CAR ACCIDENT SEVERITY

Anna Kudela

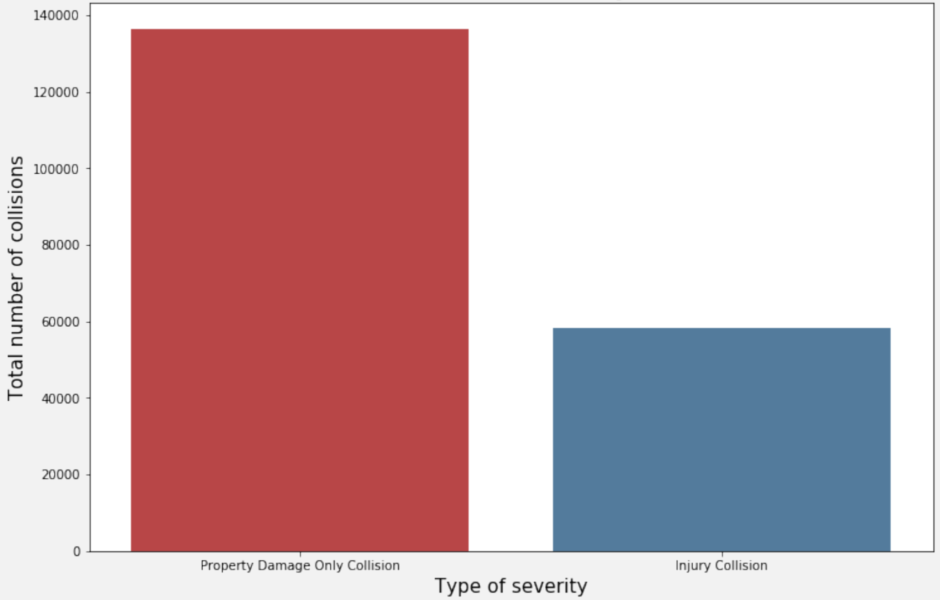
BUSINESS UNDERSTANDING

- The aim of this project is to see whether there are any actions that could be taken to avoid accidents or reduce their severity by analysing data from the Seattle Department of Transportation. It will be done by analysing the collision dataset for the city of Seattle and find patterns and determinate key factors such as weather, light and road conditions, drug or alcohol influence, driver inattention to provide the best traffic accident severity prediction. Various analytical techniques and machine learning classification algorithms will be used, such as: K-nearest-neighbours, Decision Tree Analysis, Support Vector Machine.

