PREDICTING TRAFFIC ACCIDENT SEVERITY

Seshni Govender

October 2020

WHY PREDICT ACCIDENT SEVERITY?

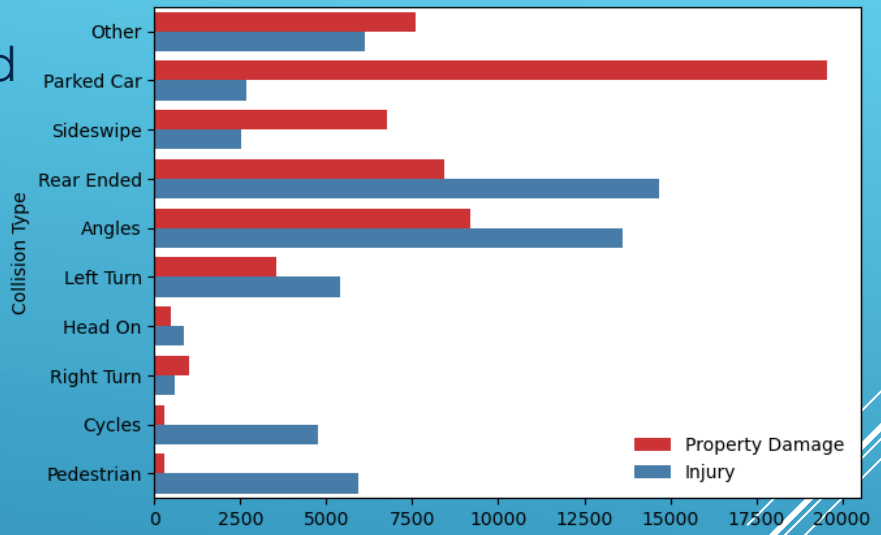
- Accidents have negative effects on the direct participants of the incident physically, emotionally and may incur considerable economic loss.
- Given several factors such as weather, road , light conditions etc. , the severity of an accident can be predicted
- Predicting an accident severity can be beneficial :
  - Traffic police may assist in improving overall road traffic safety using the predictions
  - Road users are likely to drive carefully, use different traffic routes or plan for a possible accident.
  - Insurance companies can warn road users for accident possibilities

