

Aprendizagem Supervisionada Aula 5

Revisão Trabalho Final

Bernard da Silva
Orientador: Rafael S. Parpinelli
05/12/2022

O que foi visto até agora?

- **Aula 1:**
 - Tipos de aprendizados supervisionados (regressão e classificação)
 - Transformação de dados
- **Aula 2:**
 - Análise exploratória de dados
 - Pré-processamento de dados
- **Aula 3:**
 - Seleção de algoritmos
 - Introdução RNA
- **Aula 4:**
 - Escalonamento de dados
 - Separação base de dados
 - Técnica de validação cruzada
 - Parâmetros de configuração RNA
 - Entendimento overfitting e underfitting

Objetivo dessa aula

- Revisão Trabalho Final
- Métricas de avaliação

O que será avaliado no trabalho?

Análise Exploratória de Dados



Data Cleaning



Pré-Processamento



Criação da Rede Neural



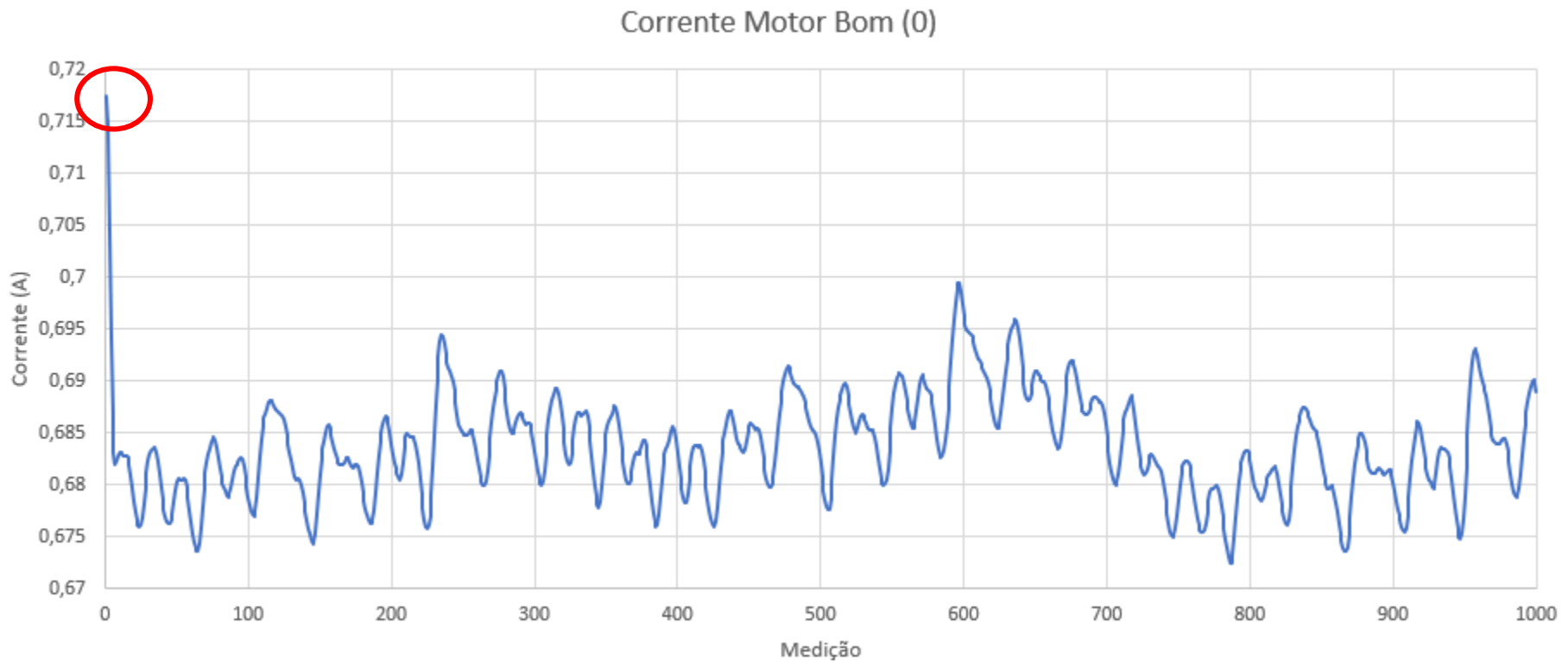
Verificação de overfit/underfit



Verificação das métricas

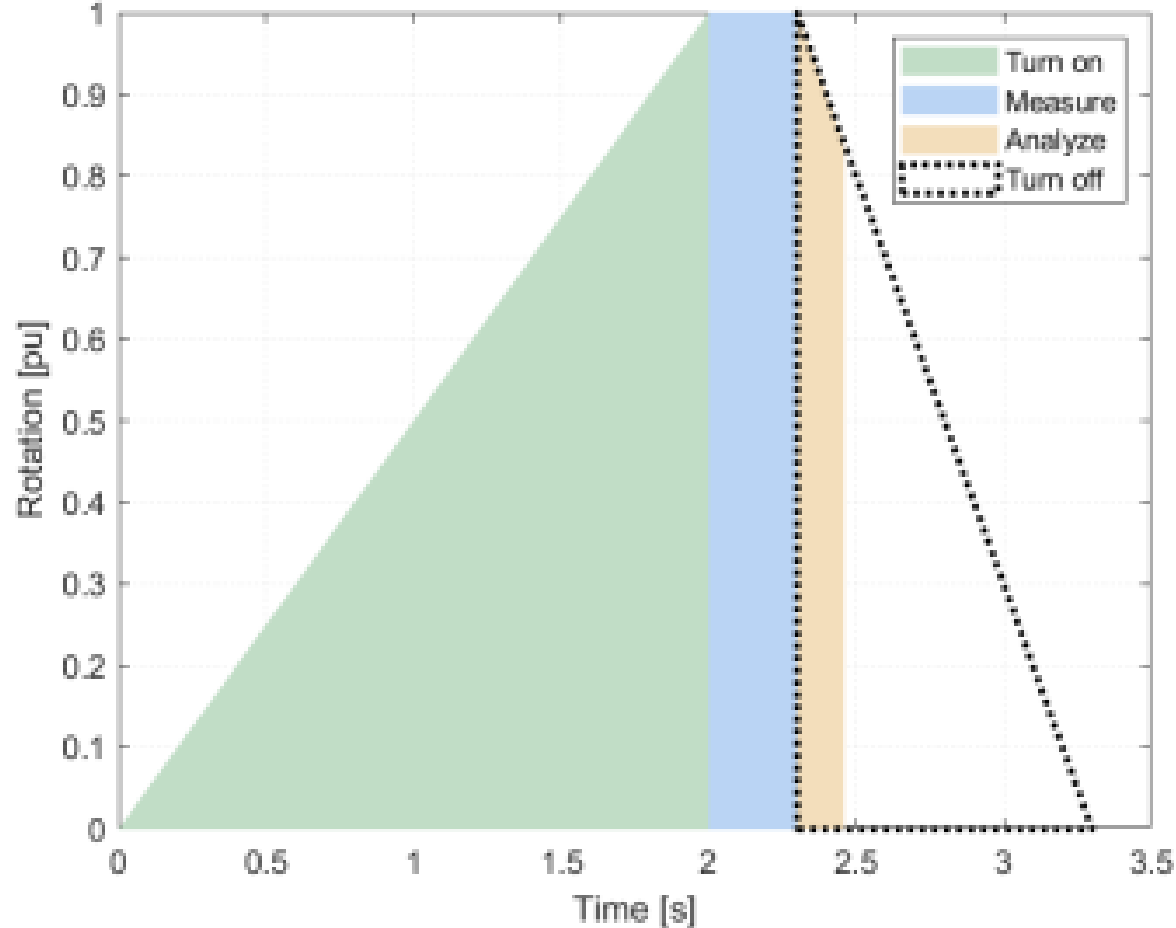
Análise Exploratória de Dados

- Outlier?



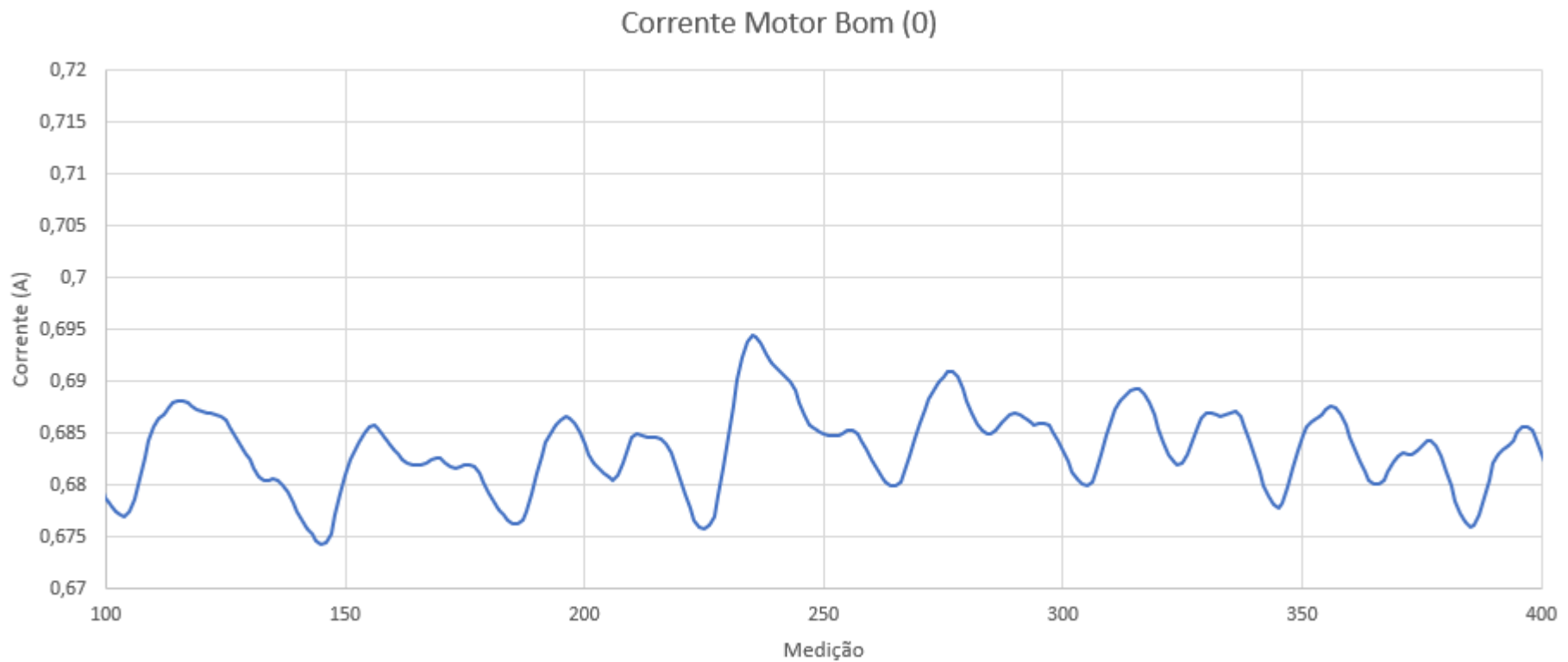
Análise Exploratória de Dados

Fig. 8. Test time



Análise Exploratória de Dados

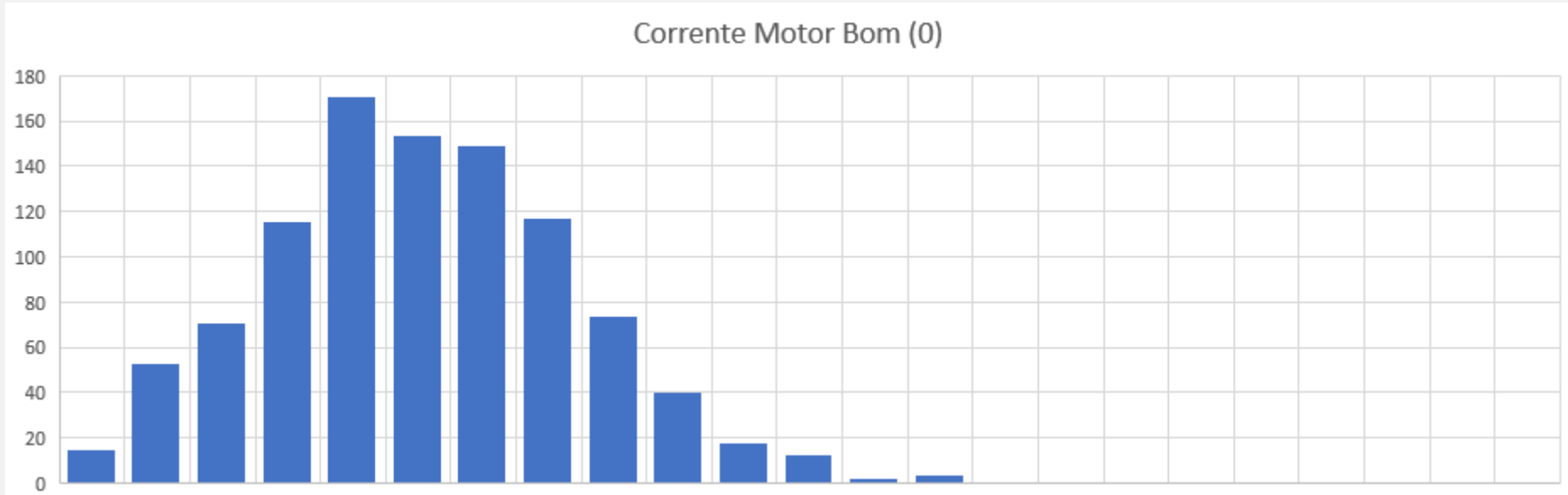
- Algum possível padrão?



