**Problem Statement**

To build a classification methodology to predict the quality of wafer sensors based on the given training data.

**Architecture**



**Data Description**

The client will send the data in multiple sets of files in batches at a specified location. Data will contain Wafer names and 590 columns of different sensor values for each wafer. The last column will have the “Good/Bad” value for each wafer.

“Good/Bad” column will have two unique values +1 and -1

“+1”: represents bad wafer

“-1“: represents good wafer

Apart from training files, we also require a “schema” file from the client, which contains all the relevant information about the training files such as:

1. Name of the files
2. Length of date value in file name
3. Length of time value in file name
4. Number of columns
5. Name of the columns with datatype

**Data Validation**

In this step, we perform different sets of validation on the given set of training files.

1. Name Validation: We validate the name of the files based on the given name in the schema file. We have created a regex pattern as per the name specified in schema file. After validating the pattern in the name, we check for the length of data in the file name as well as the length of time in the file name. if all the values are as per the requirement, we move such files to “Good\_Data” directory else move such files to “Bad\_Data”
2. Number of Columns: We validate the number of columns present in the files, and if it doesn’t match with the value given in the schema file, then the file is moved to “Bad\_Data”
3. Name of Columns: The name of the columns is validated and should be the same as given in the schema file. If not, then the file is moved to “Bad\_Data”
4. The datatype of columns: The datatype of columns is given in the schema file. This is validated when we insert the files into database. If the datatype is wrong, then move the file to “Bad\_Data”.
5. Null Values in Column: If any of the columns in a file have all the values as NULL or missing, we discard such a file and move it to “Bad\_Data”

**Data Insertion**

1. Database Creation & Connection: create a database named as “waferdb”. If the database is already created, open the connection to the database.
2. Collection Creation in Database: Collection with name – “waferCollection” create in database “waferdb” and insert all the data files from “Good\_Data”

**Model Training**

1. Data Export from DB: The data is stored in MongoDB in database “waferdb” in collection “waferCollection”, export this collection in CSV file for model training
2. Data Preprocessing:
   1. Check for NULL Value: If null value(s) present, impute the null values using KNN imputer.
   2. Columns with Zero Standard Deviation: Identify such columns and drop them as these columns do not contribute in model training.
3. Clustering: KMeans algorithm is used to create clusters in the preprocessed data. The optimum number of clusters is selected by plotting the elbow plot, and for the dynamic selection of the number of clusters, we are using “KneeLocator” function. The idea behind clustering is to implement different algorithms. To train the data in different clusters, the KMeans model is trained over preprocessed data and the model is saved for further use in prediction.
4. Model Selection: After clusters are created, we find the best model for each cluster. We are using two algorithms, “Random Forest” and “XGBoost”. For each cluster, both the algorithms are passed with the best parameters derived from GridSearch. We calculate the AUC scores for both the models and select the model with the best score. Similarly, the model is selected for each cluster. All the models for every cluster are saved for use in prediction.

**Prediction Data Description**

Client will send the data in multiple set of files in batches at a specified location. Data will contain wafer name & 590 columns of different sensor values for each wafer.

Apart from prediction files, we also require a “schema” file from client which contains all the relevant information about the training files such as:

1. Name Validation: We validate the name of the files based on the given name in the schema file. We have created a regex pattern as per the name specified in schema file. After validating the pattern in the name, we check for the length of data in the file name as well as the length of time in the file name. if all the values are as per the requirement, we move such files to “Good\_Data” directory else move such files to “Bad\_Data”
2. Number of Columns: We validate the number of columns present in the files, and if it doesn’t match with the value given in the schema file, then the file is moved to “Bad\_Data”
3. Name of Columns: The name of the columns is validated and should be the same as given in the schema file. If not, then the file is moved to “Bad\_Data”
4. The datatype of columns: The datatype of columns is given in the schema file. This is validated when we insert the files into database. If the datatype is wrong, then move the file to “Bad\_Data”.
5. Null Values in Column: If any of the columns in a file have all the values as NULL or missing, we discard such a file and move it to “Bad\_Data”

**Data Insertion**

1. Database Creation & Connection: create a database named as “prediction\_waferdb”. If the database is already created, open the connection to the database.
2. Collection Creation in Database: Collection with name – “prediction\_waferCollection” create in database “prediction\_waferdb” and insert all the data files from “Good\_Data”

**Prediction**

1. Data Export: Export the data from database “prediction\_waferdb” from collection “prediction\_waferCollection”in CSV file
2. Data Preprocessing:
   1. Check for NULL Value: If null value(s) present, impute the null values using KNN imputer.
   2. Columns with Zero Standard Deviation: Identify such columns and drop them as these columns do not contribute in model training.
3. Clustering: it’s time to load the KMeans model created during model training and find the cluster for each record.
4. Prediction: Load all the models saved for each cluster and based on the cluster number pass the data for prediction purpose.
5. Once we are ready with the predictions for each record export the predictions along with the wafer number in a CSV file and return the export location

**Deployment**

We will be deploying the model to the cloud – AWS

The folder structure for the wafer fault detection project: