U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Master in Industrial Engineering and Management

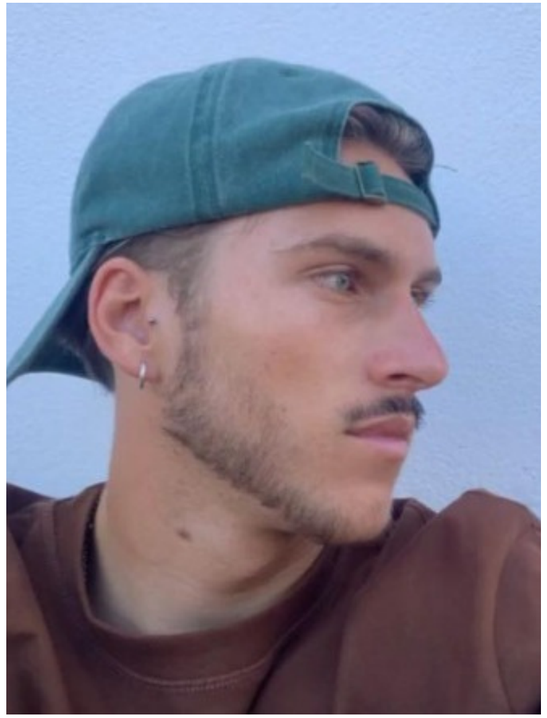
Business Analytics

Predictive Maintenance

Authors: Miguel Lopes, Rodrigo Costa

2023/2024

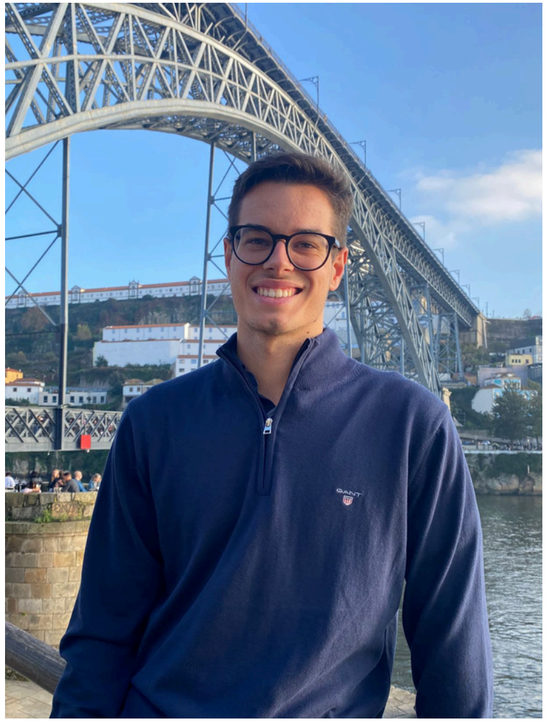
Meet the team



Miguel Lopes

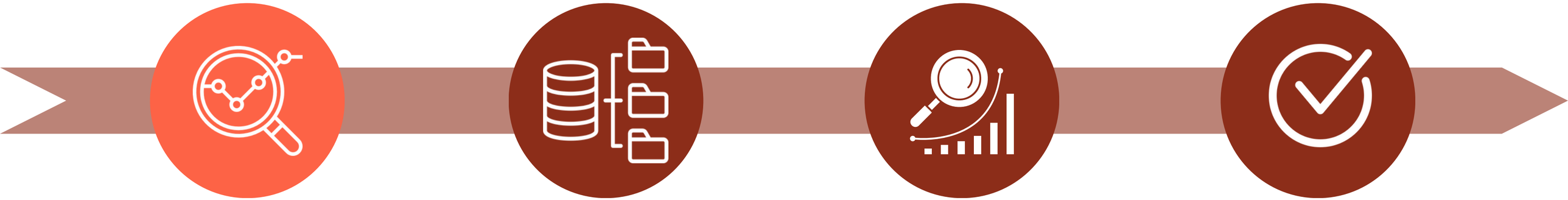
Both students at Faculdade de Engenharia da Universidade do Porto

Both of us are from the Azores Islands, Miguel is from São Miguel, Rodrigo is from Terceira