Análise Exploratória de Dados

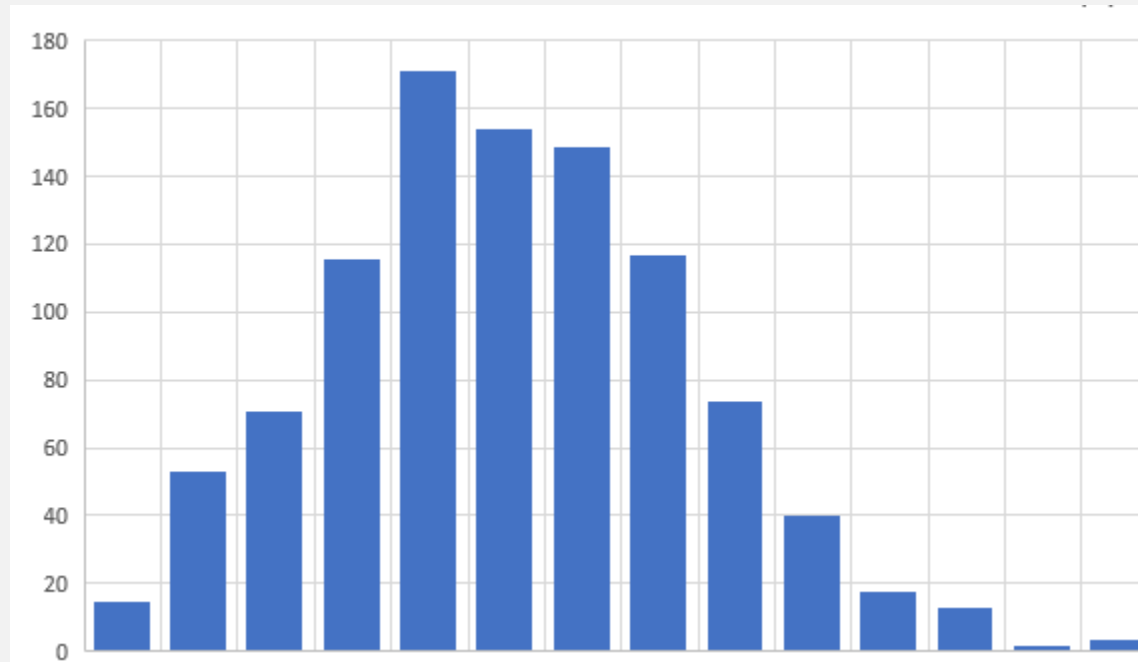
- Distribuição Normal? Unimodal ou Bimodal?

Uma curva simétrica em torno do seu ponto médio

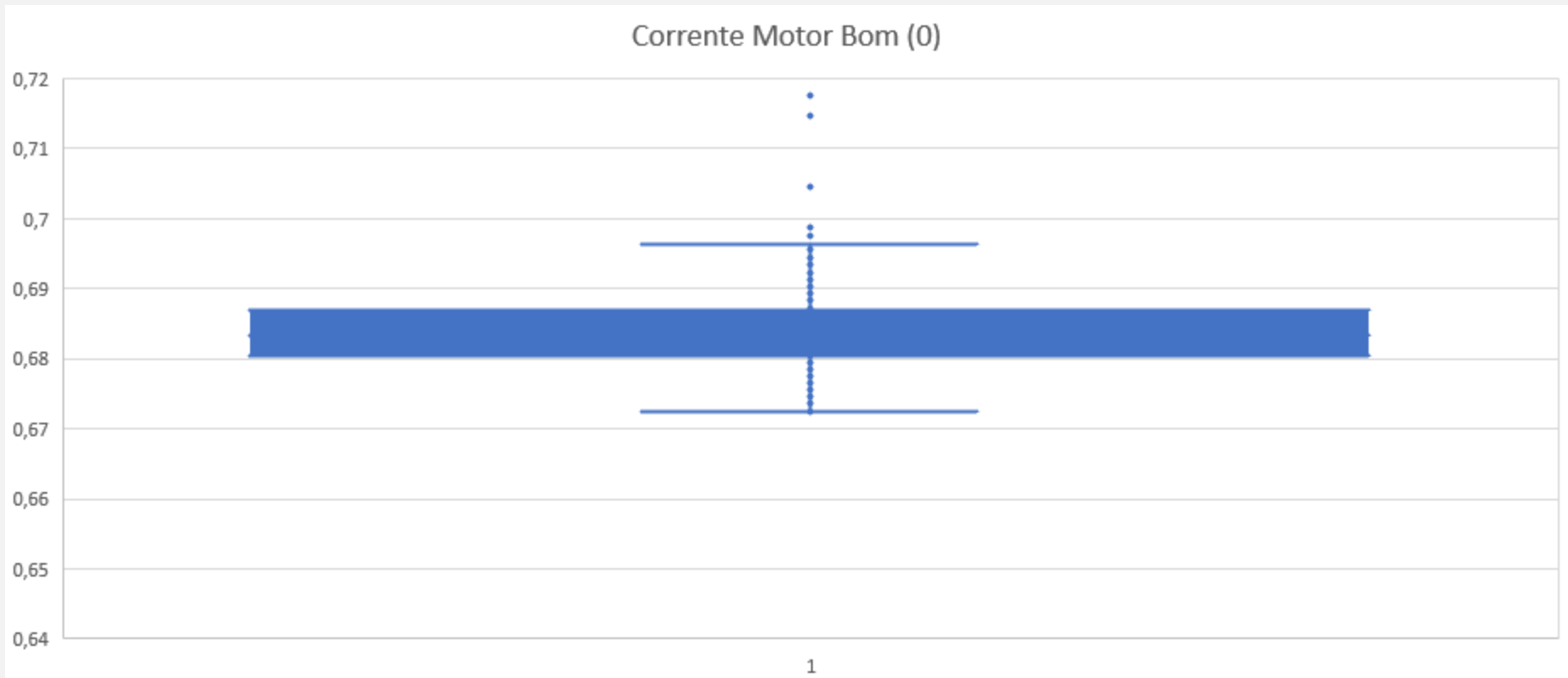


Análise Exploratória de Dados

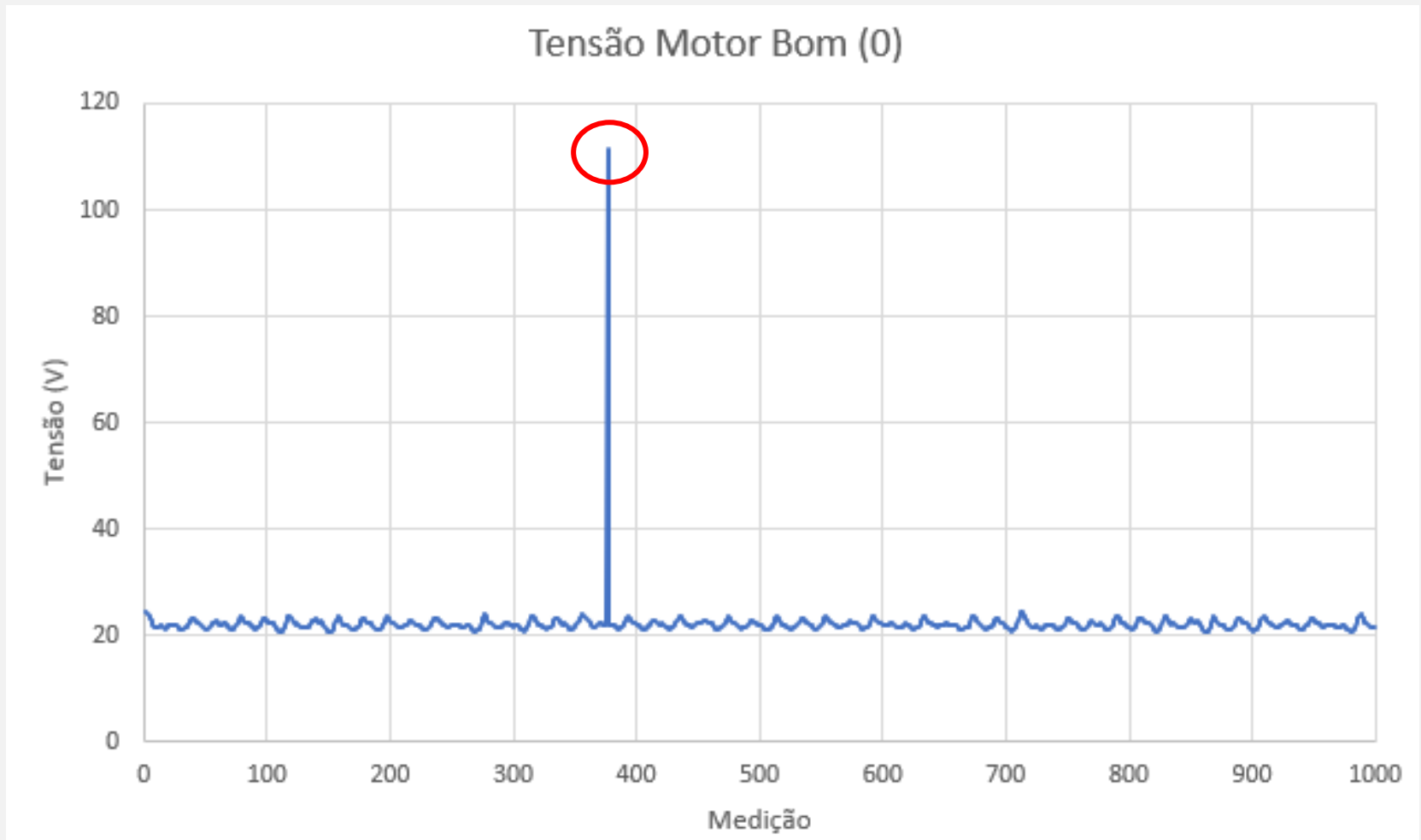
- Distribuição Normal?



