DETECTANDO POTENCIAIS FALHAS EM MÁQUINAS

RELATÓRIO

PREPARADO POR
JOÃO VICTOR SOARES SARAIVA

SOLUÇÃO PROPOSTA

.

Para resolver esse problema de classificação foi construído uma solução de ciência de dados, com base em dados extraídos através de sensores durante o processo de manufatura. A solução utiliza tecnologias como:

- Python: Linguagem de programação do projeto;
- Pandas: Para manipulação de dados;
- Seaborn, Matplotlib e Plotly: Para visualização de dados;
- Pandas-profiling: Para análise exploratória de dados;
- Scikit-learn: Para implementar algoritmos de aprendizado de máquina;
- Pycaret: Para automatizar fluxo de criação e avaliação de algoritmos de Machine Learning;
- GitHub: Para armazenar todos os códigos desenvolvidos;
- Jupyter notebook: Plataforma para desenvolvimento do projeto
- Anaconda: Gerenciador de ambientes e pacotes.

ANÁLISE EXPLORATÓRIA DE DADOS

96,52%

Maquinas não apresentaram alguma falha

3,48%

Máquinas apresentaram algum tipo falhas

Valores médios por tipo de falha

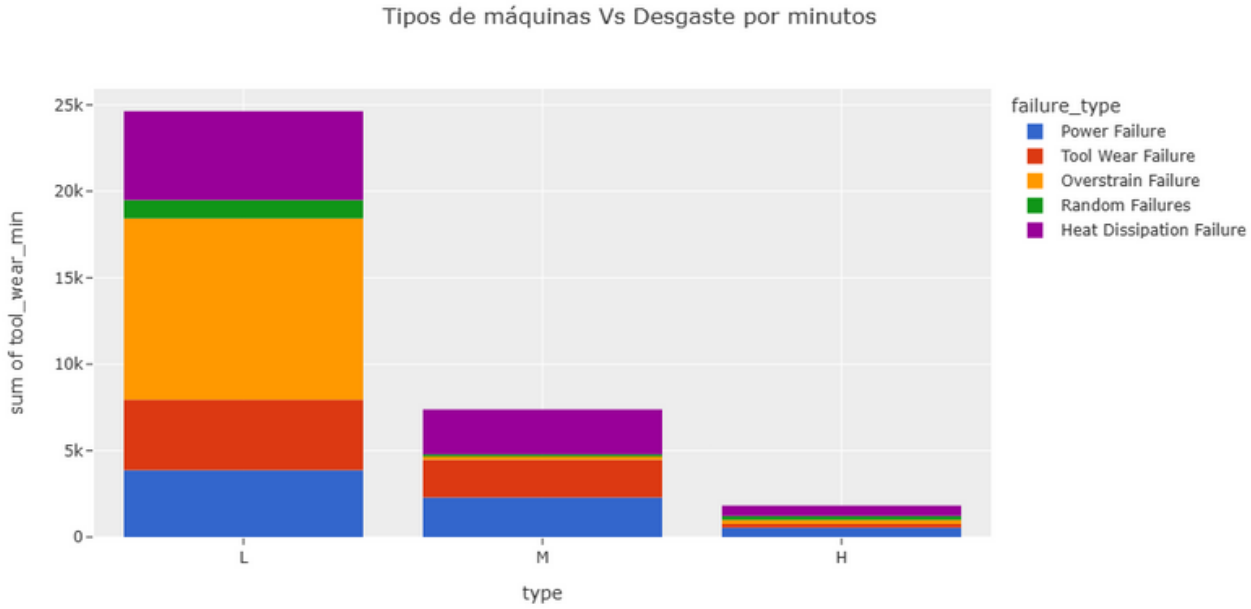
failure_type	air_temperature_k	process_temperature_k	rotational_speed_rpm	torque_nm	tool_wear_min
Heat Dissipation Failure	302.550667	310.776000	1338.946667	52.493333	110.773333
No Failure	299.961756	309.981943	1539.159751	39.693986	106.737840
Overstrain Failure	299.955769	310.073077	1358.557692	56.336538	209.865385
Power Failure	299.888889	309.873016	1712.857143	51.080952	106.587302
Random Failures	300.691667	310.691667	1492.416667	43.608333	118.583333
Tool Wear Failure	300.196667	310.156667	1619.933333	34.380000	215.766667

