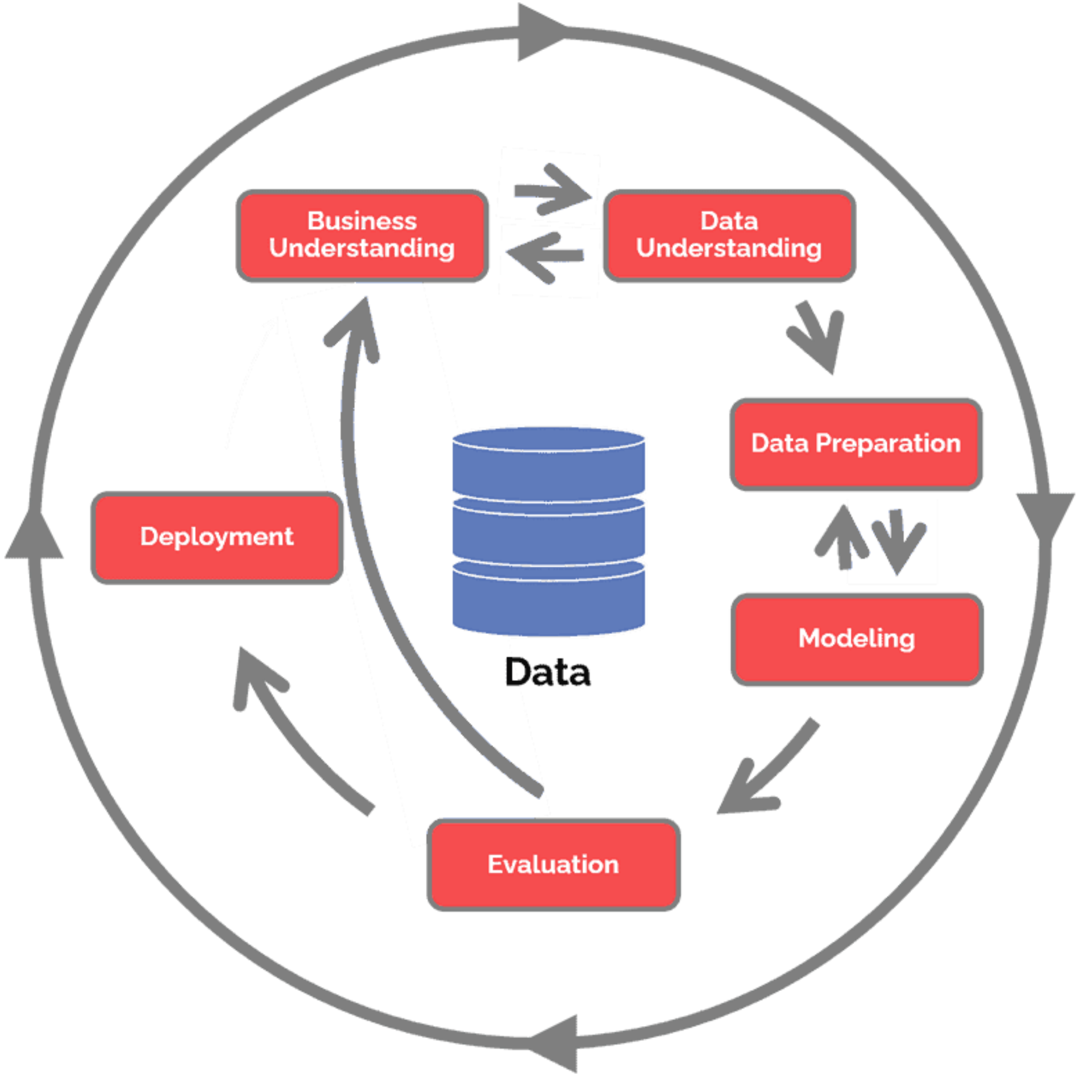
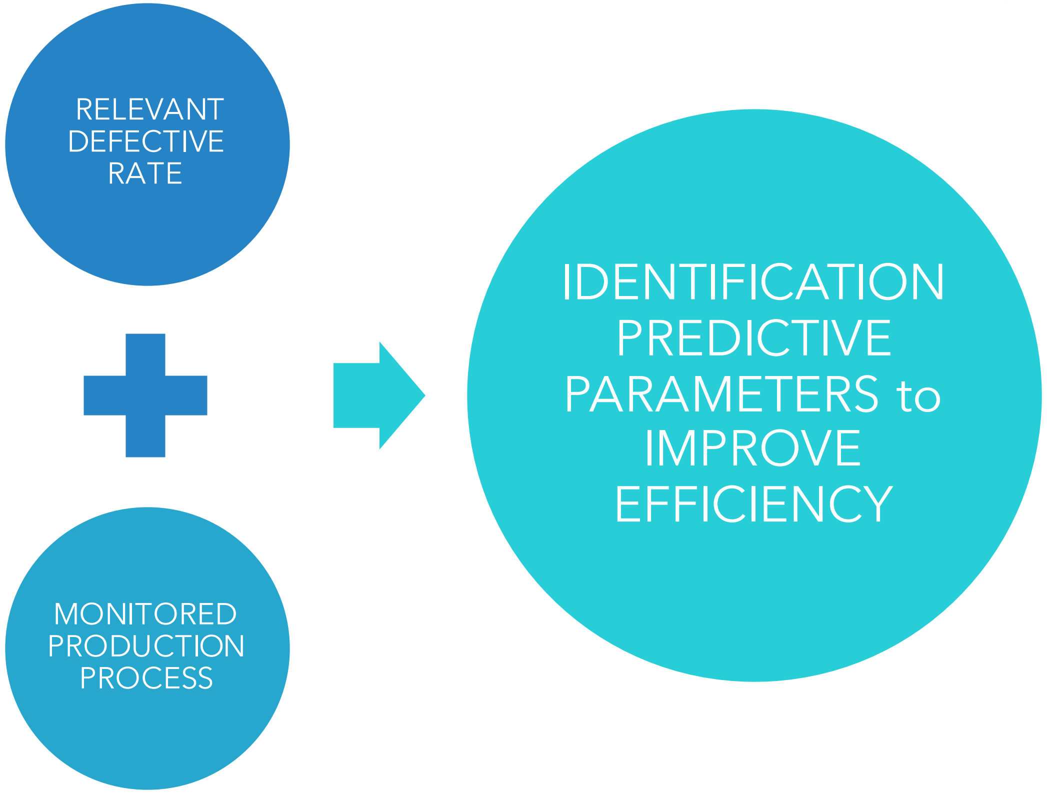
Root Cause Analysis