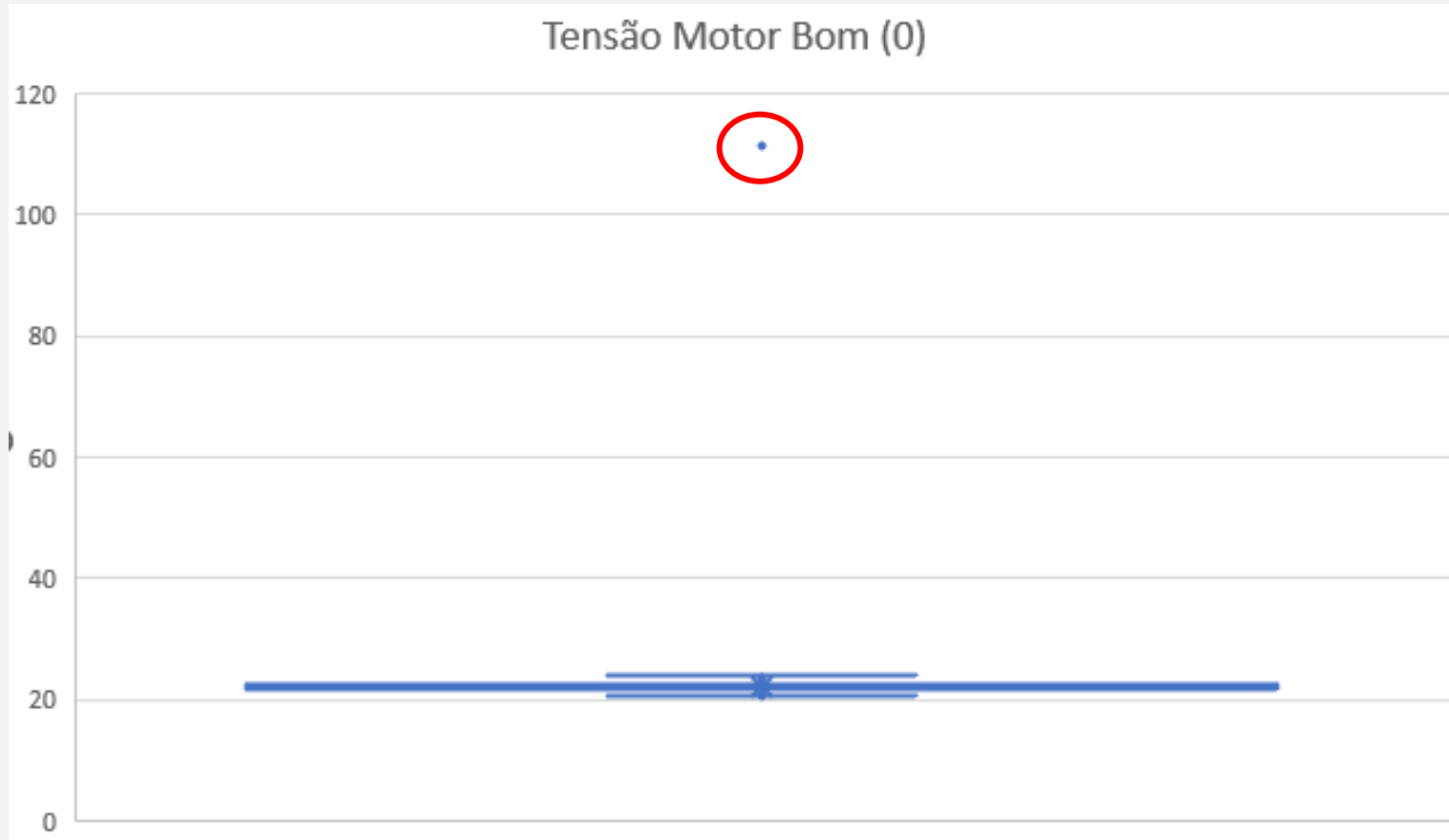
Análise Exploratória de Dados



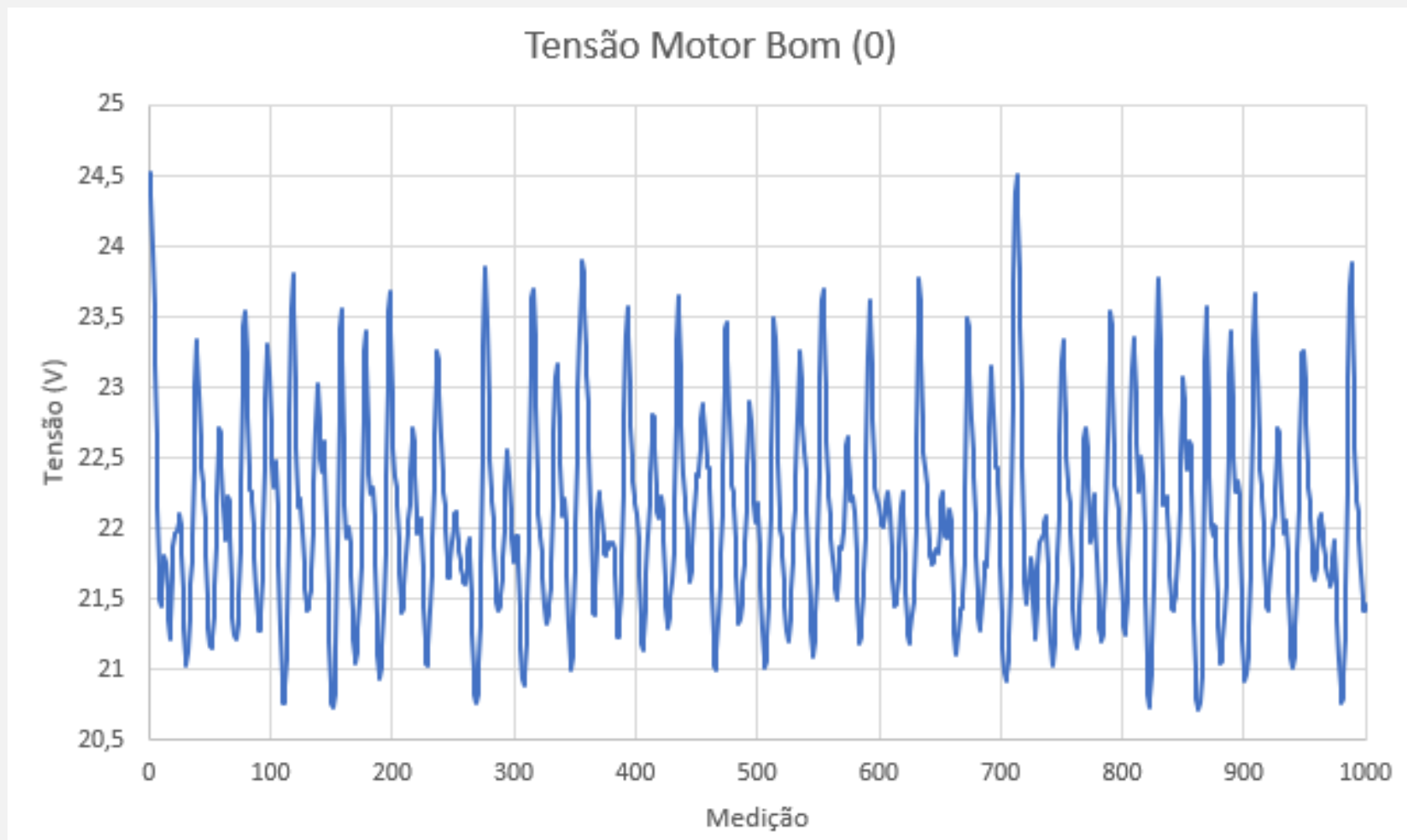
Análise Exploratória de Dados



Análise Exploratória de Dados

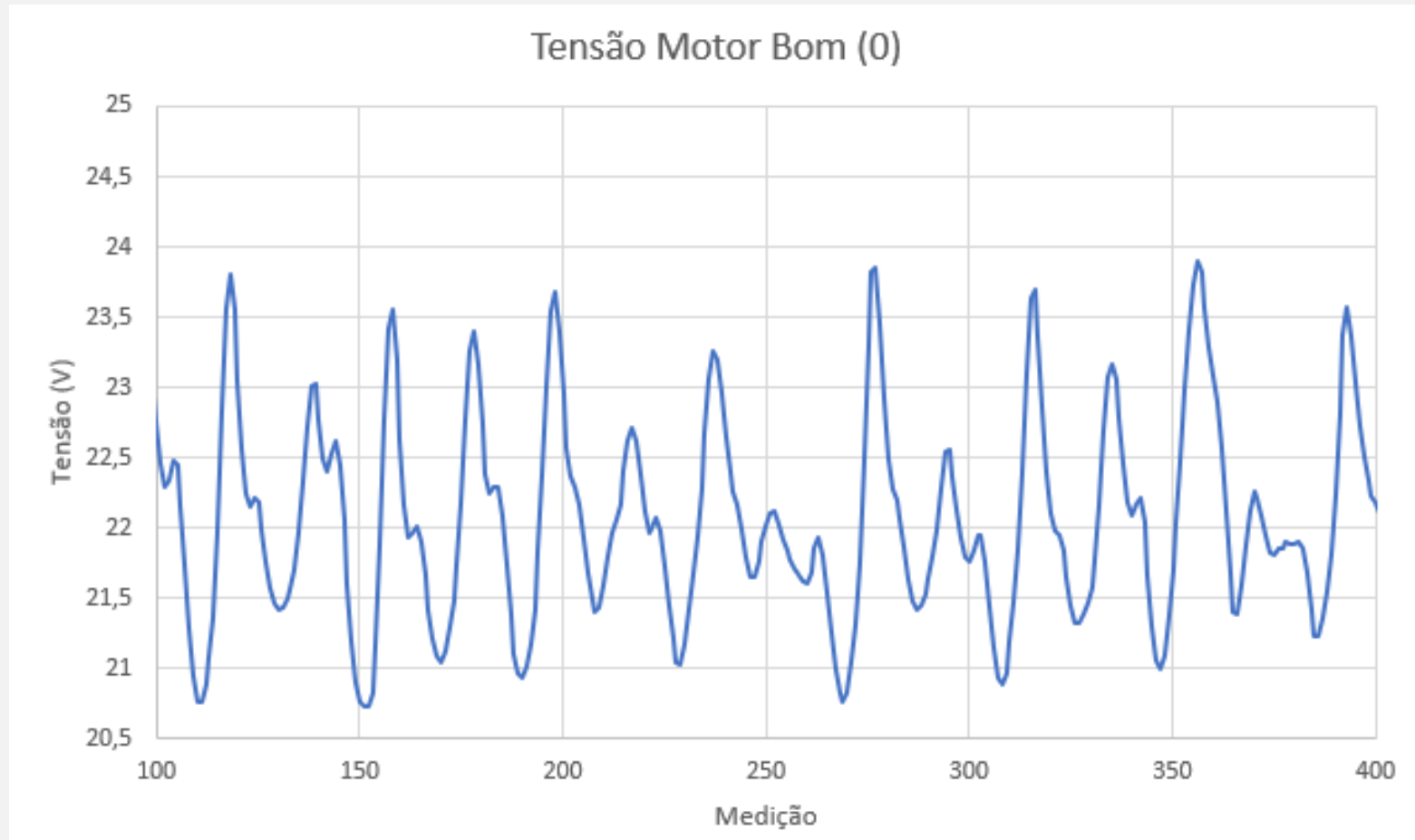


Análise Exploratória de Dados

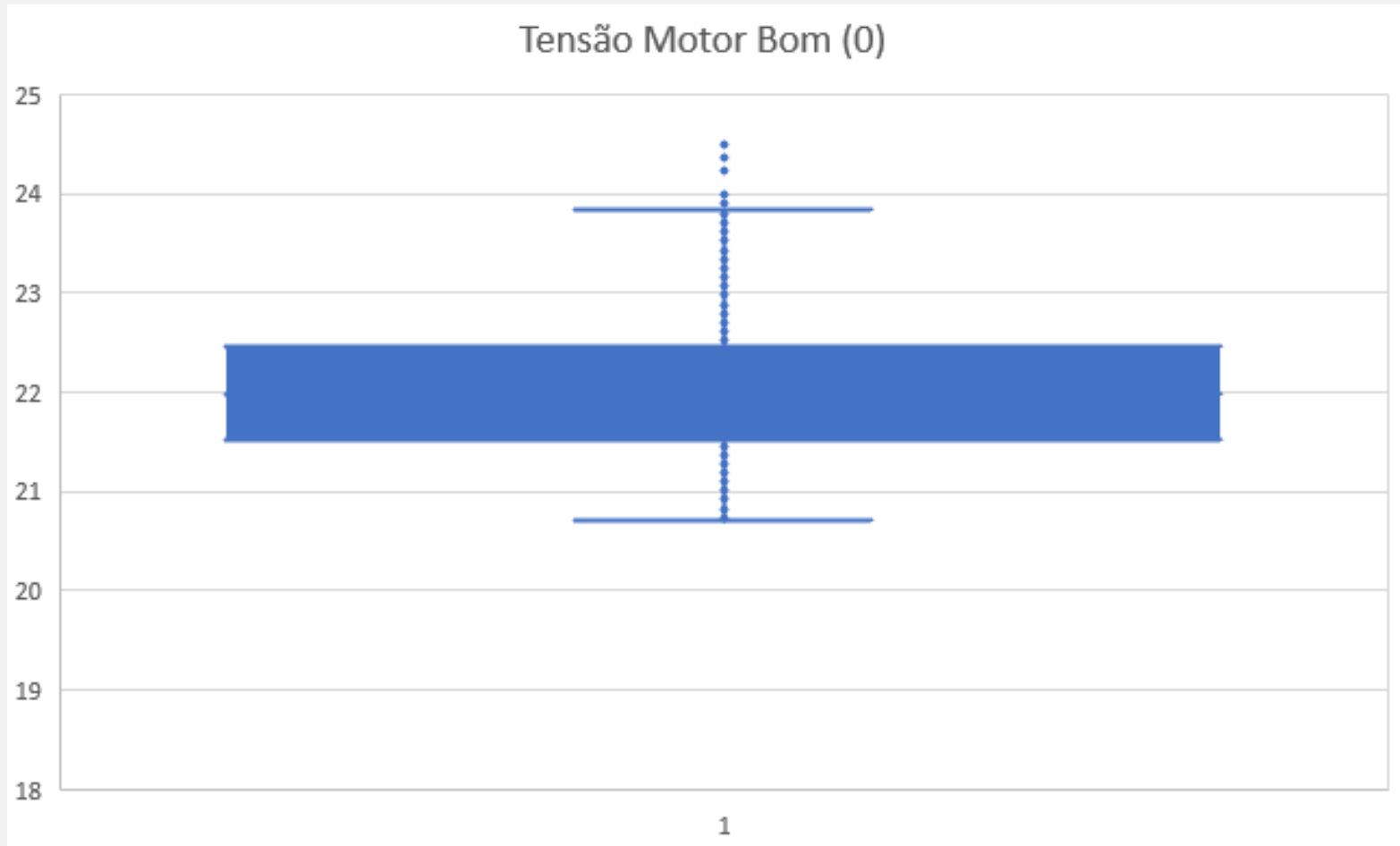


Análise Exploratória de Dados

- Algum possível padrão?

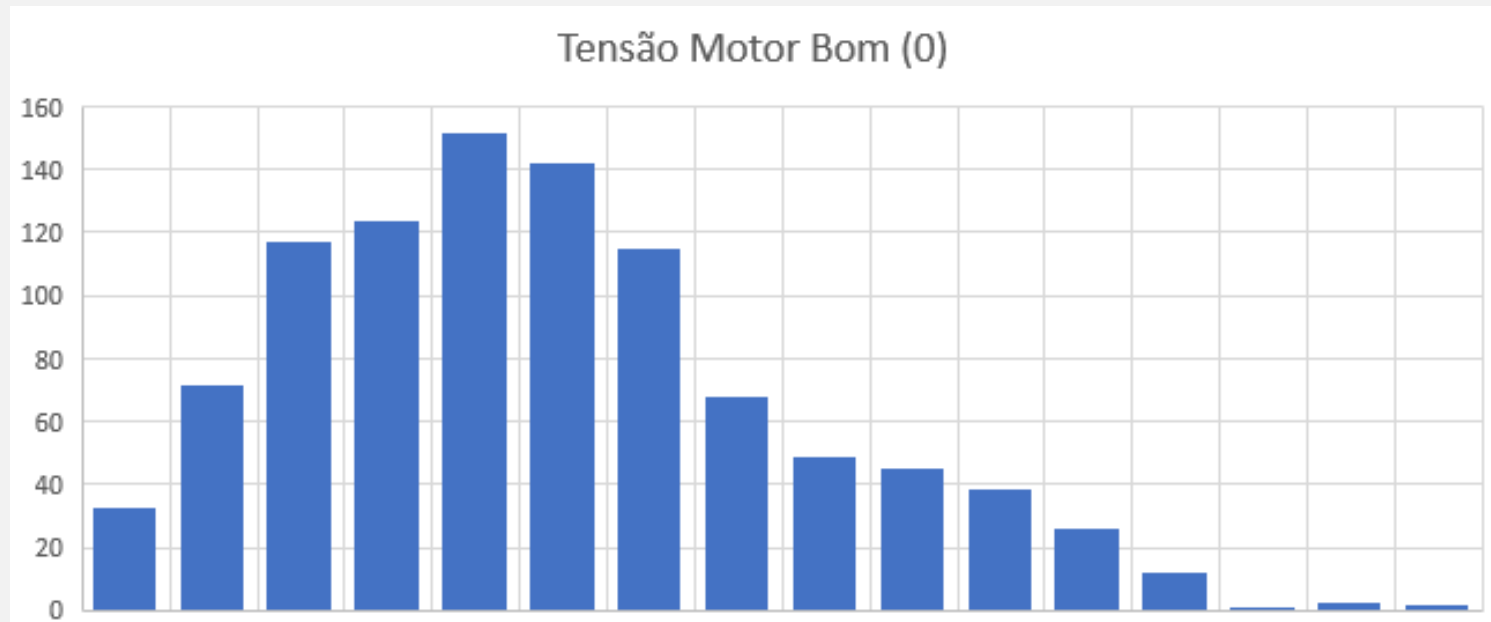


Análise Exploratória de Dados



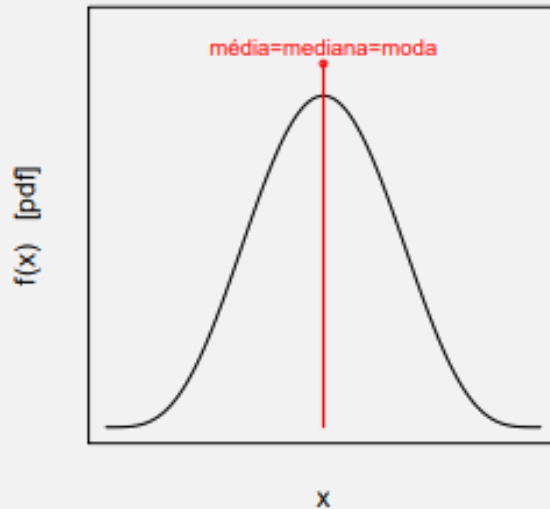
Análise Exploratória de Dados

- Distribuição Normal? Unimodal ou Bimodal?
- Ou
- Assimétrica a direita?

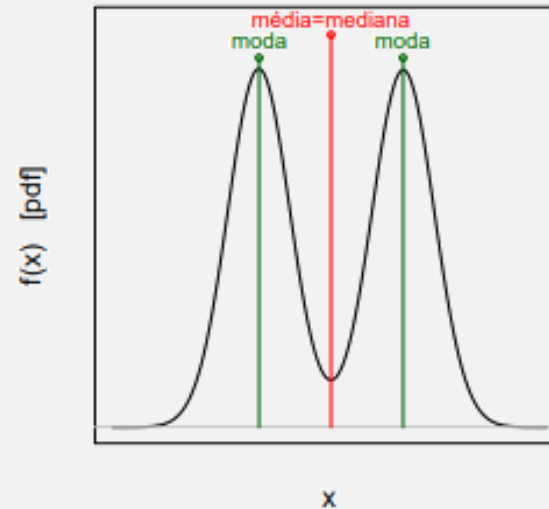


Análise Exploratória de Dados

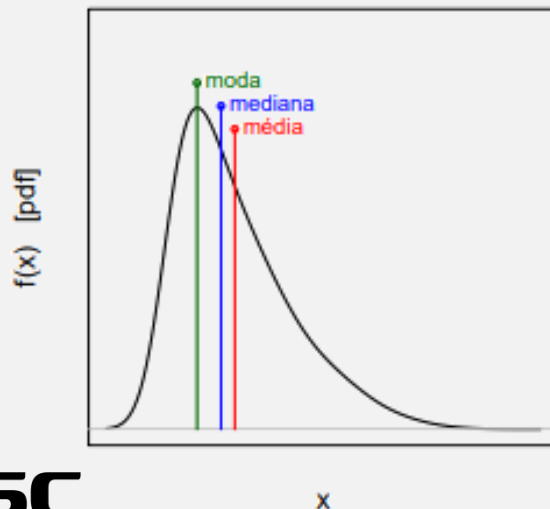
unimodal simétrica



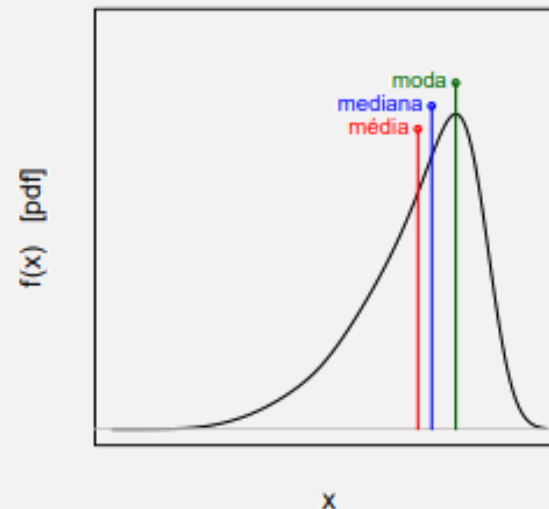
bimodal simétrica



assimétrica à direita