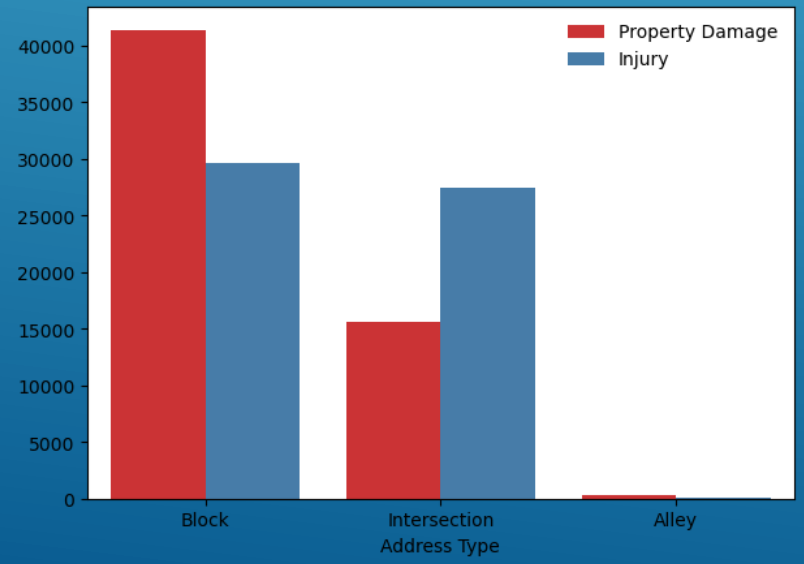
DATA ACQUISITION AND CLEANING

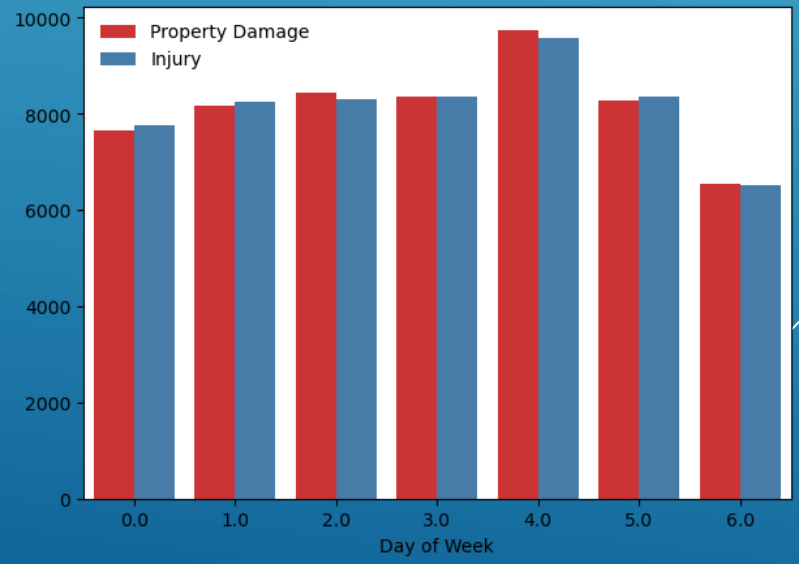
- Shared data source of traffic accidents provided by Seattle Police Department for the Seattle City from 2004 to present [1][2]
- Data-set originally contained 37 attributes and a target/label of accident “severity” by severity code.
- Unbalanced labeled data with 136 485 accidents with a ”property damage” severity and 58 188 accidents with a ”injury” severity
- Duplicated and irrelevant columns were dropped. Rows which had more than half of the values in the columns as empty (”NaN”) were dropped.
- After cleaning and balancing, a total sample size of 114 320 accident records.
- Final feature set contained 35 attributes after encoding

DAY OF WEEK, COLLISION TYPE AND LOCATION

- Day of week attribute was dropped due to balanced distribution. Accident can be equally "injury" and equally "property damage" regardless of day of the week
- Accidents are more severe (Injury) at intersections and less severe (Property damage) at blocks and alleys
- Varying distribution for collision type – good indicator for predicting severity

