Chicago City Prediction of Causes of Accidents

OVERVIEW

This presentation is divided in different parts:

- Business understanding
- Data understanding
- Modeling
- conclusion

Business understanding

Traffic congestion and accidents are a persistent problem in urban areas, particularly in large cities like Chicago. Understanding the primary causes of traffic crashes can help city planners, transportation authorities, and policymakers implement more effective measures to mitigate and reduce traffic crashes.

So, the goal is to predict the cause of crashes by using different models and to be able to understand which one of the causes are influential.

Objectives

The objectives of this project are:

- To understand the primary causes of accident for drivers.
- To merge the 3 datasets, which are the people, the crash and the vehicles. In an understandable manner by dropping column which are considered irrelevant.
- To clean the data by reducing the numbers of multiclasses for each feature.
- To test different models and understand the result of each model and choose the best performing one.

Methods

The methods used are:

- Data collection
- Data cleaning
- Data modelling
- Model evaluation

Model Building

Step 1 - Developing the predictive models

Step 2 - Testing the models used. These are;

- KNN,
- XGBoost,
- Random Forest,
- Decision Tree,
- ANN,
- Naive Bayes.

Step 3 – Encoding of the y and x which are categorical

Step 4 – Training of the data using the different models

- 80% of the data was used for training the model.

Step 5 – Prediction and evaluation of the models.

Decision

- Choose the best the best performing model for primary cause prediction.

Model Comparison

Comparing the Performance of the Models from the evaluation metrics to establish which model works best.

KNN
- Accuracy: 87%
- Weighted average:87%

Random Forest
- Accuracy: 92%
- Weighted average: 92%

Decision Tree
- Accuracy: 86.1%
- Weighted average : 70.3%

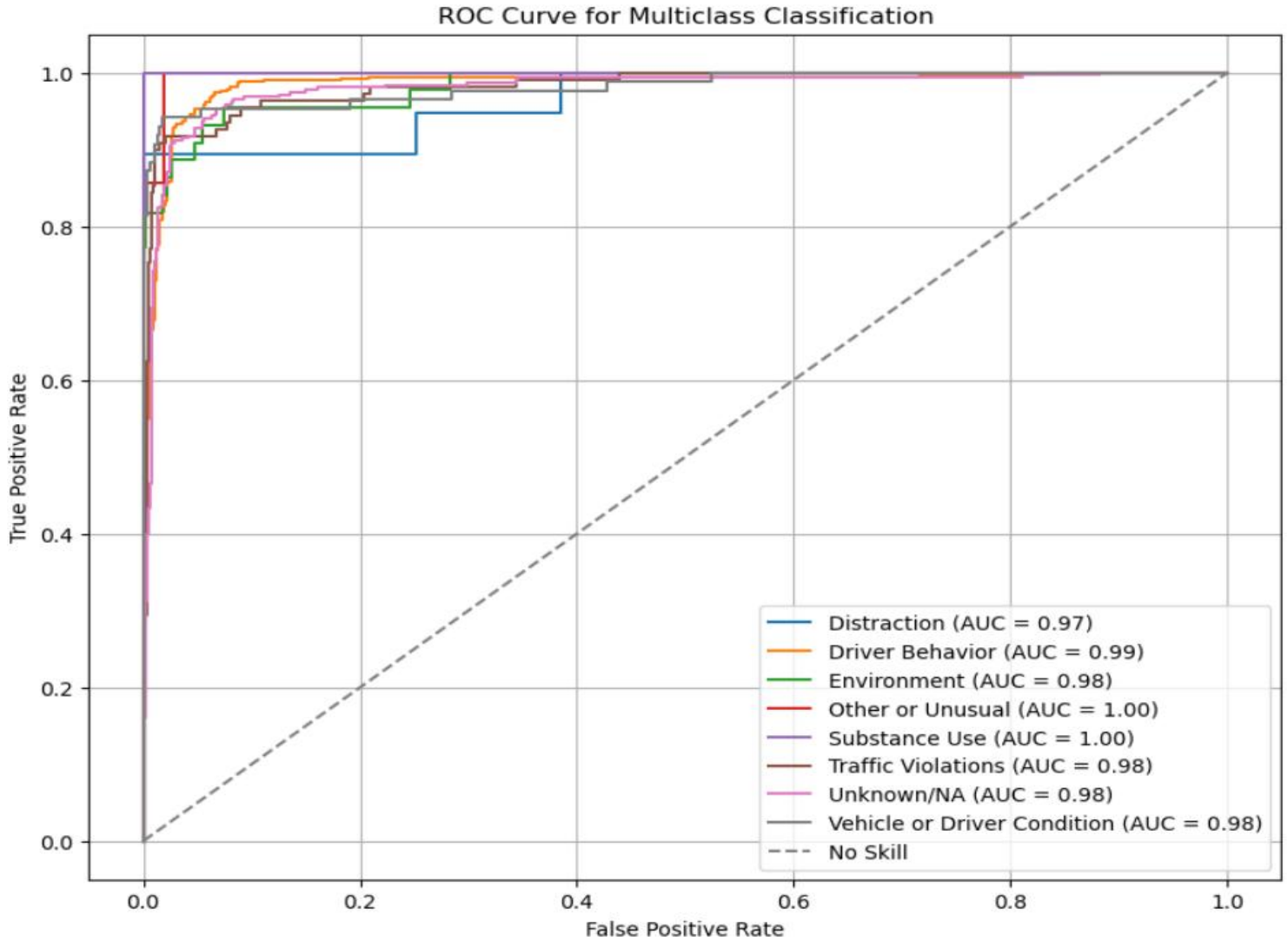
XGBoost
- Accuracy: 92%
- Weighted average : 92 %