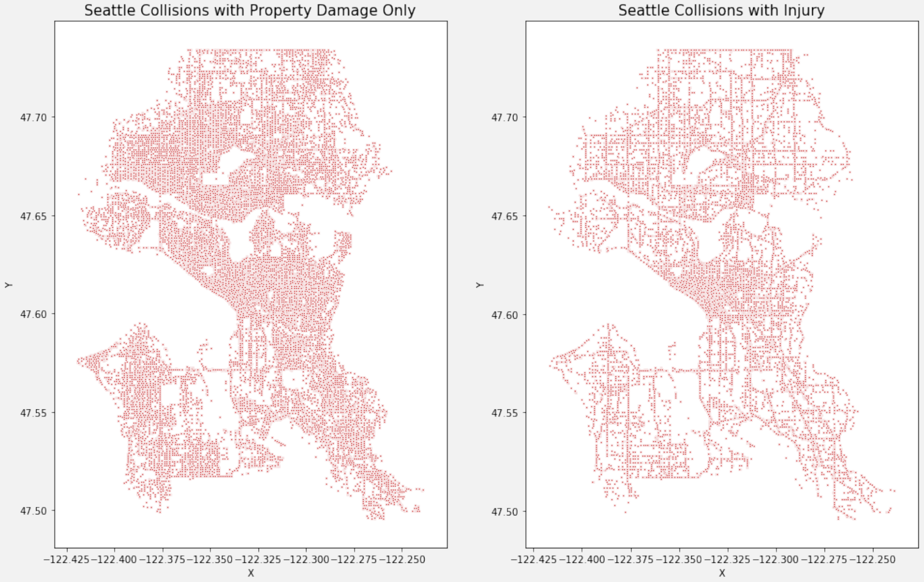
DATA ACQUISITION AND UNDERSTANDING

- The example dataset was from Seattle Department of Transportation which contains various features of collisions in Seattle from 2004 to 2020. There are almost 200,000 collisions in the dataset and 38 features.
- A few graphs using *seaborn* package were made to visualise some features from the dataset. It has helped me make some assumptions regarding the causes of car accidents:
    i. accidents, including collisions that involve injury, occurs very often between midnight and 1 am.,
    ii. although collisions with pedestrians and cyclists are very rare, the number of such accidents is still very high,
    iii. the collisions involving injury tend to happen inside and nearby the downtown area and major highways,
    iv. the number of collisions appears to be trending down in the past 16 years.

