AIML | MODULE PROJECT

AIML MODULE **PROJECT**

FEATURISATION & **MODEL TUNING**

TOTAL **SCORE** **60**

General Instructions:

1. Submission of all the parts is expected in 1 notebook only

2. Expected submission format: 1 '.ipynb' notebook and 1 '.html' notebook only

3. 50% marks will be deducted if insights/steps are missing in the corresponding questions.

4. If output for any code cell is missing, 50% marks will be deducted.

Submission Format:

1. '.ipynb' (Jupyter Notebook) and

2. '.html' (Jupyter Notebook > File > Download as > HTML)

5 Marks will be deducted if submission in any of the formats is missing.

- **DOMAIN:** Semiconductor manufacturing process
- **CONTEXT:** 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.
- **DATA 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.

Steps and tasks: [Total Score: 60 points]

1. Import and understand the data. [5 Marks]
   A. Import 'signal-data.csv' as DataFrame. [2 Marks]
   B. Print 5 point summary and share at least 2 observations. [3 Marks]
2. Data cleansing: [15 Marks]
   A. Write a for loop which will remove all the features with 20%+ Null values and impute rest with mean of the feature. [5 Marks]
   B. Identify and drop the features which are having same value for all the rows. [3 Marks]
   C. Drop other features if required using relevant functional knowledge. Clearly justify the same. [2 Marks]
   D. Check for multi-collinearity in the data and take necessary action. [3 Marks]
   E. Make all relevant modifications on the data using both functional/logical reasoning/assumptions. [2 Marks]
3. Data analysis & visualisation: [5 Marks]
   A. Perform a detailed univariate Analysis with appropriate detailed comments after each analysis. [2 Marks]
   B. Perform bivariate and multivariate analysis with appropriate detailed comments after each analysis. [3 Marks]
4. Data pre-processing: [10 Marks]
   A. Segregate predictors vs target attributes. [2 Marks]
   B. Check for target balancing and fix it if found imbalanced. [3 Marks]
   C. Perform train-test split and standardise the data or vice versa if required. [3 Marks]
   D. Check if the train and test data have similar statistical characteristics when compared with original data. [2 Marks]
5. Model training, testing and tuning: [20 Marks]
   A. Use any Supervised Learning technique to train a model. [2 Marks]
   B. Use cross validation techniques. [3 Marks]
   Hint: Use all CV techniques that you have learnt in the course.
   C. Apply hyper-parameter tuning techniques to get the best accuracy. [3 Marks]
   Suggestion: Use all possible hyper parameter combinations to extract the best accuracies.
   D. Use any other technique/method which can enhance the model performance. [4 Marks]
   Hint: Dimensionality reduction, attribute removal, standardisation/normalisation, target balancing etc.
   E. Display and explain the classification report in detail. [3 Marks]
   F. Apply the above steps for all possible models that you have learnt so far. [5 Marks]
6. Post Training and Conclusion: [5 Marks]
   A. Display and compare all the models designed with their train and test accuracies. [1 Marks]
   B. Select the final best trained model along with your detailed comments for selecting this model. [1 Marks]
   C. Pickle the selected model for future use. [2 Marks]
   D. Write your conclusion on the results. [1 Marks]