Lead University
Curso: 2025- I BCD7214 Administración de Datos (Sección 1)
Docente: Alejandro Zamora
Estudiantes: Carolina Salas, Kristhel Porras
Trabajo Grupal 1

1. Estructura y metodología

1.1 Arquitectura del Data Pipeline

El Data Pipeline está diseñado para procesar los datos de exportación de fertilizantes y transformarlos en un formato adecuado para su uso en el modelo predictivo. Este pipeline sigue un enfoque basado en ETL (Extracción, Transformación y Carga), asegurando una gestión eficiente y automatizada de los datos.

1.1.1 Extracción de Datos

La fase de extracción se realiza a partir del Sistema de Administración de Datos de Agricultura (SADA), almacenando los datos en un archivo CSV en un repositorio de GitHub. Para ello, el pipeline utiliza la función read_dataset, que permite cargar los datos desde una URL o un archivo local, identificando automáticamente el separador. Además, logging.info registra eventos importantes en un archivo de auditoría para garantizar trazabilidad en el proceso.

1.1.2 Transformación de Datos

Durante la transformación de los datos, se aplican múltiples técnicas de preprocesamiento, incluyendo limpieza, conversión de formatos, normalización y encriptación.

- Limpieza de datos: Se eliminan registros duplicados (drop_duplicates), se detectan valores nulos (check_missing_values), y se eliminan columnas irrelevantes (delete_irrelevant_values).
- Conversión de formatos: Se transforman las variables categóricas en códigos numéricos (_transform_column) o en variables dummy, dependiendo del tipo de análisis requerido.
- Normalización y escalado: Se aplica la estandarización (standardize_data) mediante StandardScaler, asegurando que todas las variables tengan una media de 0 y desviación estándar de 1.
- Seguridad y encriptación: Se utiliza el algoritmo Fernet de la librería cryptography para proteger datos sensibles mediante funciones como _load_or_generate_key y encrypt_column.

1.1.3 Carga de Datos

Una vez transformados, los datos son almacenados en distintas ubicaciones:

- save_data: Guarda los datos en GitHub para su acceso desde el modelo.
- backup_data: Genera una copia de seguridad para prevenir pérdida de información.

1.1.4 Visualización y Monitoreo

Para el análisis exploratorio, el sistema cuenta con herramientas de visualización:

- correlation_matrix: Analiza la relación entre variables numéricas.
- plot_distributions: Genera histogramas para visualizar la distribución de los datos.

- outlier_detection: Detecta valores atípicos mediante diagramas de caja.
- advanced_outlier_detection: Aplica el método IQR para identificar valores extremos.

Este pipeline garantiza que los datos sean limpios, seguros y listos para el modelo de IA, optimizando su rendimiento en la predicción del país destino de exportaciones.

1.2 Integración del Modelo de IA

El modelo seleccionado para la predicción es XGBoost Regressor, un algoritmo de gradiente boosting optimizado para trabajar con grandes volúmenes de datos y encontrar patrones complejos en las exportaciones de fertilizantes.

1.2.1 Carga y Preparación de Datos

El modelo obtiene los datos procesados a través de la función load_data(), asegurando que la información utilizada en la predicción haya pasado por las etapas de limpieza, transformación y normalización.

- Variable objetivo (y): La salida del modelo es el país destino de la exportación.
- Variables predictoras (X): Incluyen todas las columnas del dataset excepto la variable objetivo.

1.2.2 Entrenamiento del Modelo

El modelo XGBoost Regressor es entrenado utilizando 80% de los datos para entrenamiento y 20% para validación, con la función train_test_split(). Se aplican los siguientes hiperparámetros clave:

- n_estimators=100: Número de árboles en el ensamble.
- learning_rate=0.1: Controla la tasa de aprendizaje para evitar sobreajuste.
- random_state=42: Asegura la reproducibilidad de los resultados.

El entrenamiento se realiza con la función train_model(X, y), que divide los datos, entrena el modelo y genera predicciones sobre el conjunto de prueba.