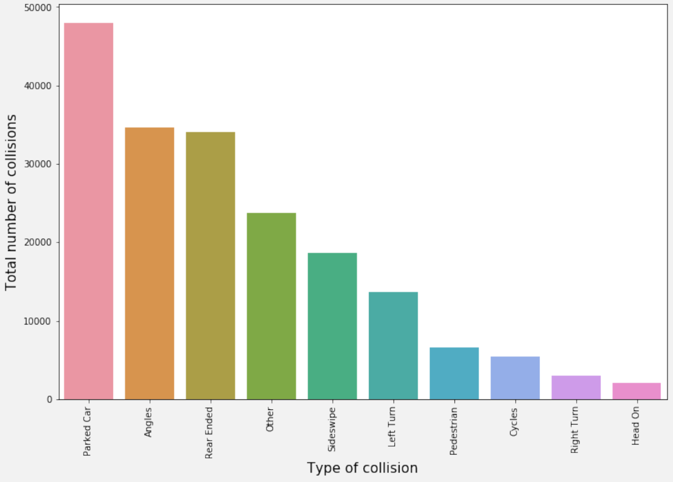
Collision Severity



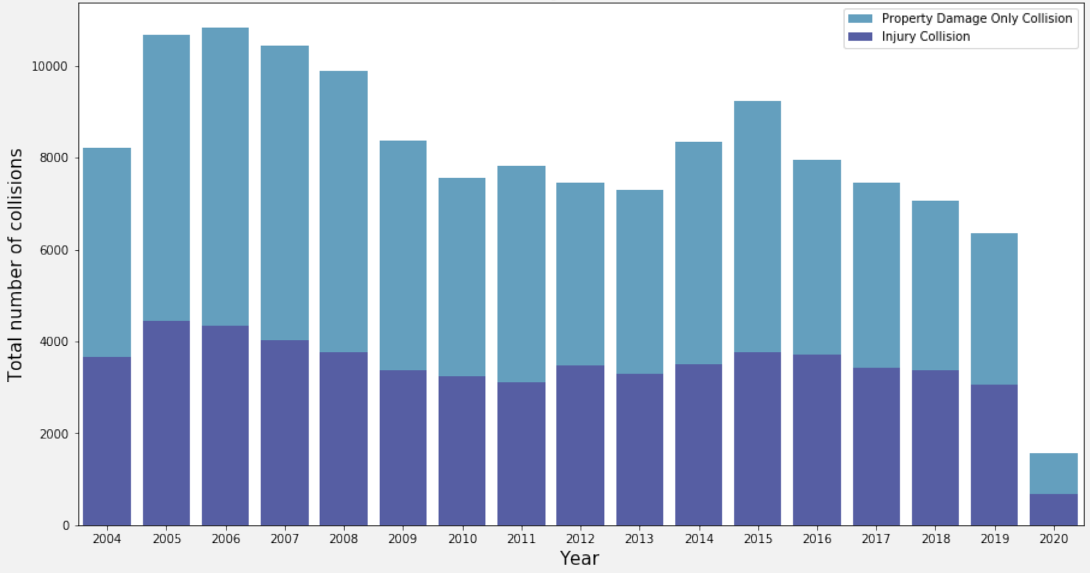
All collisions occured in Seattle


Seattle Collisions with Property Damage Only

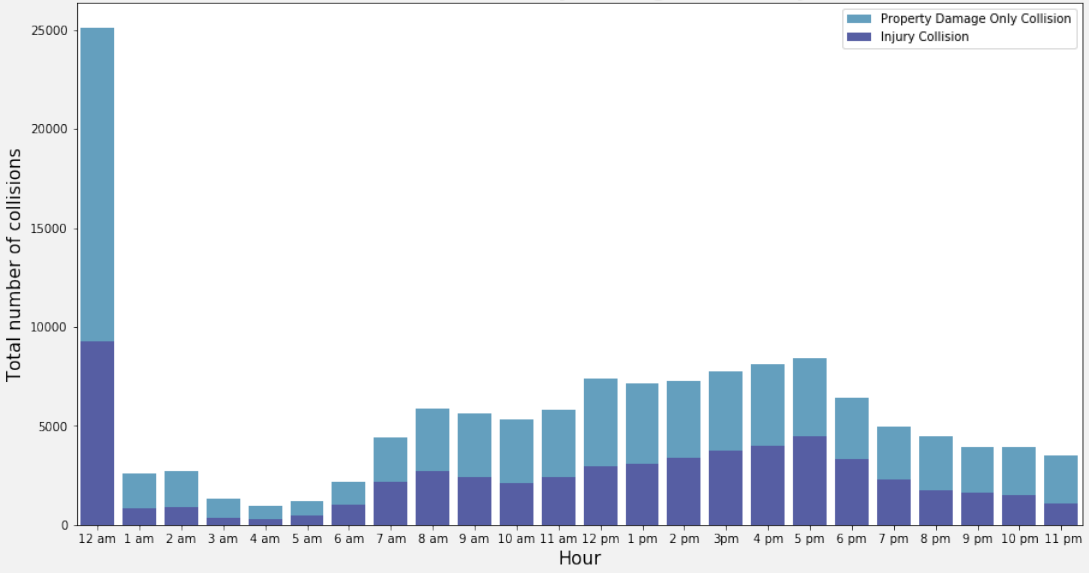
Seattle Collisions with Injury



Collision severity from 2004 to 2000



Collision severity during whole day

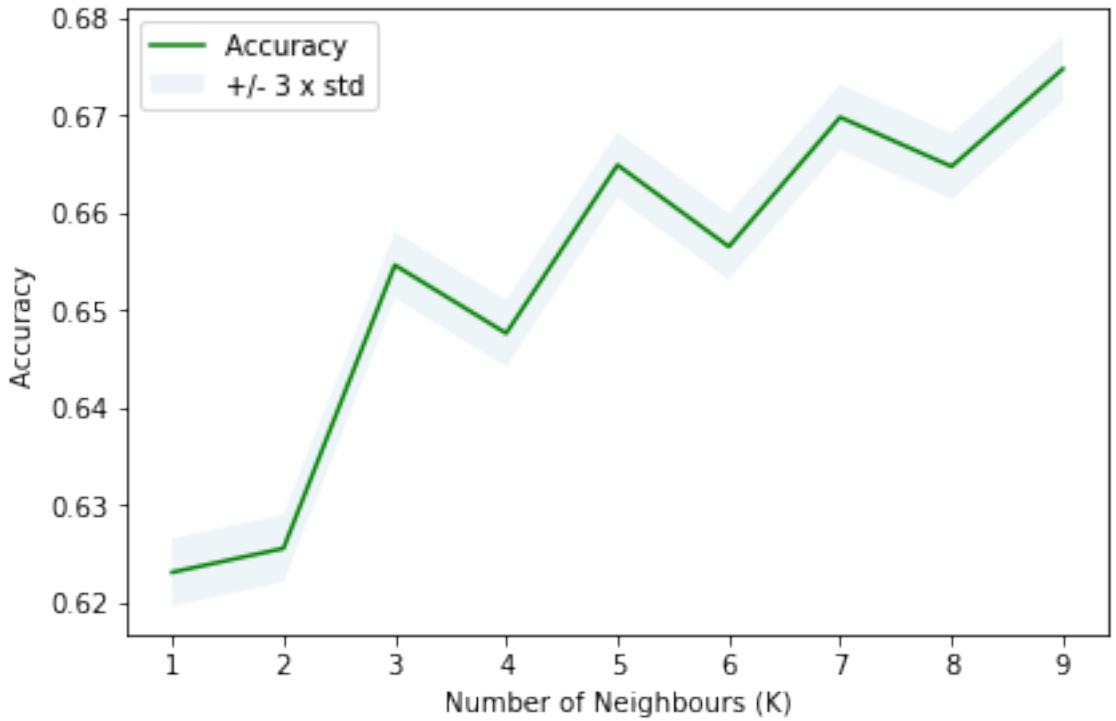


DATA PREPARATION AND CLEANING

- Raw dataset unfortunately **includes missing data, restructuring data and normalizing the data**. Since we have a large dataset, and the number of incidents with missing data is very low, we can drop the rows with missing values. Next, important issue to go through each feature to make sure it is of a type our algorithm will take, and make sure the values made sense. For some, they needed to be encoded as a '0/1' for each value. Once the data was fully prepared, it is proper to make a copy for use at the end of the project to feed into the final model.
- Second step in our analysis will be balancing the data. As there are much more collisions without injury, we can undersample the majority class (*class 1*). There are still 100,000 rows of clean, balanced data. The next step is feature selection. Most of the data is compatible for the algorithms used, except for data such as coordinates or department codes.
- In third and final step we will focus on standardizing the data to feed into the classifiers by spliting the data into training data (80 %) and testing data (20 %) first. All features' scales are equal because of the unit variance. The 3 most used classifiers have been implemented to see what each model would return.

