Alumno: Loza Ever

Profesor: Martin Mirabete

Materia: Aprendizaje Automático

Entregable 3: Análisis de Datos y Modelos de Aprendizaje Automático

---

Introducción y Origen de los Datos

- **Descripción de la Fuente de Datos**: El dataset utilizado en este proyecto fue recopilado a través de scraping de los boletines meteorológicos publicados por la Estación Astronómica Río Grande (EARG) de la Universidad Nacional de La Plata. Este conjunto de datos incluye información diaria desde enero de 2016 hasta agosto de 2024 y abarca variables meteorológicas esenciales como temperatura, humedad, velocidad y dirección del viento, presión, y radiación solar.
- **Objetivo del Proyecto**: El objetivo principal de este proyecto es desarrollar un modelo de Aprendizaje Automático que prediga la producción de energía eólica en Tierra del Fuego, teniendo en cuenta factores climáticos que afectan la generación de energía.

Análisis Exploratorio de Datos (EDA)

- **Resumen Estadístico**: Realicé un análisis estadístico preliminar del dataset para comprender la distribución de cada variable, identificar valores atípicos y evaluar posibles correlaciones entre variables meteorológicas y la producción de energía eólica simulada.
- **Gráficos y Visualizaciones**: Incluí gráficos de dispersión y diagramas de caja para visualizar la relación entre variables críticas como la temperatura y la velocidad del viento. Estos gráficos permitieron observar patrones estacionales y tendencias diarias y facilitaron la detección de outliers.
- **Conclusiones del Análisis Exploratorio**:

  Distribución de las Variables:

  Durante el análisis de las variables numéricas, se observó que algunas de ellas presentan distribuciones sesgadas, lo que puede afectar ciertos modelos de machine learning. Como la temperatura en relacion a las estaciones del año o los vientos en relacion a la ubicación geográfica.

  Es relevante observar si alguna variable tiene valores atípicos (outliers) o si hay una gran concentración de datos en ciertos rangos. Si se

encuentran estos casos, podrían ser necesarios tratamientos adicionales para la normalización o la estandarización de los datos.

Correlaciones entre Variables:

El análisis de correlación mediante matrices de correlación (como el gráfico de calor) permitió identificar cuáles de estas relaciones son más fuertes, lo que es útil para reducir la dimensionalidad del modelo si es necesario.

Desarrollo del Modelo de Aprendizaje Automático

Los modelos fueron desarrollados con los datos meteorológicos históricos. Se utilizó una técnica de escalado para mejorar el rendimiento en modelos sensibles a la magnitud de las características. La variable dependiente seleccionada fue la producción diaria simulada en kWh, mientras que la variable independiente principal fue el viento medio, por su relevancia directa en la generación eólica.

Evaluación y Rendimiento de los Modelos

Regresión Lineal

El modelo de Regresión Lineal mostró un MSE de 17,432,215 y un R^2 de 0.774, lo que indica una correlación moderada entre el viento medio y la producción eólica. Si bien el modelo capta algunas relaciones lineales, podría no ajustarse completamente debido a la posible no linealidad en los datos.

(SVM)

Para el modelo SVM con kernel lineal, los resultados muestran un MSE de 60,046,627 y un R^2 de 0.222. Este bajo desempeño sugiere que el modelo SVM no captura bien las complejidades de los datos. La baja precisión del modelo indica que un kernel lineal podría no ser adecuado para este caso, y alternativas como el kernel radial o polinomial podrían mejorar el ajuste.

Árbol de Decisión

El Árbol de Decisión alcanzó un MSE de 5,437.75 y un R^2 de 0.999, lo cual representa un ajuste excelente y captura con precisión la variabilidad en la producción eólica. Sin embargo, este resultado podría indicar sobreajuste

(overfitting), ya que el modelo parece memorizar los datos en lugar de generalizar el patrón. Este comportamiento se puede mitigar en futuros experimentos con la introducción de validación cruzada o usando técnicas de regularización.

Visualización de Resultados

Se generaron gráficos comparativos entre los valores reales y predichos, y los modelos lineal y de árbol de decisión muestran una correlación visual con la realidad, aunque el Árbol de Decisión se destaca por acercarse mucho más a los valores reales. Este ajuste visual respalda las métricas obtenidas, especialmente para el Árbol de Decisión, que reproduce los valores de manera casi exacta.

Conclusiones y Recomendaciones

El modelo de Árbol de Decisión se destacó como el mejor predictor de la producción diaria de energía eólica, aunque podría presentar problemas de sobreajuste. La Regresión Lineal logró capturar parcialmente la relación, pero puede estar limitada por la posible no linealidad de los datos. Finalmente, el modelo SVM mostró un bajo rendimiento, posiblemente debido a la elección de un kernel lineal no adecuado para los datos.

Referencias y Acceso al Proyecto

Repositorio GitHub: Todo el código, el dataset y los notebooks están disponibles en el repositorio de GitHub: Politecnico_Malvinas-Proyecto_ML.

Referencias Externas:

Boletín Meteorológico EARG

Resumen No Técnico del Proyecto Parque Eólico de Tierra del Fuego