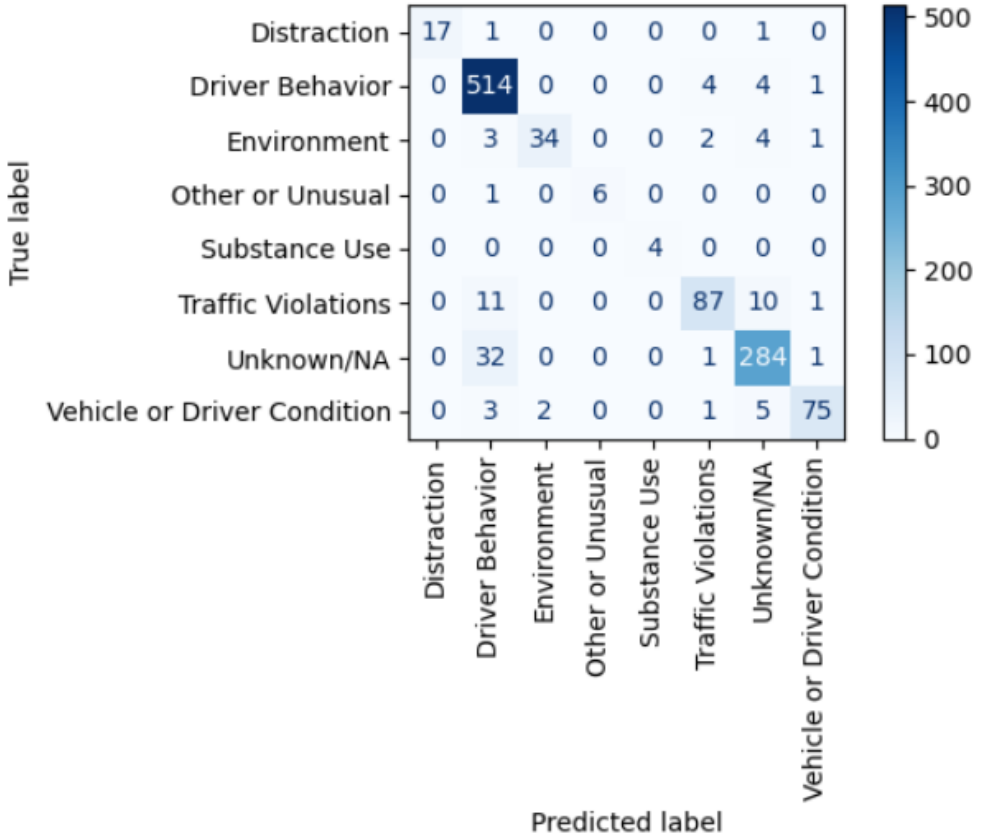
ANN
- Accuracy: 0.53
- Weighted average : 0.56

Naive Bayes
- Accuracy: 25.8%
- F1 score: 26.5%

Verdict
- XGBoost and Random Forest are the better performing model. XGBoost being the better one because of higher F1 score for the different categories.