Atributos com uma alta correlação positiva

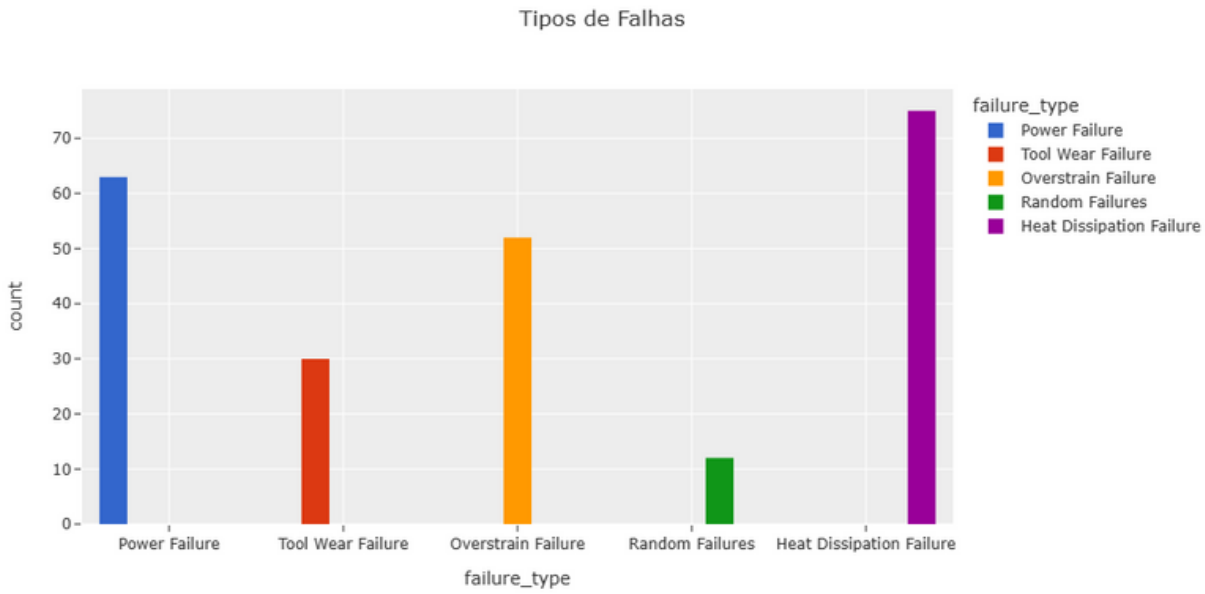
Process_temperature_k - air_temperature_k = 0.88

Atributos com correlação negativa

Torque_nm - rotational_speed_rmp = -0.88

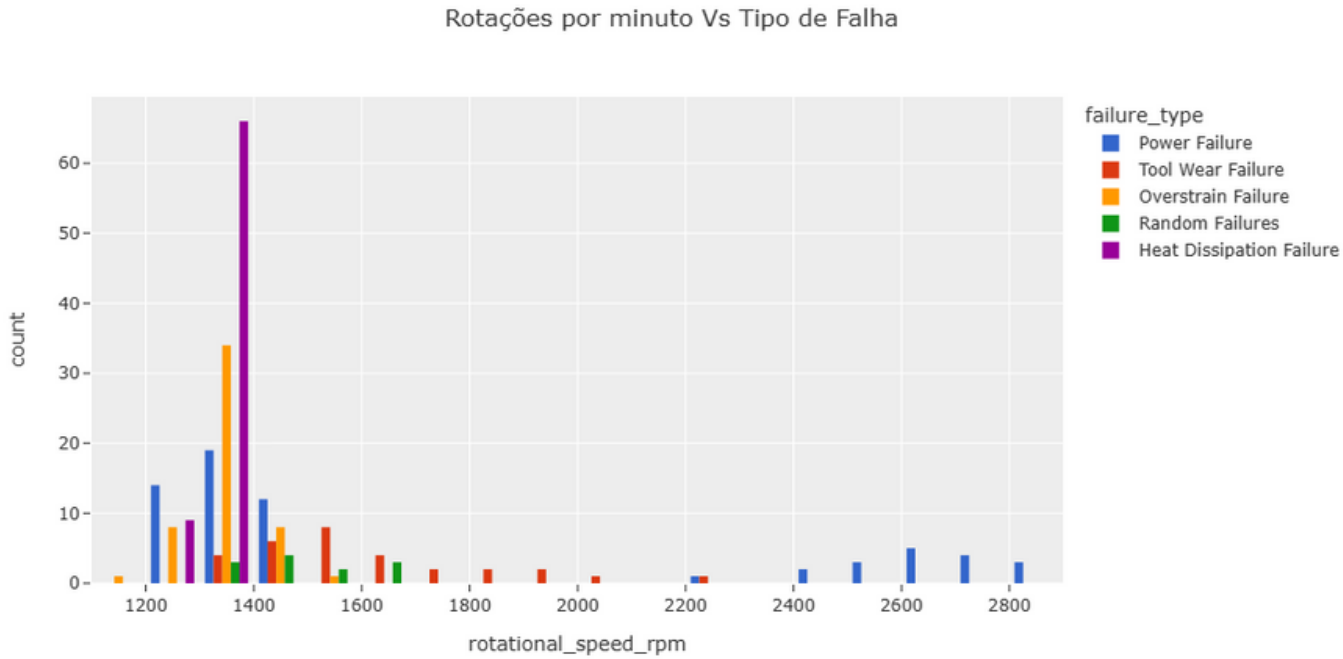


- Máquinas do tipo "L" ocorrem mais desgaste por falhas de "Overstrain Failure", "Power Failure" e "Tool Wear Failure". Povavelmente máquinas do tipo "L" precisam realizar manutenções com maior frequência.