Matteo Bonini
Mattia Campanella
Sophie Sas
Javad Toybeigibenis


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

Process Map

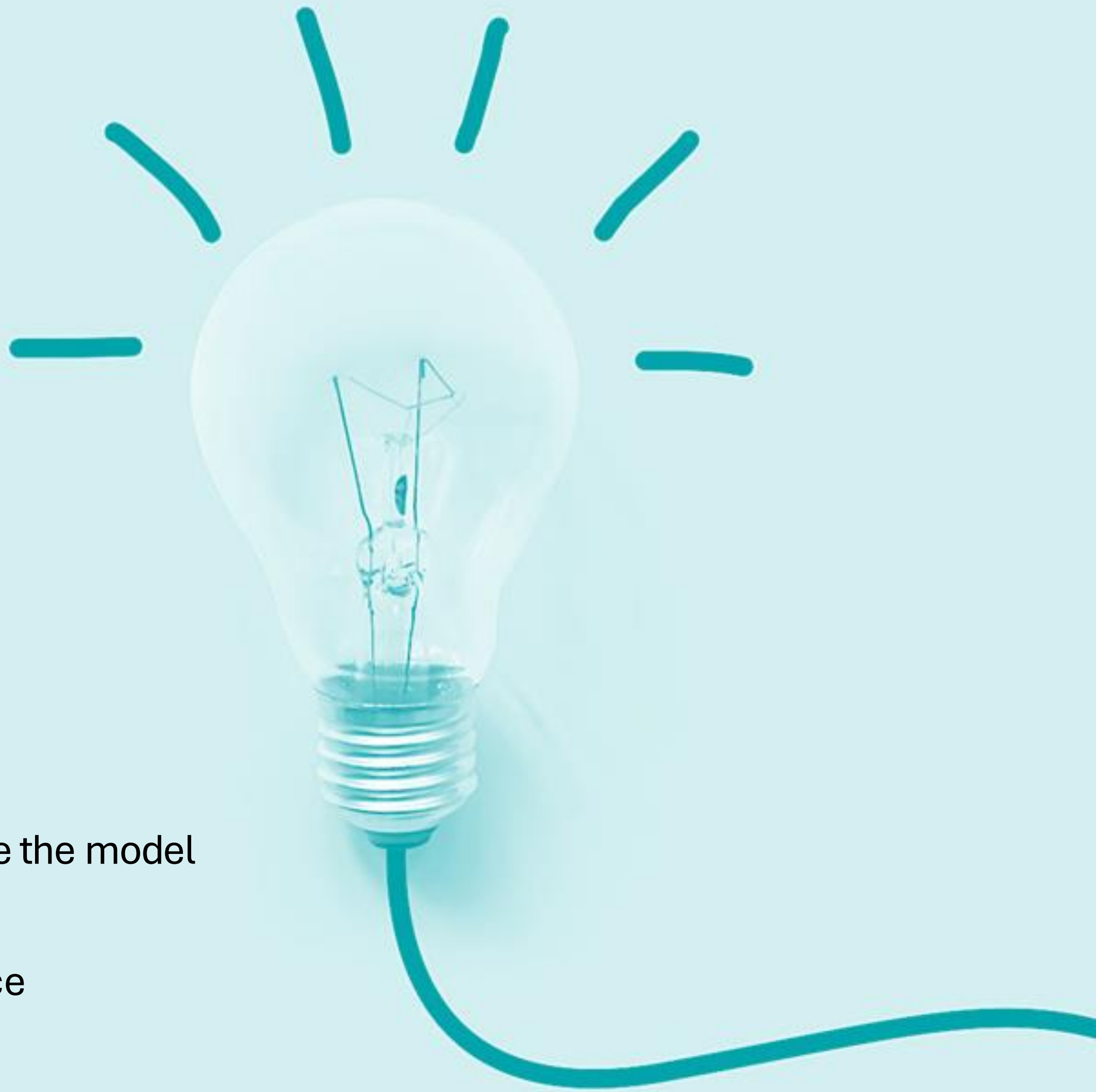
Business Understanding



Plan of Approach

Roadmap:

1. **Data Understanding and Pre-Processing.**
   Check the dataset for any irregularities.
2. **Feature Selection**
   Reduce dimensionality using a correlation heatmap.
3. **Model Training**
   Train a XGBoost (Extreme Gradient Boosting) Classifier
4. **Model Evaluation**
   Use accuracy, precision, recall and F1-score to evaluate the model
5. **Interpretation of the Results**
   Visualize the results and interpret the feature importance



Data Understanding - Datasets

PRODUCTION:

- Dataset containing PLC information on the processing settings at COIL-DATE-MT level.
- COIL ID: of the metal coil processed
- MT: meter observation of the coil (i.e. one observation every 7 meters)
- DATE: day of the year in which the processing of a given COIL-MT started
- TIME_START_PROCESS: time in which the processing started
- All the remaining fields are settings referring to the processing of a given COIL-MT.

DEFECTS:

- Dataset containing information on the defect by coil and type of defect, detected during quality control after production.
- COIL ID: of the metal coil processed
- MT_FROM: point of the coil in which a given defect start.
- MT_TO: point in which a given defect end.
- DATE: date in which the coil has been processed.
- DEF_TIPO_1 (TO 6): indicator for the kind of of defect detected.

```
Unique COIL values in 'production' dataset: 1261
Unique COIL values in 'defects' dataset: 534
Percentage of COILS with defects: 42.35%
```

