Indian Institute of Technology Bombay

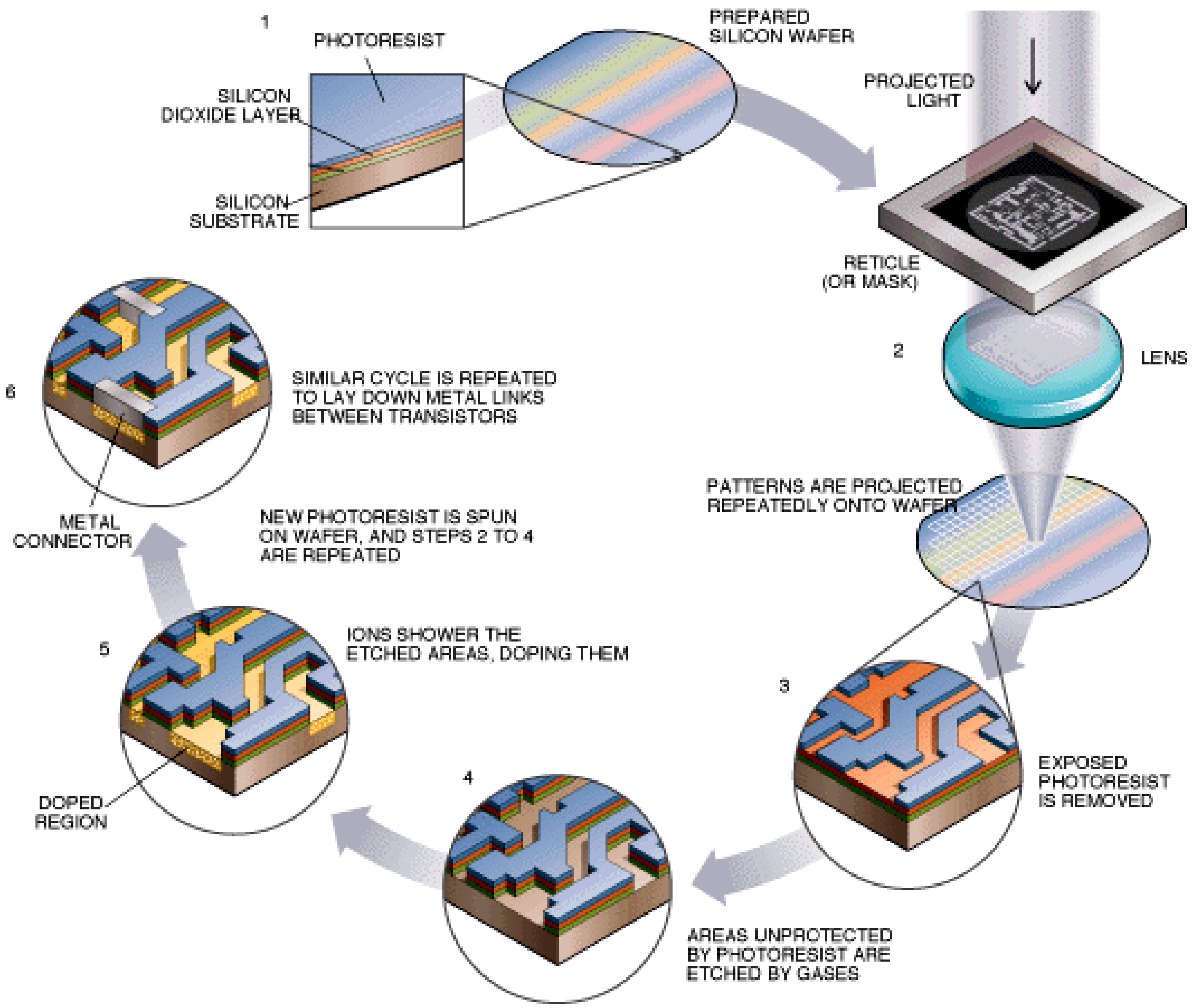
2022 April 24

Yield Prediction in Semiconductor Manufacturing Process

SUPERVISED PROJECT EXPOSITION

PRATHAM SEHGAL, TANMAY BARHARTE, SHIVPRASAD KATHANE

GUIDE: PROF ALANKAR ALANKAR



Indian Institute of Technology
Bombay

PROBLEM STATEMENT

In this project,we build a classifier to predict the Pass/Fail yield of a particular process entity and analyze whether all the features are required to build the model or not.

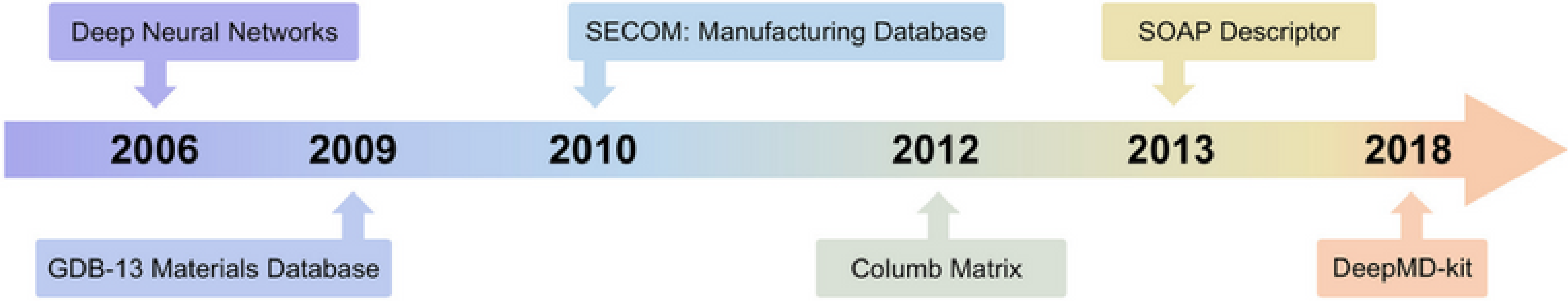


Figure 1

Papers such as the following references were reviewed to familiarise with the current/traditional approaches:
1. K Kerdprasop et al., "Feature Selection and Boosting Techniques to Improve Fault Detection Accuracy in the Semiconductor Manufacturing Process", IMECS (2011)
2. AA Nuhu et al., "Machine learning-based techniques for fault diagnosis in the semiconductor manufacturing process: a comparative study", J Supercomput 79, 2031-2081 (2023)