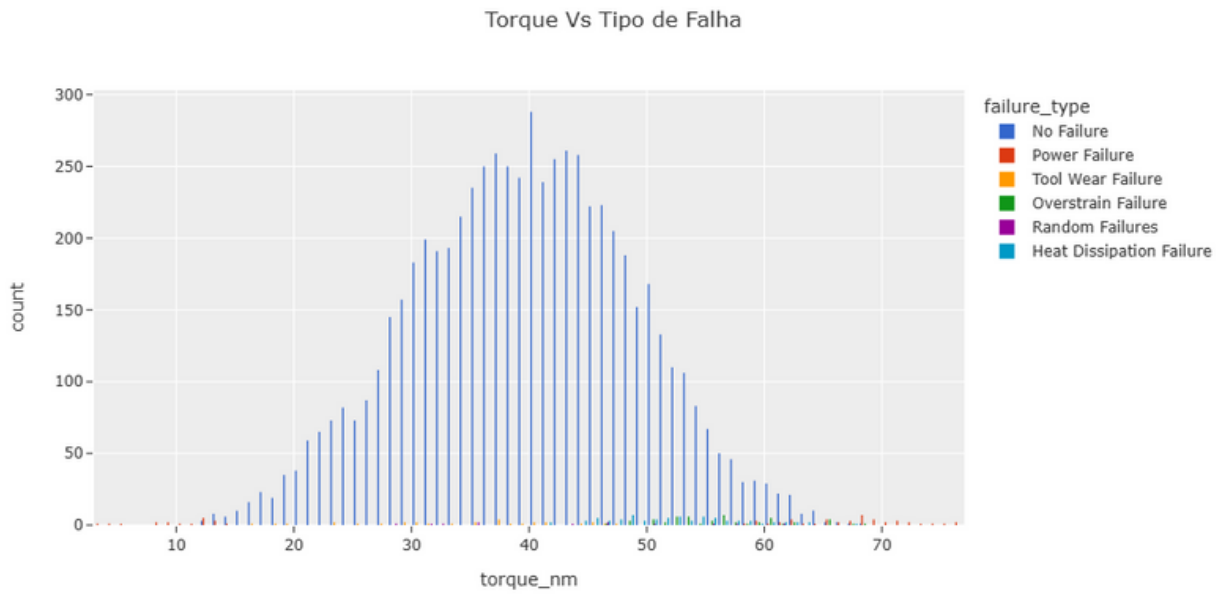


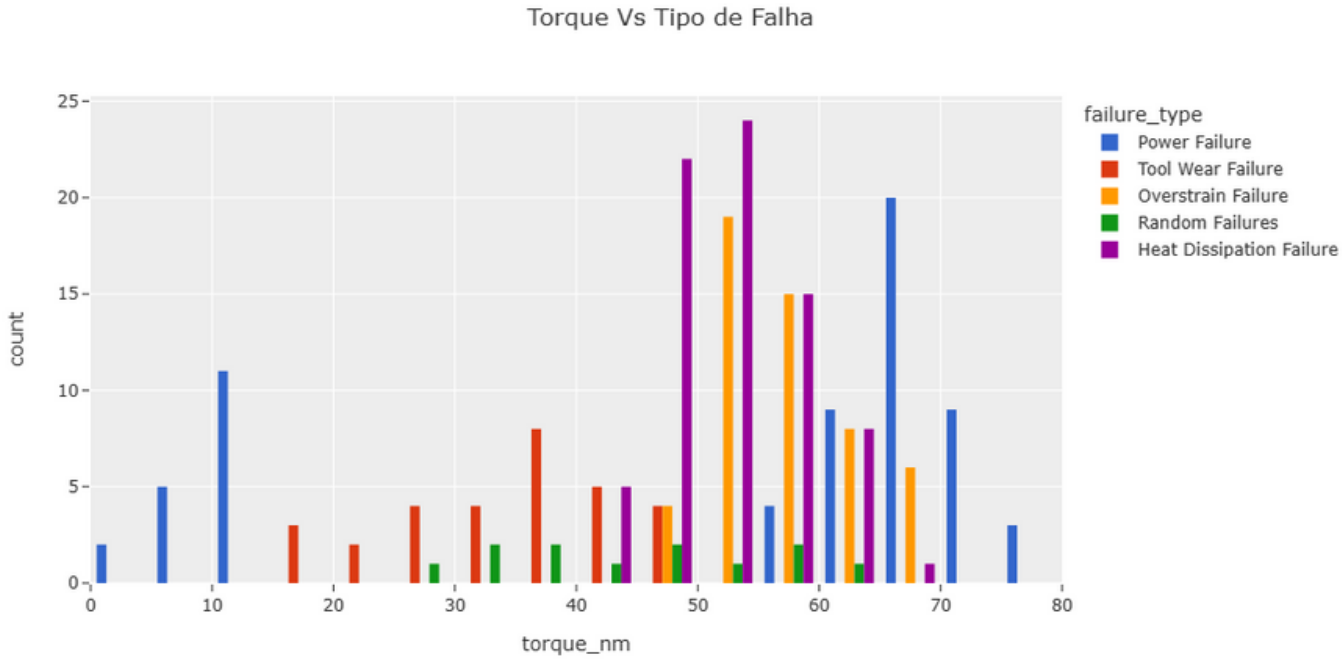
Falhas mais comuns

- As falhas mais comuns de ocorrer são por falha de energia, falha no overstrain e falha de dissipação de calor.

