STRUCTURED DATA- PROBLEM STATEMENT-1

Data Preprocessing

1.Data Jar

1.1 Data Cleaning

- Load the Dataset train.parquet file.
- To check and drop duplicate
- Check the column in wright data type
- To find and impute the missing or null Values

1.2 EDA-Exploratory Data Analysis

Necessary feature Engineering Has been done, like Frequency and Time Based procedure.

1.2.1 Positive Set

- A patient is considered eligible for a particular drug when they have taken their first prescription for that drug.
- If the patient took at least one time Target drug then the patient considers as a eligible.

1.2.2 Negative Set

- The patient who and all not taken target drug consider as a not eligible

1.3 Splitting

1.3.1 Training Data

- To develop the model (70 or 80 %)

1.2.3 Testing Data

- evaluate the ml model (30 or 20 %)

1.4 Scaling

- It is used to maintain same order or same range in all columns.
- It is not mandatory for all ml algorithm, but it is good to practice to maintain all model.

2. Task Jar

- **Labeled data-Supervised learning-binary classification**-so we can use logistic regression, knn, decision tree, random forest, xg boost.

3,4,5. Model jar, loss jar, learning jar

With the help of model, loss and learning jar, I have built a following models:

- Logistic Regression
- Knn
- Decision Tree
- Random Forest
- XG Boost

6. Evaluation Metrics jar

- With the help of Accuracy and F1 score, I have evaluated the above-mentioned model.
- NOTE: I Have added the necessary Comment in that particular step, to get a proper understanding.

NOTE: I Have added the necessary Comment in respective steps, to get a proper understanding.

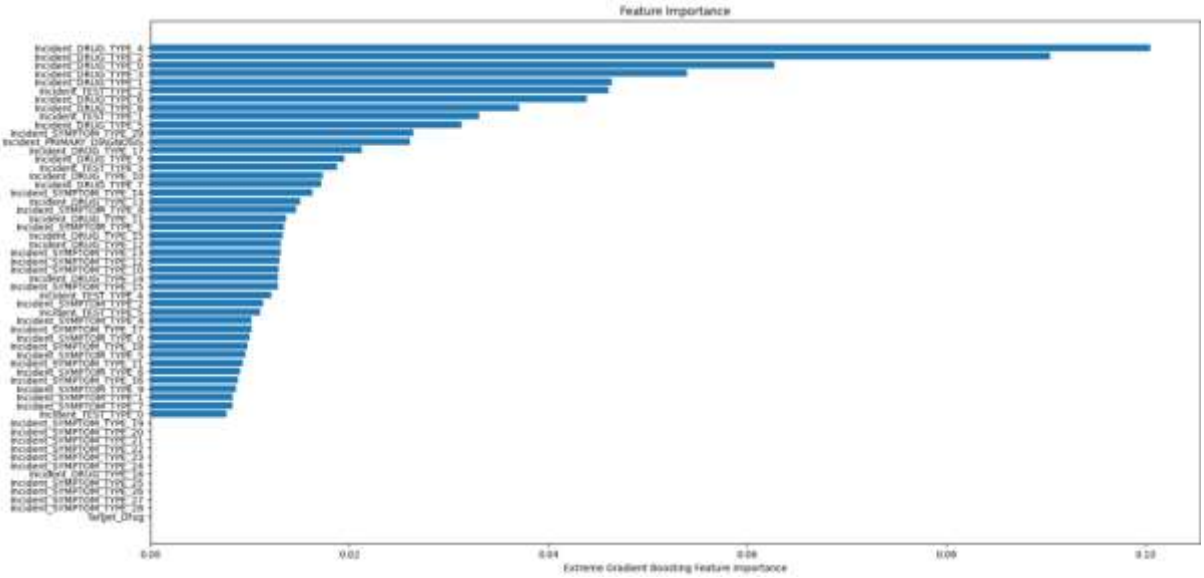
Outcome of the all models:

```
accuracy & f1_score for Logistic Regression Model : (0.7508507175617695, 0.5823412698412698)
accuracy & f1_score for K-Nearest Neighbors Regression Model : (0.7656458055925432, 0.6226774654597428)
accuracy & f1_score for Decision Tree Regression Model : (0.7538097351679243, 0.634606938954765)
accuracy & f1_score for Random Forest Regression Model : (0.8069241011984021, 0.697986577181208)
accuracy & f1_score for Extreme-Gradient Boosting Regression Model : (0.805888445036248, 0.7063563115487915)
```

Interpretation:

- By comparing all model F1-Score, XG Boost have higher f1 Score and this model will give good performance while predicting the future data.

Suggestion to Company-Solving problem statements based on Feature Importance:



Based on the feature importance given by Best Machine Learning Algorithm(Extreme Gradient Boosting) the features of order given below:

1. Incident Primary Diagnosis
2. Incident Drug Type-2
3. Incident Drug Type-4
4. Incident Test Type-2
5. Incident Drug Type-3

The above features are more impacting the target feature compare to other features