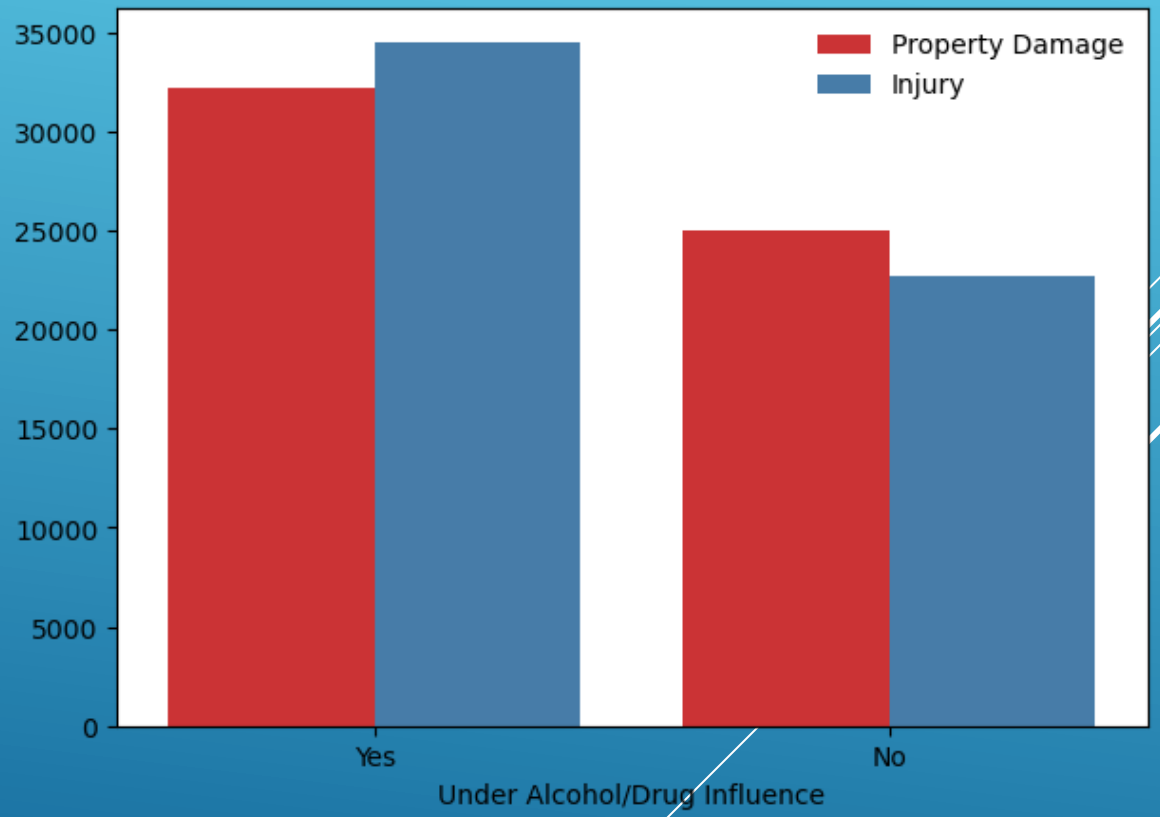






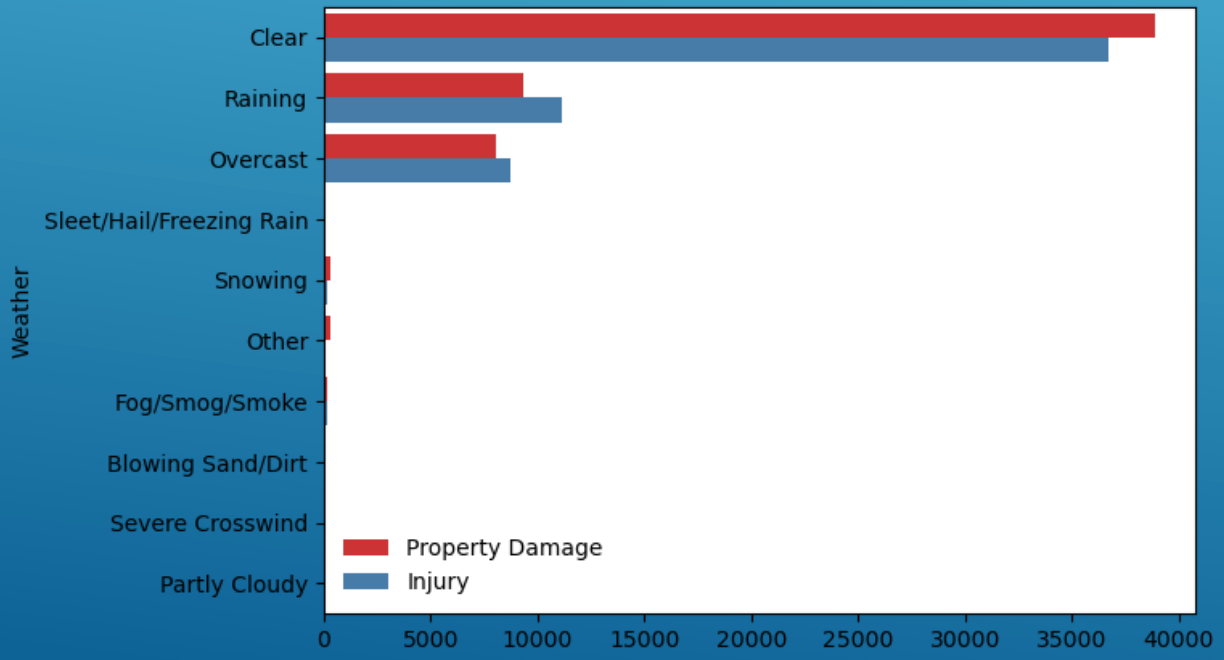
ALCOHOL AND DRUG INFLUENCE

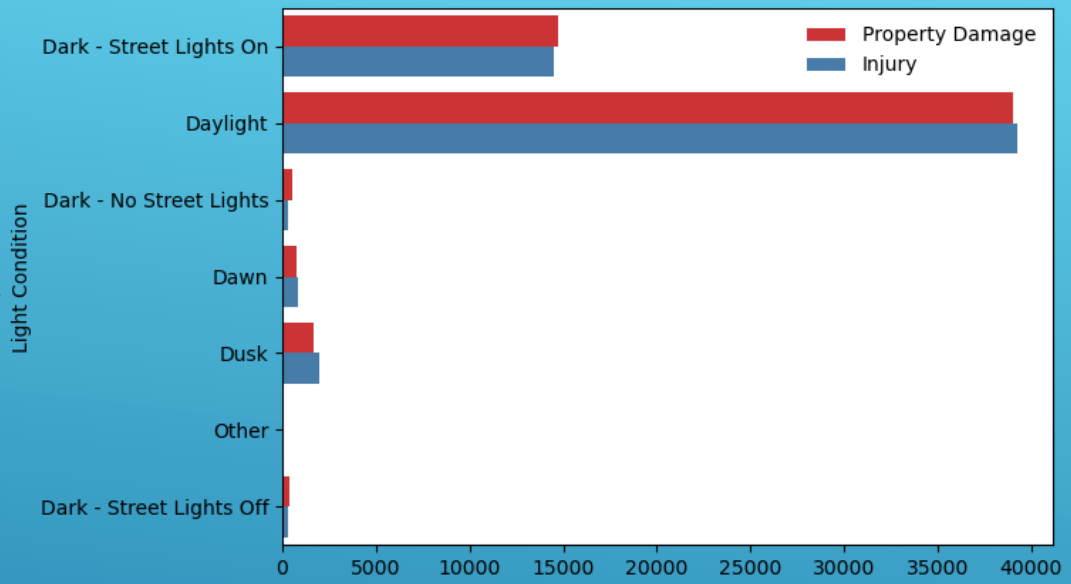
- A driver who is under the influence of alcohol or drugs may cause a more severe "injury" type accident than a driver who is not
- An intoxicated driver is not in a mental state to make safe choices on the road