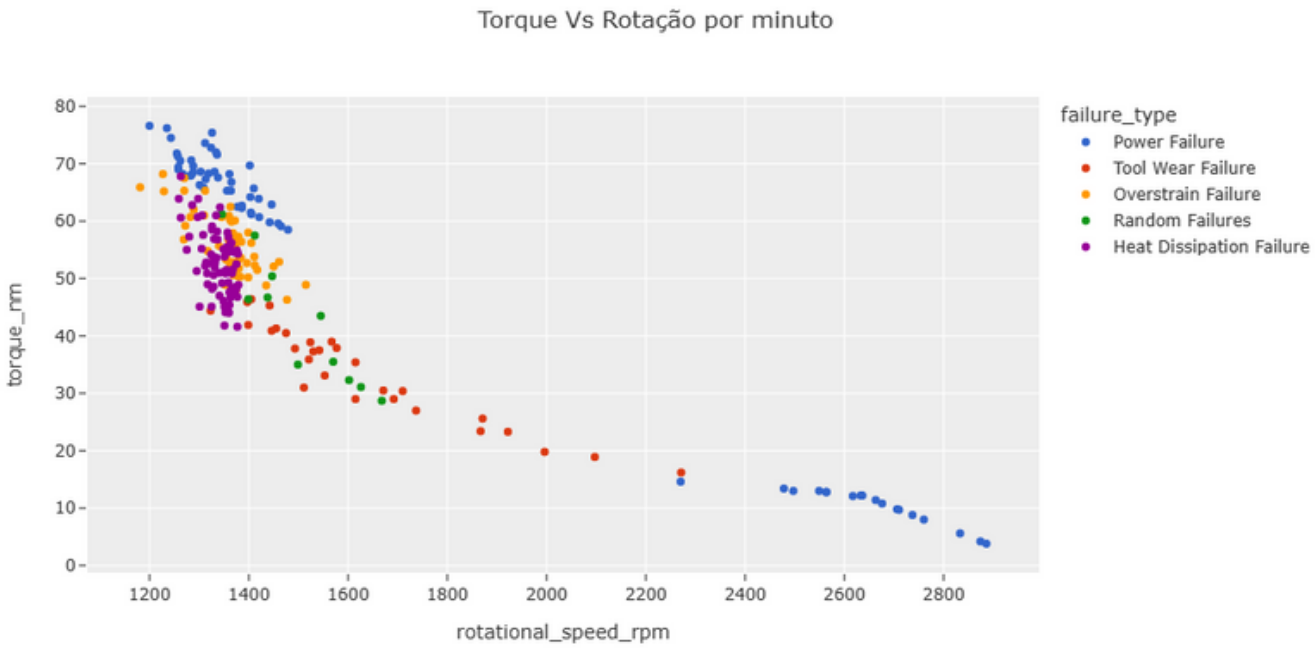


- Falhas do tipo "Heat Dissipation Failure" ocorreram em máquinas com rotação entre 1200 - 1400 rotações por minuto
- Falhas do tipo "Overstrain Failure" ocorreram em máquinas com rotação entre 1100 - 1600 rotações por minuto
- Falhas do tipo "Power Failure" ocorreram em máquinas com rotação entre 1200 - 1499 e 2200 - 2899 rotações por minuto
- Falhas do tipo "Random Failure" ocorreram em máquinas com rotação entre 1300 - 1699 rotações por minuto
- Falhas do tipo "Tool Wear Failure" ocorreram em máquinas com rotação entre 1300 - 2299 rotações por minuto