Indian Institute of Technology
Bombay

MOTIVATION

- Semiconductor manufacturing process is monitored using signals collected from sensors and measurement points.
- Feature selection can be applied to identify the most relevant signals that contribute to yield excursions downstream in the process.
- Analyzing and testing different combinations of features can identify essential signals impacting the yield type, leading to increased process efficiency and decreased production costs.

Good Yield Qualities

- Better product predictability
- Low cost per product
- Predictable schedule adherence and starts planning
- Can run the factory leaner (fewer starts)
- Better quality downstream
- No ‘firefighting’ – more resource for project work

Indian Institute of Technology
Bombay

OUTLINE

Understanding the Data

Data Cleaning

Models without PCA

Models with PCA

Conclusions

Indian Institute of Technology
Bombay

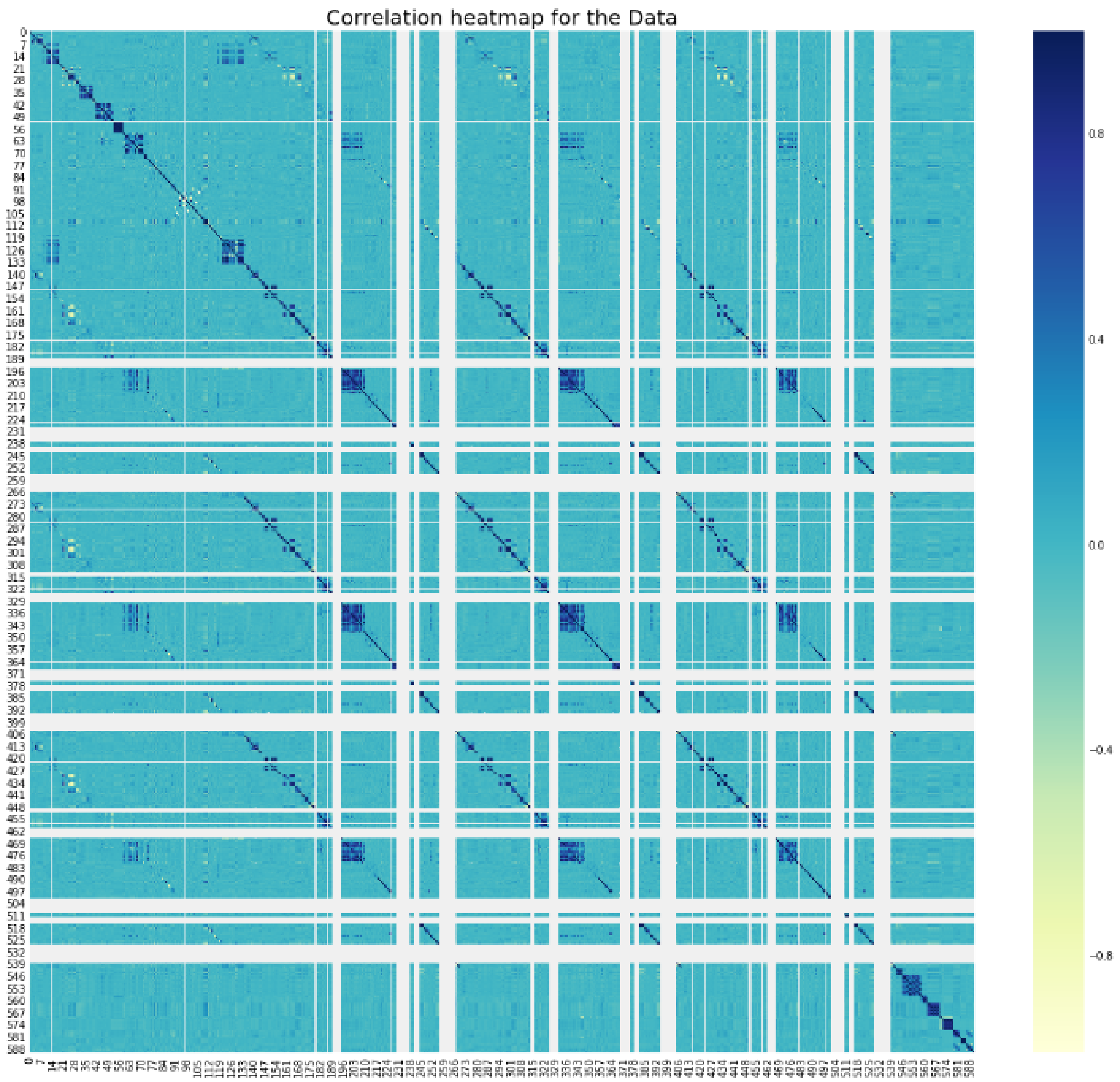
Understanding the Data

Data Visualization

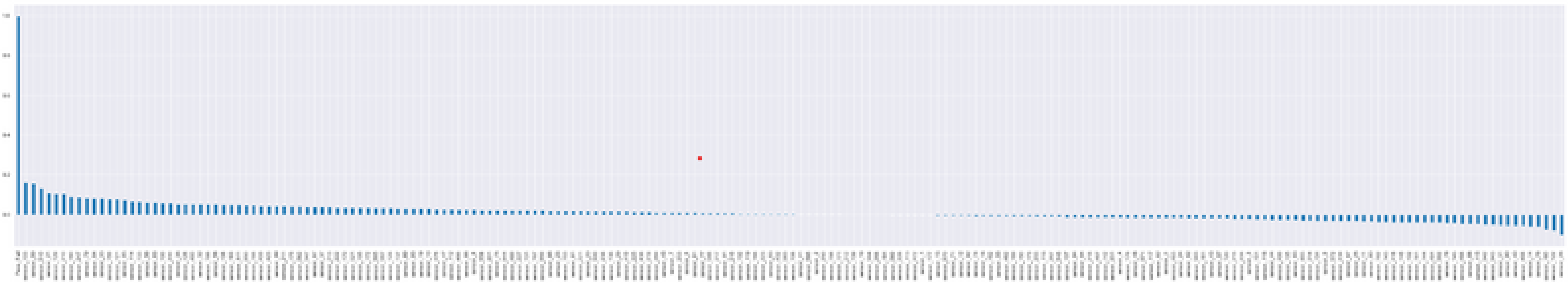
Indian Institute of Technology
Bombay

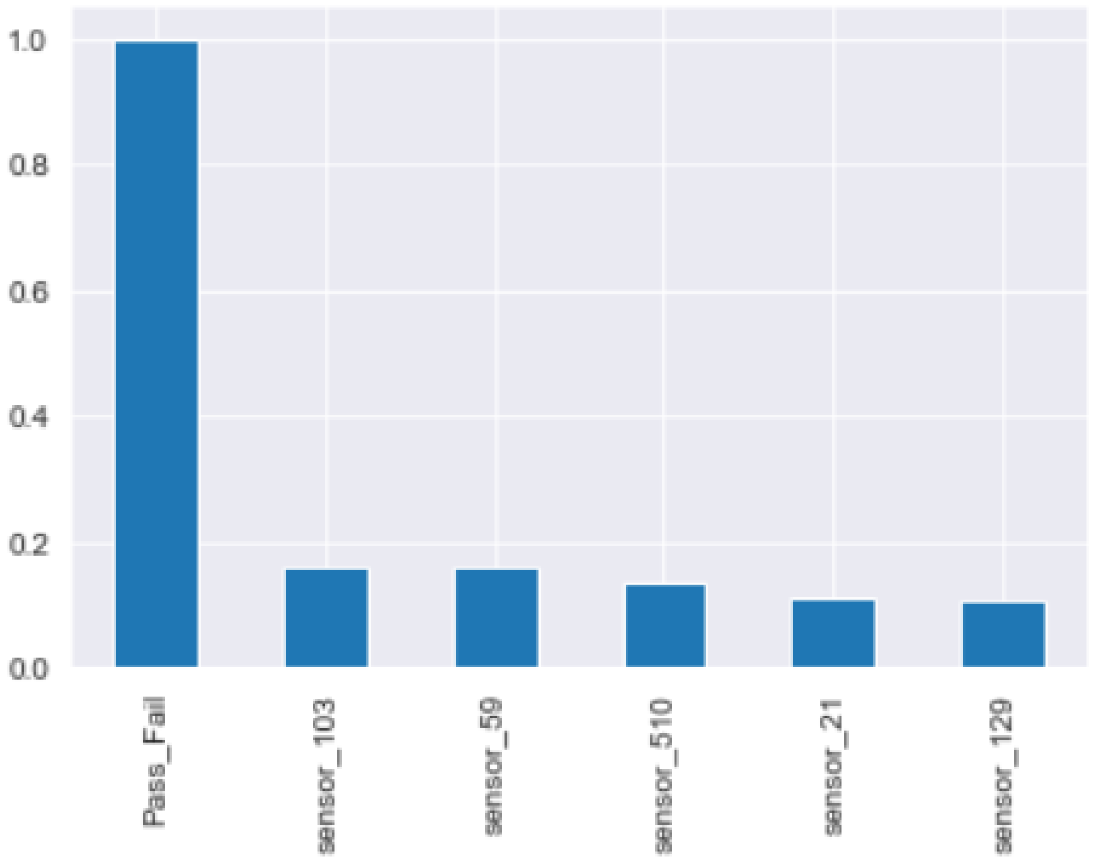
CORRELATION HEATMAP

Violet (dark) regions do appear in the heatmap implying presence of **correlated features**