assimétrica à esquerda



Análise Exploratória de Dados

- O que descobrimos até o momento?
- Existem outliers na base de dados
- Corrente e Tensão não tem o mesmo comportamento (distribuição normal e assimétrica a direita)
- Escalas diferentes (corrente 0,68~0,69, tensão 21,5~22,5)
- Padrões identificados

Análise Exploratória de Dados

- Conclusões da análise prévia?

- Existem outliers na base de dados

Necessário limpeza da base de dados. Sugestões?

- Corrente e Tensão não tem o mesmo comportamento (distribuição normal e assimétrica a direita)

Escolha das medidas estatísticas mais adequadas.

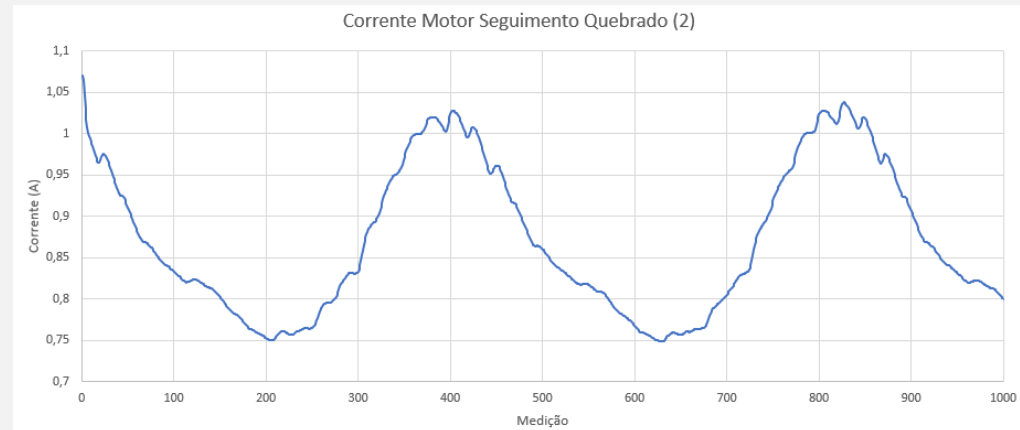
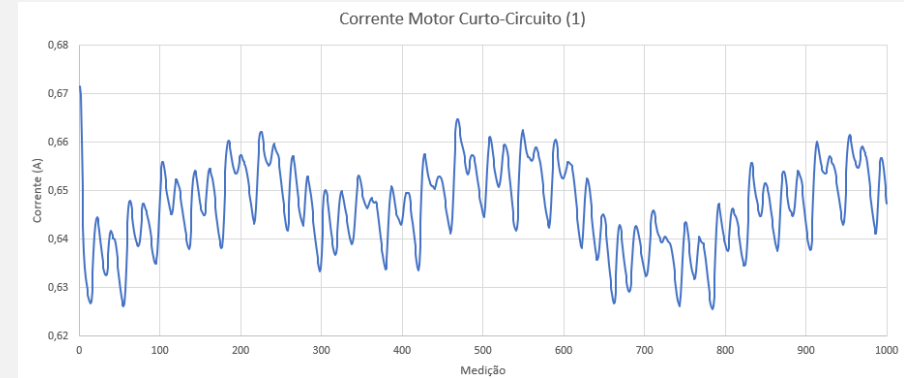
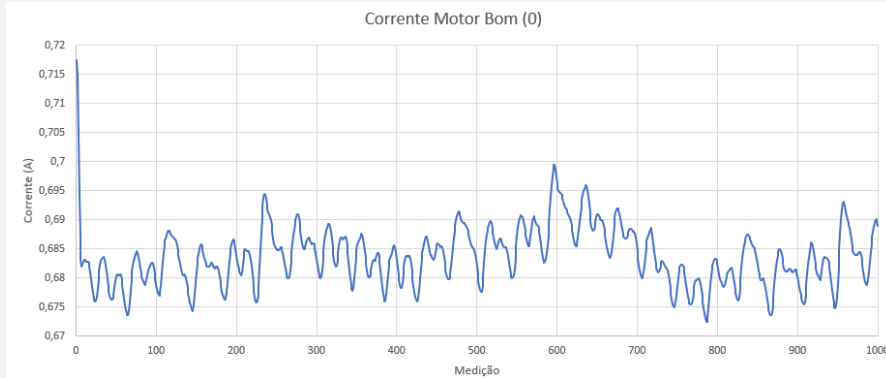
- Escalas diferentes (corrente 0,68~0,69, tensão 21,5~22,5)

Necessário escalonamento.