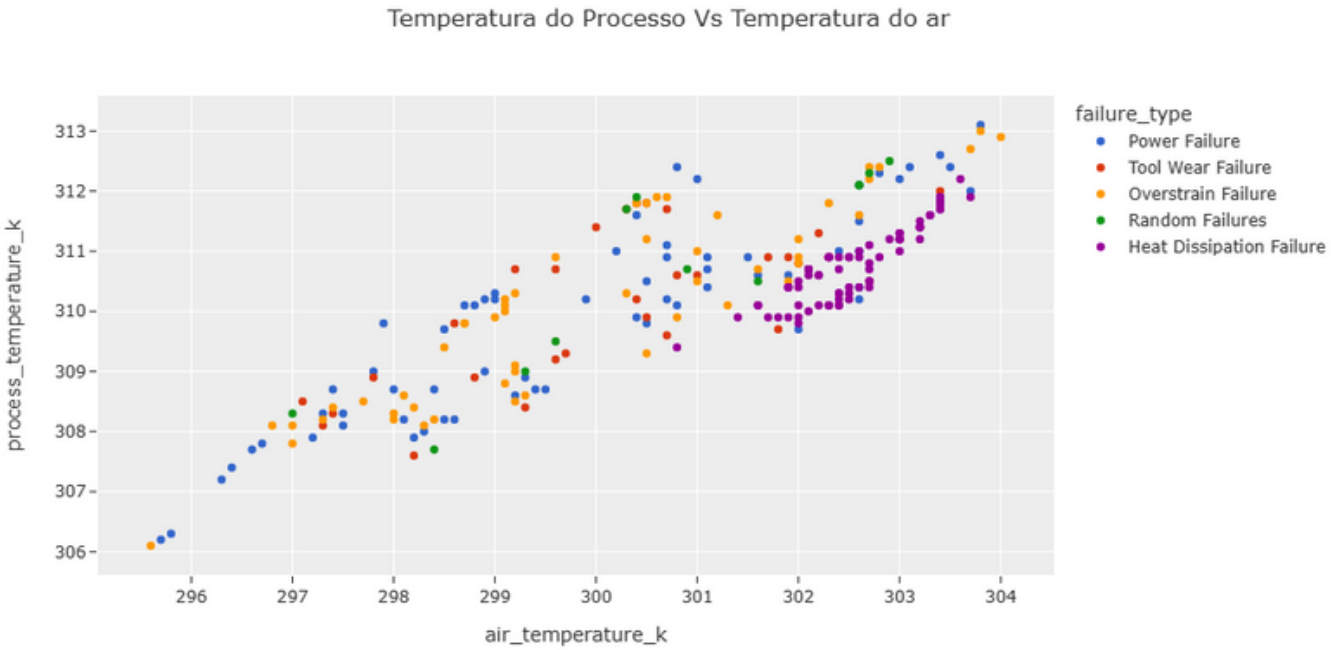




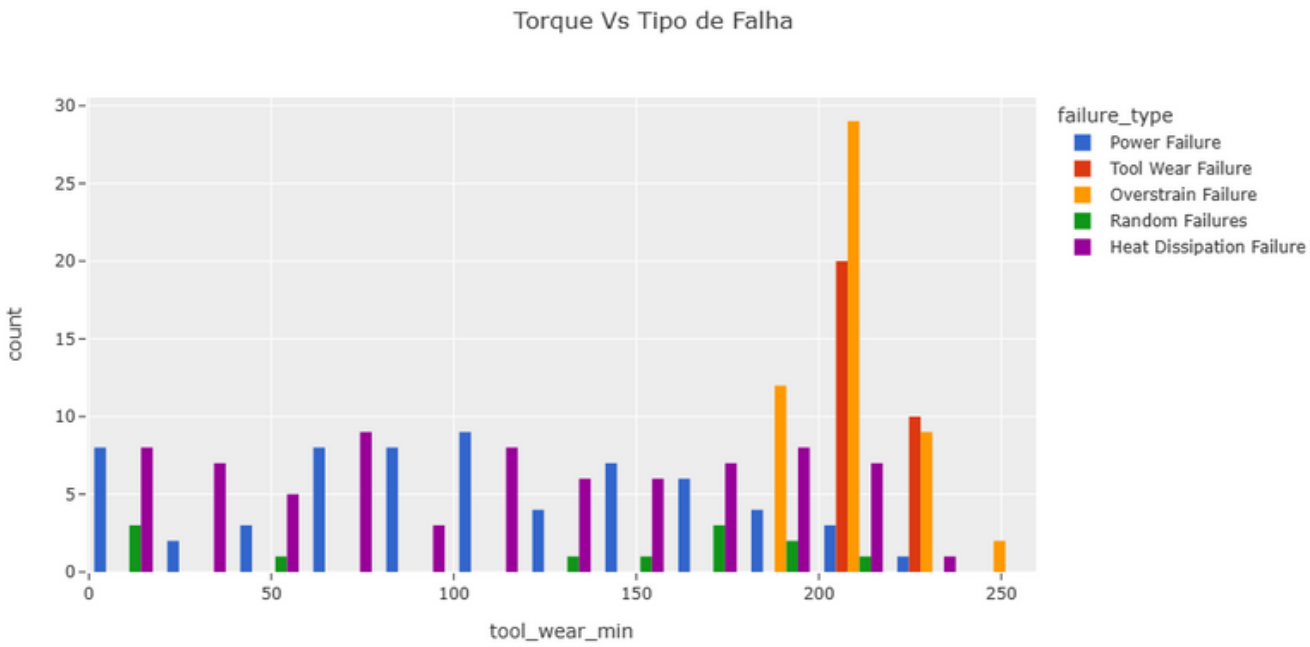
- Temos uma distribuição normal nos valores de Torque;
- Grande parte das falhas ocorreram com o torque acima dos 40;
- Falhas por dicipação de calor ocorreram com torque entre 40 - 69;
- Falhas do tipo "Overstrain Failure", ocorreram com torque entre 45 - 69;
- Falhas do tipo "Tool Wear Failure", ocorreram com torque entre 15 - 45.



- Grande parte das falhas ocorrem quando o torque está acima dos 40 e a velocidade está abaixo dos 1.600 rotações por minuto.
- Falhas do tipo "Power Failure" ocorreram quando o toque estava acima dos 55 e a velocidade de rotação estava abaixo dos 1.500 por minuto. E também ocorreram quando o torque estava abaixos dos 15 e acima dos 2.200 rotações por minuto



- Falhas por problemas de dissipação de calor se concentra quando a temperatura do ar fica acima dos 301.000, e a temperatura do processo está acima dos 309.000.



- Falhas do tipo "Overstrain Failure" implicam entre 180 - 250 minuto de desgaste.
- Falhas do tipo "Tool Wear Failure" implicam entre 200 - 240 minuto de desgaste.