Data Understanding - Challenges

- "MT" COLUMN RELIABILITY
- DEFECT TYPES DISTRIBUTION
- CORRELATED FEATURES

ORIGINAL DATA	
COIL	MT
359413	7238
359413	7245
359413	7252
359413	0
359413	7
359413	14
359413	21
359413	28
359413	35

Data Understanding

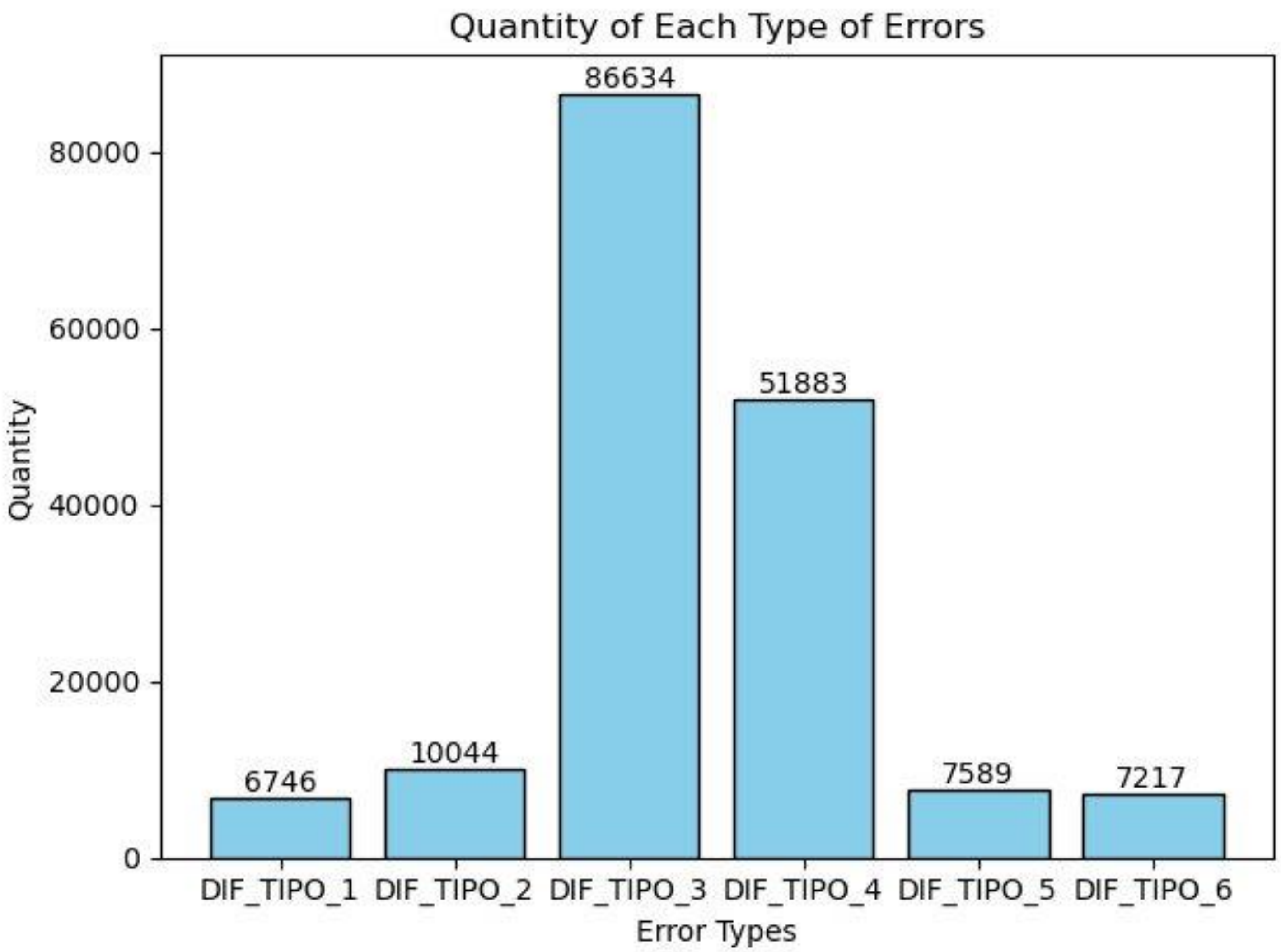
"MT" Column reliability

Data Understanding
Defect Types Distribution

Unbalanced distribution of the defects.

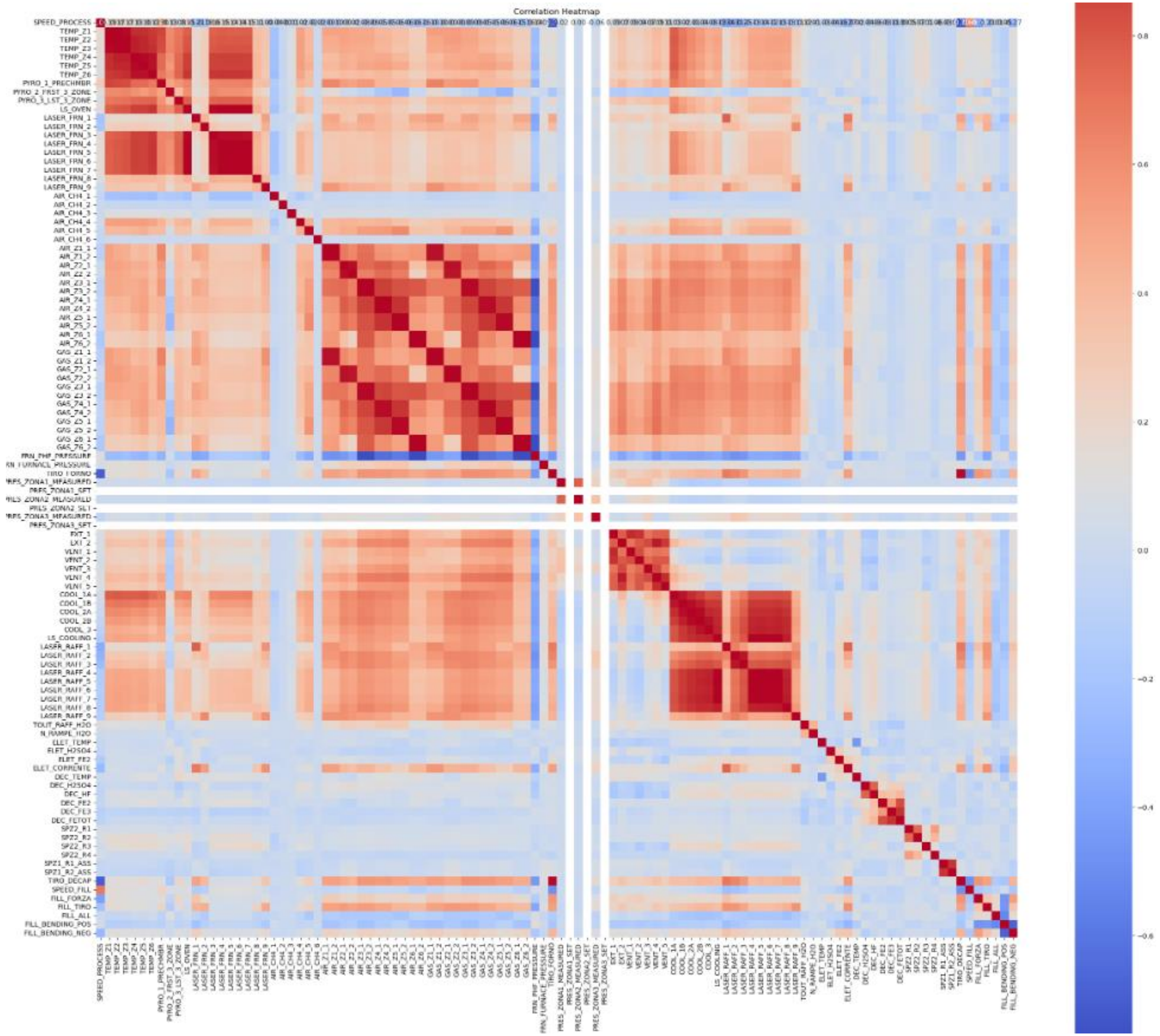
Multi-label defects

80% of the defect frequencies are of type 3 and type 4.



Data Understanding

Correlated Features



Data Preparation

Before merging:

Meters column in the production dataset showed inconsistent measures.

The reviewed meters are later needed to identify which parts of the coil contain defects.

Merging:

Merging the two datasets based on the reviewed meters column in the production dataset and the columns ‘meters from and to’ in the defect dataset.

The merged dataframe allowed to understand which parts of the coil contained any defect.

