**Analysis Report**

Archit Sood

**Executive Summary**

This report presents a comprehensive analysis of the dataset, including data exploration, data preprocessing, and the performance evaluation of a machine learning model. The primary goals were to identify any unused tables, address data type validation issues, and develop a predictive model for the given dataset.

**Methodology**

**Data Exploration**

1. **Dataset Overview:** The dataset “Classicmodels” consists of tables [customers, employees, offices, order details, orders, payments, productlines, products].
2. **Unused Tables:** Through the analysis, it was identified that the following tables, were all utilized in the machine learning model.
3. **Data Type Validation Issues:** The error indicated that one or more columns in the dataset contains string values, and these string values couldn’t be directly converted to float during the data preprocessing step. In this case, it seemed like the 'QueryType' column contains string values ('INSERT', 'SELECT', etc.), and the machine learning models requires numerical input.

**Data Preprocessing**

1. **Handling Missing Values:** I used pandas function “df.isnull().sum()” to identify any missing values in the data before feeding the data into the ML model to train it.
2. **Features and target Variable –** The features we identified for the analysis were X = ExecutionTime, QueryType and Y = Query

**Machine Learning Model**

1. **Model Selection:** The model I took we “Random forest”, and the rational behind the model selection “Ensemble learning, feature importance, robustness to overfitting, handling categorical data.
2. **Performance Metrics:** The model's performance was evaluated using the following metrics:

Classification report: precision recall f1-score support  
  
 0 0.50 0.49 0.49 4971  
 1 0.51 0.52 0.51 5029  
  
 accuracy 0.50 10000  
 macro avg 0.50 0.50 0.50 10000  
weighted avg 0.50 0.50 0.50 10000

Accuracy: 0.5025 or 50.25%

Confusion Matrix:  
[[2435 2536]  
 [2439 2590]]

**Findings**

1. **Machine Learning Model Performance:**

* **Random Forest:** **Accuracy:** The model's accuracy is close to random guessing (50%), indicating that it does not perform significantly better than chance.
* **Precision and Recall:** Both precision and recall are around 50%, suggesting a balance between false positives and false negatives.
* **F1-Score:** The F1-score is also around 50%, indicating a reasonable balance between precision and recall.
* **Support:** The dataset has an equal number of instances for both classes.
  + **Recommendations and Further Research**

1. **Feature Engineering:** Explore additional feature engineering techniques to enhance model performance.
2. **Hyperparameter Tuning:** Perform hyperparameter tuning for machine learning models to optimize their performance.
3. **Data Imbalance:** Address any class imbalance issues in the dataset for more accurate predictions.

**Conclusion**

The analysis revealed insights into the dataset, identified unused tables, and addressed data type validation issues. The machine learning models demonstrated varying performance, and further optimizations can be explored for enhanced accuracy.