MÉTRICA DE AVALIAÇÃO

Para a situação problema, foi utilizado como métrica de avaliação a Curva ROC (Receiver Operating Characteristic Curve). A Curva ROC tem o objetivo de mostrar o desempenho de um modelo de Machine Learning, com base na relação da Taxa de verdadeiro positivo(sensibilidade), e da Taxa de falso positivo (especificidade). A Curva ROC permite encontrar o ponto em que existe otimização da sensibilidade em função da especificidade. O ponto de otimização se encontra mais próximo do canto superior esquerdo do gráfico que exibe a Curva ROC.

O foco do gráfico é analisar o poder preditivo de um modelo, e assegurar que o modelo de machine learning vai detectar o máximo possível de verdadeiros positivos, enquanto minimiza os falsos positivos, quanto maior a taxa de sensibilidade e a taxa de especificidade (próximos de 100%), melhor a eficiência do modelo de machine learning.

Para a estimativa com o objetivo de predizer qual tipo de falha irá apresentar, foi implementado um modelo utilizando o algoritmo de **Regressão Logística** que atingiu uma performance de AUC em **95,14%** de eficiência.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lr	Logistic Regression	0.6450	0.9514	0.7281	0.9703	0.7615	0.1207	0.2457	5.0040
rf	Random Forest Classifier	0.9550	0.9508	0.5985	0.9748	0.9640	0.5061	0.5246	2.3140
lightgbm	Light Gradient Boosting Machine	0.9621	0.9492	0.6157	0.9790	0.9700	0.5606	0.5746	1.1580
et	Extra Trees Classifier	0.9586	0.9450	0.5394	0.9713	0.9644	0.5019	0.5107	0.8500
gbc	Gradient Boosting Classifier	0.9256	0.9396	0.6640	0.9765	0.9489	0.4026	0.4566	21.4320
lda	Linear Discriminant Analysis	0.5393	0.9336	0.7179	0.9703	0.6757	0.0882	0.2069	0.0460
nb	Naive Bayes	0.5487	0.9245	0.6556	0.9689	0.6855	0.0853	0.2003	0.0340
knn	K Neighbors Classifier	0.8789	0.8551	0.5672	0.9646	0.9164	0.2562	0.3220	0.4120
dt	Decision Tree Classifier	0.9460	0.8145	0.5570	0.9711	0.9577	0.4427	0.4649	0.1380
qda	Quadratic Discriminant Analysis	0.6745	0.7554	0.6368	0.7805	0.7189	0.2052	0.2914	0.0440
ada	Ada Boost Classifier	0.2133	0.6759	0.4832	0.7904	0.2304	0.0446	0.1050	0.9980
dummy	Dummy Classifier	0.0113	0.5000	0.1667	0.0001	0.0003	0.0000	0.0000	0.0240
svm	SVM - Linear Kernel	0.6325	0.0000	0.7202	0.9694	0.7502	0.1159	0.2377	0.1280
ridge	Ridge Classifier	0.4836	0.0000	0.7149	0.9681	0.6216	0.0766	0.1919	0.0320

Regressão Logística

Vantagens	Desvantagens
A regressão logística é um algoritmo simples de ser implementado e costuma ser bem eficiente.	A principal limitação da Regressão Logística é a suposição de linearidade entre a variável dependente e as variáveis independentes.
É rápido para classificar registros desconhecidos.	A regressão logística requer média ou nenhuma multicolinearidade entre as variáveis independentes.
Não é um algoritmo tão complexo.	É um algoritmo com um histórico conhecido de vulnerabilidade ao sobreajuste.