1.2.3 Evaluación y Visualización de Resultados

Para validar el desempeño del modelo, se utilizan métricas estándar de regresión:

- RMSE (Error Cuadrático Medio): Evalúa la magnitud del error en las predicciones.
- MAE (Error Absoluto Medio): Indica el promedio de las desviaciones absolutas.
- MAPE (Error Porcentual Medio Absoluto): Muestra el error relativo en porcentaje.
- R^2 Score: Mide qué porcentaje de la variabilidad es explicada por el modelo.

La función analyze_results(y_test, y_pred) presenta los resultados en Streamlit, mostrando métricas y gráficos clave:

- Comparación entre valores reales y predichos: Un gráfico de dispersión que permite identificar la alineación del modelo con los datos reales.
- Distribución de predicciones: Un histograma que analiza la frecuencia de los valores predichos.
- Detección de valores atípicos: Mediante diagramas de caja que muestran la variabilidad de las predicciones.
- Gráfico de residuos: Un análisis de los errores cometidos por el modelo en diferentes rangos de predicción.

2. Distribución del trabajo en base a roles de datos

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.

3. Análisis de Resultados en la Predicción de Exportación de Fertilizantes

El análisis exploratorio de datos (EDA) realizado en el contexto del pipeline de predicción de exportación de fertilizantes es clave para garantizar la calidad de los datos utilizados en el modelo XGBoost. A través de la identificación de correlaciones y valores atípicos, se pueden optimizar las variables empleadas en la predicción, eliminando redundancias y corrigiendo sesgos en los datos.

3.1 Correlaciones entre Variables y su Impacto en el Modelo

El mapa de calor de correlación revela que algunas variables presentan una fuerte relación entre sí. Un hallazgo clave es la alta correlación entre Peso y Cantidad (≈ 1.00), lo que sugiere que ambas representan información similar. Para evitar problemas de multicolinealidad en el modelo, se recomienda evaluar la eliminación o combinación de una de ellas.

3.2 Otras correlaciones relevantes

Estas incluyen la relación entre Exportador y País de Origen (≈ 0.42), lo que indica que ciertos exportadores están vinculados a países específicos. Asimismo, la variable Nombre Comercial muestra una correlación de 0.46 con el País de Origen, lo que sugiere que las marcas comerciales pueden ser un buen predictor del destino de la exportación. Estas relaciones pueden ser útiles en la selección de características y deben considerarse en el preprocesamiento.

En contraste, se detectaron variables con baja correlación con la variable objetivo. Número de Registro y País de Destino (≈ 0.05) no presentan una relación significativa, lo que sugiere que esta variable podría no aportar información valiosa para la predicción y podría ser eliminada sin afectar la precisión del modelo. De manera similar, Importador y País de Destino (-0.11) muestran una relación muy débil, por lo que su utilidad en la predicción debe ser evaluada con más detalle.

Para mejorar la selección de características, se recomienda aplicar técnicas de reducción de dimensionalidad como PCA o enfoques avanzados como SHAP (SHapley Additive Explanations) para identificar el impacto real de cada variable en el modelo XGBoost. También, el uso de embedding layers en redes neuronales podría mejorar la representación de variables categóricas como País de Origen y Exportador.

4. Valores Atípicos y su Influencia en la Predicción

El análisis de valores atípicos revela la presencia de datos extremos en múltiples variables. En particular, Cantidad y Peso presentan 957 valores atípicos cada una, lo que indica que hay envíos con volúmenes considerablemente menores o mayores al promedio. Dado que

estas variables están fuertemente correlacionadas, su inclusión simultánea podría generar inestabilidad en el modelo. Para mitigar este efecto, se recomienda aplicar transformación logarítmica o normalización, reduciendo la influencia de valores extremos en el entrenamiento de XGBoost.

Por otro lado, la variable País Destino presenta 1057 valores atípicos, reflejando un fuerte desbalance en la distribución de los países receptores de exportaciones. Esto sugiere que algunos países tienen una representación muy baja en el conjunto de datos, lo que puede afectar la capacidad del modelo para generalizar predicciones en estos casos. Para abordar este problema, se recomienda aplicar balanceo de clases mediante técnicas de oversampling o undersampling, así como ajustar los pesos en la función de pérdida de XGBoost para reducir el impacto de las clases poco representadas.

