Relatório do Trabalho de Machine Learning com o Dataset AirQualityUCI

Autor: Gabriel Queiroz Teles - 2311064 - Ciência da Computação

---

Análise Inicial dos Dados

- **Dados Nulos**: Não havia dados nulos no dataset original.
- **Valores -200.00**: Identificados em várias colunas e substituídos por NaN.
- **Tratamento dos NaNs**:
    - Coluna `NMHC(GT)` removida por estar quase toda nula.
    - Optou-se por dropar as linhas com NaNs restantes, mantendo 6941 linhas de dados completos.

Preparação e Visualização dos Dados

- **Gráficos**:
    - Boxplots indicaram a presença de outliers, que não foram removidos para evitar a redução excessiva do dataset.
    - Matriz de correlação utilizada para identificar colunas com melhor correlação com `CO(GT)`, removendo as menos relevantes.

Modelagem e Treinamento

Dividiu-se o dataset em conjuntos de treino e teste. Foram utilizados os seguintes algoritmos de regressão:

1. **LinearRegression**
2. **RandomForestRegressor**
3. **ExtraTreesRegressor**
4. **GradientBoostingRegressor**
5. **KNeighborsRegressor**
6. **SVR (kernel=”rbf”)**

- **Resultados**: Todos os algoritmos apresentaram desempenho semelhante e bom. O `ExtraTreesRegressor` se destacou levemente em 3 das 5 execuções.

Otimização

- **Métodos Utilizados**: GridSearchCV, RandomizedSearchCV e BayesSearchCV.
- **Melhor Desempenho**: BayesSearchCV com `ExtraTreesRegressor`.
    - **Métricas de Desempenho**:
        - **MSE**: 0.14357001692403945
        - **MAE**: 0.24330957115697305
        - **R^2**: 0.9330351283585044
        - **RMSE**: 0.37890634320903027

Conclusão

- **Precisão**: Alta precisão nas previsões.
- **Variância Explicada**: Modelo adequado para o problema de regressão, explicando a maior parte da variância dos dados de saída.