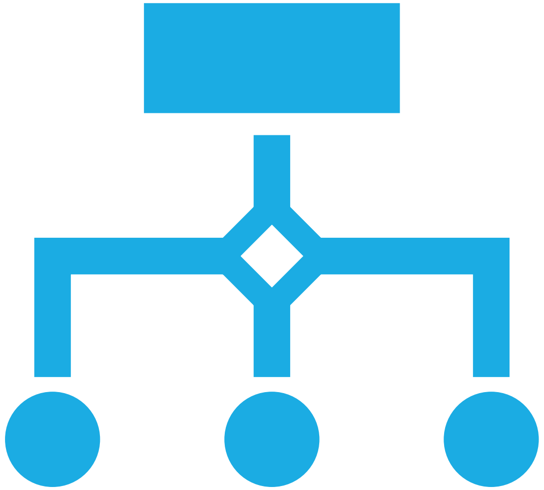
Data Preparation

Creation of a new binary variable 'Defect'

- Value 1 if any of the 6 different defects are present.
- Value 0 if no defect is present.

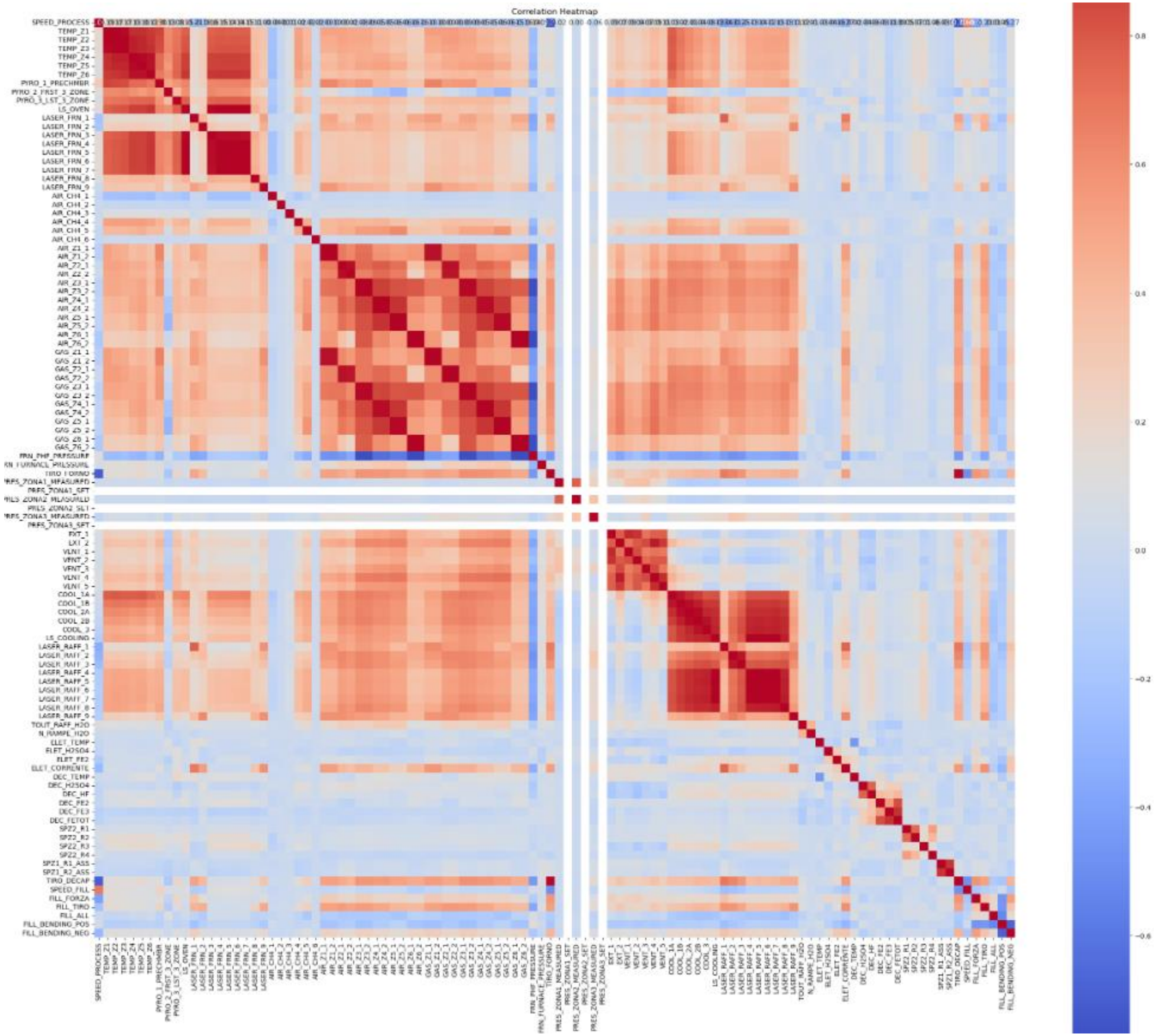
The new variable allowed to understand the root cause for a defect.

Decided not to focus on each particular defect type, since it was unknown to us what the importance of each defect type was.



