Submitted by: 1.Prasanna Kumar R – SM21MTECH14001
2.Malisetti Gayathri – BM21MTECH11005

FOML KAGGLE HACKATHON – "IS DRIVER AT FAULT DATASET"

Kaggle ID: **SM21MTECH14001_BM21MTECH11005**

Data processing is the main step to build a good ML model. A perfect ML model has many underlying processes like data collection, data processing, model creation and hyperparameter tuning etc. In the given dataset, we have tried many methods mentioned below for fine tuning of our model to get the best accuracy.

- For handling missing values we tried to use randomfill() function, replaced missing values by mode, replaced missing values by mean, assigned a string named unknown to all the missing values etc
- For encoding the categorical variables to numerical values we tried using encoding techniques such as one hot encoding, label encoding, frequency encoding etc.
- We also tried to use feature selection methods like chi square test, using Pearson correlation method etc.

Finally we are able to achieve an accuracy score of 0.86459 using Light GBM classifier.The steps we have taken in our final model are:

- Imported the train data set and understood that it has 51490 rows and 42 columns including target column
- Converted 'Crash date' column into columns of year, month, day, weekday, hour.
- Deleted the columns/features having more than 80% missing values.
- Used IQR (Inter-Quartile Range) for 'Vehicle year' column to remove outliers.
- Converted all features into string type by using astype. This step converted all NaN/missing values into string and thus all the missing values in the entire data have got filled.
- Later we have used Label Encoding for all the categorical features to convert them into numerical features.
- Used 40% of train data as validation data and checked the accuracy for different ML models. We have tried Random Forest Classifier, GB Classifier, XGBoosting Classifier, LightGBM Classifier. Here Light GBM both gave good accuracy of 0.85 compared to other classifiers.
- The same pre-processing steps were applied to test data as well.
- Finally the best two scores we got on test data are:

CLASSIFIER	ACCURACY
Light GBM	86.459%
XG Boost	86.01%

Conclusion: Light GBM is a gradient boosting framework that uses tree based learning algorithms. It can handle large datasets, has faster training speed, and uses less memory. It also does parallel processing of data which makes it more accurate to stand among many other classification models.