Analysis on Car Accident Severity using Machine Learning

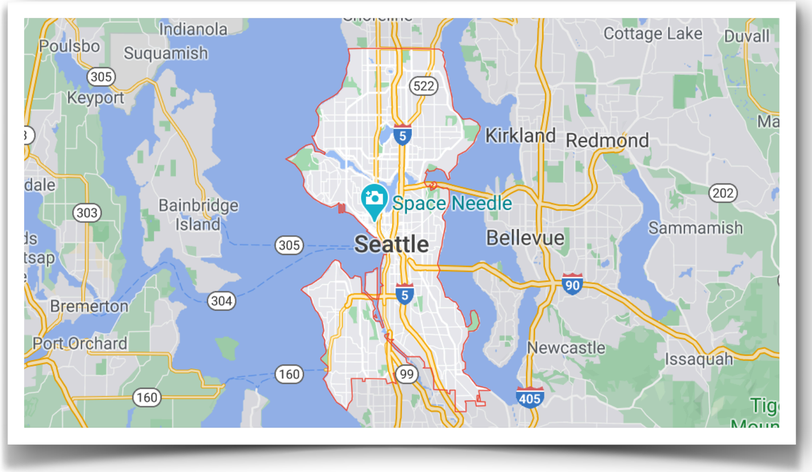
Based on Seattle Car Collision Dataset

Seohyun Lee, Oct. 2020

Introduction

Business problem

- The traffic congestion in Seattle area is among the worst in US.
- Understanding traffic problem, particularly car accident can help traffic problems.
- Stake holders
  1. Public transportation authority in Seattle
  2. Local drivers
  3. Insurance designer for car accidents



Data

Data understanding

- The data is provided by Washington State Department of Transportation 2015
- Row data comprise 194673 rows and 38 columns
- Attributes include the properties of car accidents, e.g. address type, collision type, road condition, light condition, etc.
- Individual case is labeled with severity code: 1 = property damage, 2 = injury

Data

Data cleaning

- Duplicated attributes have been removed
- Rows including NaN have been removed
- Rows including Unknown or Other elements have been removed
- Following 9 attributes remained:

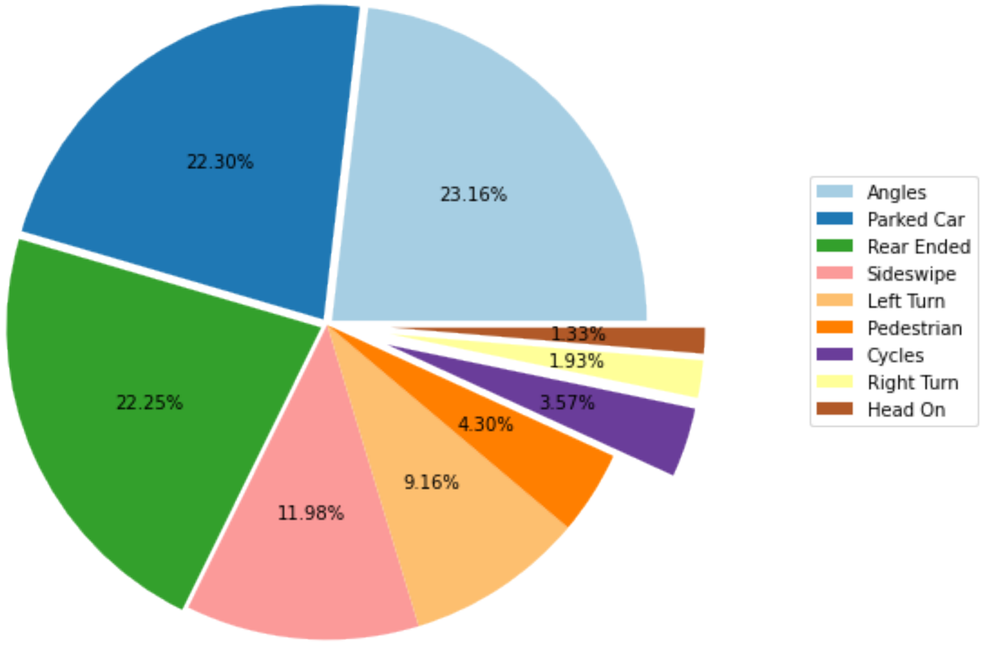
  ✓SEVERITYCODE, ADDRTYPE, COOLISIONTYPE, PERSONCOUNT, VEHCOUNT, JUNCTIONTYPE, ROADCOND, LIGHTCOND, WEATHER

