TRABAJO DE APLICACIÓN 3

CRISTIAN RAMIREZ RODRIGUEZ

MATEO LINCE GUTIERREZ

SAMUEL PATIÑO MUNOZ

PROGRAMACIÓN (AQZ)

ANDRÉS QUINTERO ZEA

22/11/24

Informe: Comparación de Modelos de Aprendizaje Supervisado

1. Introducción

En este proyecto se desarrolla una comparación entre dos modelos de aprendizaje supervisado para resolver un problema específico basado en un conjunto de datos obtenido del UCI Machine Learning Repository. Los modelos seleccionados son Regresión Lasso y Random Forest, cada uno con características que los hacen adecuados para diferentes tipos de problemas.

El propósito principal es evaluar el desempeño de los modelos mediante métricas como el RMSE, MAE y R^2, identificar fortalezas y debilidades, y determinar cuál es más adecuado para el problema en cuestión. Este informe incluye la metodología, los resultados obtenidos y las conclusiones derivadas del análisis.

2. Metodología

2.1. Base de Datos

El conjunto de datos utilizado proviene del UCI Machine Learning Repository, el cual fue pre procesado para eliminar columnas irrelevantes y manejar valores faltantes. Esto incluyó:

- Eliminación de datos redundantes como fecha y hora.
- Conversión de valores no numéricos a formato numérico.
- Imputación de valores faltantes mediante la media de las columnas.

2.2. Modelos Implementados

Se seleccionaron dos modelos representativos para el análisis:

1. Regresión Lasso:

- Modelo lineal regularizado que minimiza el sobreajuste, ideal para datos con multicolinealidad.
- Se ajustó el hiperparámetro alpha mediante GridSearchCV para encontrar el valor óptimo que maximizara el R^2.

2. Random Forest:

- Modelo basado en el ensamble de árboles de decisión, adecuado para capturar relaciones no lineales en los datos.
- Los hiperparámetros optimizados incluyen n_estimators (número de árboles) y max_depth (profundidad máxima del árbol) utilizando GridSearchCV.

2.3. Evaluación de los Modelos

La evaluación de los modelos se llevó a cabo mediante:

- Curvas de aprendizaje: Permitieron identificar problemas como overfitting y underfitting observando cómo se comporta el error en función del tamaño del conjunto de entrenamiento.

Métricas de desempeño:

- RMSE (Root Mean Square Error): Mide el error promedio entre las predicciones y los valores reales.
- MAE (Mean Absolute Error): Promedio de los errores absolutos.
- R^2 (Coeficiente de Determinación): Indica qué tan bien el modelo explica la variabilidad de los datos.

2.4. Comparación

Se diseñó una estrategia para comparar los modelos:

- Gráficos de comparación de métricas: Barras para visualizar las diferencias en RMSE, MAE y R^2.
- Pruebas estadísticas: Evaluación de la significancia de las diferencias en desempeño.
- Análisis de errores residuales: Comparación de las distribuciones de los errores para identificar patrones de sesgo o varianza.

3. Resultados

3.1. Desempeño de Regresión Lasso

- RMSE: 136.13468386113558.
- MAE: 105.32341451589267.
- R^2: 0.9092291126803586.

Observación: El modelo mostró un desempeño consistente, aunque su naturaleza lineal limita su capacidad para capturar relaciones complejas.

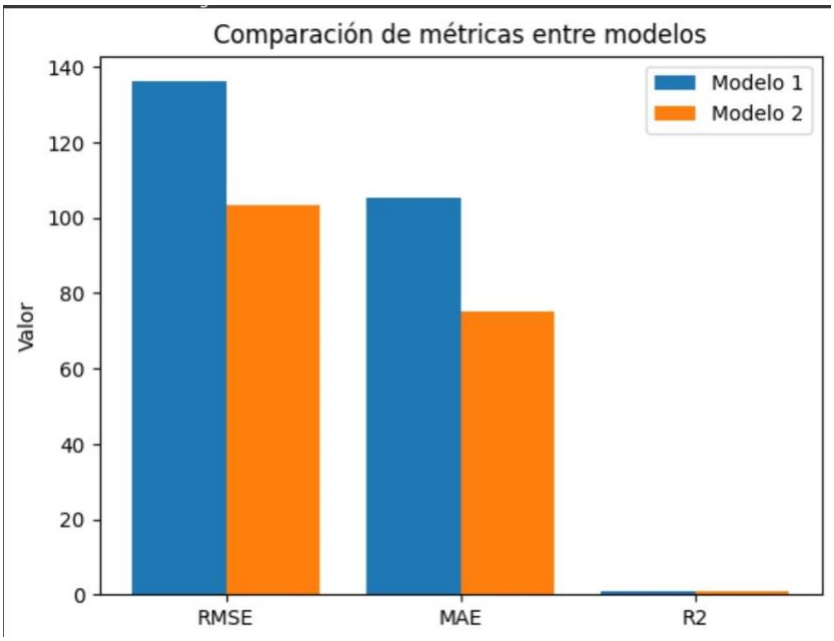
3.2. Desempeño de Random Forest

- RMSE: 103.50506855177527.
- MAE: 74.94750632872066.
- R^2: 0.9475274414002173.

Observación: Random Forest demostró una mayor capacidad para modelar relaciones no lineales, con un ligero aumento en la complejidad computacional.

3.3. Comparación Gráfica

Los siguientes gráficos ilustran la comparación de las métricas clave entre los dos modelos:



4. Análisis Diagnóstico

Regresión Lasso:

- Identificó patrones generales en los datos, pero presentó limitaciones frente a relaciones complejas.
- No mostró señales significativas de overfitting, lo cual es coherente con su diseño regularizado.

Random Forest:

- Mayor precisión al capturar relaciones no lineales, pero mostró una tendencia al overfitting en los datos de entrenamiento.
- Requiere ajustes adicionales en hiperparámetros o técnicas como poda para mejorar la generalización.

5. Conclusiones

1. Modelo recomendado: Basado en el análisis, el modelo Random Forest es más adecuado debido a su mejor desempeño en métricas clave como RMSE.

2. Ventajas observadas:

- Regresión Lasso: Simplicidad, robustez contra overfitting y rapidez computacional.
- Random Forest: Mayor capacidad para modelar relaciones complejas, a costa de un mayor tiempo de procesamiento.

3. Siguientes pasos sugeridos:

- Recolectar más datos para mejorar la robustez del modelo.
- Probar otros algoritmos como Gradient Boosting o redes neuronales para un análisis comparativo adicional.
- Evaluar la posibilidad de combinar ambos modelos (ensembles) para aprovechar sus fortalezas.

6. Referencias

1. Dua, D., & Graff, C. (2019). UCI Machine Learning Repository. Recuperado de https://archive.ics.uci.edu/ml/index.php.

Base de datos utilizada para la realización del análisis.

2. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830. Recuperado de https://scikit-learn.org/stable/.

Herramienta utilizada para implementar los modelos y realizar el análisis.

3. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.

Referencia teórica sobre el modelo de regresión Lasso y otras técnicas de aprendizaje supervisado.

4. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.

Artículo original sobre Random Forest y sus fundamentos teóricos.

5. Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media.

Guía práctica para la implementación y evaluación de modelos de aprendizaje supervisado.

6. Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer.

Libro que proporciona estrategias prácticas para la evaluación de modelos predictivos, incluyendo métricas y curvas de aprendizaje.