Data Preparation
Feature Selection

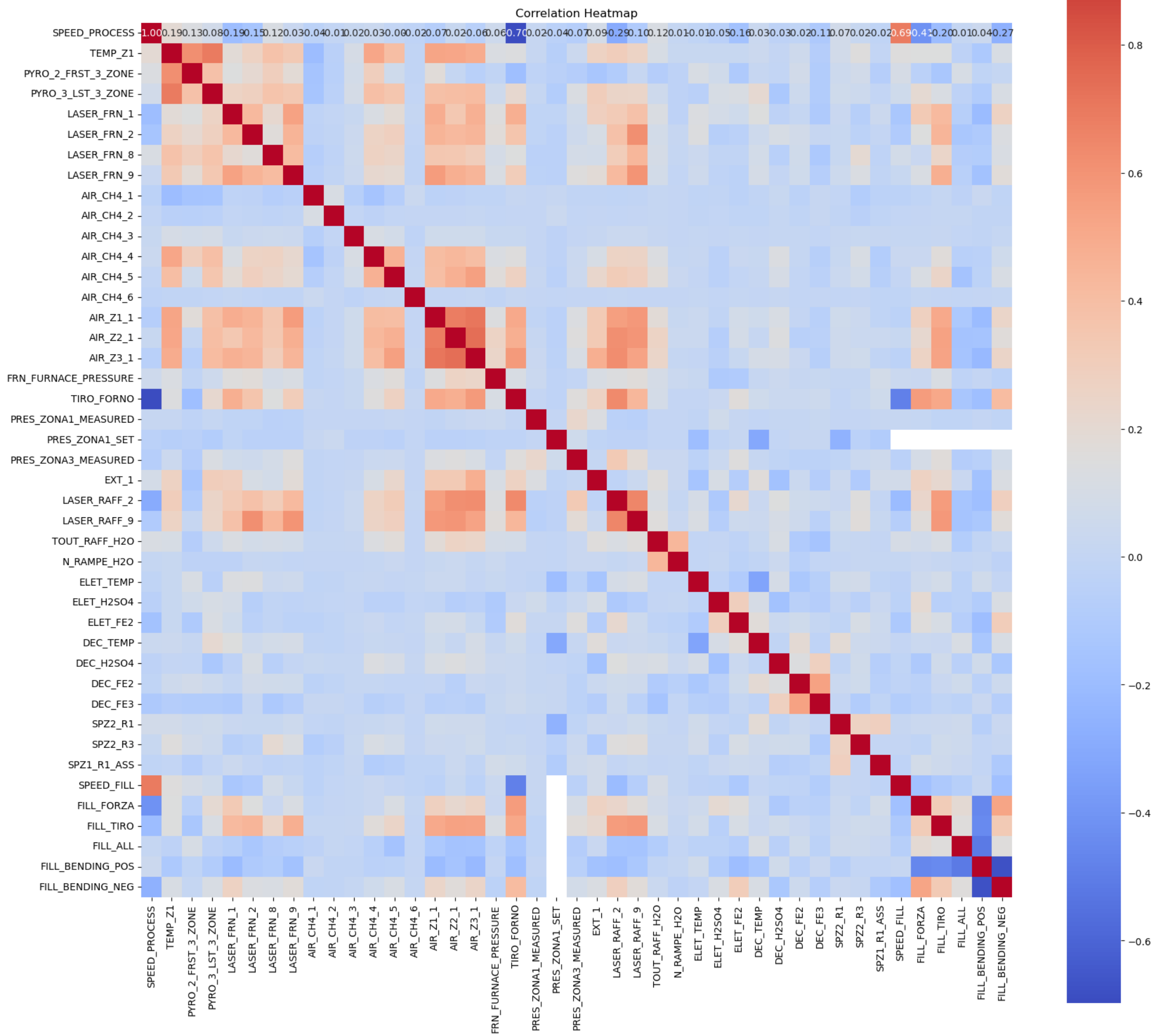
From 106 features in the merged dataset, we made a selection to avoid the inclusion of overlapping parameters in the model, computing a correlation heatmap to identify those highly related.



Data Preparation

Feature Selection

We established a correlation threshold of 0.75 and we dropped those above this ceiling. This left us with 43 features.



Data Preparation – Target Variable

X variable

Our X variable consists of the features we found after feature selection.

The X variable is used to explain the existence of a possible defect.

Y Variable (Target)

The newly created Defect variable is defined as the target variable.

The target variable allows to understand whether a defect is detected in the coil.

Modelling – Algorithm selection

Why a Decision Tree based model?

Interpretability: Easy to visualize the decision-making process.

Feature importance: Indicates the relative importance of each feature.

Scalability: Can handle large datasets efficiently.

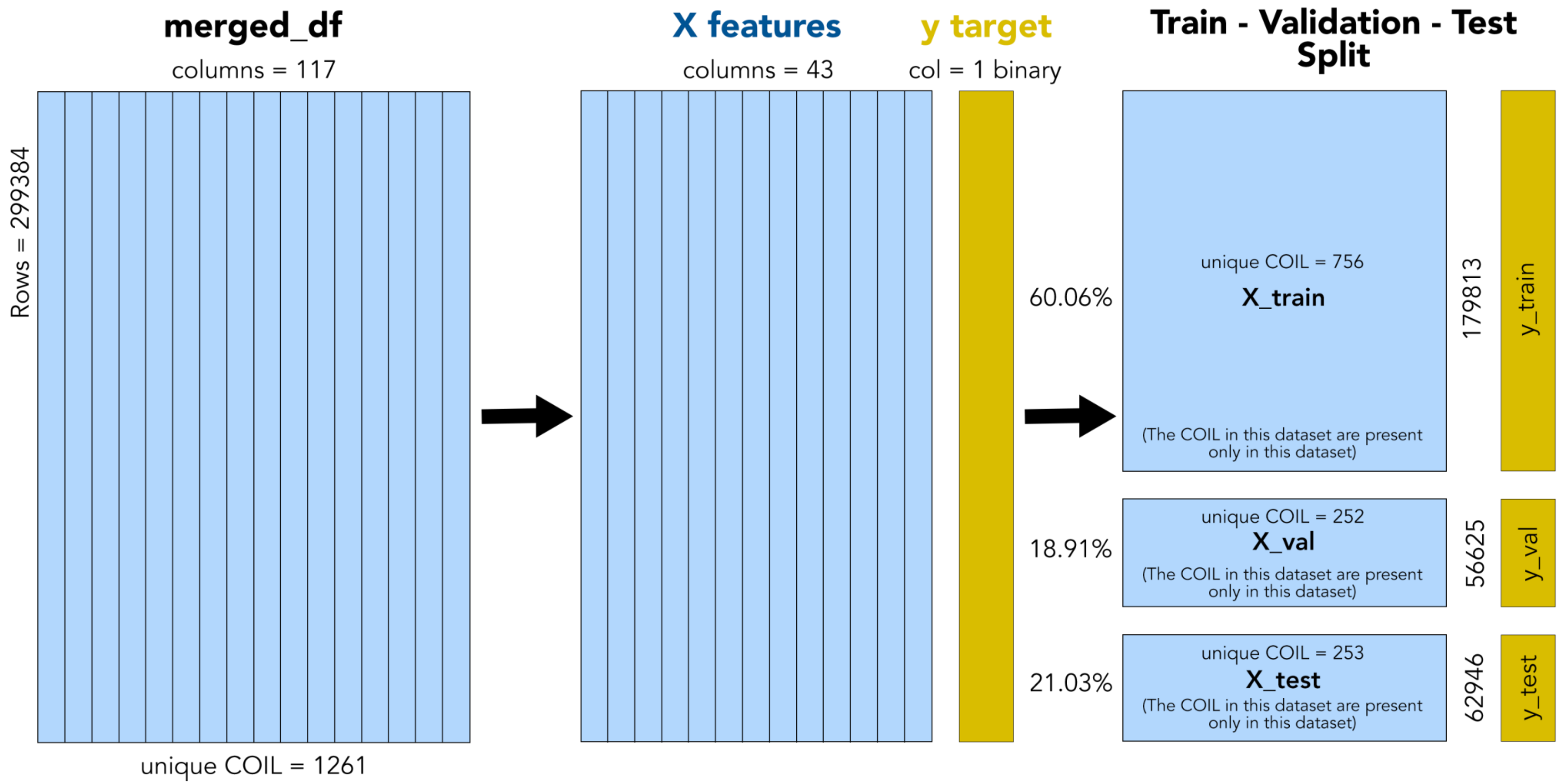
Robustness to Outliers: Partition the feature space into regions based on data splits.