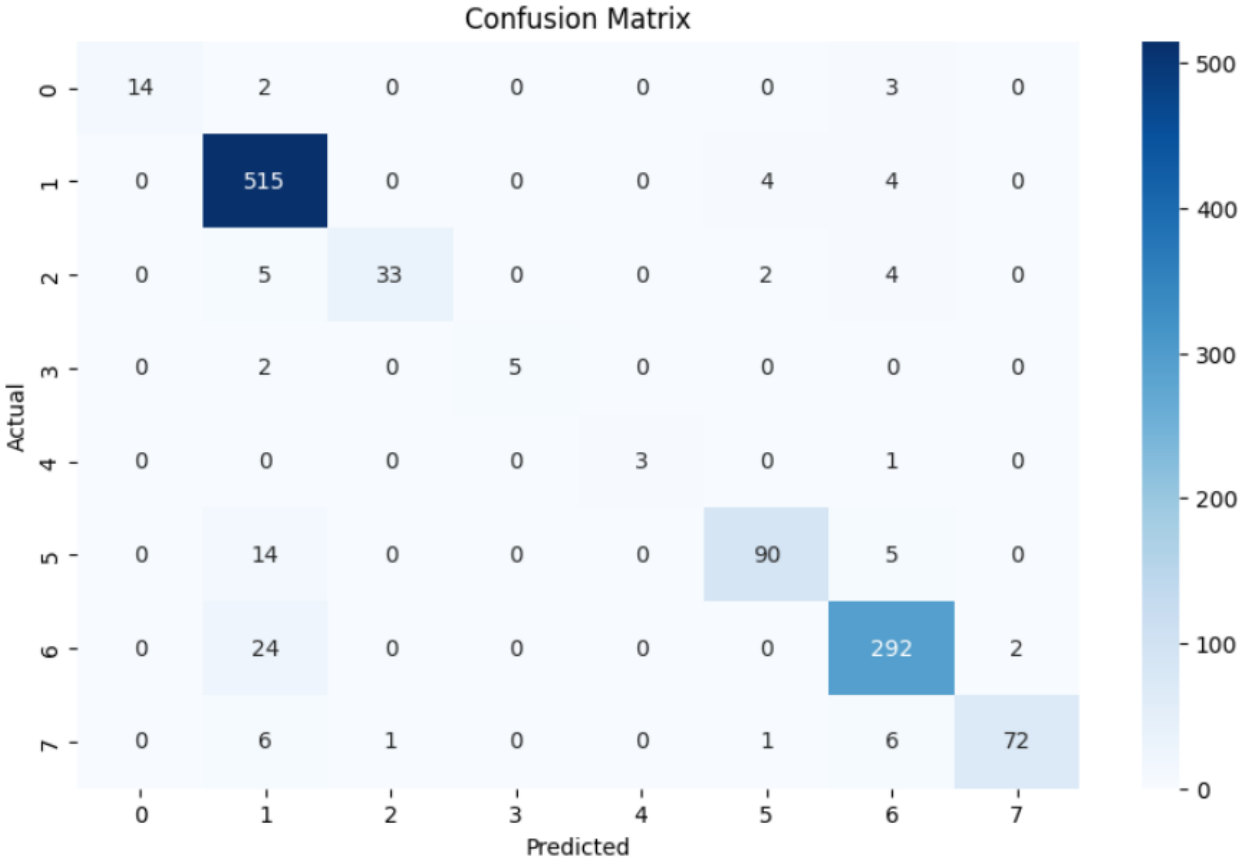
XGBoost results

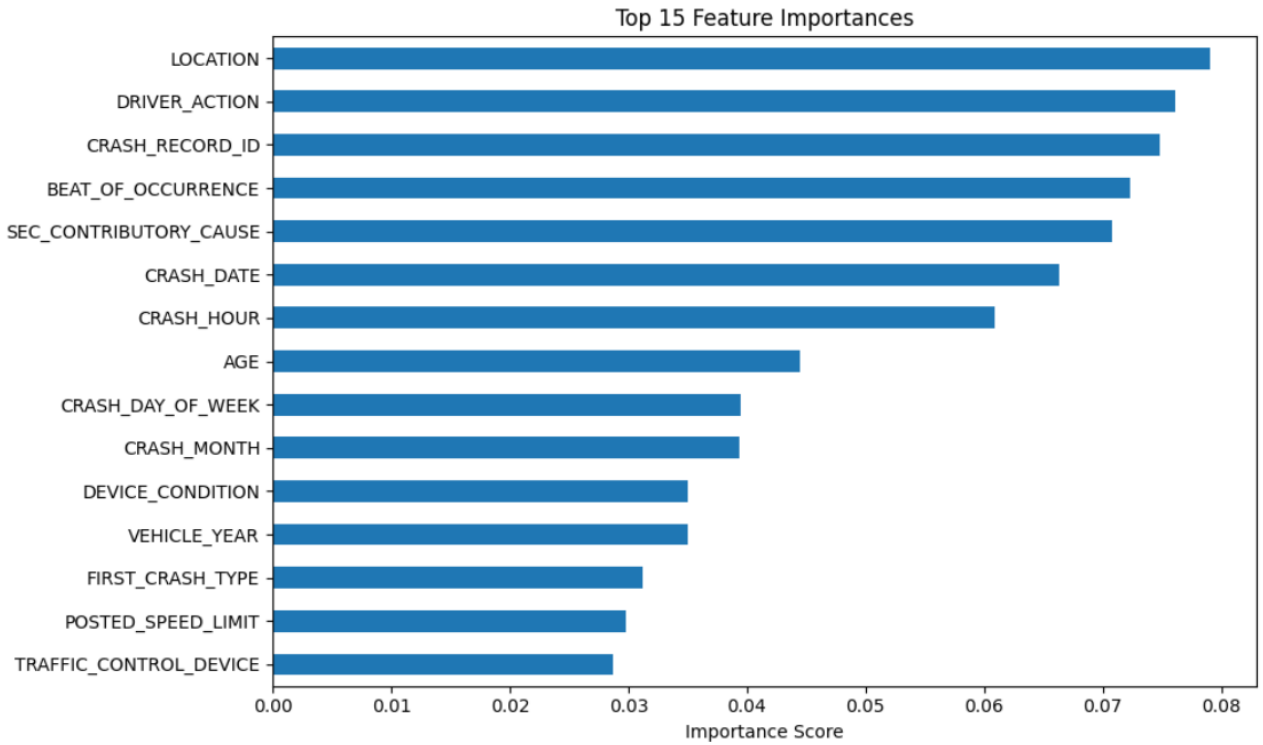




From the confusion matrix the driver behavior and the unknown as well as the traffic violation are well predicted.
from the ROC curve it can be understood that all the classes have good results, because their AUC value are very high.
Meaning that the true positive are high with increases the model accuracy.

