Project Report: Semiconductor Manufacturing Process

**Introduction**

A complex modern semiconductor manufacturing process is normally under constant surveillance via the monitoring of signals/variables collected from sensors and or process measurement points. However, not all of these signals are equally valuable in a specific monitoring system. The measured signals contain a combination of useful information, irrelevant information as well as noise. Engineers typically have a much larger number of signals than are actually required. If we consider each type of signal as a feature, then feature selection may be applied to identify the most relevant signals. The Process Engineers may then use these signals to determine key factors contributing to yield excursions downstream in the process. This will enable an increase in process throughput, decreased time to learning and reduce the per unit production costs. These signals can be used as features to predict the yield type. And by analysing and trying out different combinations of features, essential signals that are impacting the yield type can be identified.

**Dataset Description**

sensor-data.csv : (1567, 592)

The data consists of 1567 datapoints each with 591 features.  
The dataset presented in this case represents a selection of such features where each example represents a single production entity with associated measured features and the labels represent a simple pass/fail yield for in house line testing. Target column “ –1” corresponds to a pass and “1” corresponds to a fail and the data time stamp is for that specific test point.

**Project Objective**

We will build a classifier to predict the Pass/Fail yield of a particular process entity and analyse whether all the features are required to build the model or not.

**Data Pre-processing**

Data pre-processing is the initial steps when we perform data analysis and applying machine learning algorithms on the same data. In simple word, “Data pre-processing is a data mining technique which is used to transform the raw data in a useful and efficient format”.

Here below some steps that are used for data analysis:

**Data Cleaning: Remove missing values by filling Mean or median, remove duplicate values, remove unwanted values (for example: ?, ..etc)**

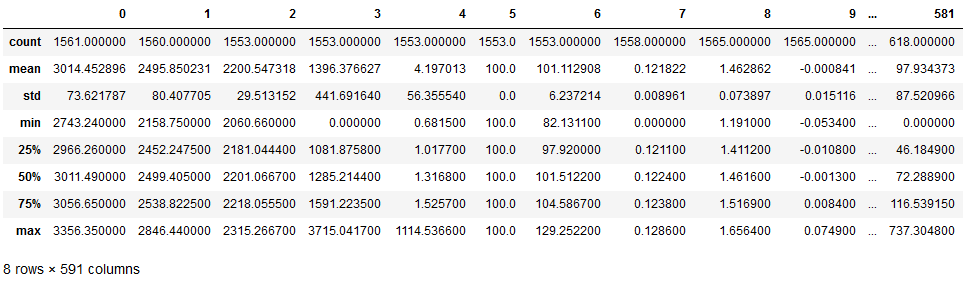
**In this project we remove all null values features that have more than 20%+ null values and impute other by mean of the features.**

**After this drop the all the features that have same values of the.**

**Summary Statistics**

Summary Statistics is summarizing the data at hand through certain numbers like mean, std etc. so it makes the data easier to understand. It used find statistic information of all numeric In machine learning we have use simple predefined function describe() to show all summary related to dataset.

See in below snippets.



**Count:** It used to count the all-feature values of dataset columns

**Mean:** This is the statistic term which used to find the mean of each numeric feature columns in dataset

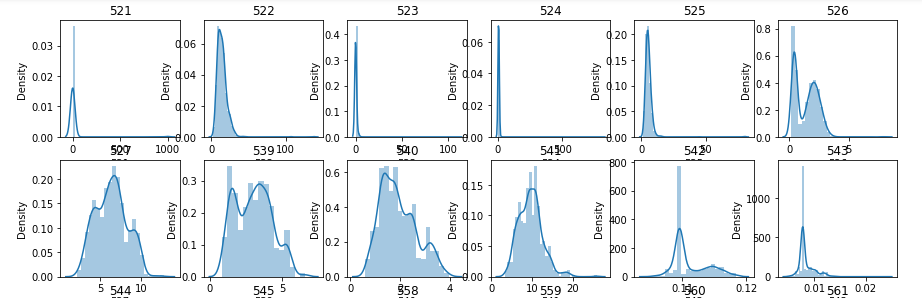
**Std:** It means standard deviation; this is also statistic term to find the standard deviation of dataset features.

**Data Visualization**

In this two types of data visualization used:

* Univariate Analysis
* Multivariate Analysis

**Univariate Analysis:** Perform a univariate Analysis with appropriate detailed comments after each analysis.



**Machine Learning Model Implementation**

**Supervised Machine Learning**

In this project we use the two supervised machine learning techniques to predict the result.

* Logistic Regression
* DecisionTreeClassifier

**Logistic Regression:** Logistic regression is commonly used for prediction and classification problems.

**DecisionTreeClassifier:** Decision Tree is a **Supervised learning technique**that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier.

**Model Evaluation Metrics**

* **Accuracy:** it compares the predicted general sentiment (positive or negative) to the real one, which was determined based on the stars or Accuracy is the most intuitive performance measure. It explains that how close a measurement is to the correct value for that measurement. It is defined as ratio of correctly predicted observation to the total observation. It can be calculated by using following expression:

Accuracy = (TP + TN)/ (TP + TN + FP + FN)

* **Precision:** this is the ration between True Positives and the sum of True Positives and False Positive reviews. It tells us how accurate we are about saying that a review is positive or Precision is how close two or more measurements are to each other. It is calculated as the number of correct positive predictions divided by the total number of positive predictions. It is defined as:

Precision = TP/ (TP + FP)

* **Recall:** this is the ration between True Positives and the sum of True Positives and False Negatives or Recall or True positive rate is calculated as the number of correct positive predictions divided by the total number of positives.

Recall = TP (TP + FN)

* **F1-score:** this is the harmonic mean of the precision and the recall or F1 Score is an overall accuracy of the model. It is harmonic mean of precision and recall. It can be defined as:

F1 Score = 2 x (Precision x Recall ) /(Precision + Recall )

**Result**

|  |  |
| --- | --- |
| Algorithm(Logistic Regression) | |
| Train, Test Accuracy | **0.85, 0.87** |
| Algorithm(DecisionTreeClassifer) | |
| Train, Test Accuracy | **1.0, 0.91** |

**Conclusion**

In this task, we have find the performance using supervised machine learning techniques (**Logistic Regression** and **Decision Tree Classifier**) for given data. Here we find the good accuracy by selecting the features from given dataset. We know accuracy or result are depending on datasets so we can say that, in future, such prediction can be make more accurate by using more features and by making large size of dataset. In this we find the Logistic Regression Model **accuracy 0.87** which is good, and for Decision Tree is **0.91.**

**As per above result we can say DecisionTreeClassification is the better choice for this dataset.**