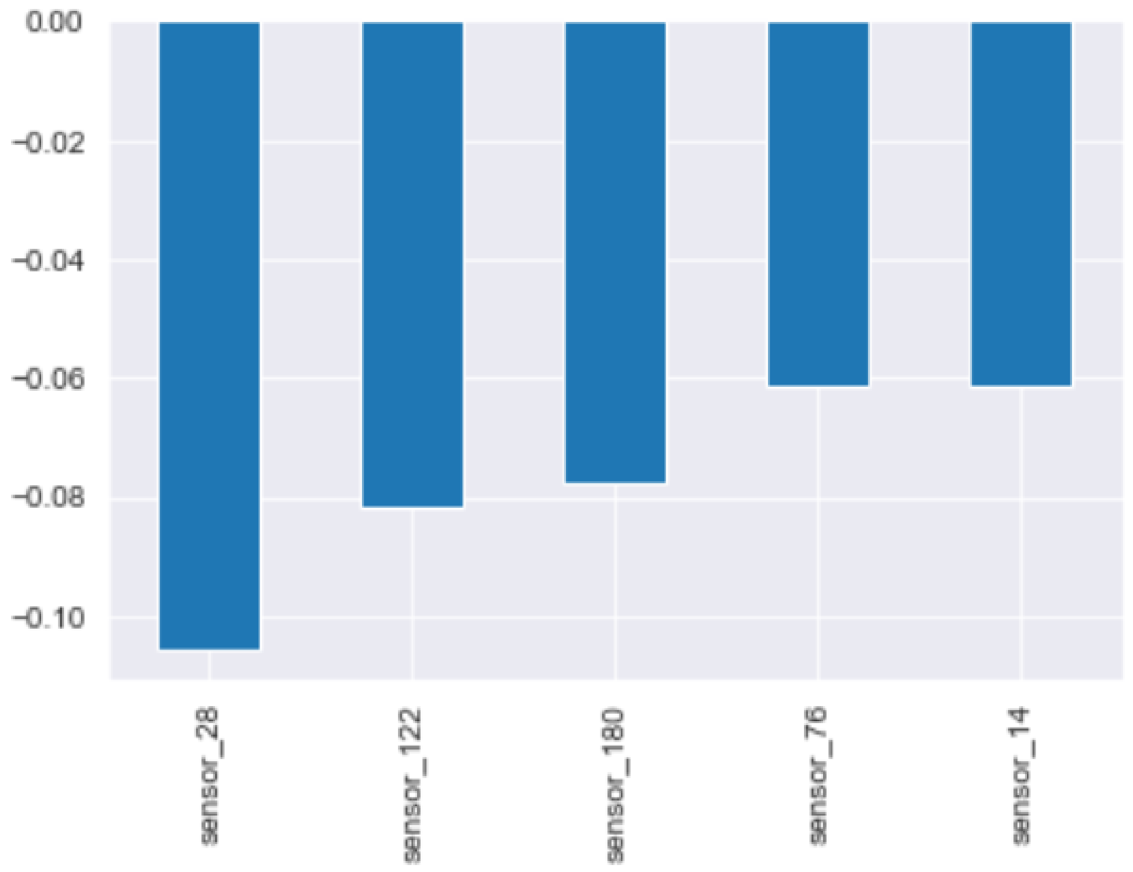


Checking how different features are correlated with target



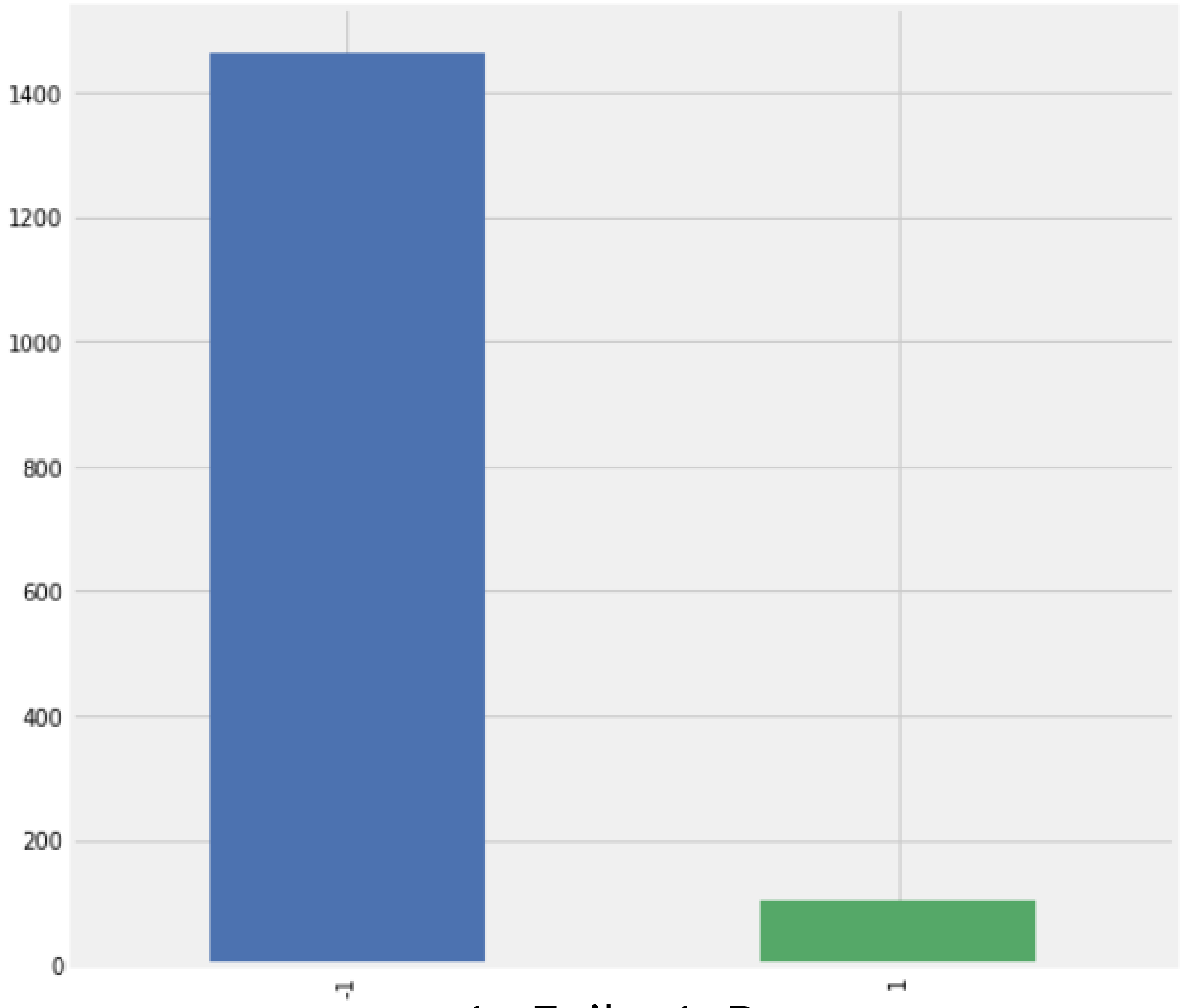


Positive Correlation