XGBoost vs RandomForest

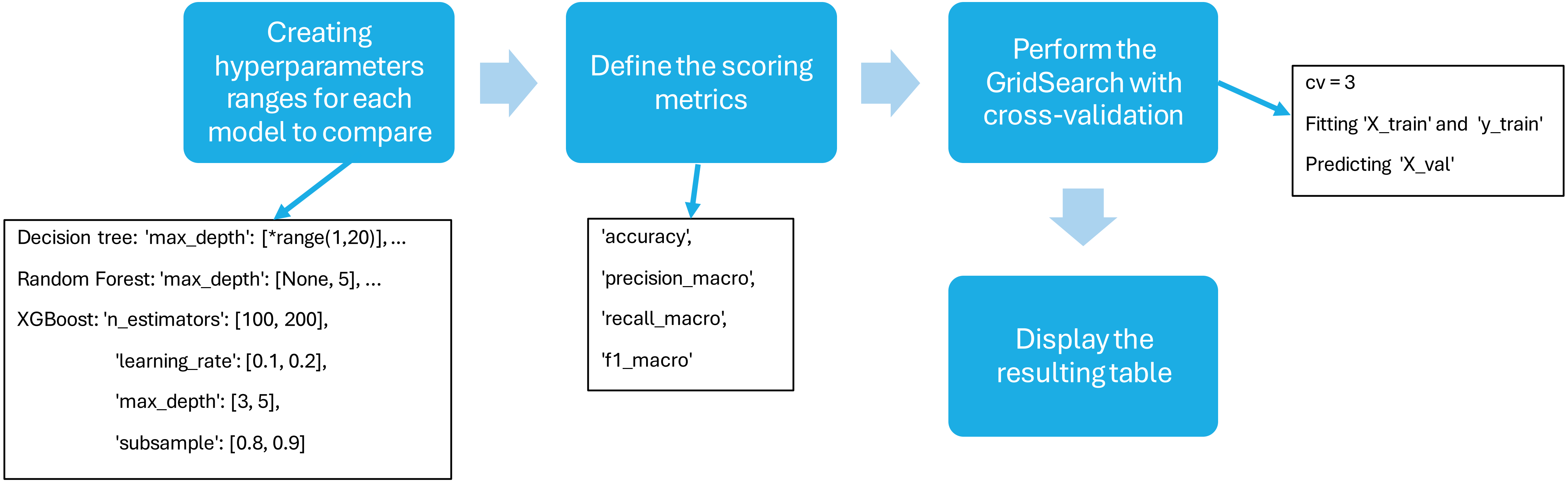
It employs a sequential strategy building decision trees one at a time, with each subsequent tree focusing on correcting the errors of the previous tree. Preventing the model from memorizing irrelevant patterns.

Produces multiple decision trees, randomly choosing features to make decisions when splitting nodes to create each tree. It then takes these randomized observations from each tree and averages them out to build a final model.

Modelling– Dataset splitting



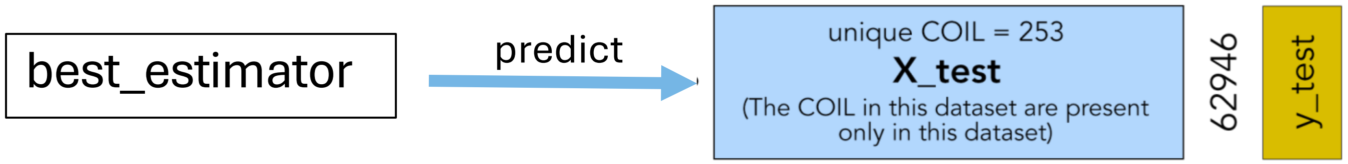
Modelling – Building the Model



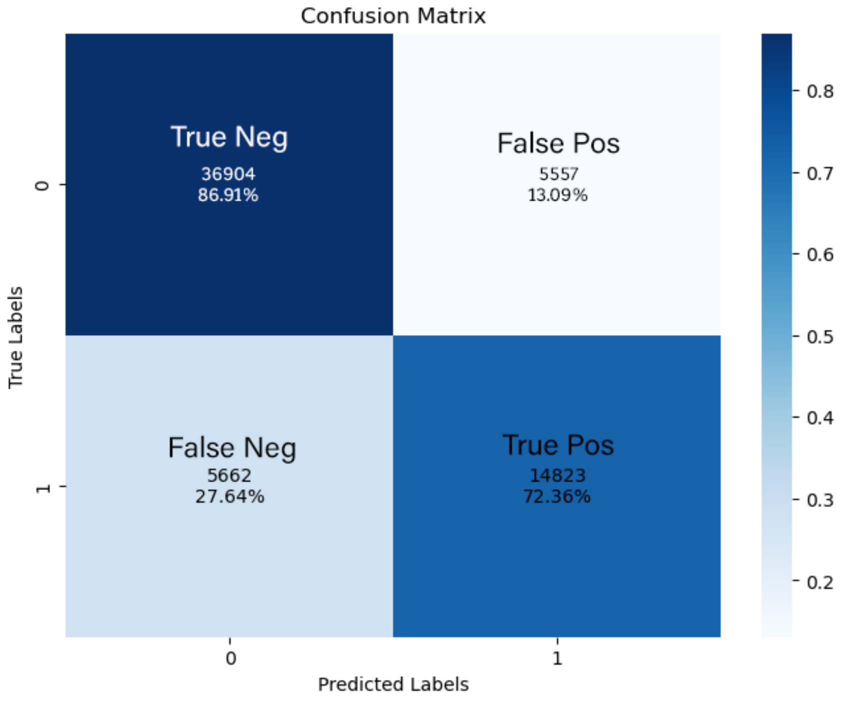
	model	best_params	accuracy	precision_macro	recall_macro	f1_macro
2	XGBoost	{'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 100, 'subsample': 0.9}	0,755	0,742	0,733	0,736
1	Random Forest	{'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 50}	0,753	0,740	0,728	0,732
0	Decision Tree	{'class_weight': 'balanced', 'max_depth': 5, 'max_features': None, 'min_samples_leaf': 1, 'min_samples_split': 2}	0,712	0,707	0,719	0,707

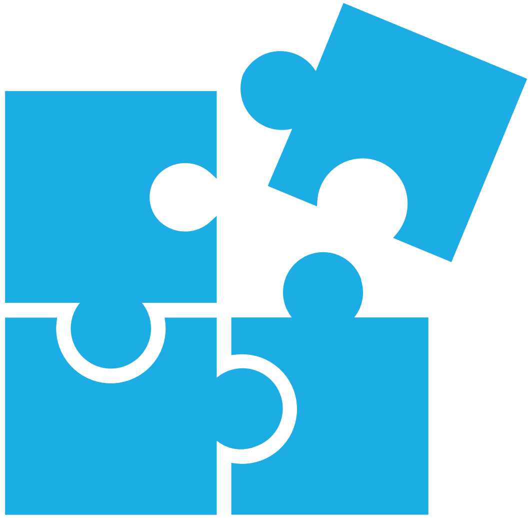
Model Evaluation

Evaluation of the Test Set



Test Accuracy:	0.8217678645187939
Test Recall:	0.7963647998208719
Test Precision:	0.7971568772470732
Test F1 Score:	0.7967577818836973





Model Evaluation
Feature Importance:

- In machine learning models, features are the individual pieces of information used to make predictions.
- Feature importance refers to how much each feature contributes to the model's predictions.
- Understanding feature importance helps identify which features are most influential for the model's decision-making process.

