Yield Prediction in Semiconductor Manufacturing

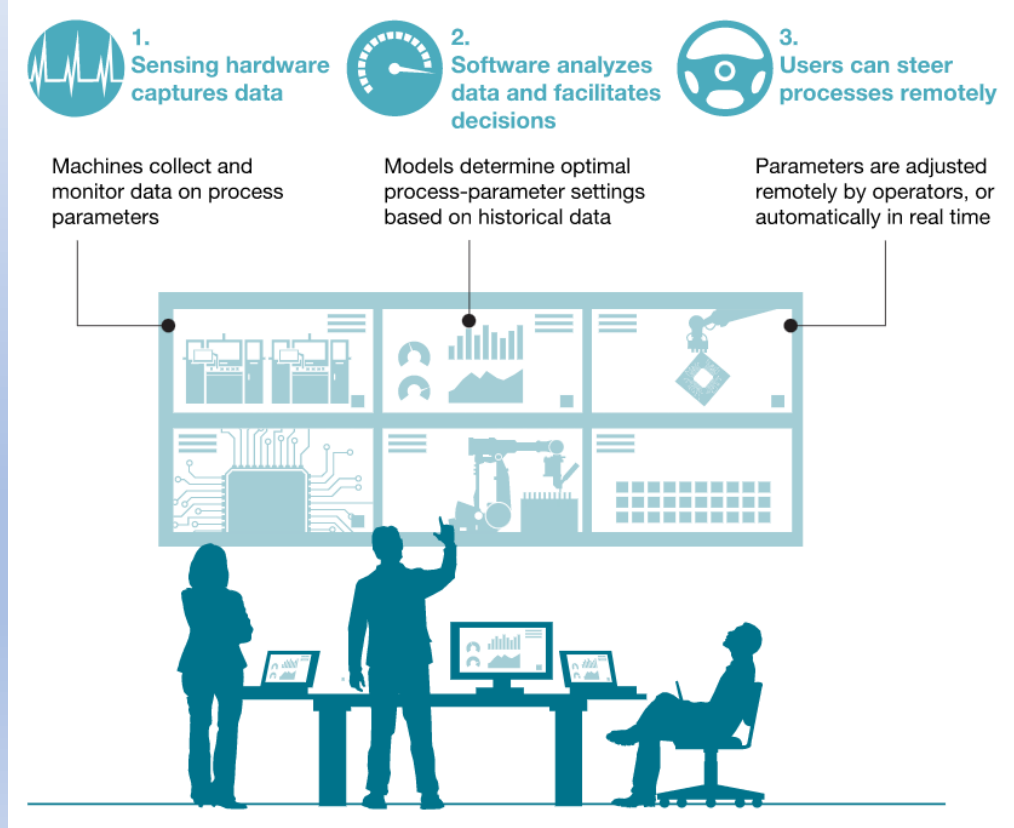
Project Proposal for course ME 793 – 2023

Team ID: 7

Shivprasad Kathane - 180110076

Tanmay Barhate - 19d170033

Pratham Sehgal - 19d170019



Motivation and Objective

Dataset: UCI SECOM available at https://archive.ics.uci.edu/ml/datasets/SECOM

Motivation:

- **Maintaining the production of high quality material is very crucial in the semiconductor industry**
- **Various process variables influence the semiconductor manufacturing process**
- **The variation in the values of these variables is recorded via sensors**
- **Each semiconductor product is tested for quality in the downstream and declared pass/fail yield**
- **It is expected that these variables influence the outcome and thus data analysis is essential**
- **A model to predict the yield beforehand based on data from sensors is desirable so as to take corrective actions thereby enabling an increase in process throughput, decreased time to learning and reduction in the per unit production costs.**

Objective:

Yield Prediction in Semiconductor Manufacturing Process via

- **Developing intelligent feature selection/transformation techniques**
- **Modelling the rare occurrence of the defective product using ML/DL**