- Padrões identificados

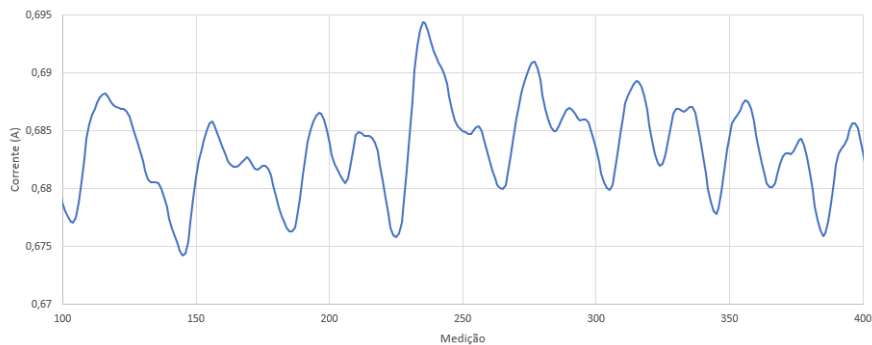
Possibilidade de redução no tamanho da base de dados.

Análise Exploratória de Dados

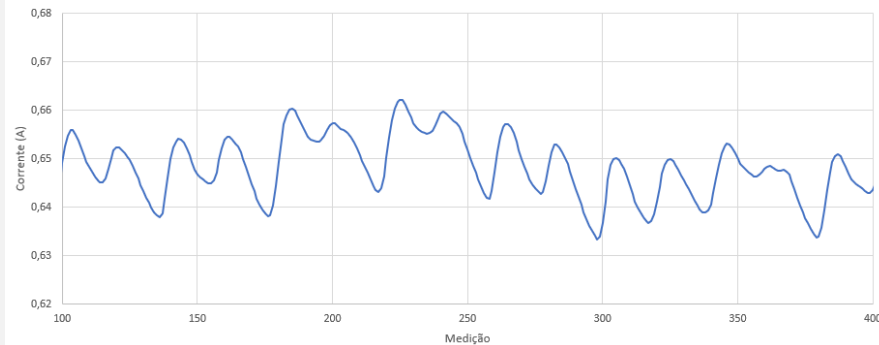


Análise Exploratória de Dados

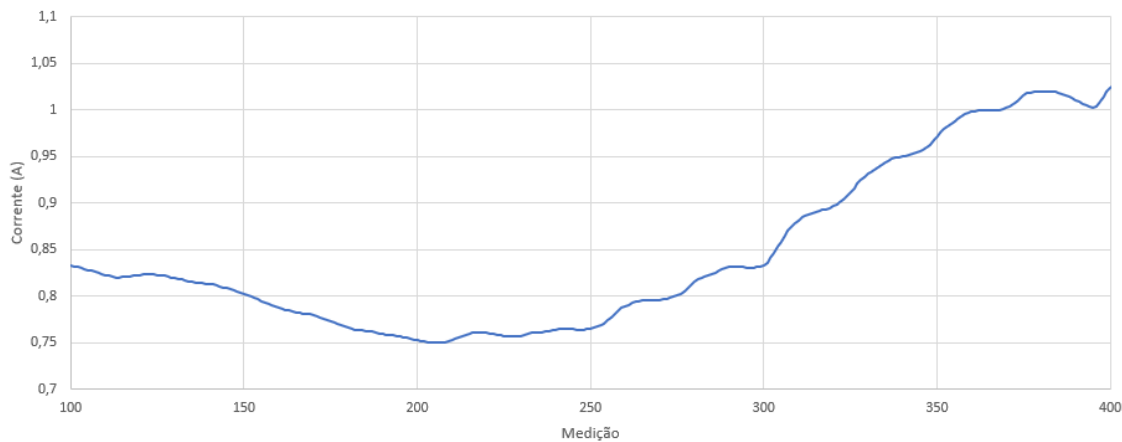
Corrente Motor Bom (0)



Corrente Motor Curto-Circuito (1)

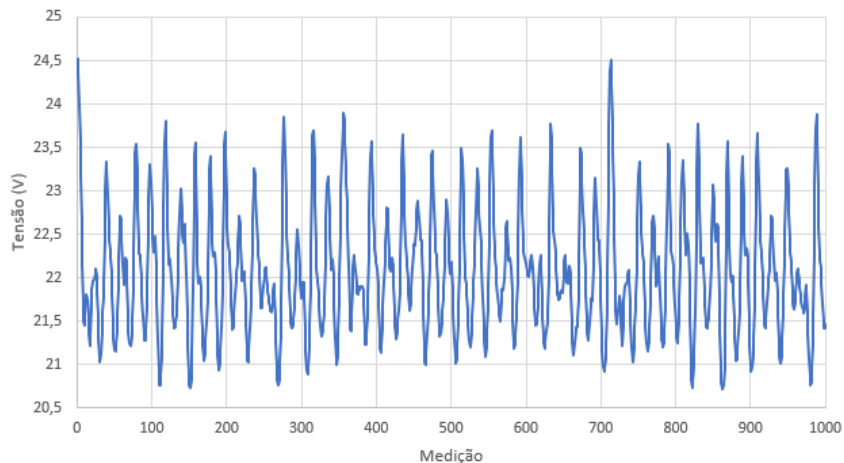


Corrente Motor Seguimento Quebrado (2)

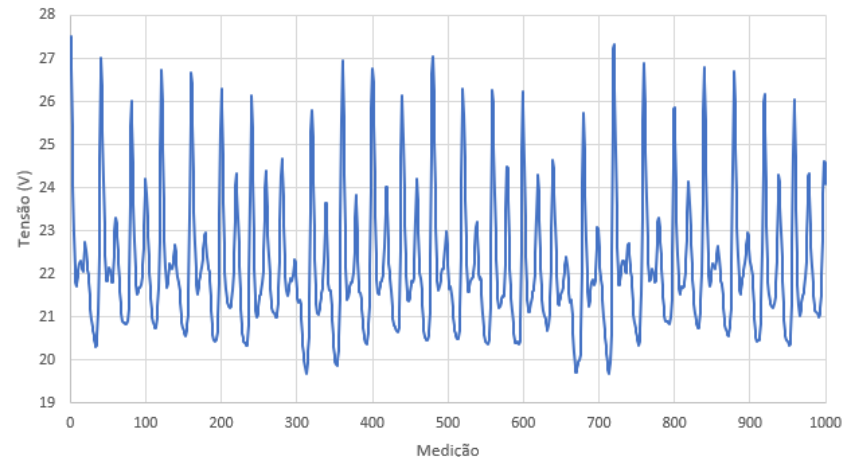


Análise Exploratória de Dados

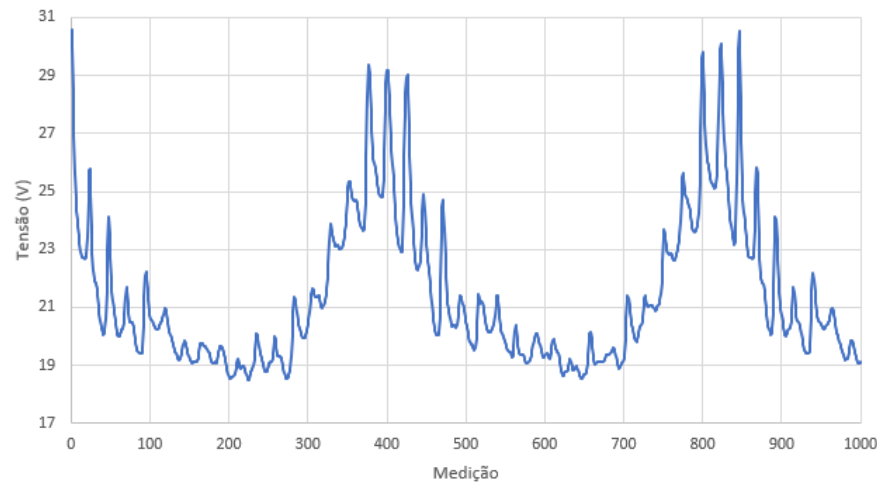
Tensão Motor Bom (0)



Tensão Motor Curto-Circuito (1)

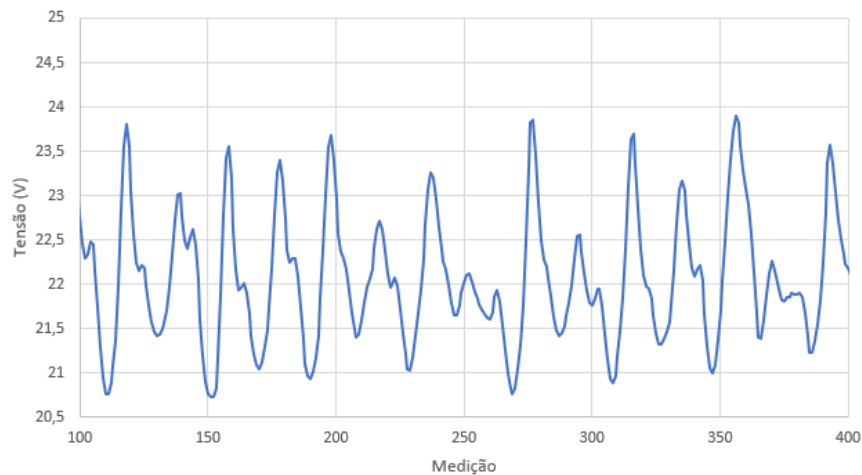


Tensão Motor Segmento Quebrado (2)

