Participante: Bruno Kenhy Higa

Relatório de Resolução do Desafio Cientista de Dados

Link do Repositório no GitHub:

Introdução

Primeiramente gostaria de pedir desculpas pois devido a uma semana extraordinariamente ocupada, infelizmente, tive que realizar todo o desafio em poucas horas e praticamente sem conhecimento prévio sobre machine learning. Contudo, estou enviando o que consegui fazer pois tenho um interesse muito grande em participar deste programa.

O desafio

O desafio proposto neste processo seletivo consiste na modelagem de um Machine Learning cujo objetivo é identificar quais máquinas apresentam potencial de falha. Para isso foram fornecidas duas bases de dados, uma para treino do modelo e outra para teste.

A resolução

Estamos resolvendo um problema de classificação utilizando o Machine Learning. As escolhas dos métodos e parâmetros de medida são mencionadas e brevemente explicadas no passo a passo à seguir:

Os passos seguidos para a resolução do case foram os seguintes:

1) Entendimento das colunas presentes no *dataset*
2) Importar o *dataset* no Jupyter Notebook
3) Verificar se existem NaN values para possível tratamento de dados
4) Identificar possíveis correlações entre as colunas de dados. Para isso, foi calculada a correlação entre as colunas de dados e representadas em uma matriz.
    a) Foi constatado que não existe forte correlação entre a grande maioria das variáveis.
5) Não existindo grande correlação, foi optado pelo modelo de “Decision Tree” por ser um modelo vantajoso para um grande número de variáveis independentes.
6) O modelo foi testado, primeiramente, utilizando 80% dos dados para treino e 20% para teste, alcançando uma acurácia de aproximadamente 97%. A métrica acurácia foi escolhida pois nos mostra a quantidade de máquinas classificadas corretamente.

7) Após isso, o modelo foi treinado com 100% dos dados da base de dados de treino com o objetivo de atingir uma maior acurácia para a utilização da base de dados destinada para o teste.
8) Em seguida foi importada a base de dados para teste, verificada a existência de NaN value.
9) Após rodar o modelo, a array resultante foi adicionada à uma cópia do dataset de teste como uma coluna nomeada “predictedValues”
10) Por último, esta coluna junto com o índex foram exportados para uma planilha csv nomeada “predicted.csv”.

Comentários

Gostaria de ter feito mais análises neste desafio. A minha rotina fora do comum nesta semana realmente foi bastante prejudicial para a execução do desafio, principalmente por praticamente não praticamente não possuir conhecimento técnico prévio na área.

Uma análise visando encontrar possíveis outliers seria importante também para maior qualidade dos dados que alimentam o modelo.

Uma das análises que gostaria de ter feito é a comparação dos resultados obtidos pelo modelo e os “requisitos” para as falhas ocorrerem fornecidas pelo enunciado. Idealmente utilizaria pares de dataframes para comparar a eficácia do modelo.
Por exemplo, no caso do “heat dissipation failure (HDF)” :

- Criaria um dataframe filtrando por diferença absoluta entre “air_temperature_k” e “process_temperature_k” menor que 8,6 e “rotation_speed_rpm” menor que 1380.
- Criaria outro dataframe filtrando por “predictValues” == “Heat Dissipation Failure”
- Em seguida, utilizaria um *join inner* e um *join outer* para levantar a informação de quantas máquinas apresentariam problema segundo as condições e não foram previstos pelo Machine Learning e quantas máquinas foram previstas “erroneamente”

Além disso, o maior problema da minha resolução é que não consegui realizar testes para entender a confiabilidade do modelo aplicado à base de dados teste. O único valor referente à acurácia está relacionado ao teste com uma amostragem da base utilizada para o treinamento. Porém, entendo que este valor de acurácia seja consistente e confiável desde que haja uma boa qualidade dos dados da base de treinamento, sendo representativa,

aleatória e variada. Além disso, entendo que a base apresente um boa quantidade de valores, aumentando assim a acurácia do modelo.

Conclusão

O modelo poderia ser aprimorado e muitas outras análises estatísticas poderiam ter sido realizadas. Análises visando entender melhor a confiabilidade do modelo seriam muito bem-vindas.