Las variables Importador y Exportador también muestran valores atípicos significativos, lo que indica que ciertas empresas manejan volúmenes de exportación muy superiores al resto. Si estos valores representan casos excepcionales y no errores de captura, podrían mantenerse en el modelo con técnicas de clustering o winsorization para reducir su impacto en el entrenamiento.

5. Análisis del Modelo XGBoost Regressor en la Predicción de Exportación de Fertilizantes

5.1 Evaluación del Desempeño del Modelo

Los resultados obtenidos en la evaluación del modelo reflejan un rendimiento sólido. El coeficiente de determinación (R^2) de 0.9034 indica que el modelo explica más del 90% de la variabilidad en los datos, lo que sugiere una alta capacidad predictiva. Además, el MAE de 0.0896 refleja un error promedio bajo en las predicciones, lo que significa que el modelo realiza estimaciones cercanas a los valores reales en la mayoría de los casos.

El RMSE de 0.2858 confirma que los errores no son extremos y que el modelo mantiene una estabilidad en sus predicciones. Sin embargo, el MAPE de 13.77% sugiere que, en ciertos segmentos de datos, el modelo aún tiene margen de mejora. En general, estos resultados indican un buen desempeño, aunque se identifican valores atípicos que podrían estar afectando la precisión en algunos casos específicos.

5.2 Comparación entre Valores Reales y Predichos

El análisis del gráfico de dispersión entre valores reales y predichos muestra que el modelo sigue la tendencia esperada, con una alineación clara entre predicciones y valores observados. No obstante, se detectan algunos puntos alejados de la línea de regresión, lo que indica errores sistemáticos en ciertos segmentos.

Estos errores son más notorios en valores extremos, lo que sugiere que el modelo podría beneficiarse de un preprocesamiento más detallado de los datos o de ajustes en la regularización. Para mitigar estas diferencias, se recomienda aplicar técnicas de detección y tratamiento de outliers, así como ajustar hiperparámetros como learning rate y max_depth, con el fin de mejorar la adaptabilidad del modelo a casos atípicos.

5.3 Distribución de Predicciones y Posible Sesgo

Revela una concentración alta en torno a cero, lo que sugiere que el modelo podría estar favoreciendo ciertos valores dentro de un rango limitado. Este patrón puede indicar que el modelo tiene dificultades para capturar completamente la variabilidad de los datos.

5.4 Diagrama de Caja de Predicciones y Gráfico de Residuos

Confirma que las estimaciones del modelo se concentran en un rango específico, con presencia de valores atípicos en los extremos. Si bien la mayoría de las predicciones se encuentran dentro de los límites esperados, la presencia de valores extremos sugiere que el modelo puede estar sobreestimando o subestimando ciertos casos.

6. Conclusión y recomendaciones

El modelo XGBoost Regressor ha demostrado una alta precisión (R^2 = 0.9034) en la predicción del país destino de exportaciones de fertilizantes, con errores bajos (MAE = 0.0896, RMSE = 0.2858). Sin embargo, se han identificado sesgos en la distribución de predicciones y presencia de valores atípicos, lo que sugiere la necesidad de ajustes para mejorar su capacidad de generalización.

Para optimizar el rendimiento del modelo, se recomienda ajustar los hiperparámetros, incluyendo learning_rate, max_depth y n_estimators, además de aplicar regularización (alpha, lambda) para reducir la influencia de valores extremos. También es clave balancear la variable objetivo mediante pesos en la función de pérdida, asegurando una mejor representación de países con menor frecuencia en los datos.

Adicionalmente, se sugiere utilizar SHAP values para identificar las variables más relevantes y evaluar modelos complementarios como Random Forest o LightGBM para comparar su desempeño con XGBoost. Finalmente, técnicas de stacking o blending pueden mejorar la capacidad predictiva al combinar múltiples modelos.

En conclusión, aunque XGBoost ha mostrado un desempeño sólido, la implementación de estas mejoras permitirá reducir errores, mejorar la estabilidad y garantizar predicciones más confiables en el contexto de la exportación de fertilizantes.