Rodrigo Costa

Agenda



01. Data Understanding

02. Data Preparation

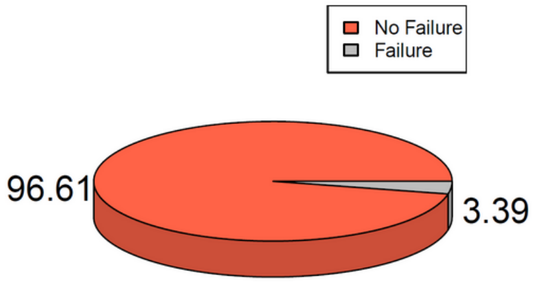
03. Modeling

04. Results & Comparison

01. Data Understanding

Initial Data Analysis

→ **10 000** observations:
- **9661** - No Failure
- **339** - Failure



10 attributes:
- UID
- Product ID
- Type
- Air Temperature
- Process Temperature
- Rotational Speed
- Torque
- Tool Wear
- Target
- Failure Type

Types of Failure:
- Heat Dissipation
- Power
- Tool Wear
- Overstrain
- Random

Some considerations:

1. **Fail** tends to occur, on average:
   - Under significantly **higher air temperature** conditions.
   - When the equipment is subject to **higher levels of tool wear.**
   - When machines have **high torque** values

2. Failures related to **heat dissipation** are associated with **higher air temperature** conditions.

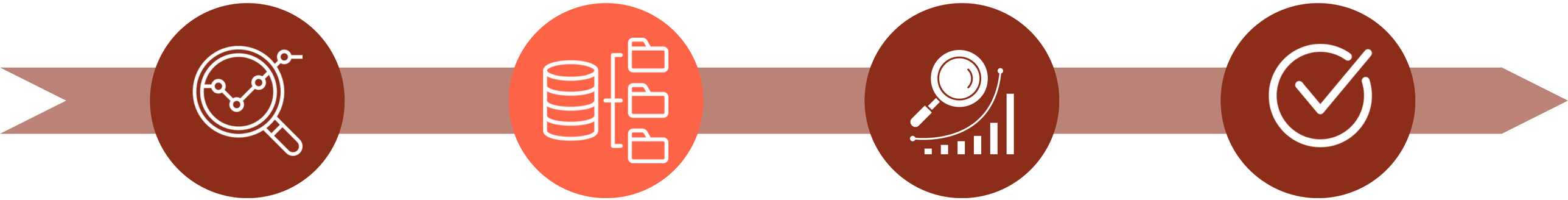
3. **Highly correlated** variables:
   1. Torque and Rotational Speed
   2. Air Temperature and Process Temperature

New attributes

Power = Rotational Speed x Torque

Temp_dif = Air Temperature - Process Temperature

Agenda



01. Data Understanding

02. Data Preparation

03. Modeling

04. Results & Comparison

02. Data Preparation

Data Cleaning

1. Variables **UID** and **productID** were **removed**, since they are unique to each observation, not adding any information to our model, and being bad predictors for further analysis.

2. **Eighteen observations** from the inconsistencies found were registered as "**Random Failures**", but had the target variable as 0. We assumed that this was related to some sort of failure that also affects the failure detection system, and was afterward reported. In these cases, the **target variable was established as 1**.

3. **Nine observations** from the inconsistencies found had the **target variable defined as 1**, but the failure type was reported as "No Failure". Since we can't clearly define what could lead to this inconsistency, and it won't add much information to the model, these **observations were dropped**.

4. **Z-Score Approach**
   1. **Calculate the Z-Score** for all observations - **numeric variables** (ProcessT, RotSpeed, Torque, and ToolWear);
   2. Calculate the **absolute value** of each of them;
   3. **Define a threshold** = **3** for ≈ 99.7% of the data falls inside three standard deviations;
   4. Identify the **strong outliers**
   5. Check if they are **generating failures**

189 outliers found

Variable Selection

Inconsistency Management

Outlier Detection

02. Data Preparation

Data Cleaning

Inter Quartil Range (IQR) Approach

5

- Divides the dataset into quartiles, resulting in four equal parts
- Deviation of 1.5
- Used to identify outliers in the **Torque** and **RotSpeed** features

488 outliers found

Local Outlier Factor (LOF) Approach

6

- Computes the degree of outlyingness for each data point by comparing its local reachability density to the minimum density of its nearest neighbors
- **Threshold** = 2 and **k** = 2

66 outliers found

Outlier Detection

However...