- Final shape of the data is 145369 x 9

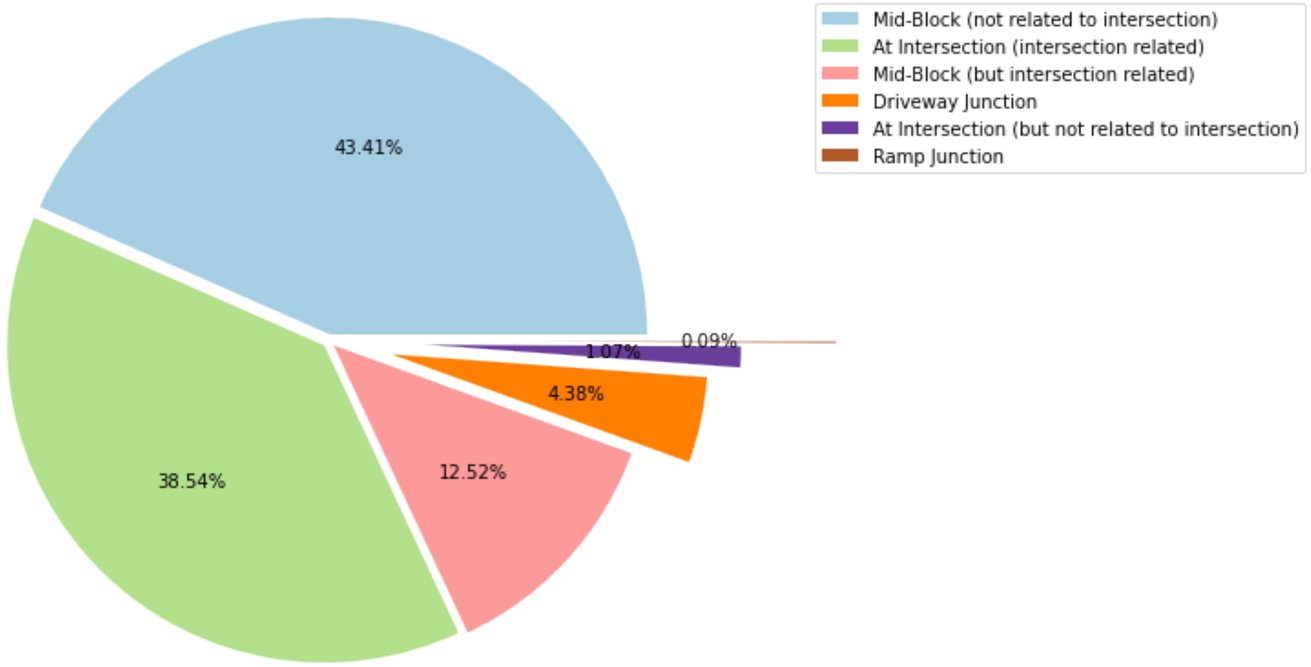
Data analysis

Distribution of attributes

COLLISIONTYPE

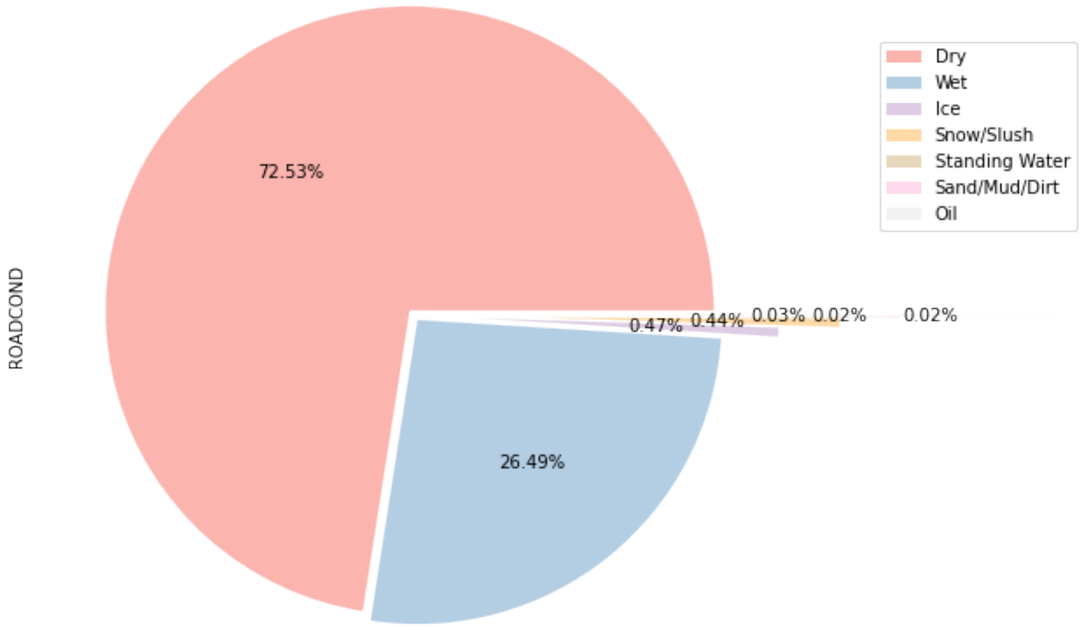


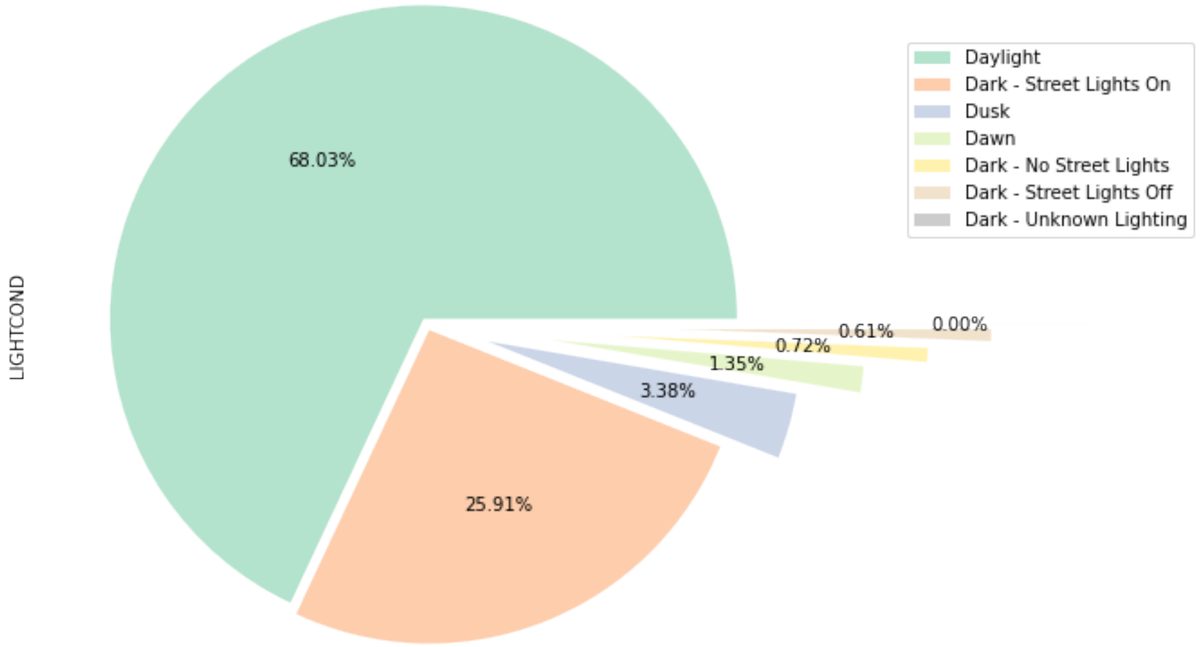