in comparison to a sober driver who may be able to make choices to avoid accidents or at least reduce the impact.

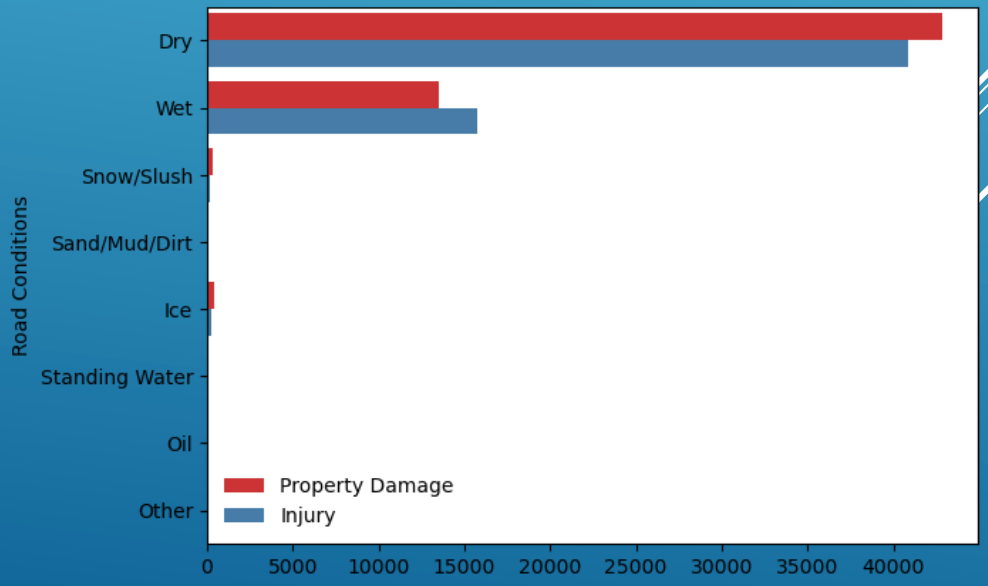


WEATHER, ROAD CONDITIONS AND LIGHT CONDITIONS

- Darker lighting conditions generally result in more “injury” type accidents
- Wetter roads and rainy weather conditions may also cause more severe accidents