- Graphical analyses
- No additional information about **wrong measurements**

No Real Outliers

0
Outliers Eliminated

02. Data Preparation

Data Construction & Formatting

Encoding

Variable *Type*:

Categorical → Numerical

Normalization

- **Z Score** Normalization Method

Feature Selection

- **MRMR** (Minimum Redundancy Maximum Relevance)
- **Stepwise Backwards** Feature Selection

Feature Scenarios Created

Split

70% **Training Data** + **30%** **Testing Data**

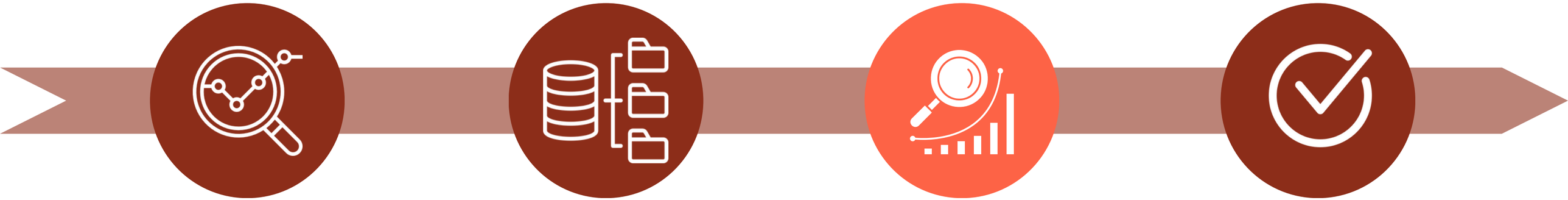
Balancing

Dataset **unbalanced** → **SMOTE** **Oversampling**

Synthetic observations

Increased **nr of failures**

Agenda



01. Data Understanding

02. Data Preparation

03. Modeling

04. Results & Comparison

03. Modeling

Models & Metrics

Models

One Rule

Multinomial Logistic Regression

K-Nearest Neighbours

Decision Trees

Random Forest

Neural Networks

Support Vector Machines

Extreme Gradient Boosting

Naïve-Bayes

Metrics:

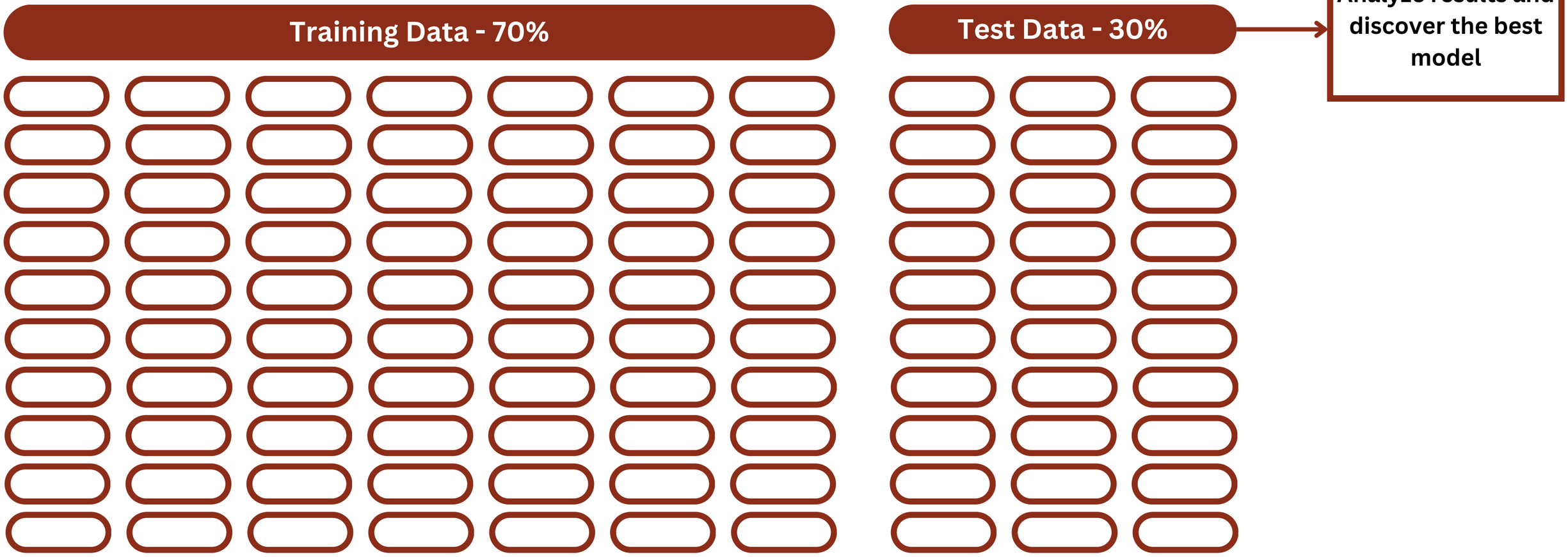
- **Accuracy:** percentage of correct predictions in total predictions.
- **Recall:** proportion of true positive instances that are correctly identified by the model.
- **AUC Score:** probability that a model ranks a positive higher than a negative, i.e. the trade-off between true positives and false alarms.
- **Weighted F1 score**: weighted mean of the precision and recall scores for each type of failure, taking into account their impact in the dataset to assign different weights.
- **Precision**: Precision is calculated as the proportion of true positives among all instances predicted.