Negative Correlation

Indian Institute of Technology
Bombay

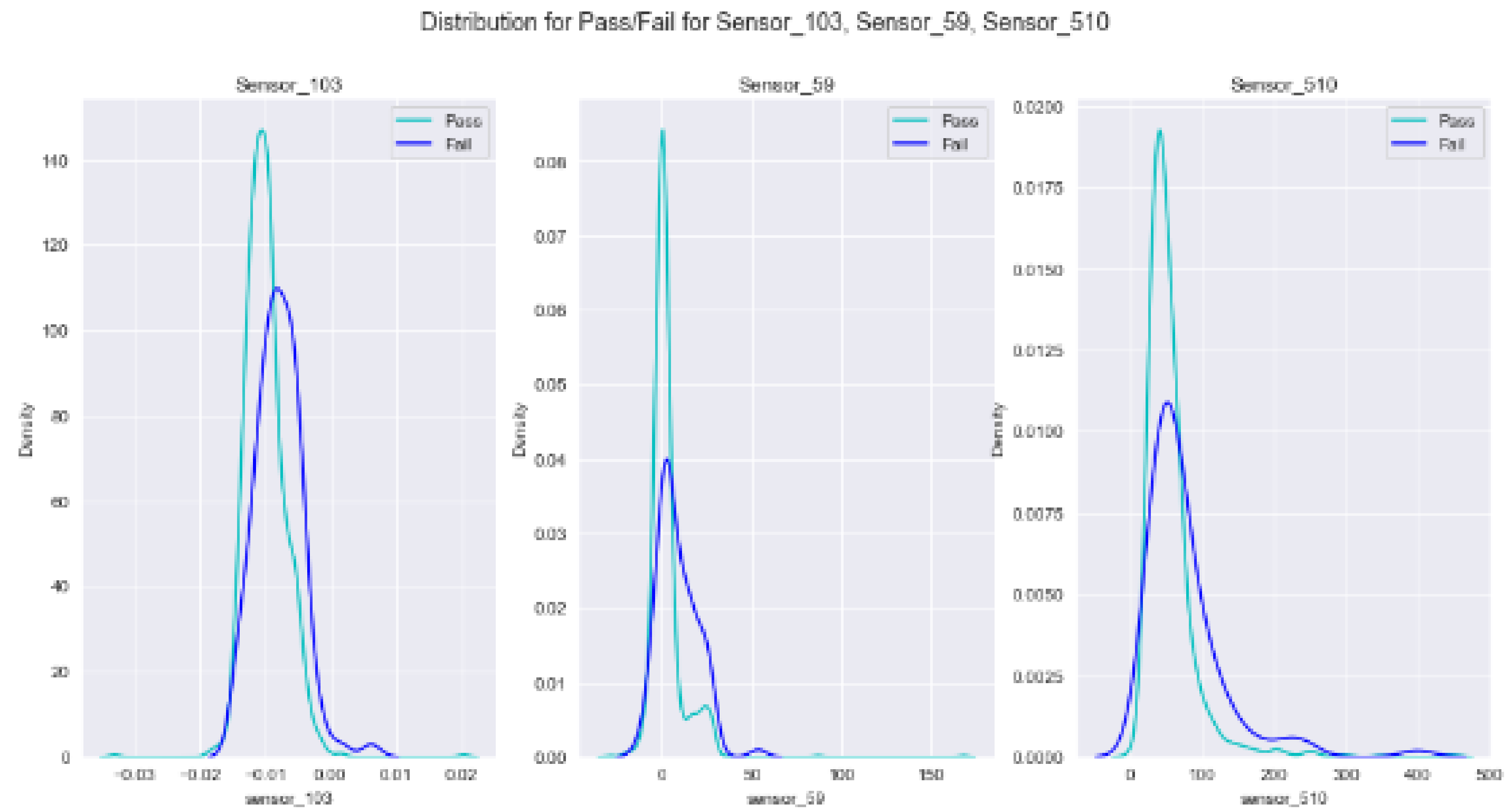


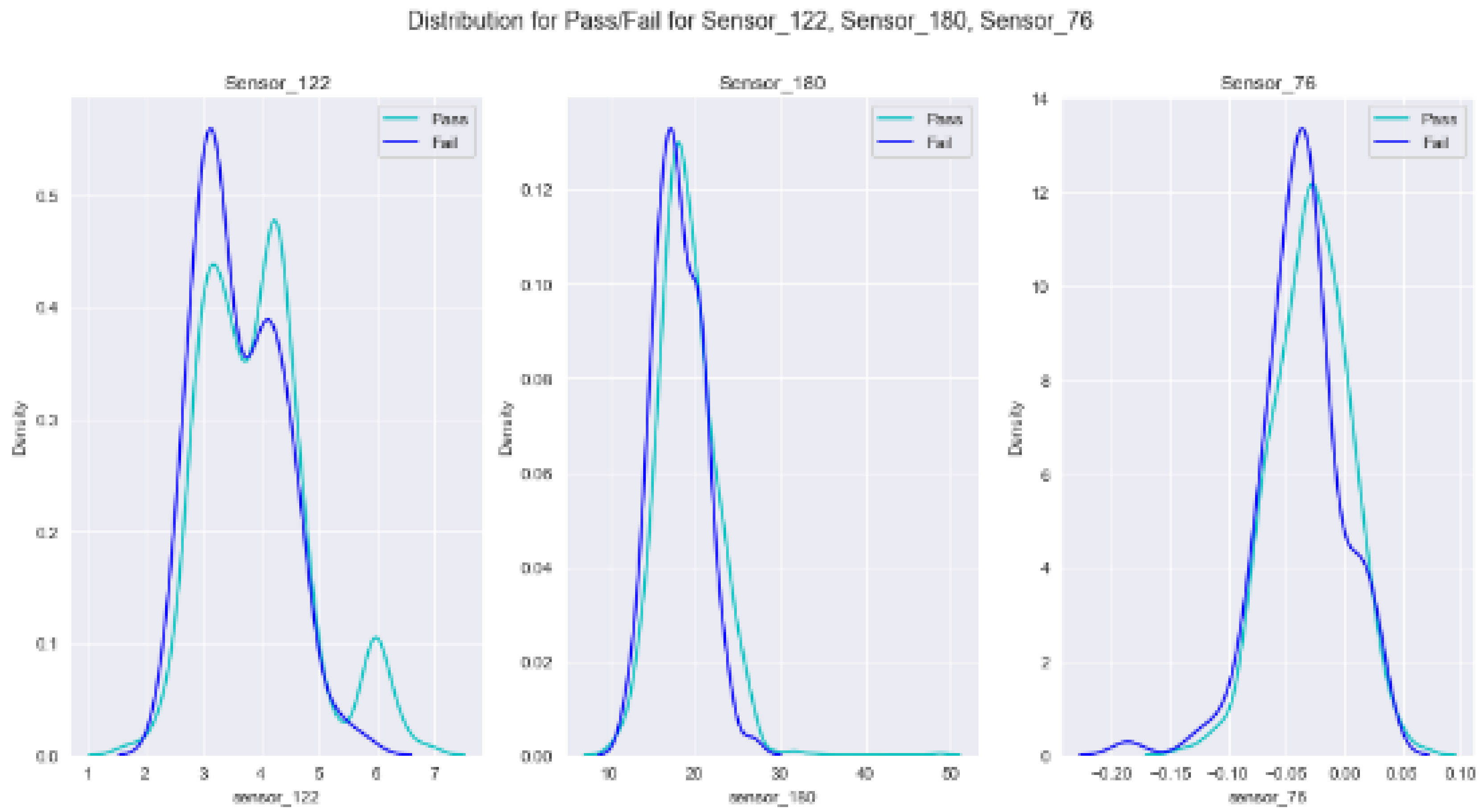
-1 : Fail :: 1: Pass

