AIQoD ML Assessment

Domain Introduction:

- The project is set in the realm of predictive analytics, harnessing machine learning to forecast outcomes based on historical data patterns.

Problem Statement:

- Develop a predictive model to accurately forecast future events based on historical data.

Data Cleaning and Preprocessing:

- Null values, representing missing or undefined data, can hinder a machine learning model's accuracy by skewing its understanding. Dropping these values during data cleaning ensures the model receives only valid data, with each feature contributing meaningful information.

EDA Findings:

- **Univariate Analysis:** Pair plots visualized data distributions and potential correlations.
- **Bivariate Analysis:** Heatmaps highlighted strong correlations within the dataset.

Feature Engineering:

- Feature extraction is based on the transformation of hash values to integers and encoding of categorical variables. The features are then used to train the machine learning model.
- Hash values in object-type columns are converted to integers.
- Features and labels are separated from the merged training data.
- Yes/No columns are encoded using `LabelEncoder`'

Model Selection & Tuning :

Random Forest: Chosen for its ability to handle complex data and resist overfitting. Tuning involved optimising the number of decision trees (n_estimators) and the information gain criterion.

K-Nearest Neighbors (KNN): Selected as a contrasting model, relying on feature similarity for predictions. Tuning focused on the optimal number of neighbours (n_neighbors) to consider.

Model Evaluation Metric:

Model Name	Accuracy	Precision	Recall	F1 score
Random Forest	0.79	0.90	0.79	0.82
KNN Classifier	0.60	0.62	0.58	0.58

Random forest classifier is give better Accuracy compared to KNN Classifier

Conclusion - Feature Importance:

- The Random Forest model provides a robust solution for predicting quality defects, enabling proactive quality control measures.
- Feature importance analysis identified the most influential factors driving predictions, guiding targeted interventions to address root causes of defects.

Business Solution:

1. **Targeted Quality Interventions:** The model's ability to identify high-risk products or processes allows for targeted quality interventions. Instead of blanket inspections, resources can be focused where they are most needed, optimizing efficiency and reducing costs.

2. **Root Cause Analysis for Process Improvement:** By pinpointing the features most predictive of defects, the model provides actionable insights into the root causes of quality issues. This enables targeted process improvements to prevent recurring defects, leading to long-term quality enhancement.

Code file :
https://drive.google.com/file/d/1wiHPjUQJnV0GfZ4Ux5cdI1CtmL7n5R0G/view?usp=sharing.

Output Csv :
https://drive.google.com/file/d/1Kz8Ba2LvO0beEdV9dv-kcANPwBqDHvDV/view?usp=sharing.