Model Evaluation

Feature Importance

Output Interpretation:

- The output confirms that the sum of all feature importances is close to 1, indicating a valid set of importance scores.
- The sum of the top 6 features' importance is 0.36, which means roughly one-third of the total importance is concentrated in these features.
- By examining the feature_importances_df.head(10) section, we can see the names of the most important features and their corresponding importance scores. This helps identify which features have the strongest influence on the model's predictions.

```
The sum of all the features is 1.0000001192092896
The sum of importances for the top 6 features is 0.36138463020324707
```

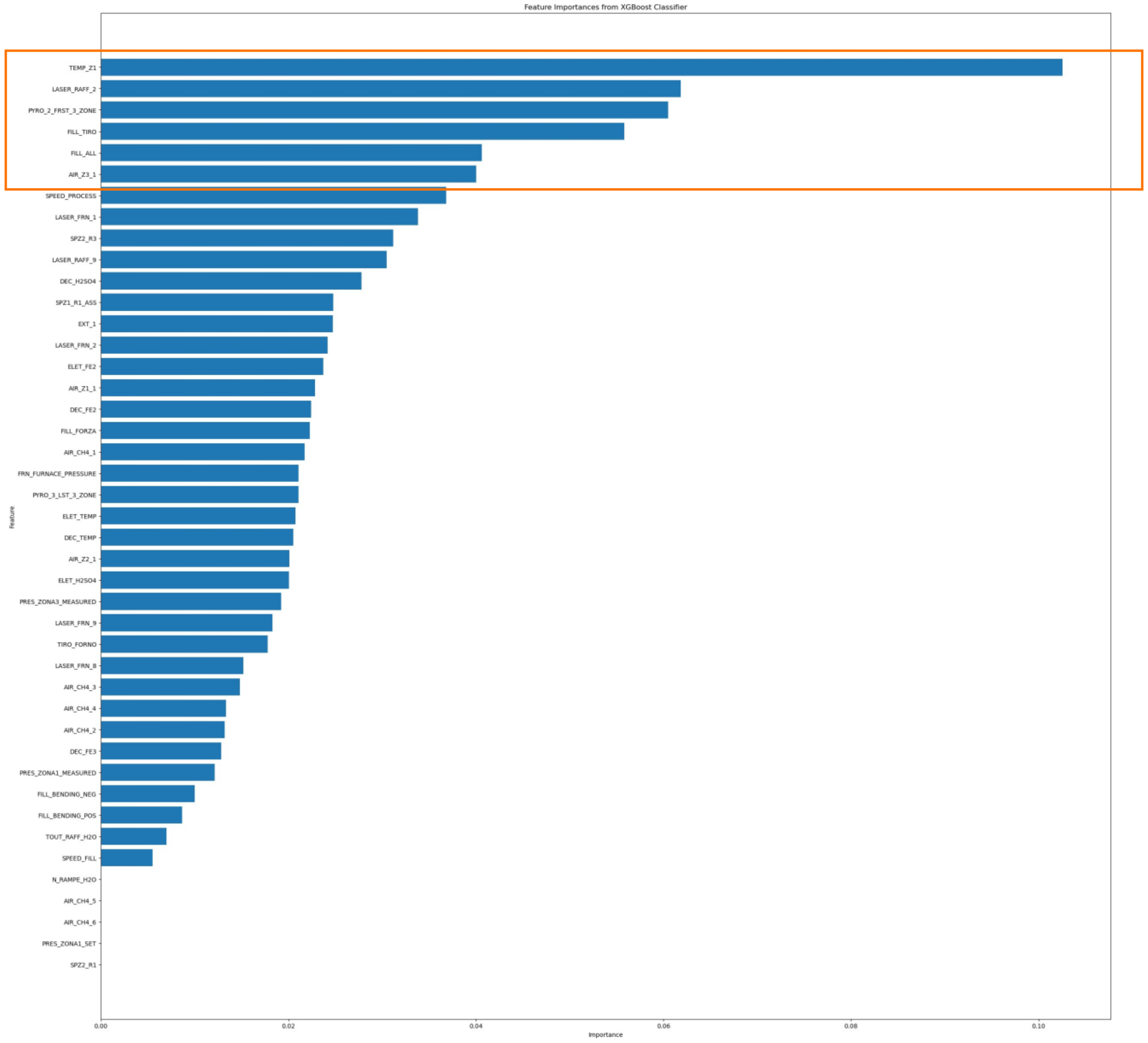
	Feature	Importance
1	TEMP_Z1	0.102561
23	LASER_RAFF_2	0.061827
2	PYRO_2_FRST_3_ZONE	0.060497
39	FILL_TIRO	0.055822
40	FILL_ALL	0.040647
16	AIR_Z3_1	0.040030
0	SPEED_PROCESS	0.036840
4	LASER_FRN_1	0.033820
35	SPZ2_R3	0.031163
24	LASER_RAFF_9	0.030477

Model Evaluation

Feature Importance

Among the 10 most important features, we selected 6 parameters to analyse more in detail:

1) TEMP_Z1
2) LASER_RAFF_2
3) PYRO_2_FRST_3_ZONE
4) FILL_TIRO
5) FILL_ALL
6) AIR_Z3_1



Model Evaluation - Interpretation

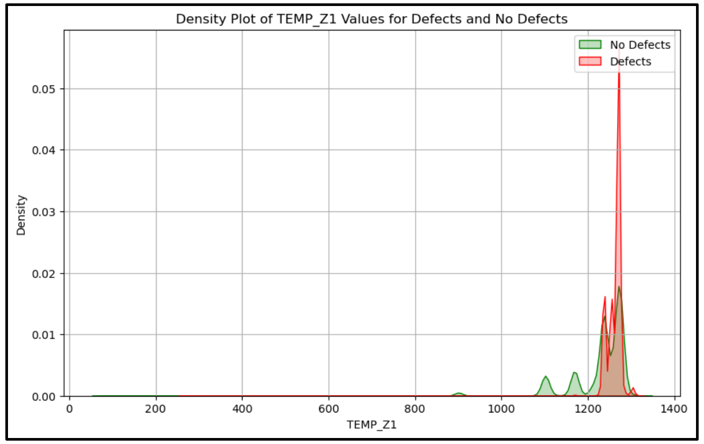
Statistical metrics help decipher the significance of individual features.

Comparing the statistical values of each feature in scenarios where defects are detected and where they aren't enables us to grasp the importance of each feature in discerning defects.