A typical semiconductor manufacturing process Ref

The data and the problem was discovered during the literature review on ML applications for semiconductors
Ref: DY Liu et al. “Machine learning for semiconductors”. *Chip* **1** (2022) https://doi.org/10.1016/j.chip.2022.100033

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)

Most focus on feature selection (say PCA, RFE) + class balancing (say SMOTE) + Tree based ML Details Here

	0	1	2	3	4	5	6	7	8	9	...	580	581	582	583	584	585	586	587	588	589
0	3095.78	2465.14	2230.4222	1463.6606	0.8294	100	102.3433	0.1247	1.4966	-0.0005	...	0.006	208.2045	0.5019	0.0223	0.0055	4.4447	0.0096	0.0201	0.006	208.2045
1	2932.61	2559.94	2186.4111	1698.0172	1.5102	100	95.4878	0.1241	1.4436	0.0041	...	0.0148	82.8602	0.4958	0.0157	0.0039	3.1745	0.0584	0.0484	0.0148	82.8602
2	2988.72	2479.9	2199.0333	909.7926	1.3204	100	104.2367	0.1217	1.4882	-0.0124	...	0.0044	73.8432	0.499	0.0103	0.0025	2.0544	0.0202	0.0149	0.0044	73.8432
3	3032.24	2502.87	2233.3667	1326.52	1.5334	100	100.3967	0.1235	1.5031	-0.0031	...	NaN	NaN	0.48	0.4766	0.1045	99.3032	0.0202	0.0149	0.0044	73.8432
4	2946.25	2432.84	2233.3667	1326.52	1.5334	100	100.3967	0.1235	1.5287	0.0167	...	0.0052	44.0077	0.4949	0.0189	0.0044	3.8276	0.0342	0.0151	0.0052	44.0077
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1561	2899.41	2464.36	2179.7333	3085.3781	1.4843	100	82.2467	0.1248	1.3424	-0.0045	...	0.0047	203.172	0.4988	0.0143	0.0039	2.8669	0.0068	0.0138	0.0047	203.172
1562	3052.31	2522.55	2198.5667	1124.6595	0.8763	100	98.4689	0.1205	1.4333	-0.0061	...	NaN	NaN	0.4975	0.0131	0.0036	2.6238	0.0068	0.0138	0.0047	203.172
1563	2978.81	2379.78	2206.3	1110.4967	0.8236	100	99.4122	0.1208	NaN	NaN	...	0.0025	43.5231	0.4987	0.0153	0.0041	3.059	0.0197	0.0086	0.0025	43.5231
1564	2894.92	2532.01	2177.0333	1183.7287	1.5726	100	98.7978	0.1213	1.4622	-0.0072	...	0.0075	93.4941	0.5004	0.0178	0.0038	3.5662	0.0262	0.0245	0.0075	93.4941
1565	2944.92	2450.76	2195.4444	2914.1792	1.5978	100	85.1011	0.1235	NaN	NaN	...	0.0045	137.7844	0.4987	0.0181	0.004	3.6275	0.0117	0.0162	0.0045	137.7844