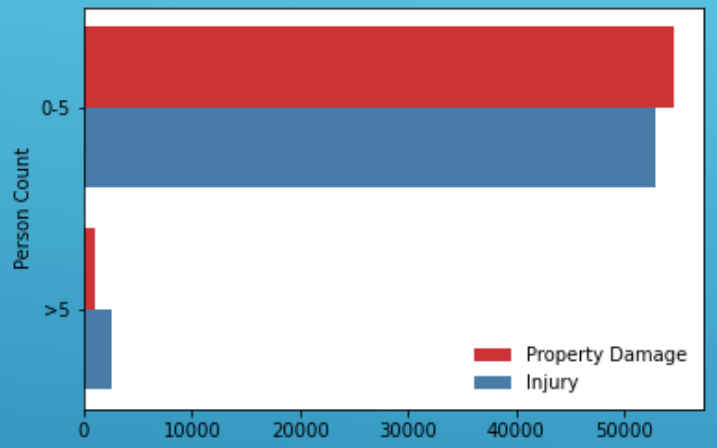


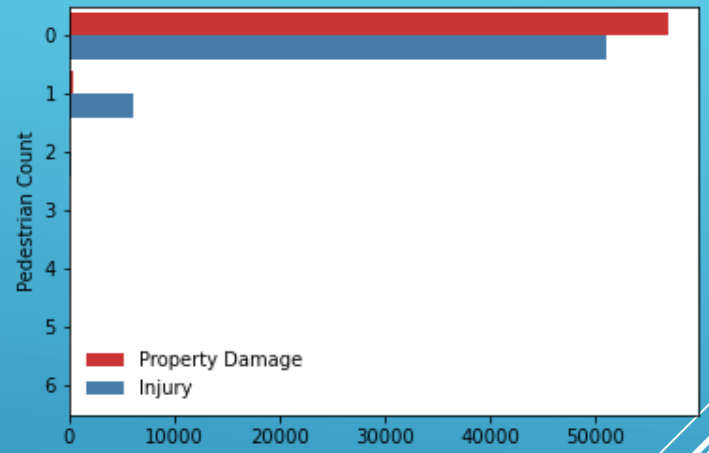


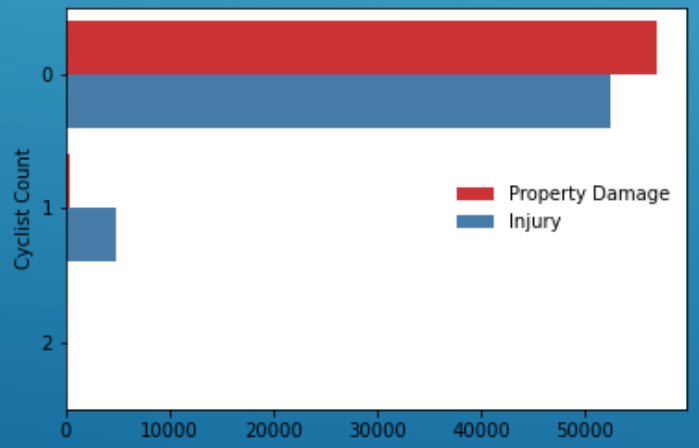


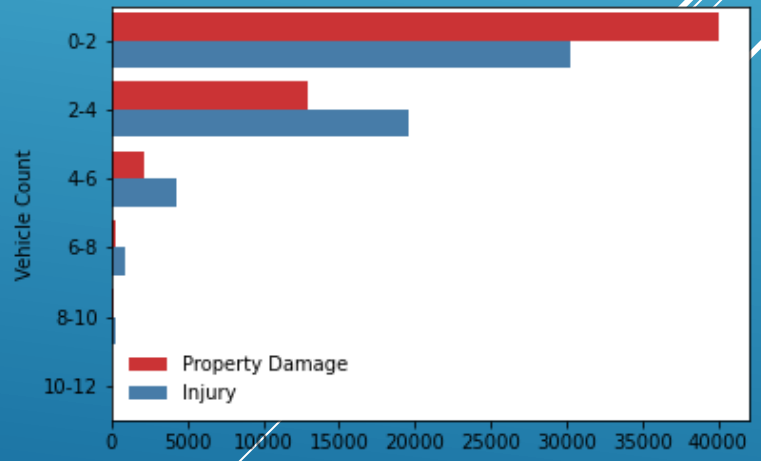
PERSON, VEHICLE, PEDESTRIAN AND CYCLIST COUNT

- Accidents involving more than 5 persons are likely to involve injury