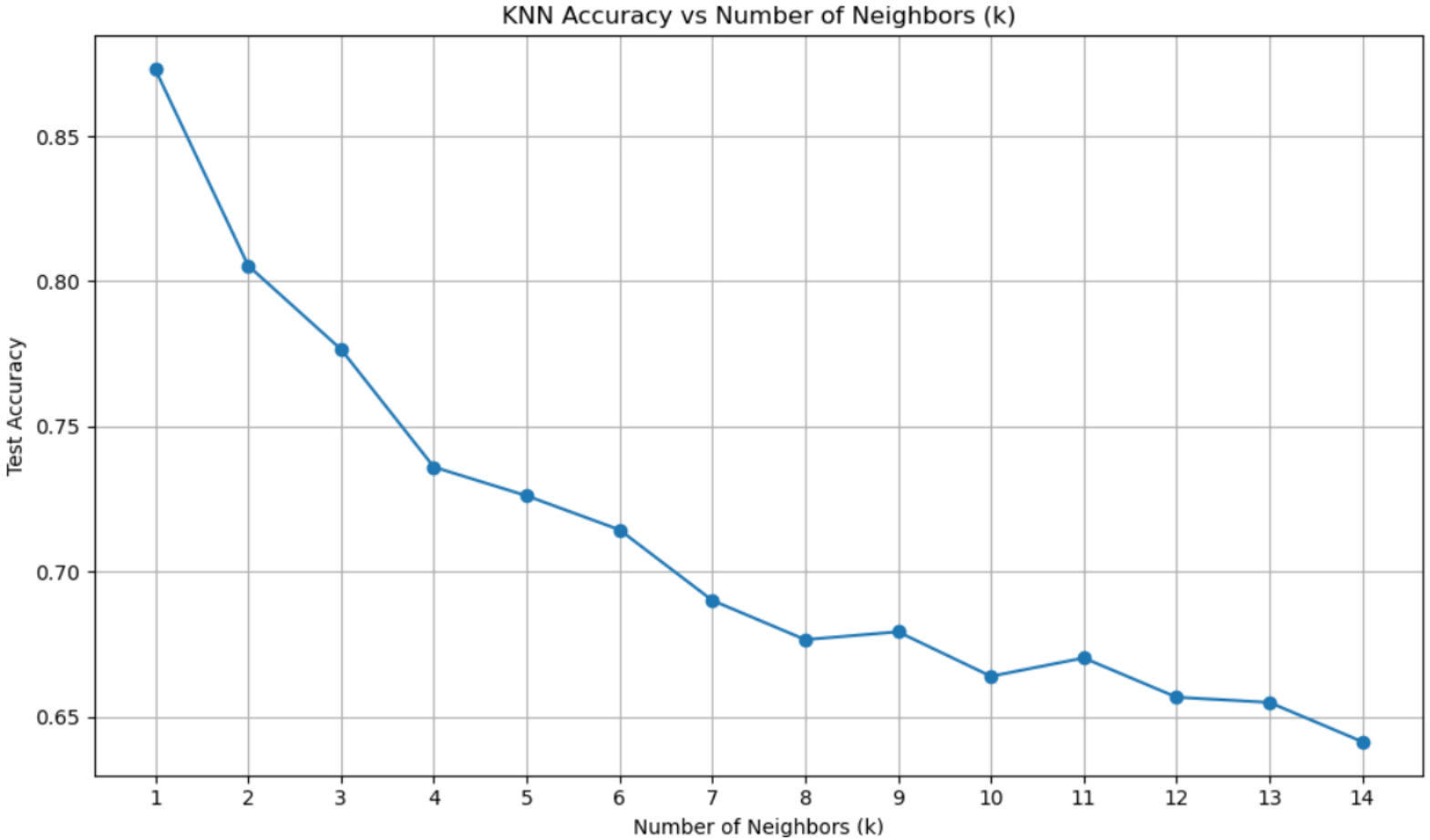
Random Forest results

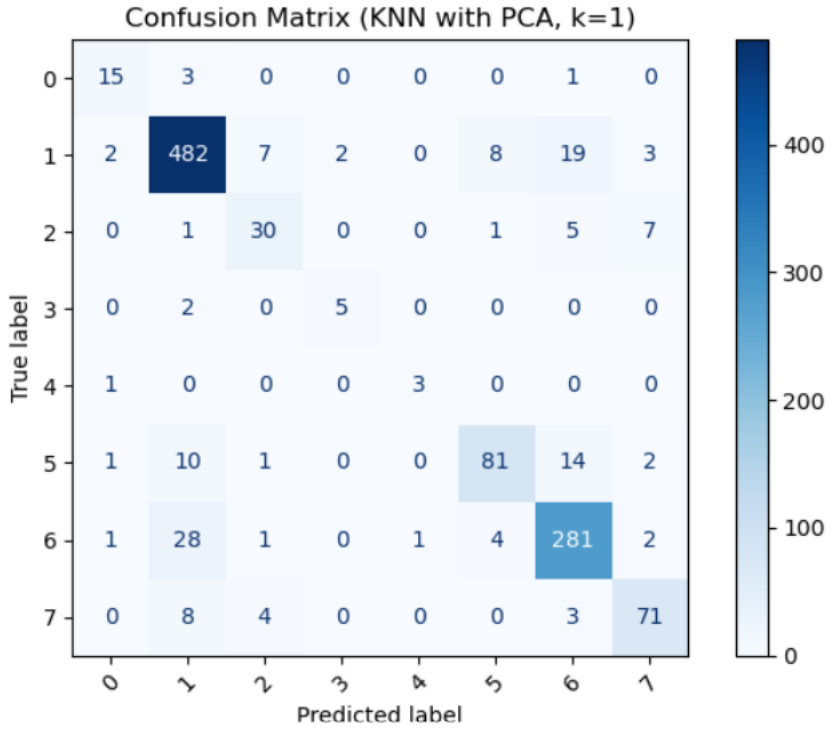




From the feature importance the driver action and location are considered to be the most influential in the model prediction.