1566 rows × 590 columns

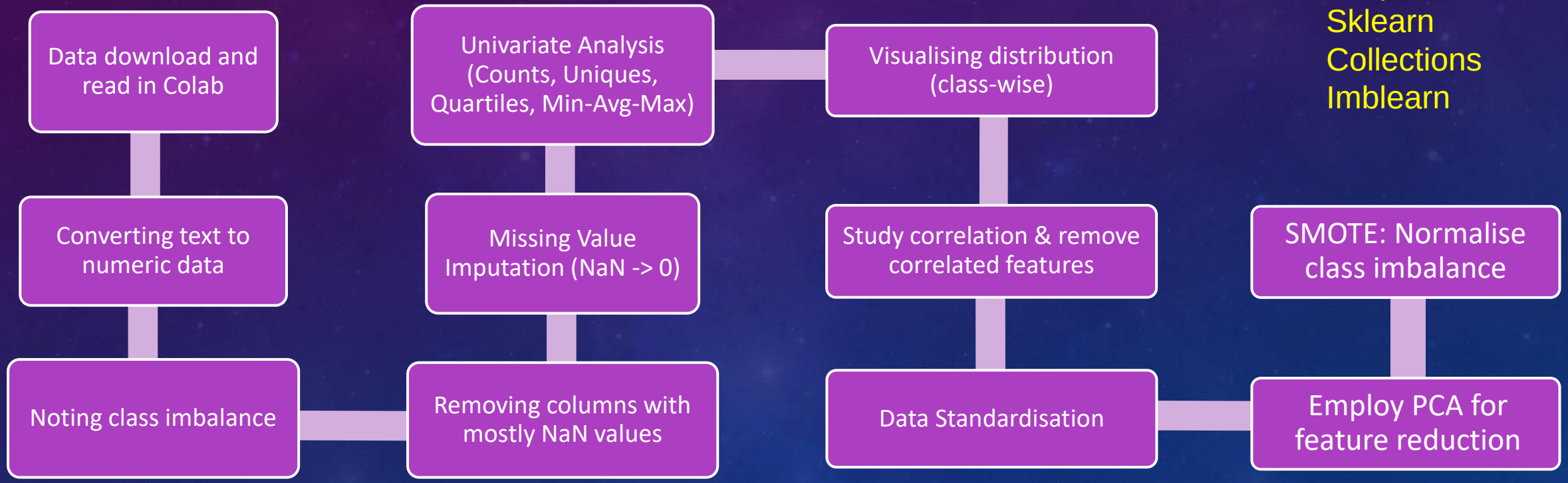
Data

0	-1 "19/07/2008 12:32:00"
1	1 "19/07/2008 13:17:00"
2	-1 "19/07/2008 14:43:00"
3	-1 "19/07/2008 15:22:00"
4	-1 "19/07/2008 17:53:00"
...	...
1561	-1 "16/10/2008 15:13:00"
1562	-1 "16/10/2008 20:49:00"
1563	-1 "17/10/2008 05:26:00"
1564	-1 "17/10/2008 06:01:00"
1565	-1 "17/10/2008 06:07:00"

Methodology (Implemented till now)

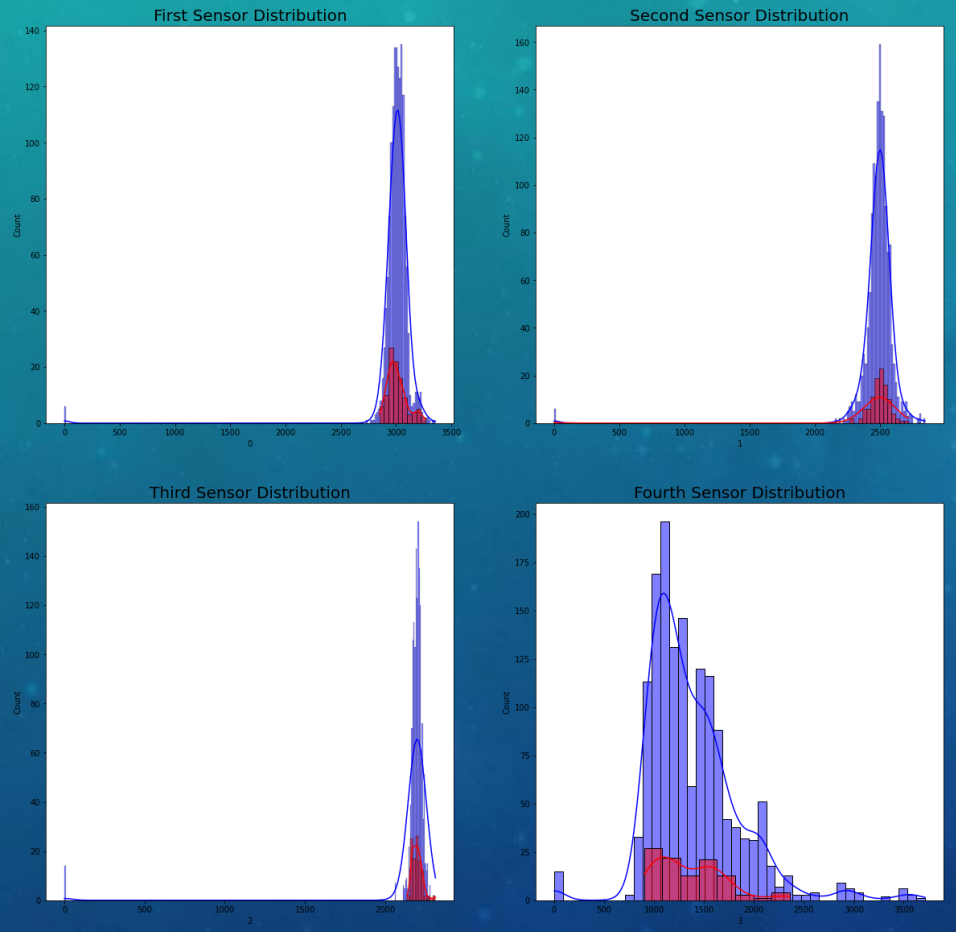
Libraries:

