Universidade Federal do Paraná

Curso de pós-graduação em Informática

Disciplina: Aprendizagem de Máquina

Professor: Luiz Eduardo S. Oliveira

Aluna: Angelita Rettore de Araujo Zanella

**LABORATÓRIO REGRESSÃO**

Para a execução desse laboratório, foram selecionadas as características F5 e F6 para fazer a predição para F3 e F4. As características foram escolhidas por serem as que apresentam maior correlação com as variáveis de predição.

Antes de definir os parâmetros para cada regressor, foram analisados os desempenhos com algumas variações, afim de encontrar aquela que resultasse em um menor MSE. A Tabela 1 apresenta os regressores e seus respectivos parâmetros, com exceção do regressor Linear, que não possui parâmetros para otimização.

|  |  |  |
| --- | --- | --- |
| Regressor | Parâmetros avaliados | Valores escolhidos |
| KNN | n\_neighbor  weight  metric | 5  distance  manhattan |
| Decision Tree | splitter  max\_depth | best  3 |
| SVR | kernel  C  gamma  degree  epsilon  coef0 | poly  100  auto  3  .1  1 |
| MLP | hidden\_layer\_sizes activation  solver  learning\_rate  max\_iter  learning\_rate\_init  alpha | (5,3)  logistic  adam  invscaling  1000  0.01  0.01 |
| Random Forest | max\_depth  n\_estimators | 3  50 |
| Gradient Boost | n\_estimators  random\_state | 20  0 |
| Tabela 1: Padrâmetros utilizados pelos regressores | | |

Antes de iniciar a simulação a base de dados foi dividida em duas partes, utilizando o *random state* com valor 4. Em seguida a base de treinamento foi dividida novamente, sendo 20 para validação e 80 para treinamento. Os dados de entrada foram normalizados com a função MinMaxScaler. O resultado geral pode ser observado nas Figuras 1 e 2, que apresentam os erros obtidos por cada regressor para as variáveis F3 e F4. De acordo com esses dados, os Random Forest, Gradient Boosting e Decision Tree foram os que obtiveram melhor desempenho, sendo que o Gradient Boosting obteve melhor desempenho geral.

A close up of a logo

Description automatically generated

Figura 1: Erro quadrado médio para a variável F3

A close up of a logo

Description automatically generated

Figura 2: : Erro quadrado médio para a variável F4

Ao analisar os gráficos pode-se perceber que os erros para predição das variáveis F3 e F4 são bastante diferentes, o que pode ser respondido pelas correlações entre as variáveis e as características do modelo. Em especial, a variável F3 possui menor correlação com todas as demais características e F4 possui melhor correlação com as características utilizadas para predição. A Figura 3 apresenta uma comparação do desempenho dos regressores para ambas predições. A figura mostra que os melhores desempenhos foram alcançados pelos algoritmos Decision Tree, Random Forest e Gradient Boosting.

A screenshot of text

Description automatically generated

Figura 3: Desempenho dos classificadores

Uma vez que o Gradient Boosting foi o regressor que apresentou melhor desempenho, este foi escolhido para análise dos erros. Os erros foram obtidos pela diferença entre o resultado da predição e o *label* na respectiva base. As Figuras 4 e 5 apresentam o erro quadrado médio para a variável F3. Pode-se perceber que na base de validação mais de 95% dos erros ficaram dentro do desvio padrão. Porém, quando analisado o erro obtido na base de testes, percebe-se um grande volume de erro ocorrendo fora do desvio padrão. Isso pode ser explicado pela baixa correlação da variável F3 com as características utilizadas para predição. No entanto, a utilização de mais características não melhoraram o desempenho, ao contrário, aumentaram a taxa de erro. Já para a variável F4 o desempenho alcançado foi muito bom e mais de 95% dos erros se encontram dentro do desvio padrão. Os resultados alcançados pelo Random Forest foram muito similares aos do Gradient Boosting, por isso optou-se por apresentar neste relatório apenas os resultados de um dos algoritmos.

A room with white walls

Description automatically generated

Figura 4: Erro quadrado médio para F3 na base de validação

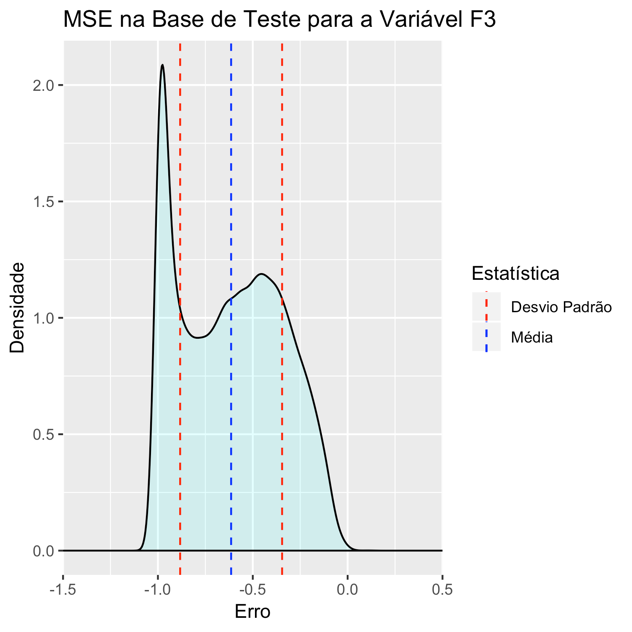
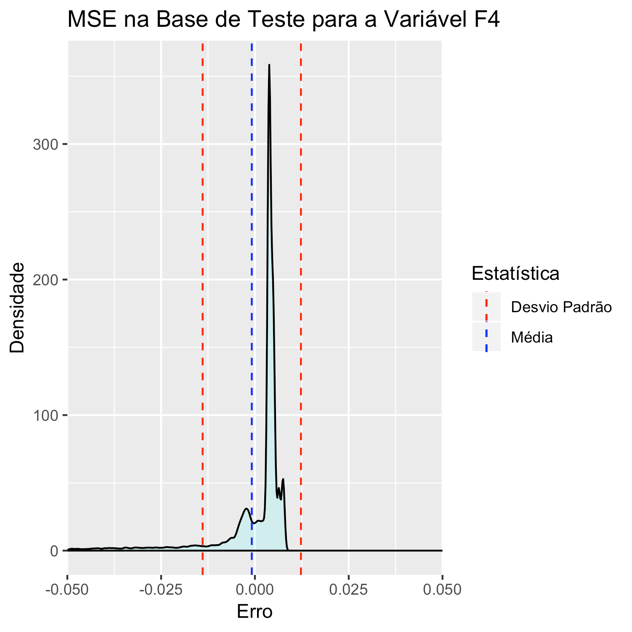


Figura 5: Erro quadrado médio para F3 na base de teste

A screenshot of a cell phone

Description automatically generated

Figura 6: Erro quadrado médio para F4 na base de validaçãoFigura 7: Erro quadrado médio para F4 na base de validação