JUNCTIONTYPE



Data analysis

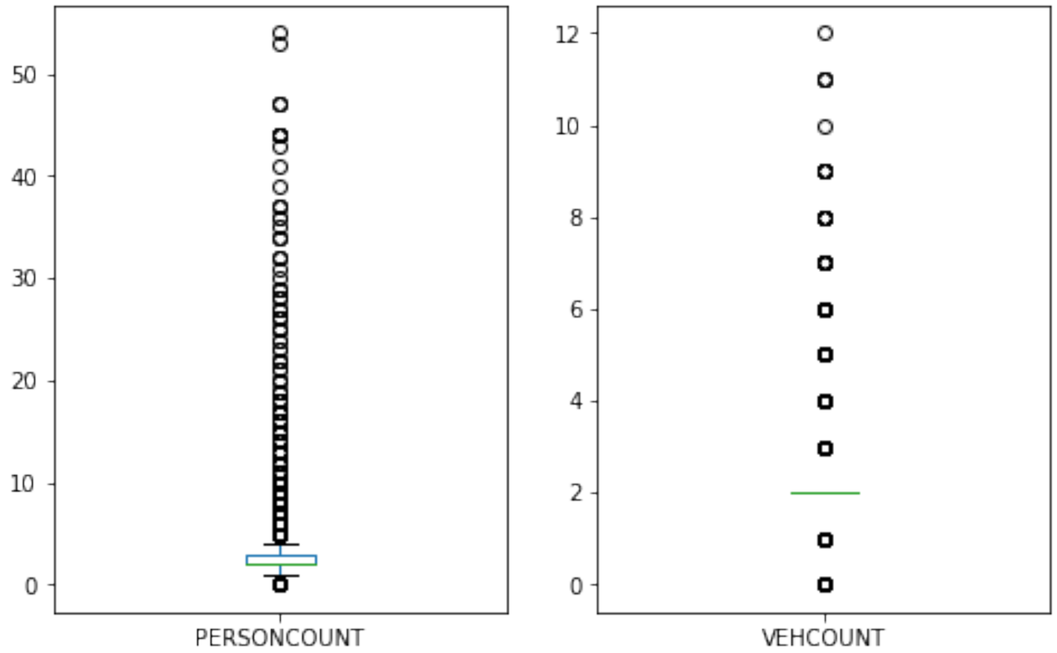
Distribution of attributes

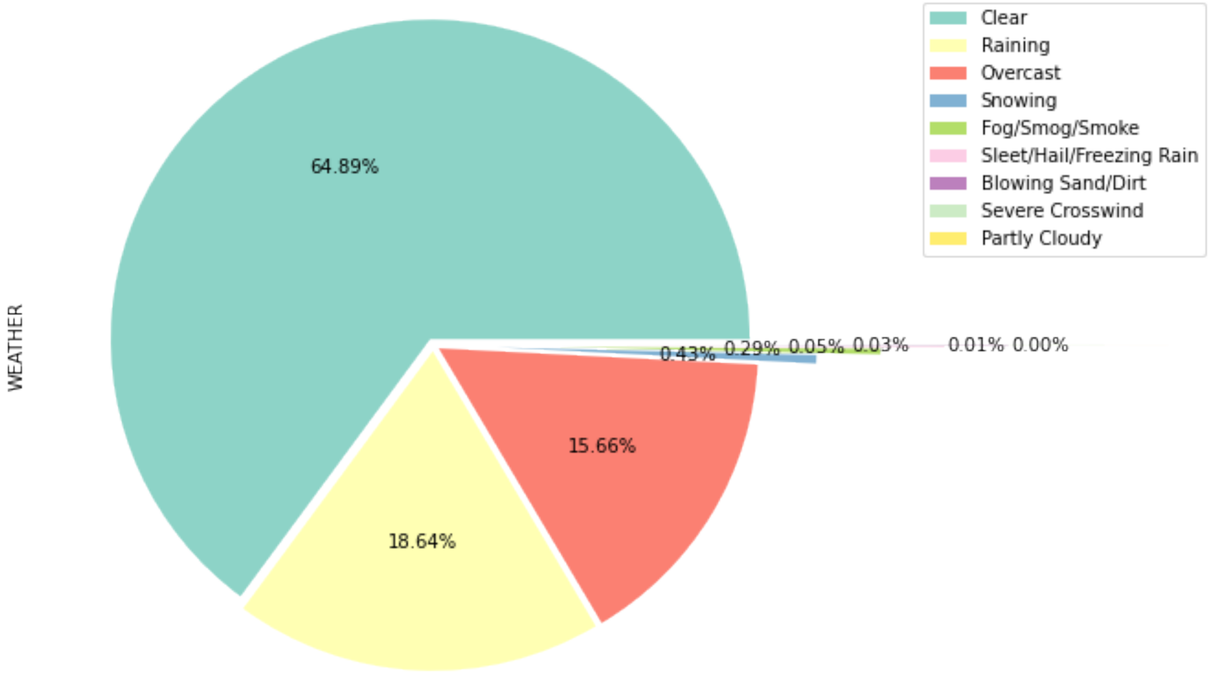




Data analysis

Distribution of attributes





Methodology