- Accidents with more than 2 vehicles are likely to involve injury
- A higher number of cyclists and pedestrians in accidents are more likely to result in injury





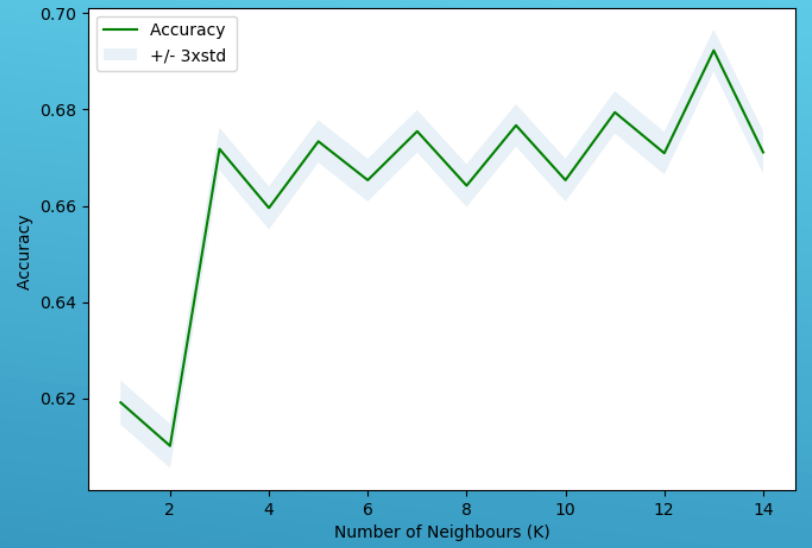




CLASSIFICATION: K NEAREST NEIGHBOR AND DECISION TREE

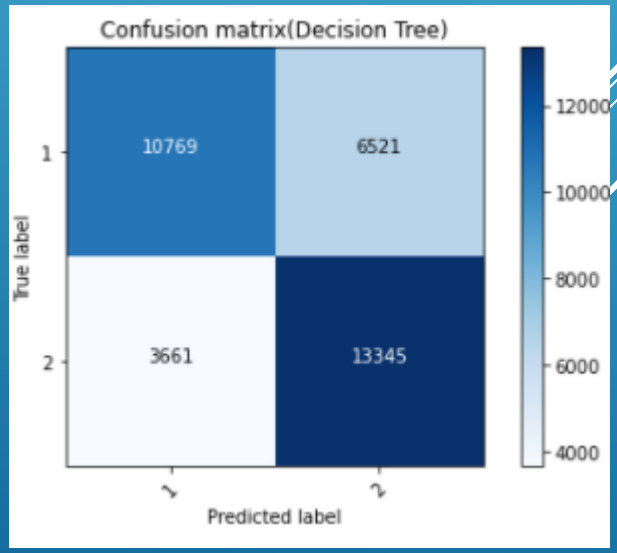
KNN:

- Data was split into 10% test data and 90% training data with a random state of 4.
- Best accuracy of 69% with k=13 neighbors
- 74% of the total number of type 2 accidents in the test set predicted correctly
- 65% of the total number of type 1 accidents in the test set predicted correctly



Decision Tree:

- Data was split into 30% test data and 70% training data with a random state of 3
- Best accuracy of 70% with a max depth of 13
- 78% of the total number of type 2 accidents in the test set predicted correctly
- 62% of the total number of type 1 accidents in the test set predicted correctly



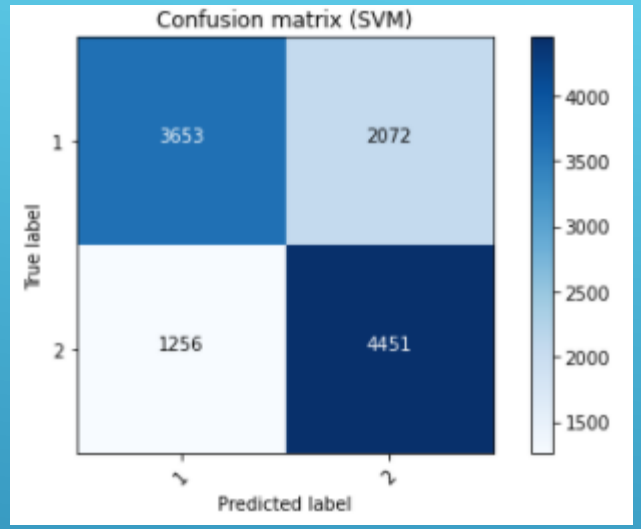
CLASSIFICATION MODELS: SVM AND LOGISTIC REGRESSION

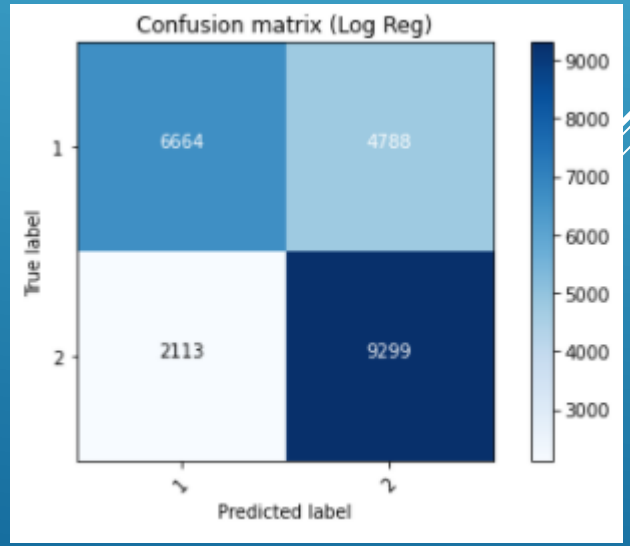
SVM:

- Data was split into 10% test data and 90% training data with a random state of 4.
- Trained using ‘rbf’ kernel
- Accuracy of 71% and best performing model
- 78% of the total number of type 2 accidents in the test set predicted correctly
- 64% of the total number of type 1 accidents in the test set predicted correctly