Numpy
Pandas
Seaborn
Matplotlib
Sklearn
Collections
Imblearn

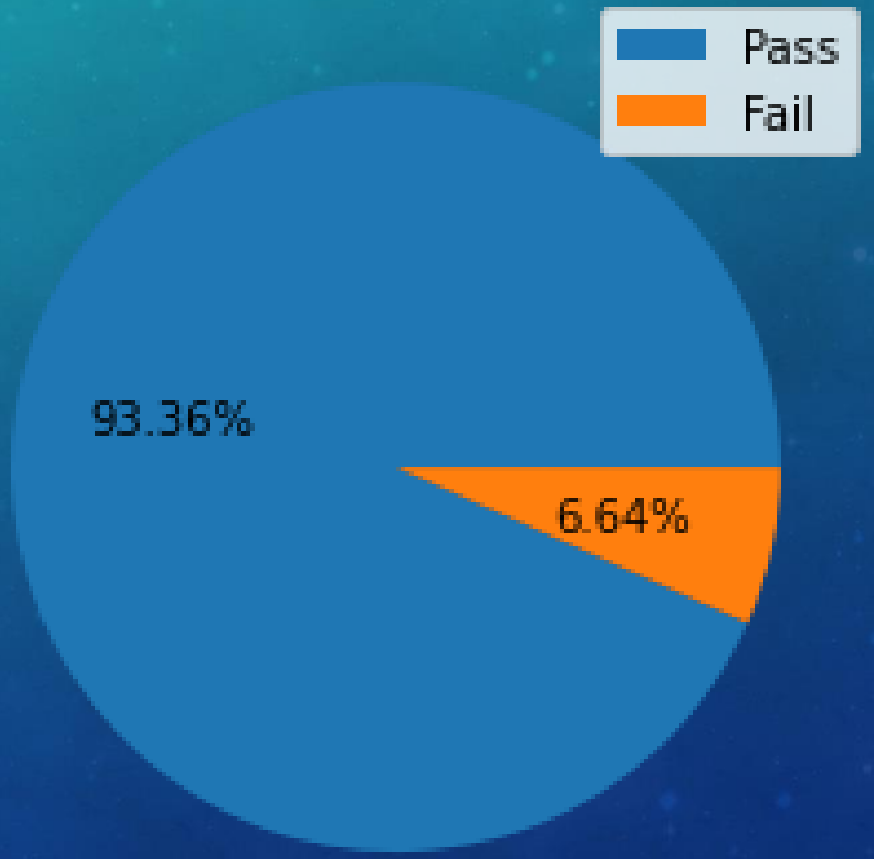


Data Exploration

Distribution of first 4
sensor measurements

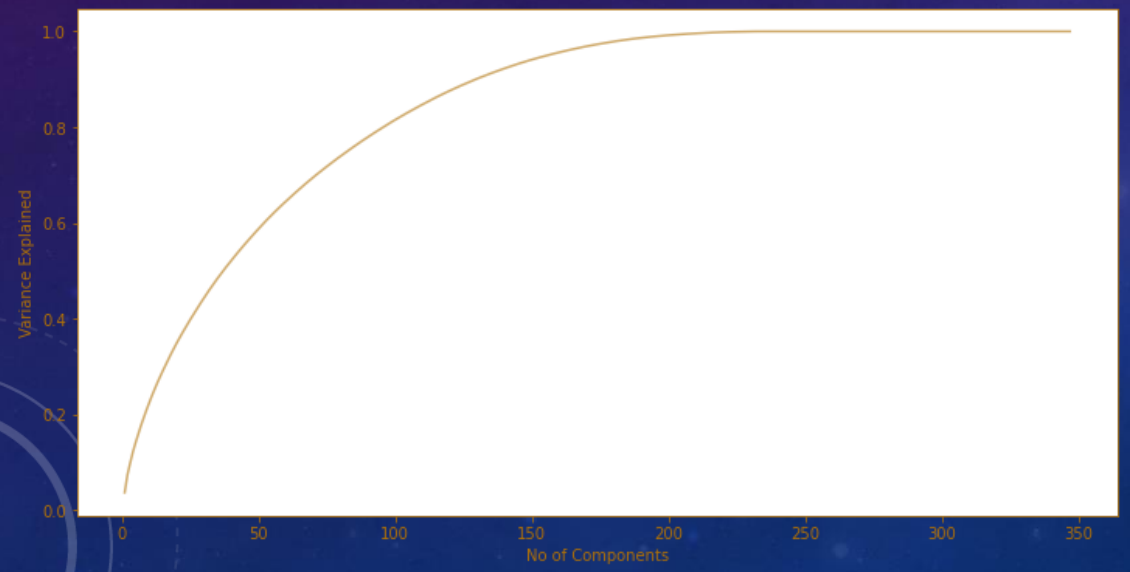


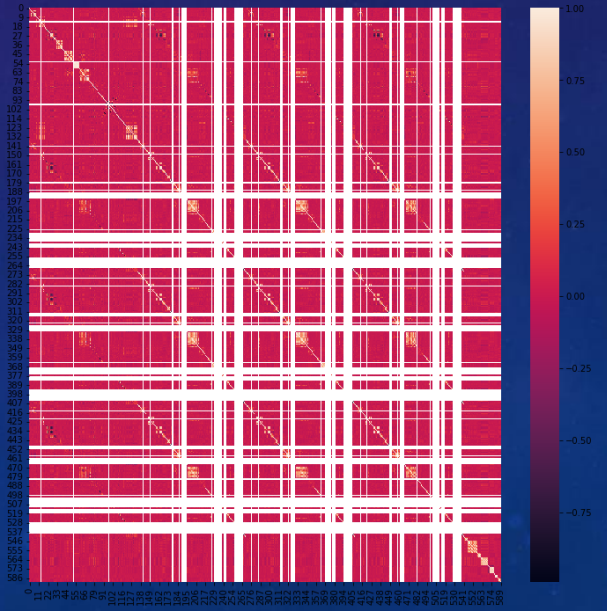
Existence of
Class Imbalance



Feature Selection

- Removed 28 features which had more than half data as null values
- Presence of correlated features found in correlation matrix / heatmap
- Removed 215 variables with correlation between a pair of features > 0.9
- Employed PCA ---> Top 200 components explain almost all of the variance
- Transformed data has 200 features down from initial 590 sensor variables





Challenges & Future Plan

- Applying ML Classification Algorithms (Ex: XG Boost)
- Cross-Validation for various hyperparameters
- Taking Care of Class Imbalance
  - Oversampling using SMOTE
  - Evaluation using F1 score
- Experimenting with Deep Learning techniques
- Playing with various feature selection techniques

Q (f)

1.
Pratham = Analysing existing research work and preparing project plan and methodology
Tanmay = Data cleaning and correlation analysis
Shivprasad = Data exploration, feature selection and slides

2.
Pratham = 30%, Tanmay = 15%, Shivprasad = 55%