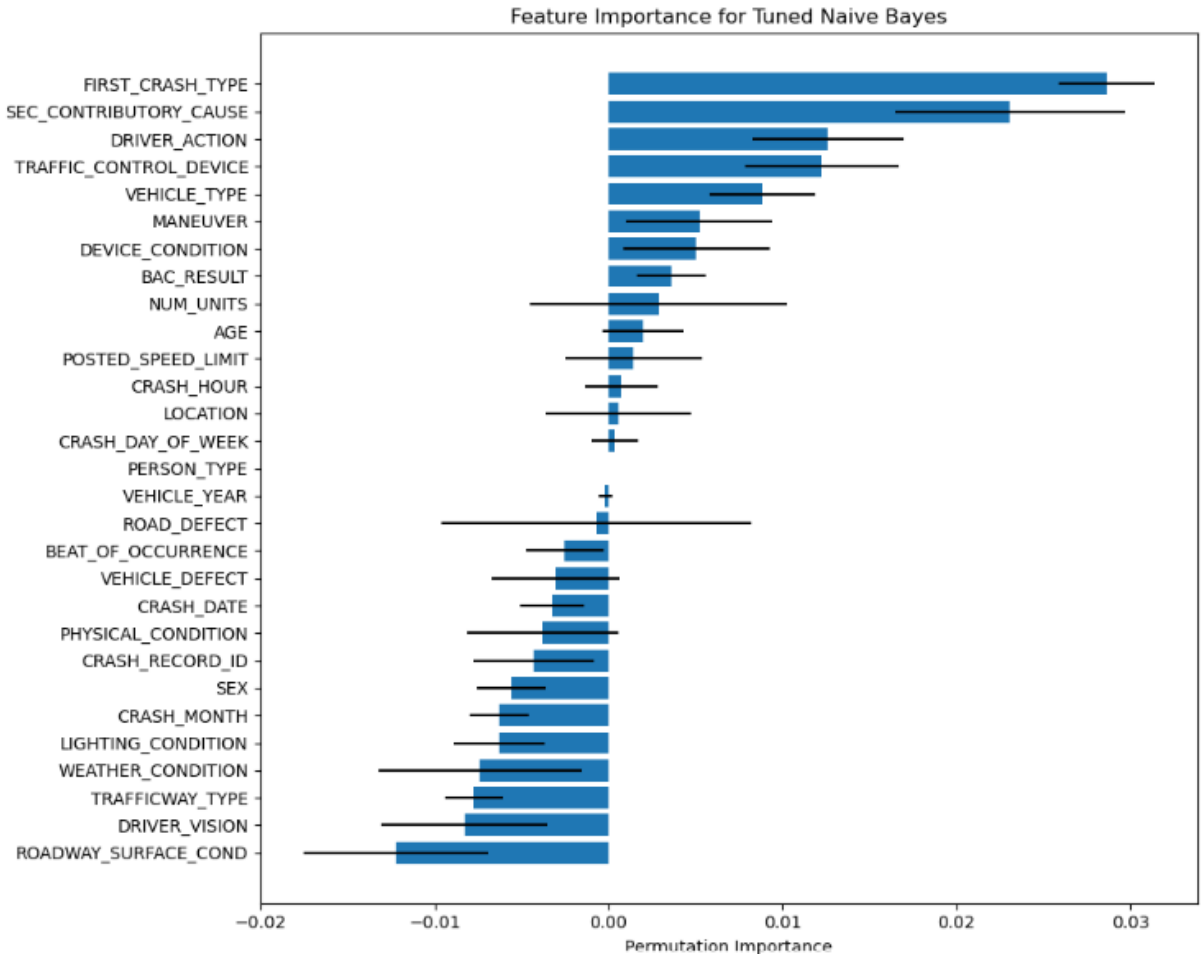
KNN results

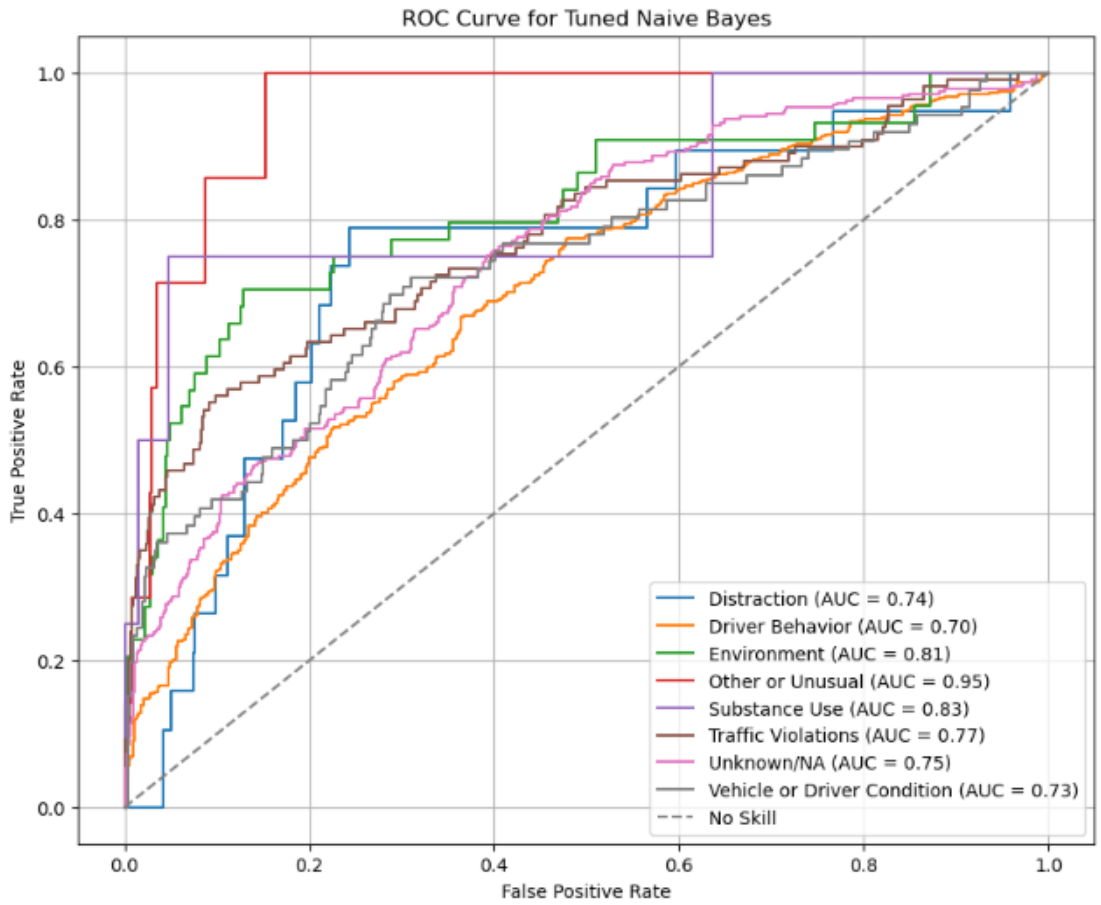




From the graph it can be understood that the K=1 has high accuracy. According to the confusion matrix the class 1 has high predictability.

Naive Bayes results





From the Feature importance, the crash type and the secondary cause for crash seem to have more influence in the prediction. From the ROC graph it can be seen that some features like unusual and substance use have high AUC value.

Conclusion & Recommendation

Conclusion

- **XGBoost and Random Forest are the best performing model** with the best level of primary cause prediction out of the evaluated models.
- The feature importance was analyzed for Random Forest to understand which causes are considered as major accident causes.
- Location and Driver action are considered the major causes for accidents that happen upon drivers.
- **For XGBoost the model performed well and all the different causes resulted in a high AUC**. Indicating that the model was able to learn and based on the data differentiate the different causes.

Thank You!
Any Question?