# Capstone Project

*Accident Severity Predictor*

# Predicting Accident Severity

- This is valuable to several stakeholders.
- Drivers can use this information to understand how their driving can contribute to an increased accident severity risk.
- Emergency services can allocate resources based on when they are more likely to be needed.
- Insurance companies can tailor insurance using the predictors.

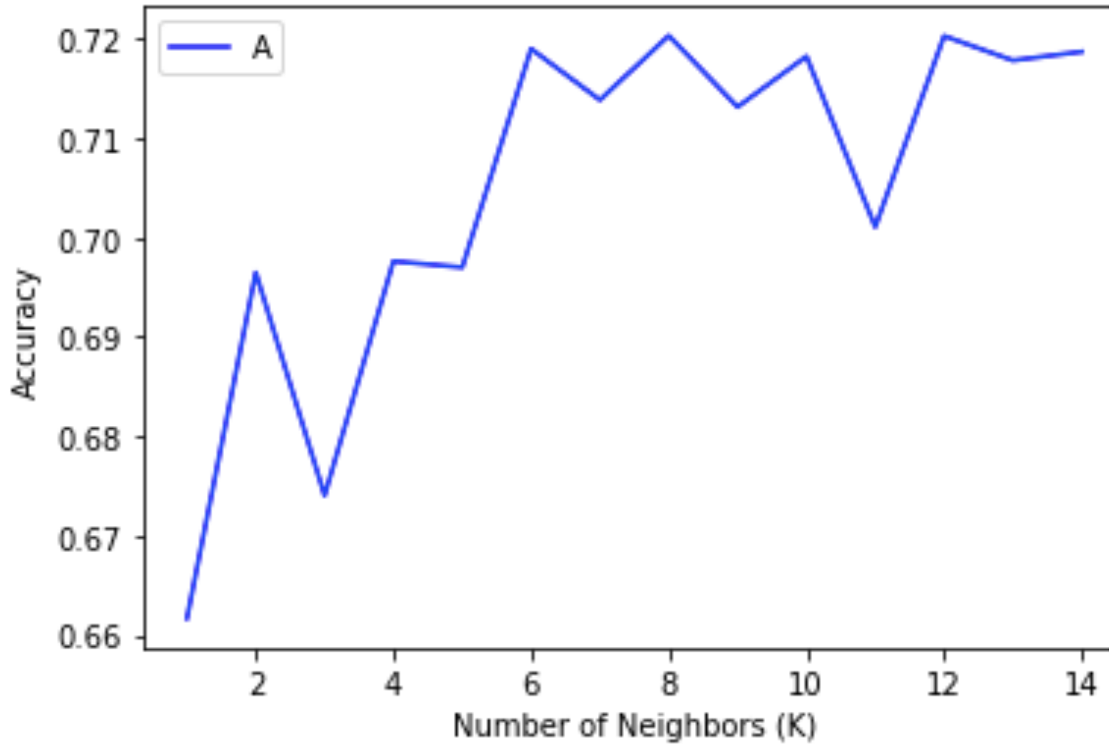
# Data: Information and Cleansing

- The data is Seattle traffic accidents from 2004 onwards. The raw dataset had 38 features and slightly over 194,000 rows.
- Many rows have incomplete data as a result the data was unbalance.
- Rows containing null values were removed and categorical data were encoder into numerical values.
- 5 features were selected for the model:
  - Weather
  - Light Condition
  - Speeding
  - Collision Type
  - Vehicle Count

# Classification Models

- Due to the nature of the data, Severity was either property damage or injury.
- Classification models were used.
- KNN, Decision Tree, SVM and logistic Regression were built to determine which would be the most accurate.