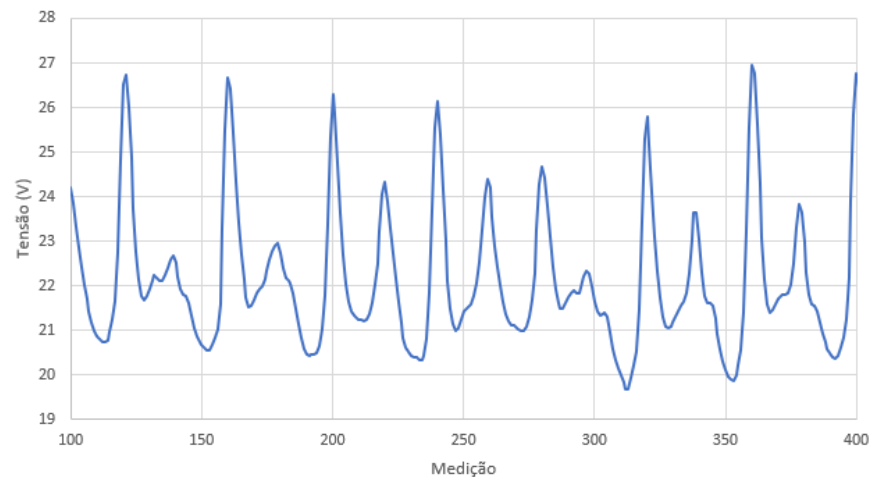


Análise Exploratória de Dados

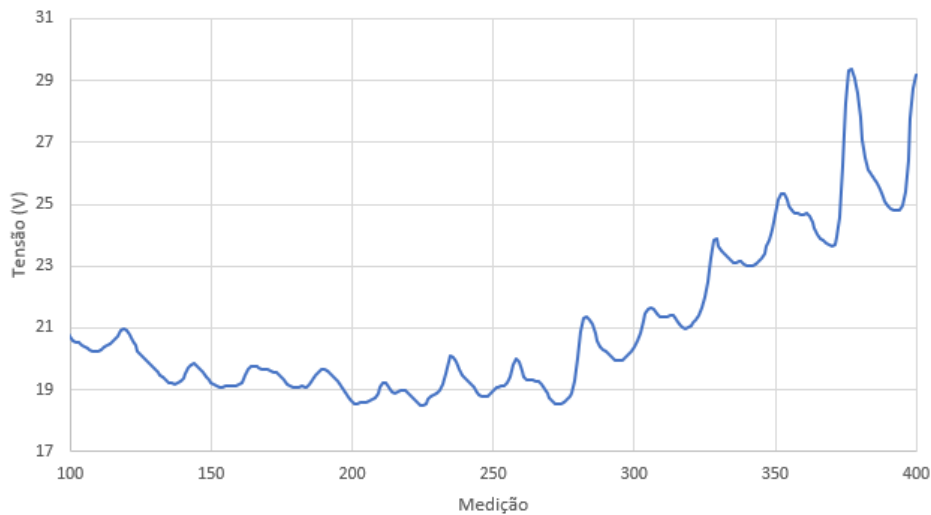
Tensão Motor Bom (0)



Tensão Motor Curto-Circuito (1)



Tensão Motor Segmento Quebrado (2)

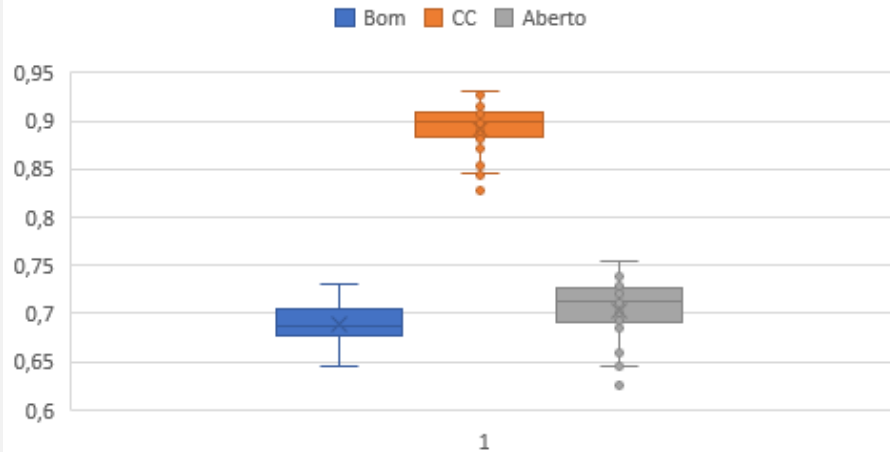


Análise Exploratória de Dados

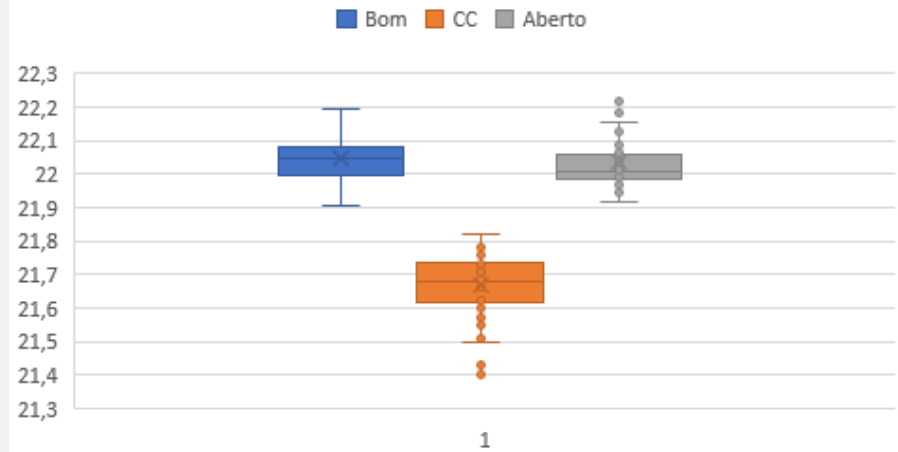
- O que descobrimos até o momento?
- Padrões bem definidos para tensão?
- Padrões bem definidos para corrente?

Análise Exploratória de Dados

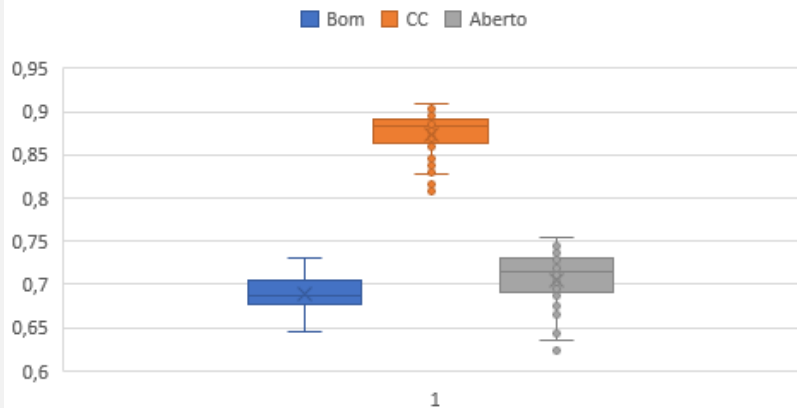
Média Aritmética Corrente



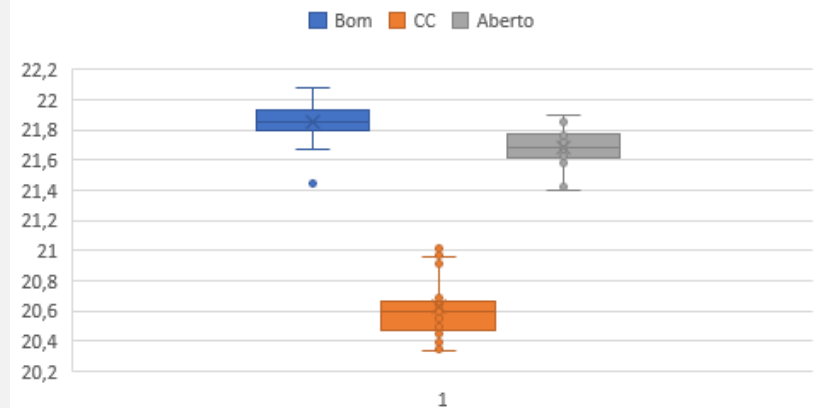
Média Aritmética Tensão



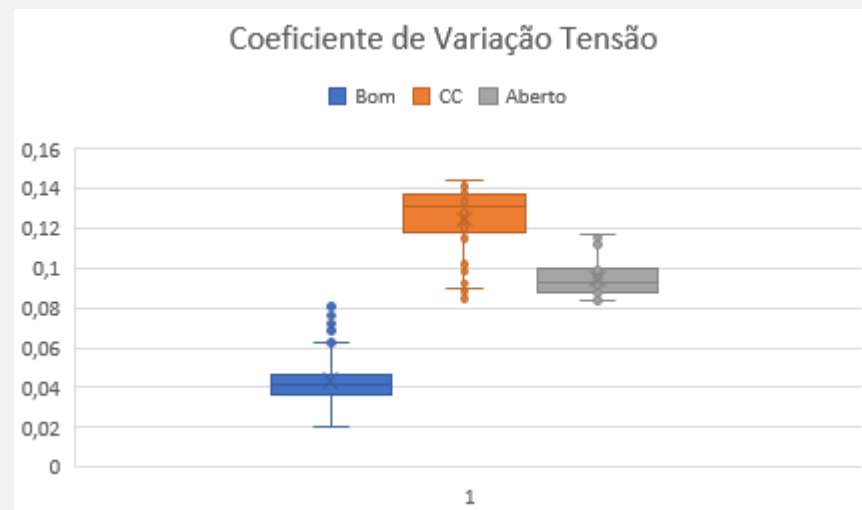
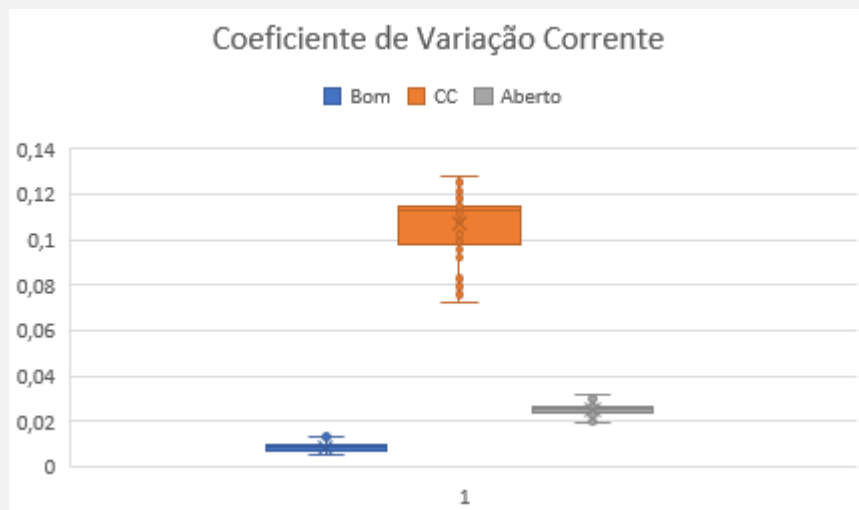
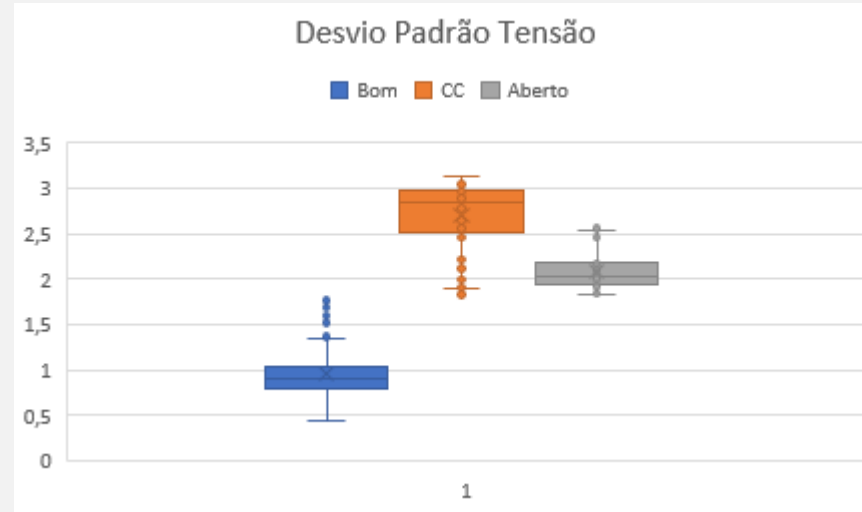
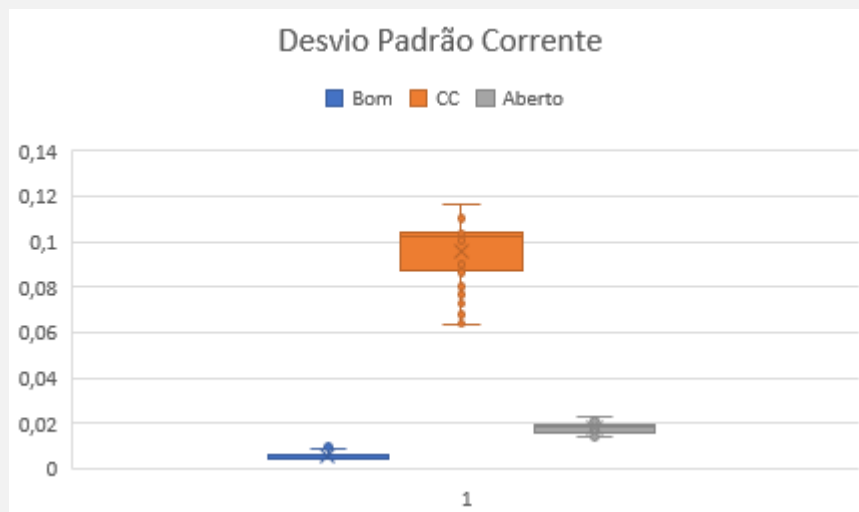
Mediana Corrente