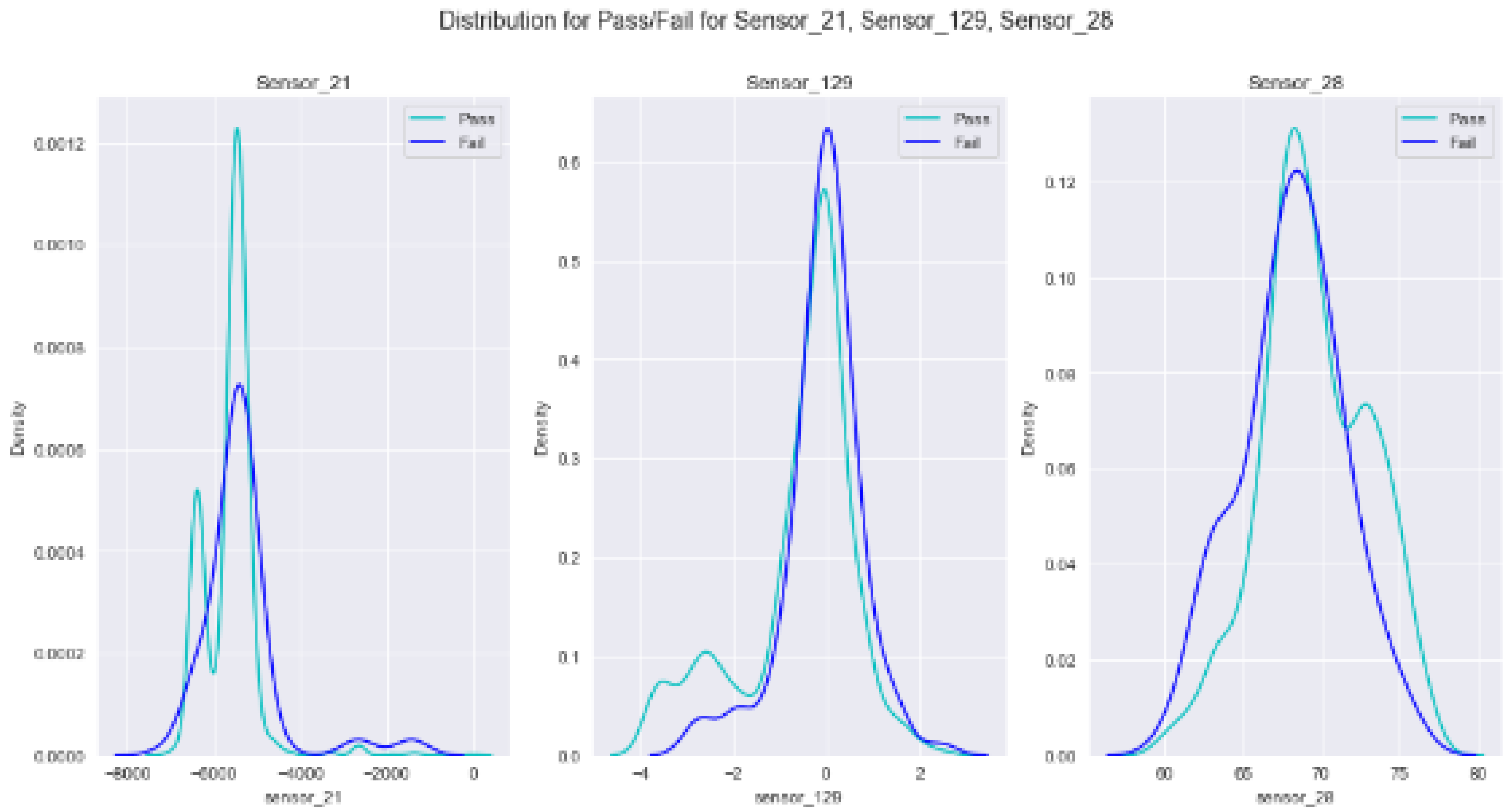
Fail percentage is just 6.65% and pass percentage is 93.35%

It is apparent from the plots that the dataset exhibits class imbalance.