# KNN



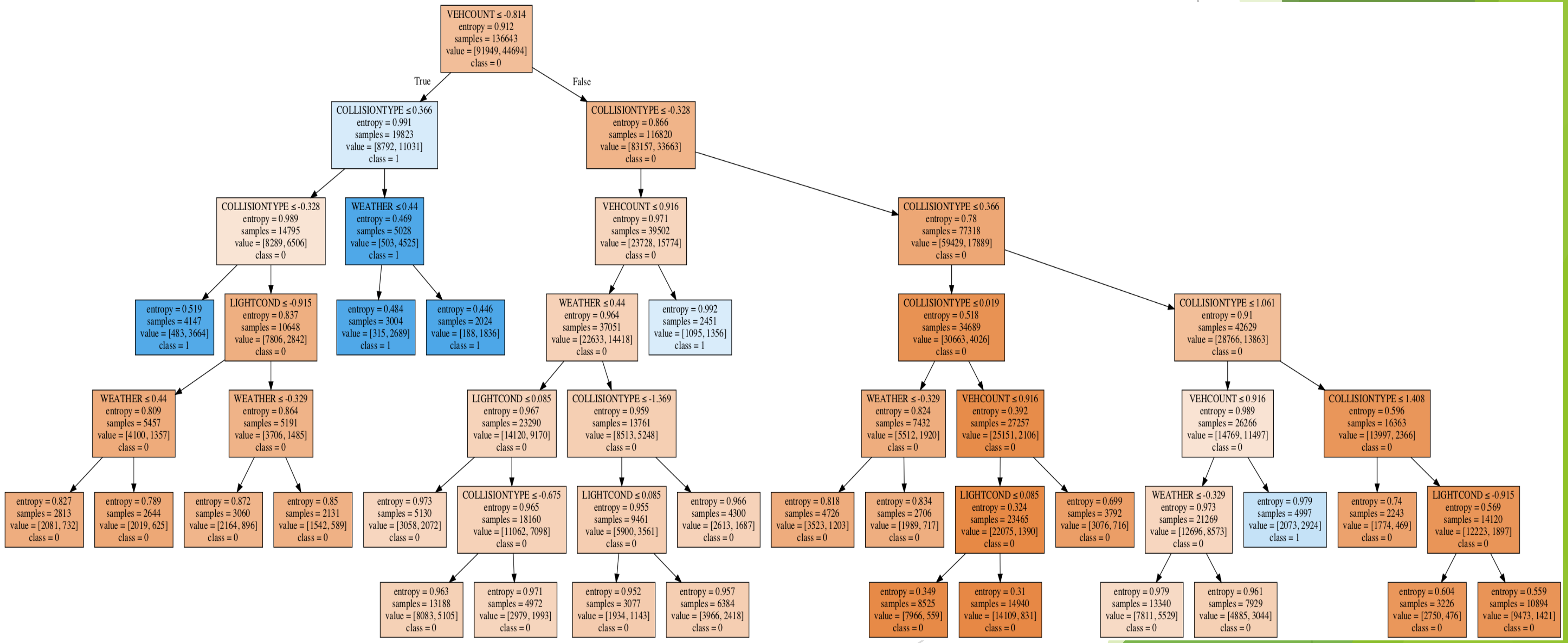
- Most Accurate K value was 8.
- This produced an accuracy of 0.72.

# Decision Tree

- Decision Tree was most accurate when max depth was 6. The accuracy scores were one of the highest tested.

- Although decision tree can become large when using large dataset.

- For those not familiar with decision trees this could be difficult to interpret.



# SVM and Logistic Regression

- SVM had issues fitting the whole training set. Time to process this was too long. Smaller sample had to be used as a result.
- SVM had good accuracy scores compared to other models.
- Logistic Regression – Accuracy scores were lower than other models. Probability for datapoints class was predicted.
- Max probability - 0.96
- Min probability – 0.04
- Mode probability – 0.58 & 0.42
- Logistic probability mode highlights model did not predict values with certainty.

# Model Evaluation Comparison

| Model | Jaccard Score | F1-Score | Log-Loss |
|---|---|---|---|
| KNN | 0.67 | 0.69 | N/A |
| Decision Tree | 0.71 | 0.78 | N/A |
| SVM | 0.71* | 0.78* | N/A |
| Logistic Regression | 0.68 | 0.55 | 0.62 |

* Small dataset was used to train model

# Conclusions + Recommendations

- Several Models have been built that can predict the severity of an accident.
- Decision Tree is recommended as this was most accurate.

- Future Recommendations:
- Using target that has more classes for severity. This could be fatalities, serious injury, minor injury ect.
- Accuracy of models could be improved.
- Using data from other cities. Comparing different cities could provide a different angle of insight.