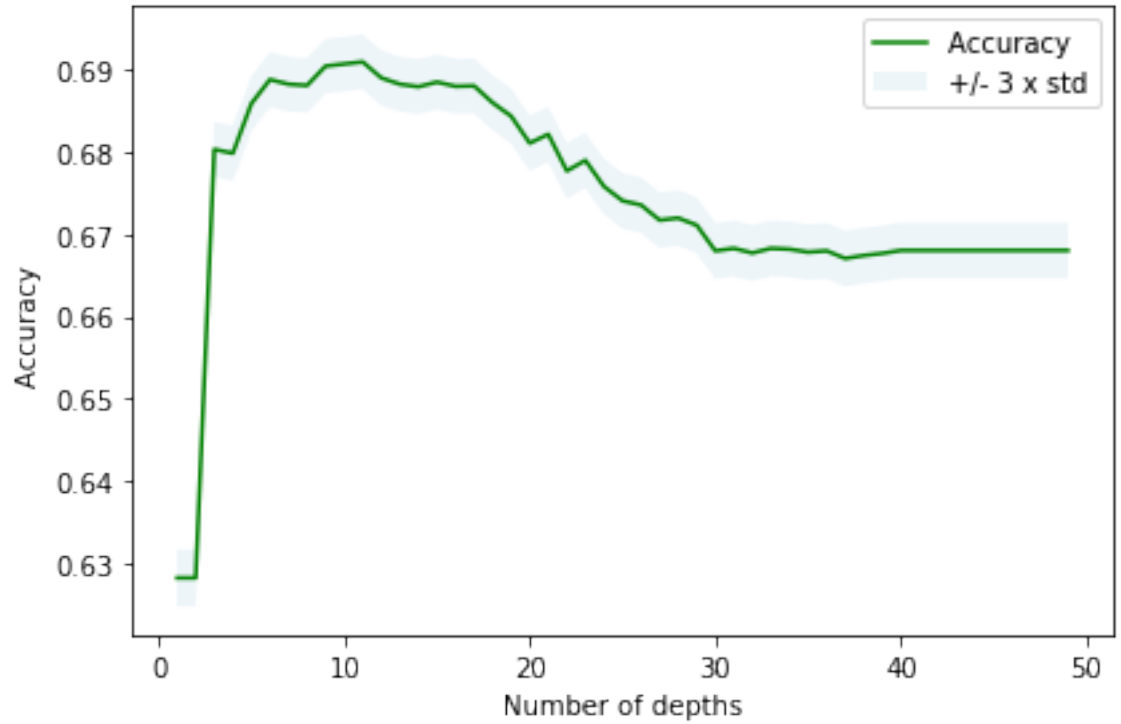
DATA MODELLING

- **K-Nearest Neighbours algorithm**
- Using a *for* loop, the accuracy of prediction using all samples in the test set was calculated and the right value for K was chosen.



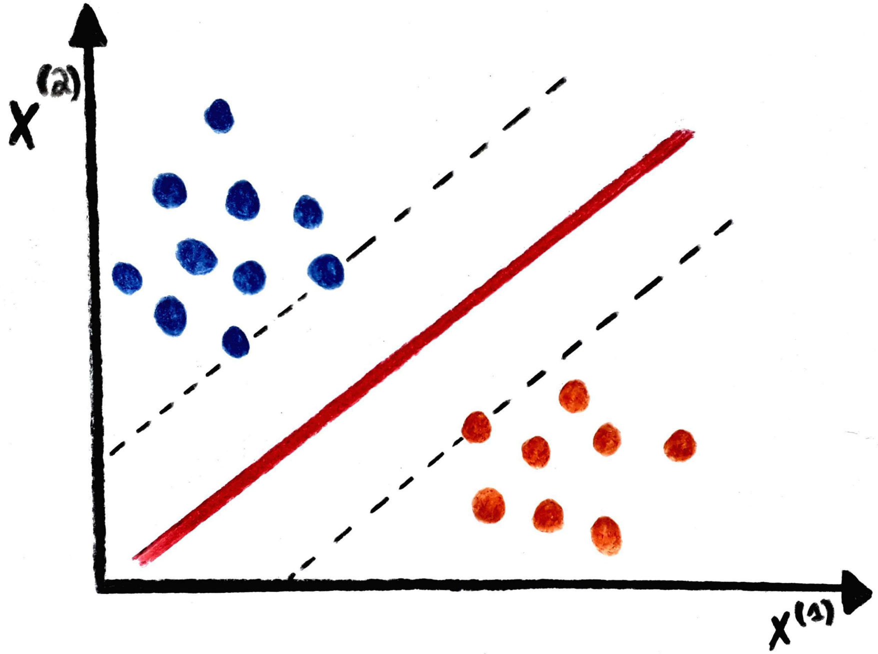
DATA MODELLING

- **Decision Tree algorithm**
- To notice the information gain of each node, the *criterion="entropy"* was used inside the classifier. It calculates the homogeneity of the samples in that node.
- The depth of a decision tree is the length of the longest path from a root to a leaf.