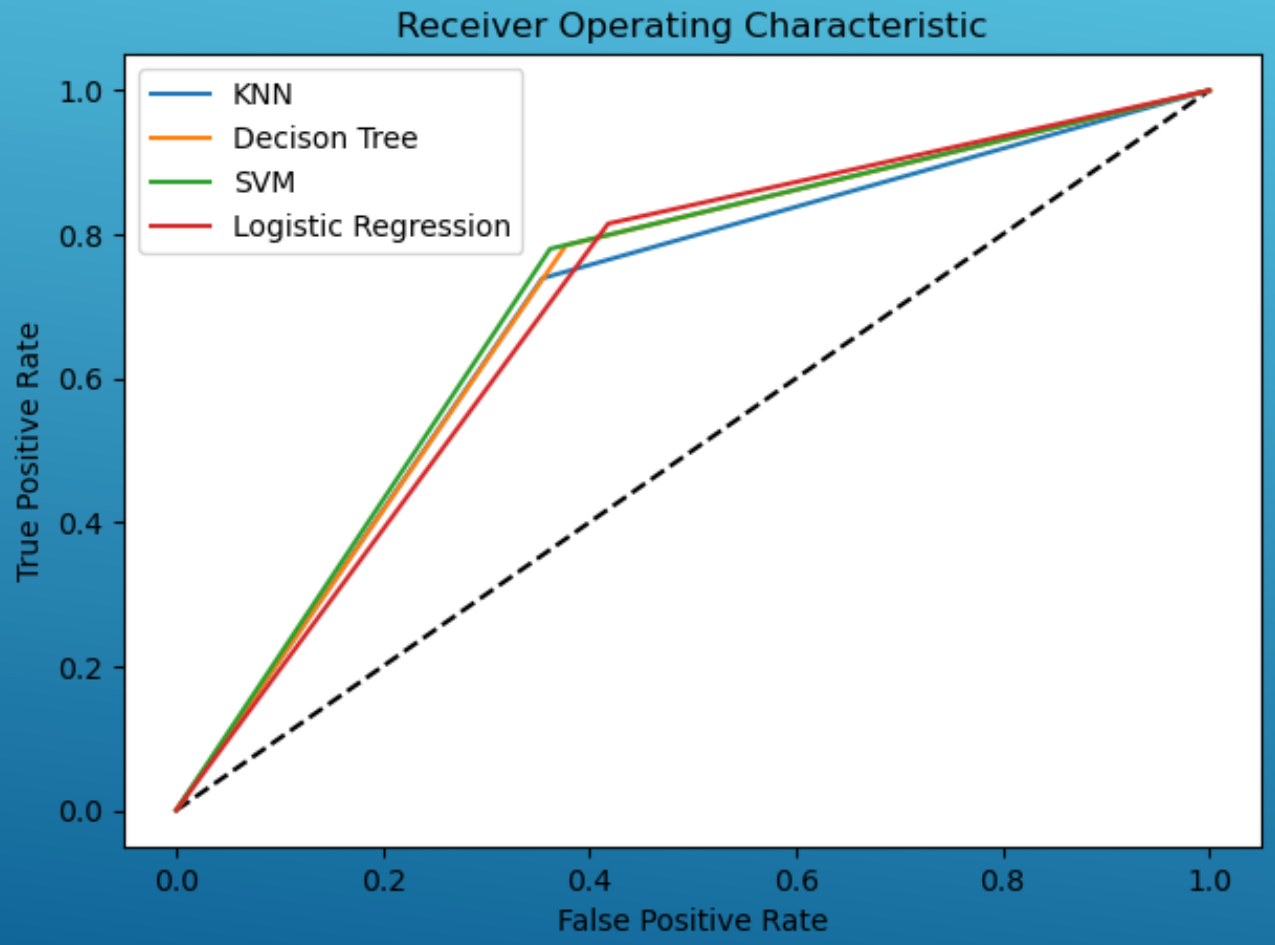
Logistic Regression:

- Data was split into 20% test data and 80%
- Trained using ‘liblinear’ solver and C=0.01
- Accuracy of 70%
- 81% of the total number of type 2 accidents in the test set predicted correctly
- 58% of the total number of type 1 accidents in the test set predicted correctly





SUMMARY OF MODELS



	Algorithm	Jaccard	F1-score	LogLoss
0	KNN	0.692092	0.691432	NA
1	Decision Tree	0.703260	0.701400	NA
2	SVM	0.708887	0.707430	NA
3	LogisticRegression	0.698172	0.694047	0.548357

CONCLUSIONS AND RECOMMENDATIONS

- Four different models were used to classify accident severity. The models performed well with 0.5% difference between each
- SVM model performed the best with 71% accuracy.
- More attributes such as whether the driver was speeding can help improve model performance
- Use of data with different types of accident severity outcomes for variance
- Optimize models by changing classifier parameters

REFERENCES

1. https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv
2. https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf