Cantidad**

**Análisis Técnico**

El histograma muestra la distribución de la variable Cantidad de fertilizantes exportados. Se observa un pico en un valor central, lo que indica que muchas exportaciones tienden a concentrarse en ciertas cantidades específicas.

**Hallazgos Clave**

* La distribución presenta una asimetría ligera, con una acumulación en ciertos valores.
* Esto puede indicar restricciones comerciales o patrones estándar en la cantidad de fertilizantes exportados.

**Implicaciones para el Modelo**

* Si esta variable es significativa para el modelo de XGBoost, es importante verificar la presencia de valores atípicos que puedan afectar el rendimiento.
* En caso de detectar valores extremos, se pueden aplicar técnicas de transformación logarítmica o detección de outliers (IQR o DBSCAN).
* Es fundamental evaluar si la cantidad influye en la precisión de la predicción del país de destino.

**Conclusión General del EDA**

El análisis exploratorio realizado a través de estas imágenes es fundamental para optimizar el pipeline de datos y mejorar el rendimiento del modelo de predicción basado en XGBoost.

**Acciones Recomendadas Antes del Entrenamiento del Modelo**

1. Evaluar la multicolinealidad y eliminar variables redundantes basándose en el mapa de correlación.
2. Balancear los datos, en particular la variable de país de destino, para evitar sesgos en la predicción.
3. Aplicar técnicas de transformación adecuadas, como normalización para variables numéricas y one-hot encoding para variables categóricas.
4. Identificar valores atípicos y determinar su impacto en el rendimiento del modelo.
5. Evaluar métricas más allá de la precisión global, asegurando que las clases minoritarias sean consideradas en la optimización del modelo.

**Aporte de las integrantes**

Carolina asumió los roles de Data Architect y Data Custodian gracias a sus aptitudes técnicas para estructurar repositorios, creando la arquitectura en GitHub y administrando la rama principal con los documentos Python. Kristhel, por su parte, desempeñó como Data Steward y Data Governance Manager aprovechando sus capacidades en documentación, desarrollando el README y archivos de texto explicativos. Ambas colaboraron en el código Python, estableciendo un modelo efectivo de gobernanza de datos con responsabilidades complementarias.