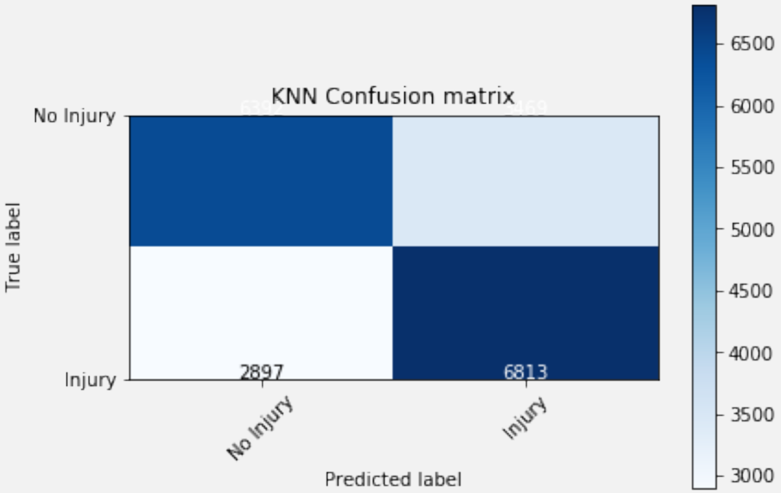
DATA MODELLING

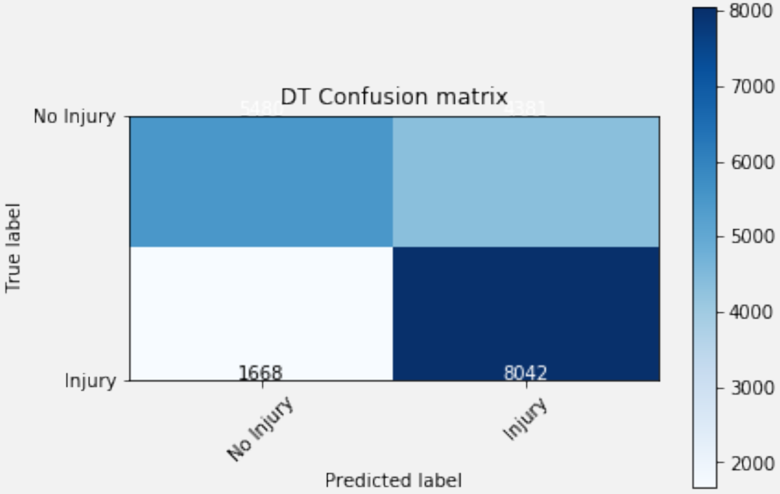
- **Support Vector Machines algorithm**
- Since SVM does not performance well with large datasets only '*rbf*' kernel was used in this model.

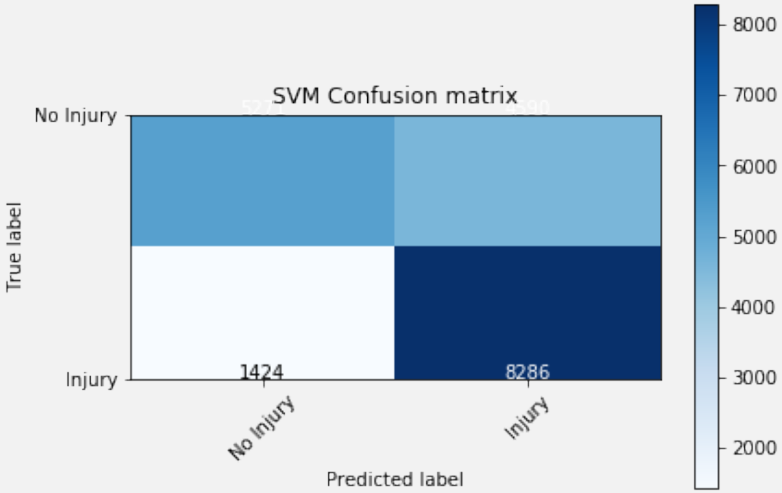


https://medium.com/@sweetai/a-brief-introduction-to-support-vector-machine-ba2ae5ea1d4e