03. Modeling

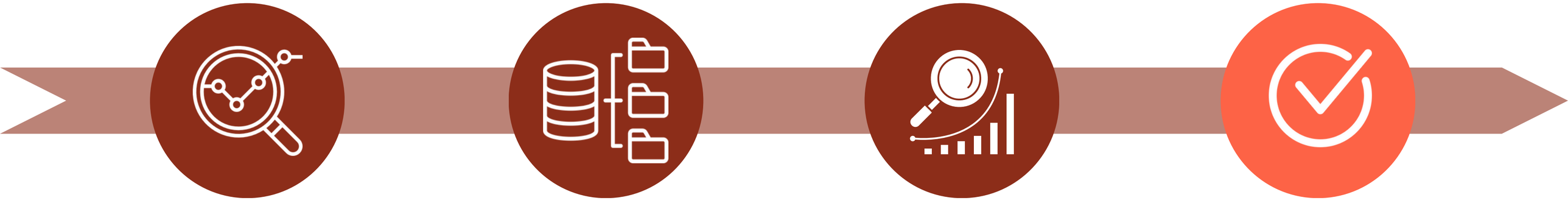
Hyper-Parameter tuning

Performed using a 10-Fold Cross-Validation process to:

- Optimize algorithmic parameters – **better model performance**.
- Avoid overfitting and underfitting.



Agenda



01. Data Understanding

02. Data Preparation

03. Modeling

04. Results & Comparison

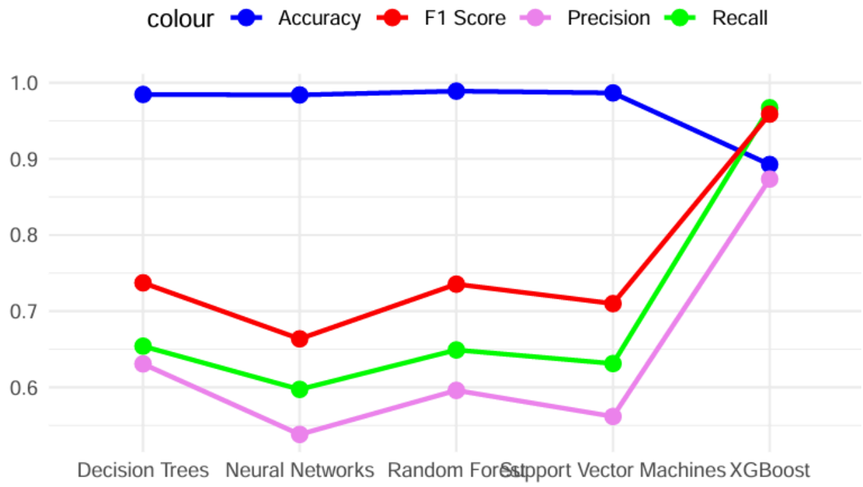
04. Results

Best Performance Values

MODEL	DATA	SCENARIO	NORMALIZED	WEIGHTED_F1
K-NN	Balanced	2	YES	0,5194097372
K-NN	Unbalanced	7	YES	0,5140356683
DT	Unbalanced	7	YES	0,7371882086
DT	Balanced	2	YES	0,6980752888
MLR	Unbalanced	1	YES	0,615857261
MLR	Balanced	2	YES	0,4994220924
Naive Bayes	Unbalanced	2	YES	0,4437940379
Naive Bayes	Balanced	1	NO	0,3365664228
Random Forest	Unbalanced	2	YES	0,7354497354
Random Forest	Balanced	2	YES	0,7137087493
SVM	Unbalanced	2	YES	0,7099080945
SVM	Balanced	7	YES	0,5788444376
Neural Networks	Unbalanced	1	YES	0,6635796918
Neural Networks	Balanced	1	YES	0,566751258
XGBoost	Balanced	6	YES	0,9588339083
XGBoost	Unbalanced	7	YES	0,8100840336

Scenario	Variables
1	AirT, ProcessT, RotSpeed, Torque, Type, Power, Temp_Diff, Toolwear, FailType
2	RotSpeed, Torque, Type, Power, Temp_Diff, Toolwear, FailType
3	AirT, ProcessT, Type, Power, Temp_Diff, Toolwear, FailType
4	AirT, ProcessT, RotSpeed, Torque, Type, Temp_Diff, Toolwear, FailType
5	AirT, ProcessT, RotSpeed, Torque, Type, Power, Toolwear, FailType
6	Type, Power, Temp_Diff, Toolwear, FailType
7	RotSpeed, ToolWear, Temp_Diff, Power, Torque, FailType
8	Type, RotSpeed, ToolWear, Temp_Diff, FailType
9	Type_Num, Torque, ToolWear, Temp_Diff, FailType

Metrics Graphical Comparison



04. Results

Best Hyperparameters

Hyperparameters Tune Grid

Model	Hyperparameters Grid
NB	**threshold**: 0, 5, 10, 15, 20 **gamma**: 0.1, 0.2, 0.3, 0.4, 0.5
KNN	**number of neighbors**: 1 to 40
DT	**min. number of observations per leaf**: 2 to 20 **criterion**: information, gini **complexity**: 0.01, 0
RF	**number of trees**: 250, 500, 750, 1000 **number of features**: 1, 3, 5, 7
SVM	**gamma**: 1, 0.1, 0.01, 0.001, 0.0001 **cost**: 0.01, 0.1, 1, 10, 100, 1000
XGB	**depth**: 4, 8 **number of rounds**: 10000 **minimum child weight**: 1, 5 **eta**: 0.001, 0.05 **gamma**: 0, 5
NN	**hidden layers**: 1, 3 **learning rate**: 0.1 **stepmax**: 10 000 000

Optimal Hyperparameters & Scenarios

Naive Bayes

MODEL	DATA	SCENARIO	NORMALIZED	LAPLACE	THRESHOLD
Naive Bayes	Unbalanced	2	YES	0	0,1
Naive Bayes	Balanced	1	NO	0	0,1

K-NN

MODEL	DATA	SCENARIO	NORMALIZED	K
K-NN	Balanced	2	YES	1
K-NN	Unbalanced	7	YES	5

Decision Trees

MODEL	DATA	SCENARIO	NORMALIZED	CRITERION	CP	MIN_OBJ
DT	Unbalanced	7	YES	gini	0	2
DT	Balanced	2	YES	gini	0	10

Support Vector Machines

MODEL	DATA	SCENARIO	GAMMA	COST
SVM	Unbalanced	2	0,01	1000
SVM	Balanced	7	0,1	1000

Random Forest

MODEL	DATA	SCENARIOS	NORMALIZED	MTRY	NTREE
Random Forest	Unbalanced	2	Yes	3	250
Random Forest	Balanced	2	Yes	1	250

Neural Networks

MODEL	DATA	SCENARIOS	NORMALIZED	LAYERS
Neural Networks	Unbalanced	1	Yes	3
Neural Networks	Balanced	1	Yes	3

XGBoost

MODEL	DATA	SCENARIO	DEPTH	CHILD_WEIGHT	ETA	GAMMA
XGBoost	Balanced	6	8	1	0,001	5
XGBoost	Unbalanced	7	4	1	0,05	0

04. Results

MODEL	ADVANTAGES	DISADVANTAGES
K-NN	KNN is relatively straightforward and easy to implement	The choice of the value of k can significantly affect the performance of KNN
DT	Decision trees produce models that are easily interpretable and understandable	May not work well in unbalanced data sets
MLR	Includes a variety of data pre-processing techniques	Easily outperformed (for complex relationships)
Naive Bayes	It´s quick to train and make predictions	Independence assumption
Random Forest	High accuracy	Training time is computationally intensive
SVM	Effective handling of high dimensionality	Parameter sensitivity
Neural Networks	Adaptive Learning Capacity	Propensity for overfitting, especially in small data datasets
One-R	Simplicity and ease of interpretation	Sensitivity to the presence of irrelevant or correlated attributes
XGBoost	Speed, performance, regularization techniques, flexibility in handling different data types	Complexity in implementation and tuning, high computational requirements, limited interpretability, and risk of overfitting

04. Results

Conclusions & Recommendations

Company's Goal	MODEL SELECTION	BEST-IN METRIC
Accurate Classification Across Multiple Failure Types	**Random Forest (Balanced)**	**Accuracy**
Balanced True Positive Rate and False Positive Rate	**Support Vector Machines (Balanced)**	**AUC**
Minimizing False Negatives Across Multiple Classes:	**XGBoost (Balanced)**	**Mean Recall**
Minimizing False Positives While Correctly Identifying Instances for Each Class		**Mean Precision**
Robust and Reliable Performance in Identifying Failure Types:		**Weighted F1**

Thank You!