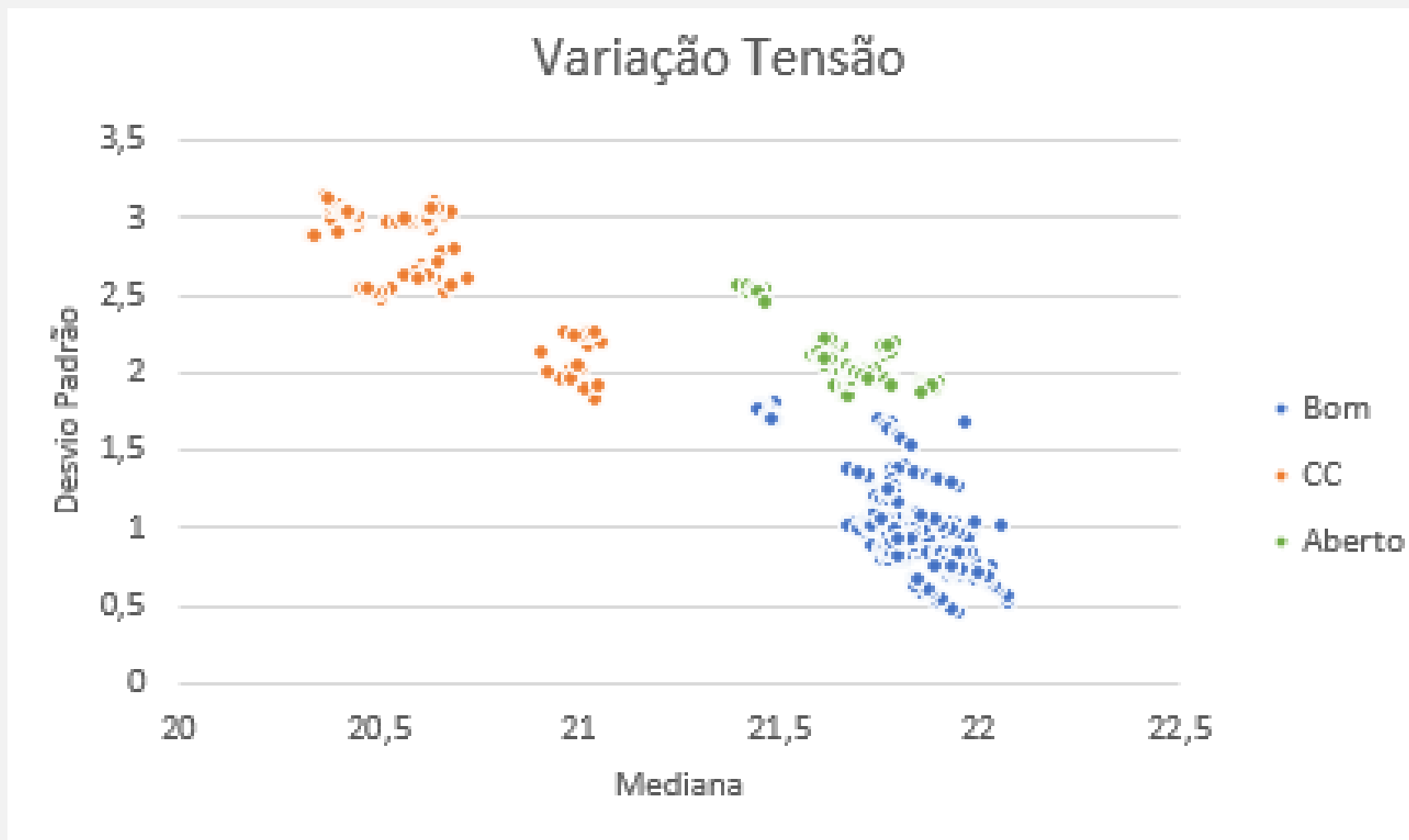
Mediana Tensão



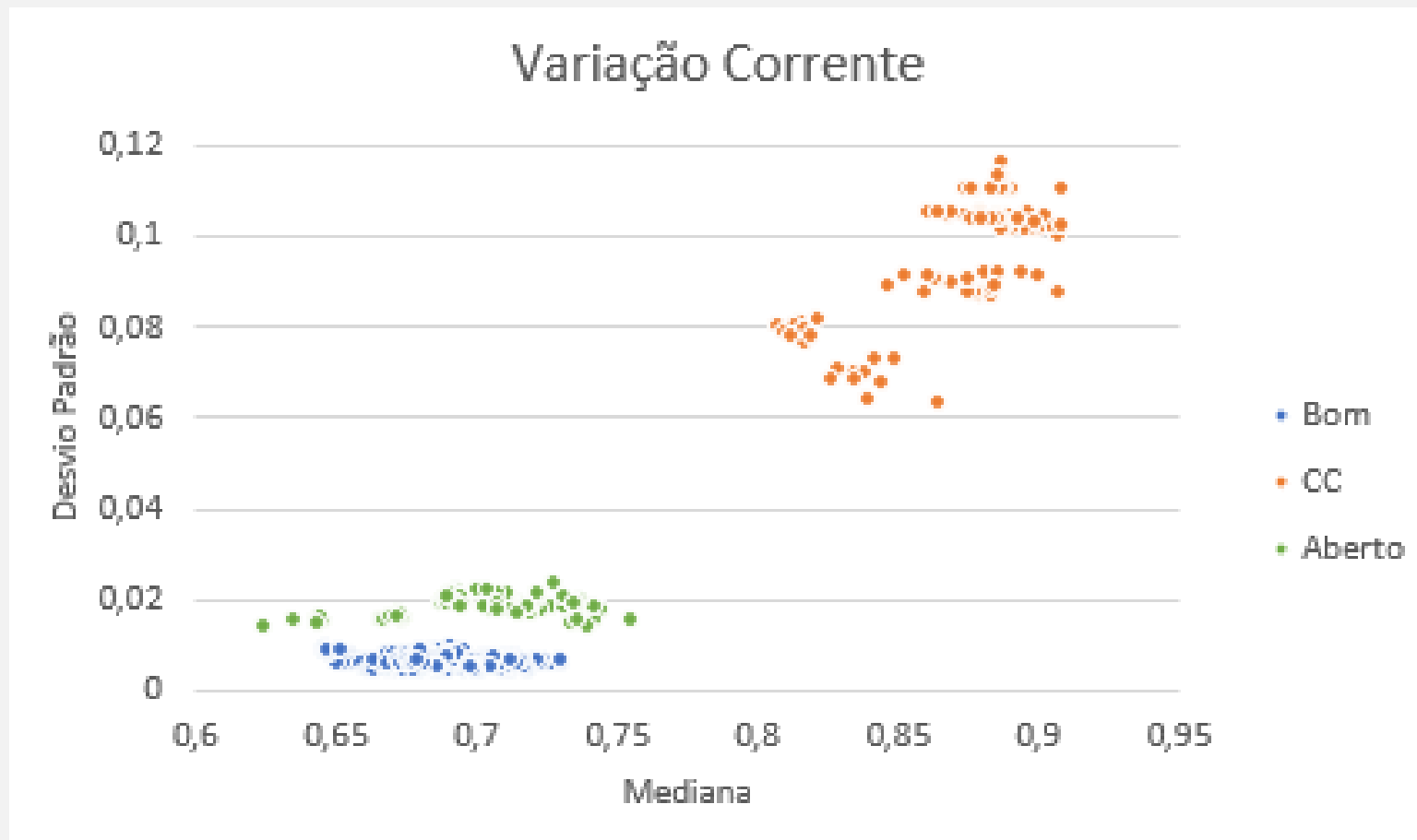
Análise Exploratória de Dados



Análise Exploratória de Dados



Análise Exploratória de Dados



Pré-Processamento

- Feature Engineering (decomposição, cruzamento, reinterpretação, discretização)?
- Feature Selection (redução de dimensionalidade)?
- Feature Encoding (substituir um dado categórico por um dado numérico)?
- Qual escalonador? (Normalizer, MinMaxScaler, StandardScaler, RobustScaler, QuantileTransformer, PowerTransformer)

Pré-Processamento

Método		Dados em distribuição normal	Dados não estão em distribuição normal	É desejado que os dados estejam em distribuição normal	É desejado eliminar a influência dos outliers
1	Normalizer	×	✓	×	×
2	MinMaxScaler	×	✓	×	×
3	StandardScaler	✓	✓	✓	×
4	RobustScaler				✓
5	QuantileTransformer				✓
6	PowerTransformer				

Pré-Processamento

- Método de divisão da base de dados?

Hold-out validation

Bootstrap

Cross validation (k-fold, Leave-one-out)

- Base de dados equilibrada?

- Como resolver?

```
from sklearn.model_selection import train_test_split  
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.8)
```

- Como definir os Rótulos?

LabelEncoder, OneHotEncoder

Estado	Característica
0	Motor sem defeito
1	Motor com curto-circuito
2	Motor com segmento quebrado

Tabela 1 - Estado dos motores em teste

Criação da Rede Neural (MLP)

- Sugestão de biblioteca:

```
import tensorflow as tf
```

```
from tensorflow import keras
```

Sugestão de início:

1 – 1 Camada oculta (quantidade de entradas / 2) = neurônios,
função de ativação Relu)

2 – 1 Camada de saída (3 neurônios, função de ativação Sigmoid)

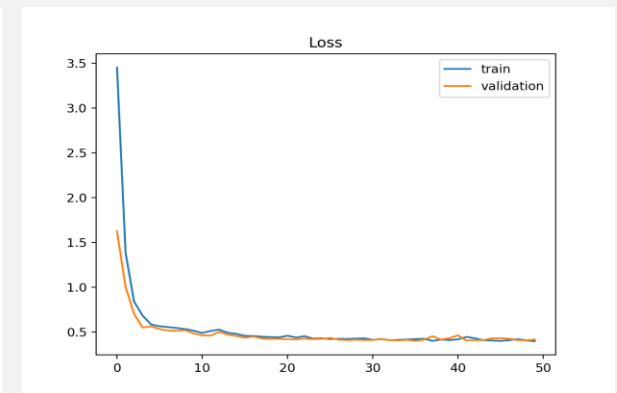
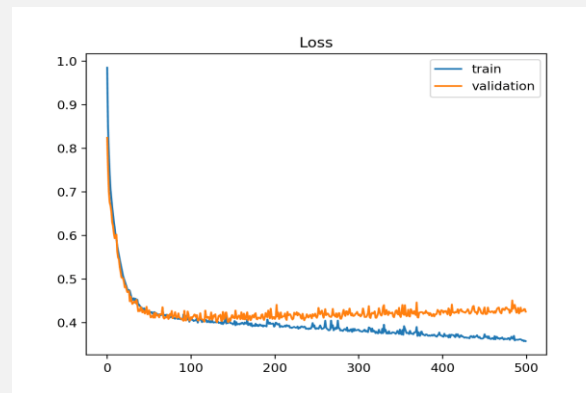
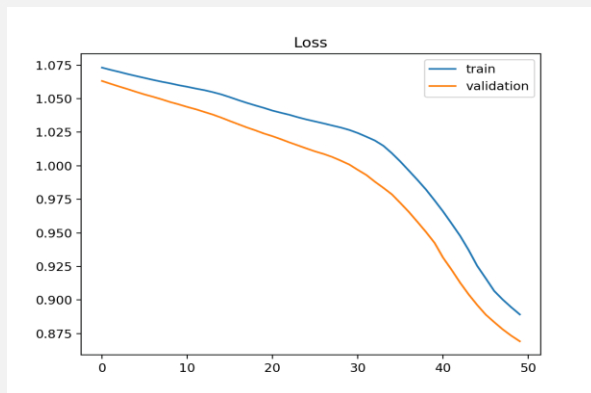
Otimizador = Adam

Função de perda = Categorical_crossentropy (quando existem 2 ou mais rótulos, geralmente utilizado quando aplicado método HotEncoder)

Métrica = Accuracy

Verificação de overfit e Underfit

- Essa divisão gera “curvas” de loss
- Essas curvas indicam como a rede neural está aprendendo, se está ocorrendo underfit ou overfit



Métricas de Classificação

- Acurácia
- Indica a taxa de acerto de uma classificação.
- Exemplo:
 - gabarito = [0, 2, 1, 3]
 - predição = [0, 1, 2, 3]
 - acurácia = 0.5



Métricas de classificação

Tabela confusão / Confusion Matrix

- Uma tabela que “resume” os casos de classificações corretas e incorretas.

	Predição: FALSO	Predição: VERDADEIRO
Real: FALSO	True Negative (TN)	False Positive (FP)
Real: VERDADEIRO	False Negative (FN)	True Positive (TP)

Exemplo:

- TP = Mulher grávida, preveu mulher grávida.
- TN = Mulher não está grávida, preveu que mulher não está grávida
- FP = Mulher não está grávida, preveu que mulher está grávida
- FN = Mulher está grávida, preveu que mulher não está grávida

Métricas de classificação

Tabela confusão / Confusion Matrix

- Exemplo:

		Valores preditos	
		Grávida	Não Grávida
Valores reais	Grávida	3	1
	Não grávida	2	4

- Previu grávida 3 vezes corretamente
- Previu não grávidas 4 vezes corretamente
- Previu grávida 1 vez incorretamente
- Previu não grávida 2 vezes incorretamente

Métricas de classificação

Tabela confusão / Confusion Matrix

- Exemplo:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} = \frac{\text{predições corretas}}{\text{todas as predições}}$$

$$recall = \frac{TP}{TP + FN} \quad precision = \frac{TP}{TP + FP} \quad 2 * \frac{precision * recall}{precision + recall}$$

- Accuracy = 0,7 (quantidade de acertos)
- Recall = 0,6 (quando o modelo prediz VERDADEIRO, com que frequência ele está correto?)
- Precision = 0,75 (positivos identificados corretamente)
- F-score = 0,66 (mostra o balanço entre precisão e recall)
- Taxa de erro = Com que frequência o modelo está errado?
(FP+FN)/total = (1+2)/10 = 0.3 ou 1-Acurácia

Métricas de classificação

Tabela confusão / Confusion Matrix

- Várias outras métricas podem ser obtidas da tabela:

Diagnostic testing [\[edit \]](#)

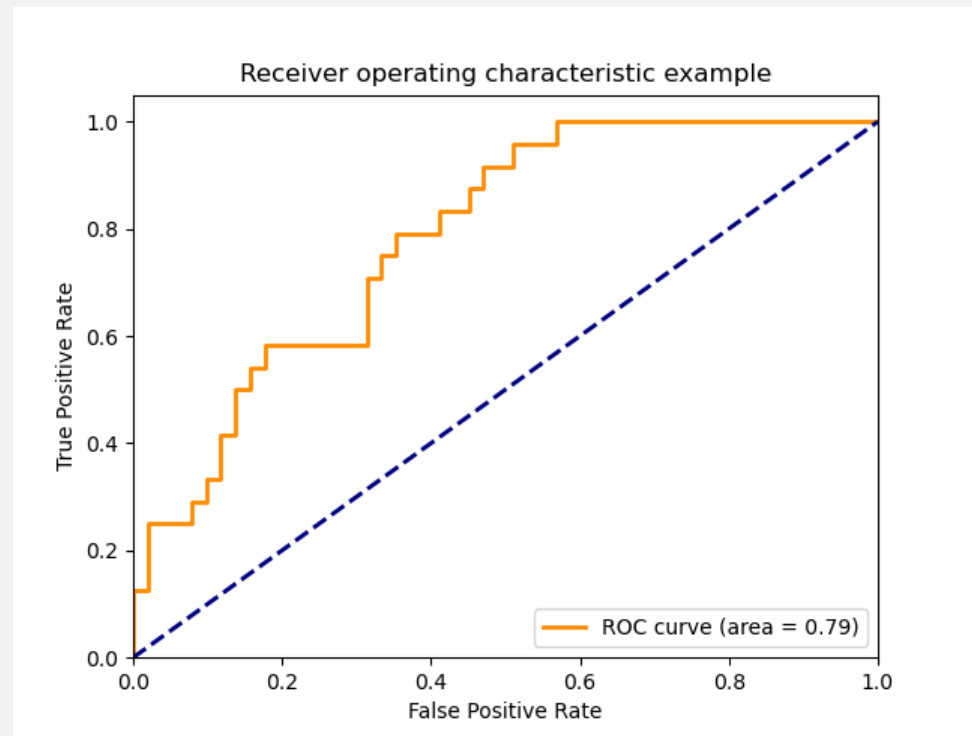
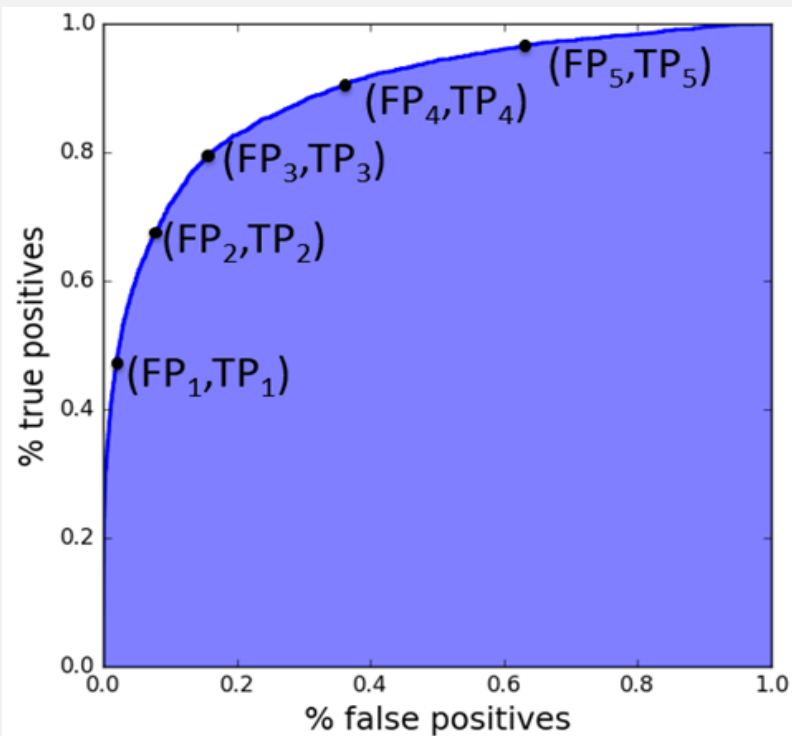
This is related to the field of [binary classification](#) where recall is often termed "sensitivity".

		Predicted condition		Sources: [3] [4] [5] [6] [7] [8] [9] [10] view · talk · edit	
Total population = P + N		Predicted condition positive (PP)	Predicted condition negative (PN)	Informedness, bookmaker informedness (BM) = TPR + TNR - 1	Prevalence threshold (PT) = $\frac{\sqrt{TPR \cdot FPR} - FPR}{TPR - FPR}$
Actual condition	Actual condition positive (P)	True positive (TP), hit	False negative (FN), Type II error, miss, underestimation	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power = $\frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate = $\frac{FN}{P} = 1 - TPR$
	Actual condition negative (N)	False positive (FP), Type I error, false alarm, overestimation	True negative (TN), correct rejection	False positive rate (FPR), probability of false alarm, fall-out = $\frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity = $\frac{TN}{N} = 1 - FPR$
Prevalence = $\frac{P}{P + N}$		Positive predictive value (PPV), precision = $\frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) = $\frac{FN}{PN} = 1 - NPV$	Positive likelihood ratio (LR+) = $\frac{TPR}{FPR}$	Negative likelihood ratio (LR-) = $\frac{FNR}{TNR}$
Accuracy (ACC) = $\frac{TP + TN}{P + N}$		False discovery rate (FDR) = $\frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) = $\frac{TN}{PN} = 1 - FOR$	Markedness (MK), deltaP (Δp) = PPV + NPV - 1	Diagnostic odds ratio (DOR) = $\frac{LR+}{LR-}$
Balanced accuracy (BA) = $\frac{TPR + TNR}{2}$		F_1 score = $\frac{2 \cdot PPV \cdot TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$	Fowlkes–Mallows index (FM) = $\sqrt{PPV \cdot TPR}$	Matthews correlation coefficient (MCC) = $\frac{TPR \cdot TNR \cdot PPV \cdot NPV - \sqrt{FNR \cdot FPR \cdot FOR \cdot FDR}}{1}$	Threat score (TS), critical success index (CSI) = $\frac{TP}{TP + FN + FP}$

Métricas de classificação

ROC

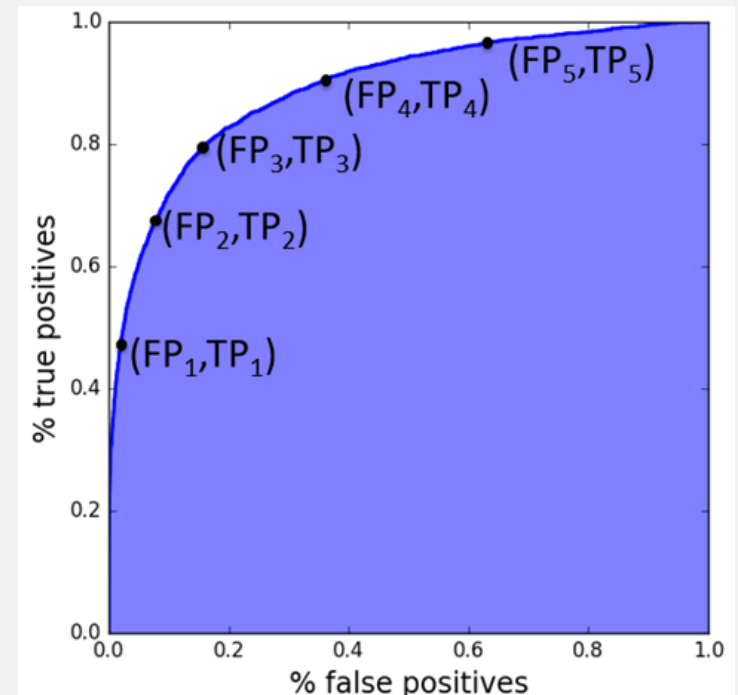
- A curva Receiver Operating Characteristics (ROC) mede o Recall em relação ao especificidade (TP x FP).



Métricas de classificação

AUC

- De forma bem simplista, Area Under Curve é a área abaixo da curva ROC
 - Valor = 1 -> Ótimo
 - Valor > 0.8 -> De forma geral, é um bom resultado (mas isso pode depender de problema pra problema)
 - Valor = 0.5 -> Aleatório



Métricas de Regressão

- R-Quadrado
- R-Quadrado Ajustado
- MSE
- RMSE
- MAE
- MAPE
- RMSLE

Métricas de Regressão

- R-Quadrado
- R-Quadrado Ajustado
- MSE
- RMSE
- MAE
- MAPE
- RMSLE

Resumo

- Métricas de avaliação são importantes para garantir que o modelo possui boa interpretabilidade.
- Em regressão geralmente se faz uso de múltiplas métricas.

Trabalho Final (parte 3)

Implementação Métricas de Avaliação

Data de entrega do relatório final: 07/12

Dúvidas e envio do trabalho:

e-mail: bernarddss62gmail.com