Top 9 features with target Pass/fail comparison.



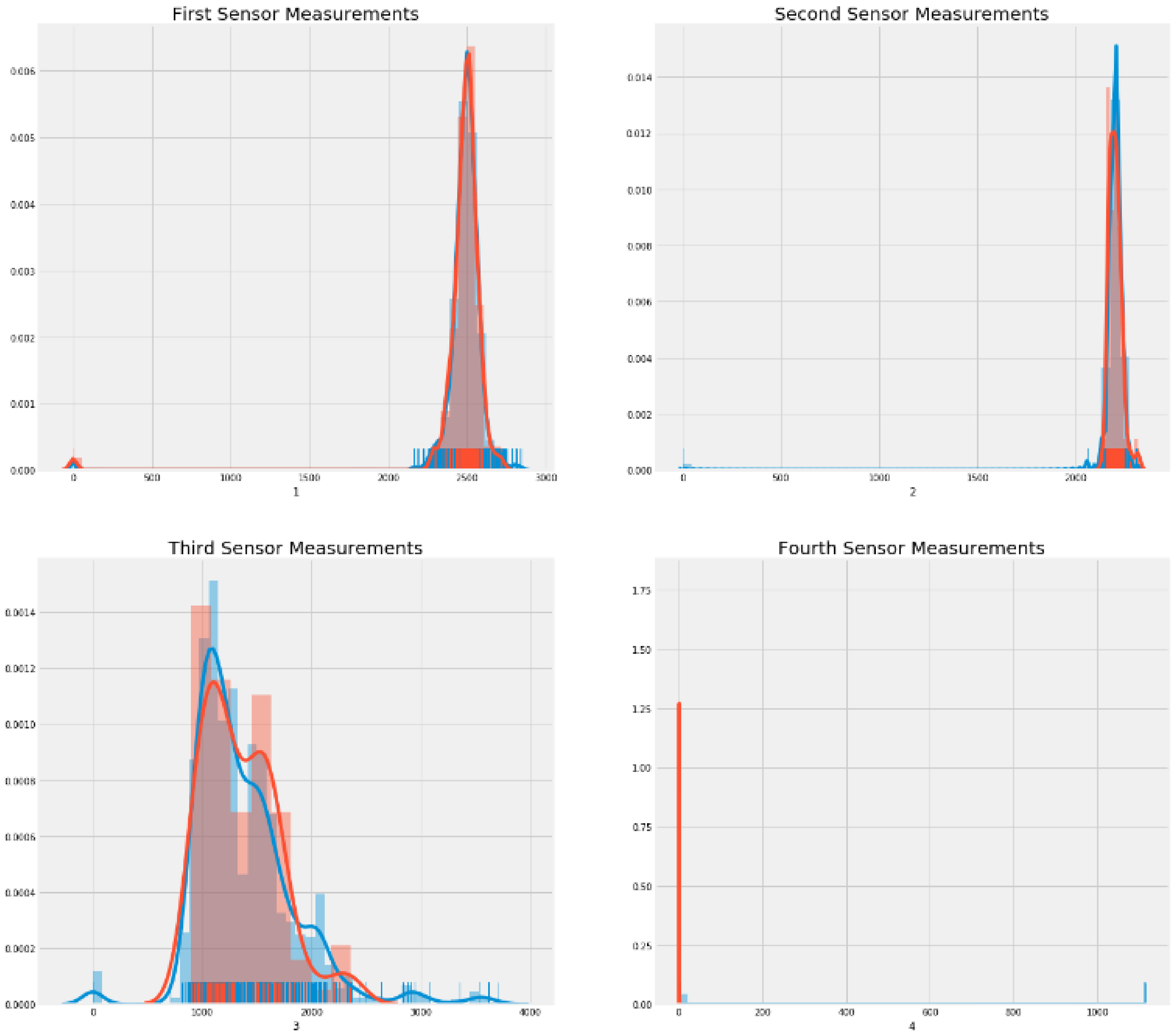




Indian Institute of Technology
Bombay

DATA VISUALIZATION

CHECKING (CLASS-WISE) DISTRIBUTION FOR FIRST 4 SENSOR MEASUREMENTS



Indian Institute of Technology
Bombay

Data Preprocessing

Oversampling, z-test

Procedure

Data Standardization

Handling Imbalance

Balance the target variable with SMOTE technique

Check if the train and test data have similar statistical characteristic

Use One-Sampled Z test to compare a sample mean with the population mean.