MODEL EVALUATION





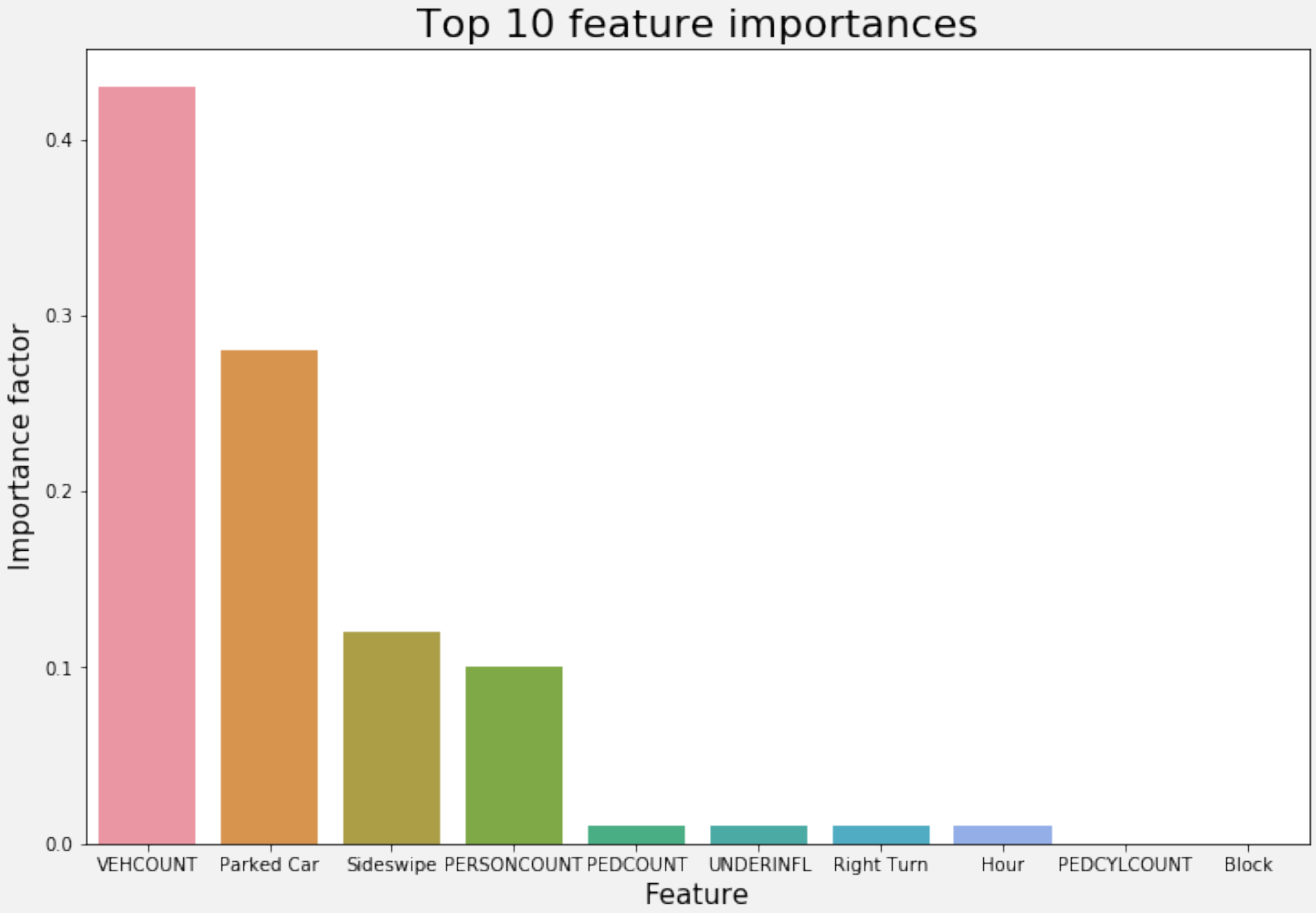


MODEL EVALUATION

From all the classifiers, **Support Vector Machine** has the best score of accuracy. The F1 score (the weighted average of the precision and recall) is highest with the **K-Nearest Neighbours** classifier. F1 Score and the recall score just the injury class are important scores, if there is a strong need to predict more of the injuries correctly. For a higher recall score in the injury class, the classifier will also mis-classify more of the 'property damage only' categories as 'injuries'. It would be a waste ambulance resources to false alarms when a real injury is happening elsewhere. That's why in my opinion K-Nearest Neighbours would be the best option since it scored the highest F1 score and it is sensitive to both the 'property damage only' class and the 'injury' class.