Gráfico do desempenho da curva ROC usando regressão logística

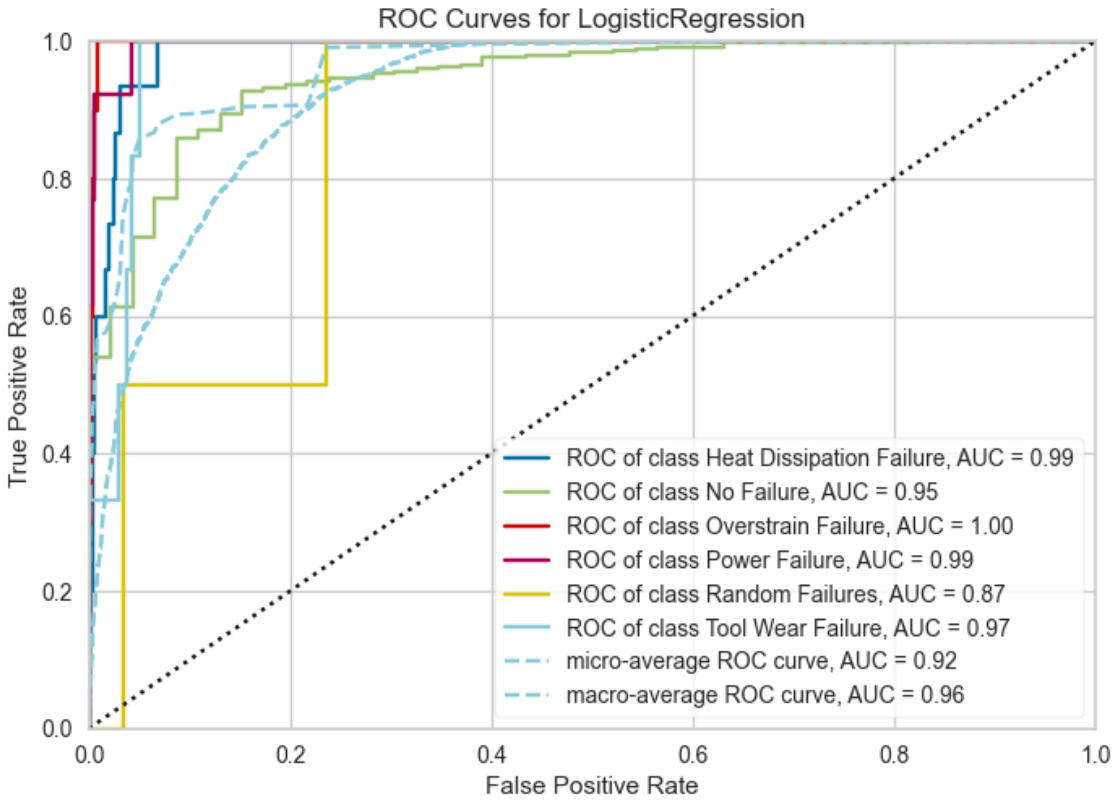
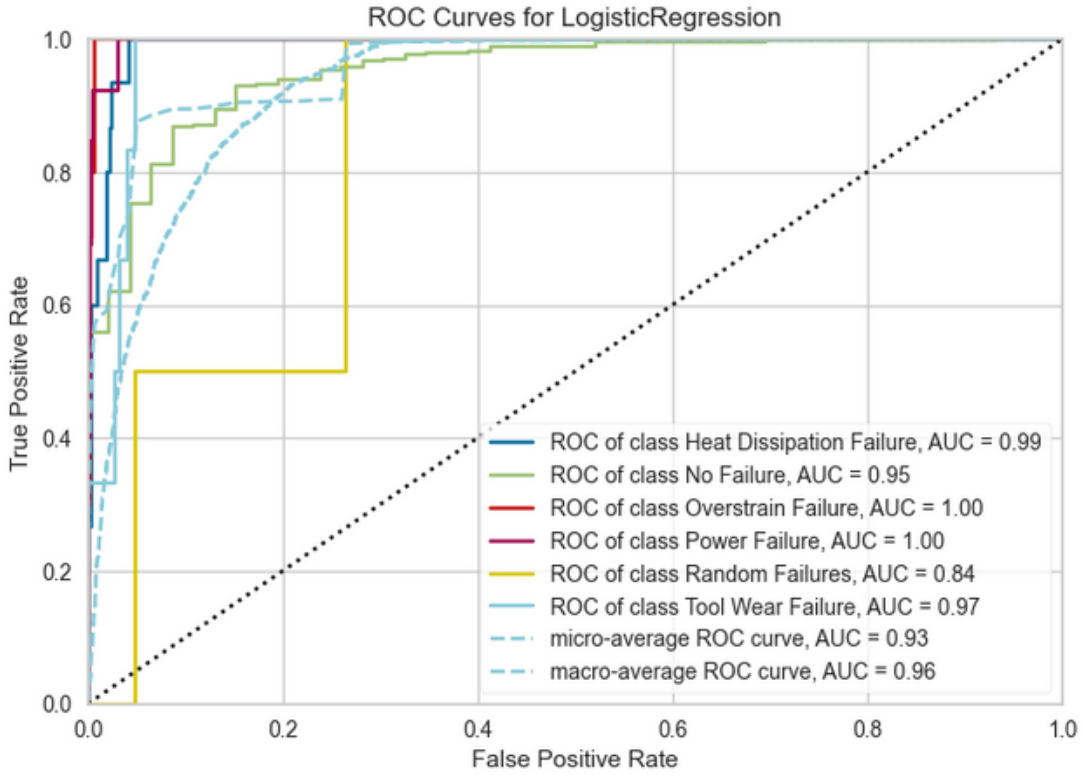


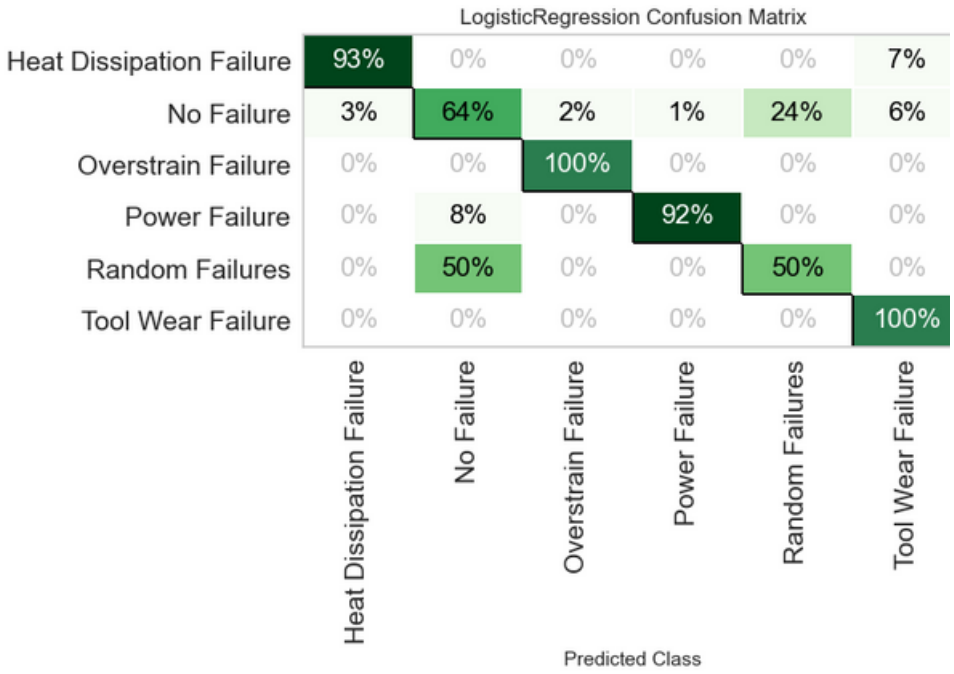
Gráfico do desempenho da curva ROC usando regressão logística com modificações nos parâmetros