Model Fitting

With/ Withoutout PCA and hypertuning

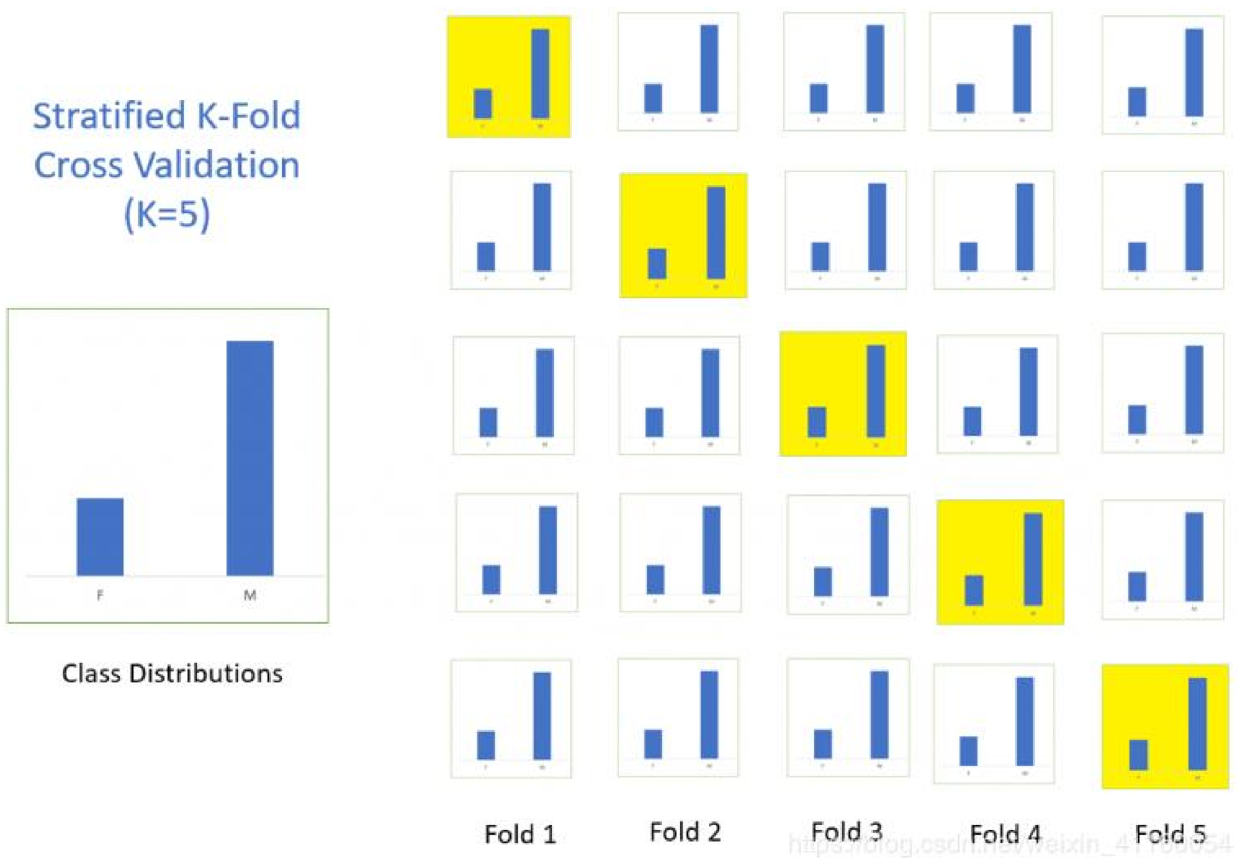
Model	Train_Accuracy	Test_Accuracy	Precision	Recall	F1 Score
XGBClassifier	100.000000	92.841649	0.934066	0.992991	0.962627
CatBoostClassifier	100.000000	92.624729	0.932018	0.992991	0.961538
RandomForest	100.000000	92.407809	0.928105	0.995327	0.960541
BaggingClassifier	99.950199	91.540130	0.933185	0.978972	0.955530
GBClassifier	99.302789	90.238612	0.944316	0.950935	0.947614
AdaBoostClassifier	97.609562	87.635575	0.944844	0.920561	0.932544
DecisionTree	100.000000	83.731020	0.937965	0.883178	0.909747
BernoulliNB	84.511952	79.175705	0.951087	0.817757	0.879397
RidgeClassifier	91.484064	78.308026	0.953039	0.806075	0.873418
Logistic Regression	72.260956	69.414317	0.947040	0.710280	0.811749
KNeigbors	85.308765	66.160521	0.938710	0.679907	0.788618
SVM	71.862550	59.002169	0.928315	0.605140	0.732673
GaussianNB	63.147410	31.887202	0.945312	0.282710	0.435252

RF ,XGB and CBC are considered as best models as the test Accuracy, precision, recall and F1-score are almost very near for all three models.

They have different accuracies for training and test data, implying overfitting but there are other accuracy measures that we can consider to identify the best model.,

Hyper Parametertuning using Grid Search CV

We have appplied Hyperparameter tuning for Random Forest, XGBoost and Catboost Classifier.
The Cross-Validation technique used is Stratified K-fold with K=5.



Overview of SK-Fold Sampling

The Best Parameters obtained are

Criterion	Gini
Max_depth	9
Max_features	12
n_estimators	100
Min_samples_leaf	4
Min_Samples_Split	10

FITTING TOP 3 MODELS WITH TUNING WITHOUT PCA

Model	Train Accuracy	Test Accuracy	K-Fold Mean Accuracy	SK-Fold Mean Accuracy	Precision	Recall	F1 Score
CatBoostClassifier	100.000000	93.275488	98.207214	98.456471	0.944072	0.985981	0.964571
XGBClassifier	100.000000	92.624729	98.506965	98.755847	0.935841	0.988318	0.961364
RandomForest	99.850598	92.624729	98.357463	98.655724	0.939732	0.983645	0.961187

FITTING TOP 3 MODELS WITH TUNING WITH PCA

Model	Train Accuracy	Test Accuracy	K-Fold Mean Accuracy	SK-Fold Mean Accuracy	Precision	Recall	F1 Score
CatBoostClassifier	100.000000	93.1590	93.398641	93.3170	0.931596	1.000	0.964587
XGBClassifier	100.000000	93.159609	93.3170	93.398706	0.931596	1.000	0.964587
RandomForest	99.850598	93.159609	92.828202	92.8286	0.9315956	1.000	0.964587

CONCLUSIONS

- SVM model using principal component analysis performs the best, evidence from above results.
- SVM model able to predict the test daya with 93% accuracy with 100% recall score.
- Tuning hyperparameters yielded/did not yield in an improvement.
- SVM model performance can be improved by repeating PCA steps further.
- There is no feature/sensor that highly attributes with the output.
- The features were reduced from 591 to 203 by using many techniques such as repetition checking, correlation checking etc.
- There are 156 principal components which explains 95% of variance, and are sufficient to predict the pass/fail yield of a process.
- Achieved test and train accuracies remains unchange if different sample population used.

THANKYOU

Link to Code:
https://colab.research.google.com/drive/1T0NPpi3HDqZYlUkDIq2c1bAwUGQXG24n