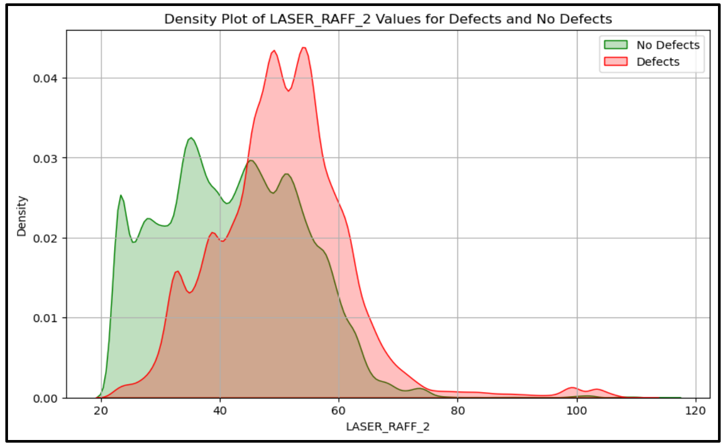
Statistic	TEMP_Z6 No Defects	TEMP_Z6 Defects	COOL_1B No Defects	COOL_1B Defects	TEMP_Z5 No Defects	TEMP_Z5 Defects
Min	79.205925	267.988500	22.500000	22.500000	78.567278	268.998750
Max	1362.465000	1362.536250	888.621750	893.760975	1367.288438	1357.942500
Mean	1272.617338	1297.454358	672.682146	733.207100	1282.360498	1314.116866
Median	1292.369891	1311.682500	697.459179	736.327731	1293.555536	1332.056250
Mode	1293.243750	1316.643750	22.500000	759.988125	1337.861250	1333.293750

Statistic	GAS_Z3_2 No Defects	GAS_Z3_2 Defects	TEMP_Z2 No Defects	TEMP_Z2 Defects	TEMP_Z1 No Defects	TEMP_Z1 Defects
Min	-1.495564	-6.243480	74.225453	257.701050	79.722742	262.855125
Max	336.219750	336.357000	1344.982500	1339.344000	1322.268750	1320.543750
Mean	180.586039	239.405926	1243.609142	1283.077681	1223.142224	1260.480439
Median	179.385188	252.017888	1260.443411	1293.126890	1237.871250	1269.641250
Mode	142.672500	332.311500	1259.921250	1293.795000	1236.712500	1270.383750

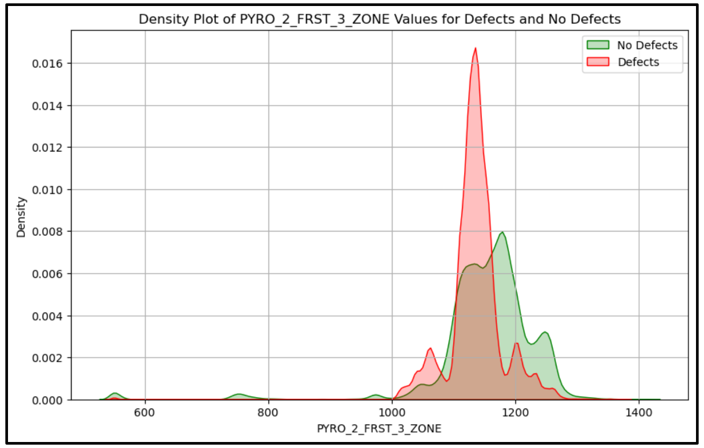
Feature importance: 0.102561



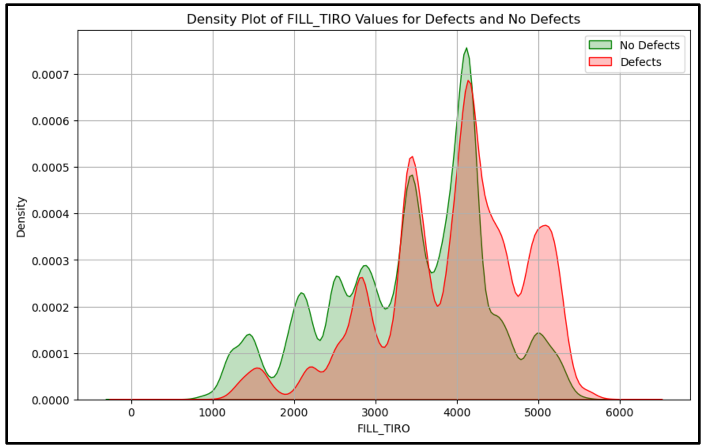
Feature importance: 0.061827



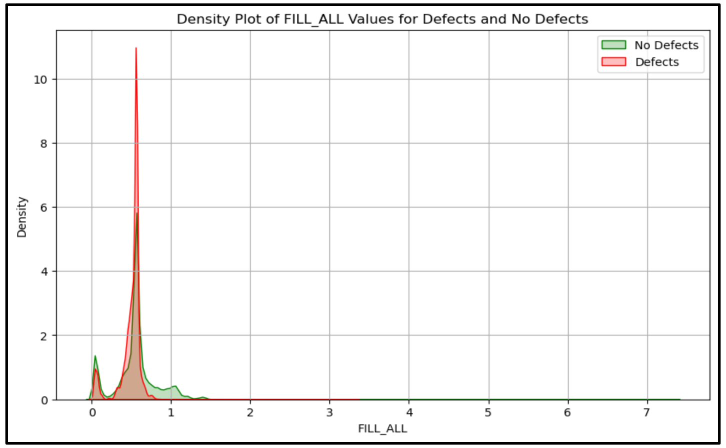
Feature importance : 0.060497



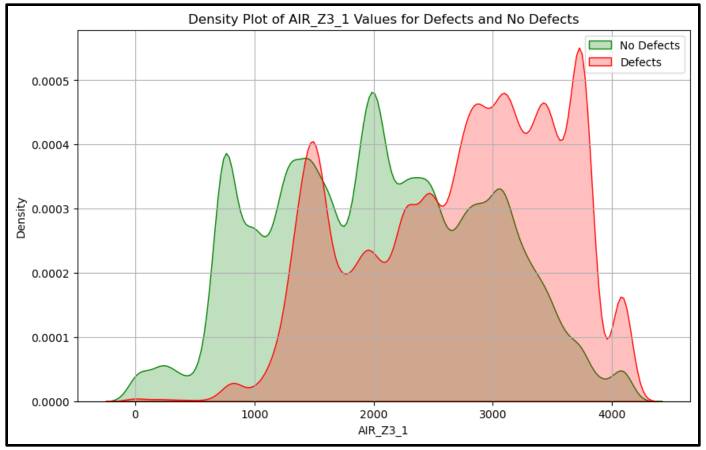
Feature importance: 0.055822



Feature importance: 0.040647



Feature importance: 0.040030



Deployment

Recommendations for the company:

- Plan monitoring and maintenance to control each of the features.
- Monitoring in quality control process

- Future research could take into account the different types of defect, to get more precise outcomes.
  - To do so, it is necessary to understand the importance of each defect.

Thank You For Your Attention!