- Podemos concluir que não houve uma melhora significativa



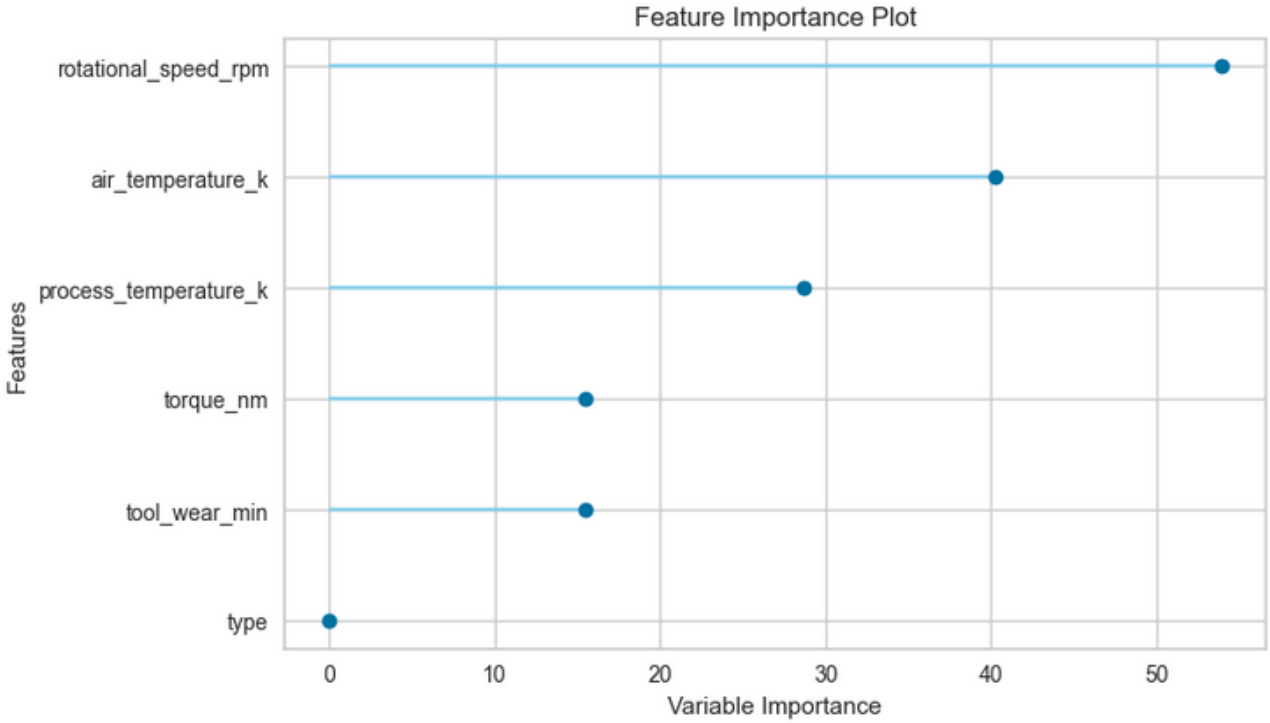
Matriz de confunsão

- A classe "Random Failures" foi classificada 50% das vezes errada, indica que precisa passar por trabalhos futuros;
- A classe "No Failure" foi classificada 36% das vezes errada, indica que precisa passar por trabalhos futuros.



O gráfico abaixo apresenta a importância de cada feature. Quanto maior, mais o modelo leva em consideração para precer determinada classe.

- As features que o modelo está levando mais em consideração são "rotational_speed_rpm", "air_temperature_k" e "Process_temperature_k"



CONCLUSÃO

Através desse projeto foi possível praticar e implementar conceitos importantes da Ciência de Dados, e propor uma solução para um problema latente e recorrente de em várias empresa, através dados de sensores.

Como um processo de melhoria contínua podemos desenvolver uma automação para executar o pipeline de transformação de dados e automatizar a etapa de Machine Learning.

Contato


João Victor Soares Saraiva


Victorjoaosoares88@gmail.com


https://bit.ly/3hNBU9t


https://bit.ly/36k88nF