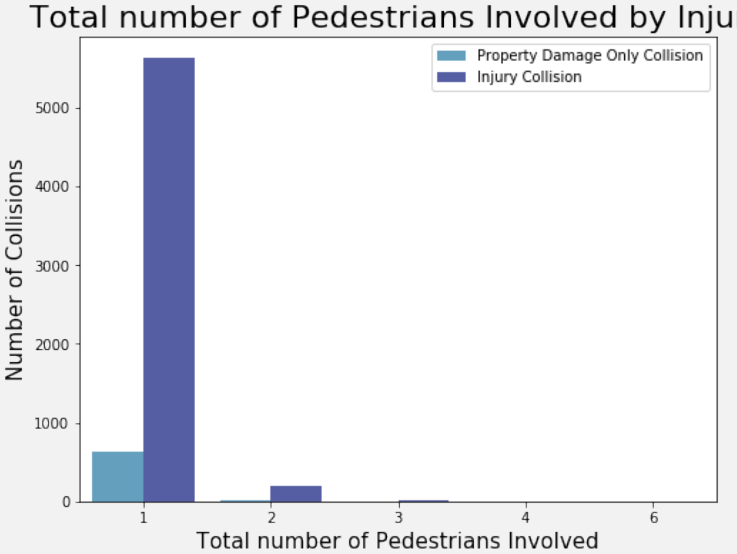
However, to see which features have the biggest influence in determining whether or not injury occurs when there's a collision the data through the Decision Tree classifier will be run.

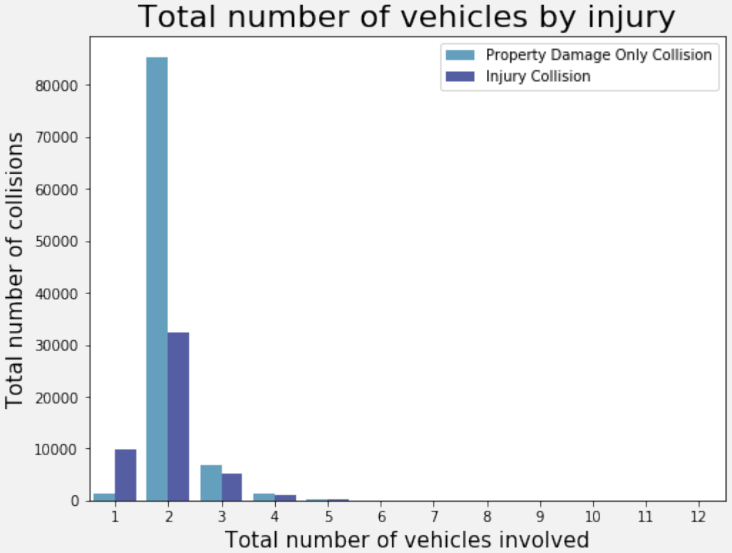
	K-Nearest Neighbours	**Decision Tree**	**Support Vector Machine**
Jaccard Score	0.6747	0.6909	0.6927
F1 Score	0.6745	0.6852	0.6849
Injury Class F1 Score	0.6816	0.7267	0.7337
Injury Class Recall Score	0.7016	0.8282	0.8533


Top 10 feature importances
Importance factor
0.4
0.3
0.2
0.1
0.0
VEHCOUNT
Parked Car
Sideswipe
PERSONCOUNT
PEDCOUNT
UNDERINFL
Right Turn
Hour
PEDCYLCOUNT
Block
Feature

DISCUSSION

- Whereas, '*parked car*' and '*sideswipe*' categories result in few injuries, the number of vehicles turns out to be important. When it is three or more of them, the chance of injury is nearly 50%. The chance for injury far outweighs property damage if there is only one car.
- Any collision with a pedestrian is likely to cause injury. Chance of injury also increases when the number of people involved goes up.


Total number of Pedestrians Involved by Injury
Property Damage Only Collision
Injury Collision
Number of Collisions
Total number of Pedestrians Involved


Total number of vehicles by injury
Property Damage Only Collision
Injury Collision
Total number of collisions
Total number of vehicles involved