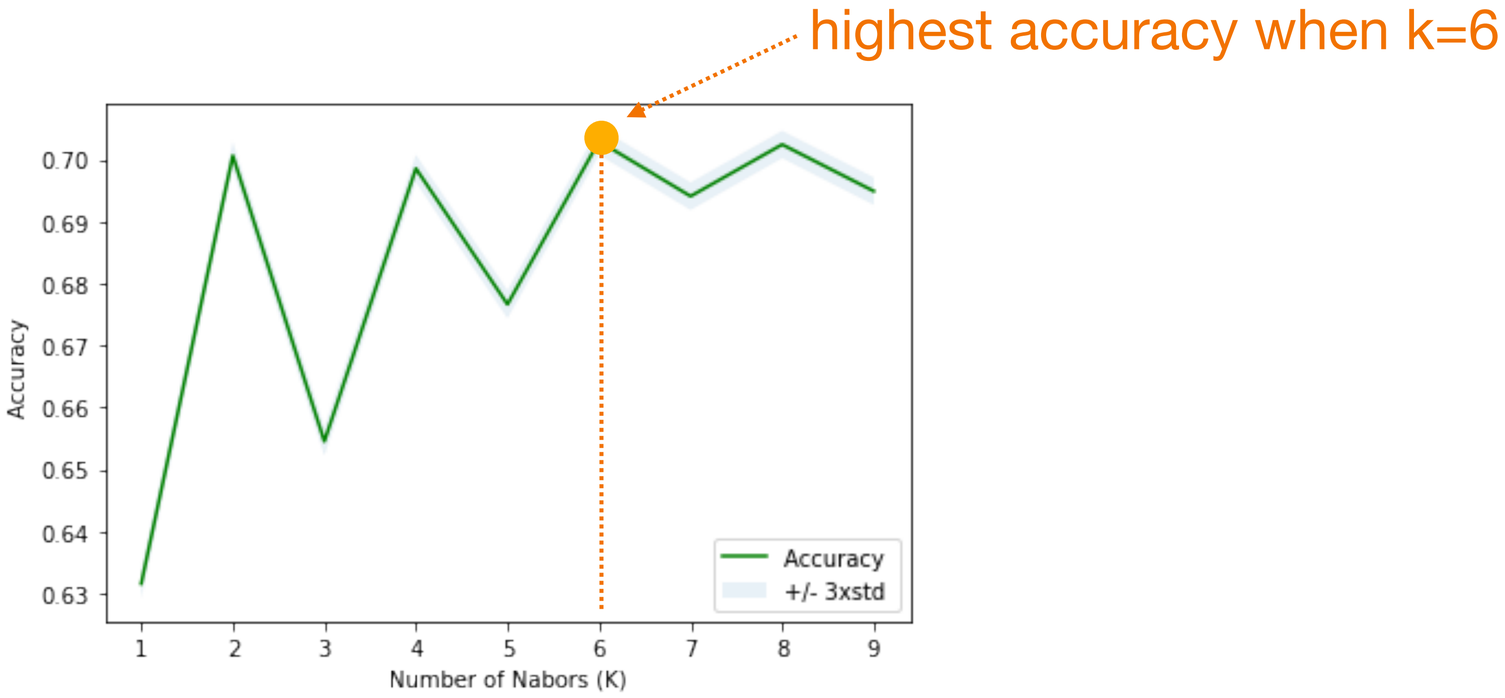
Training and test dataset preparation

- For applying machine learning to categorical variables, each category in attributes is encoded in integer value
- The encoded value except SEVERITYCODE is prepared as training data X
- The SEVERITYCODE is prepared as training data y
- The dataset is divided by 7:3 where test size is 30 % of total dataset

Analysis

K-nearest neighbors

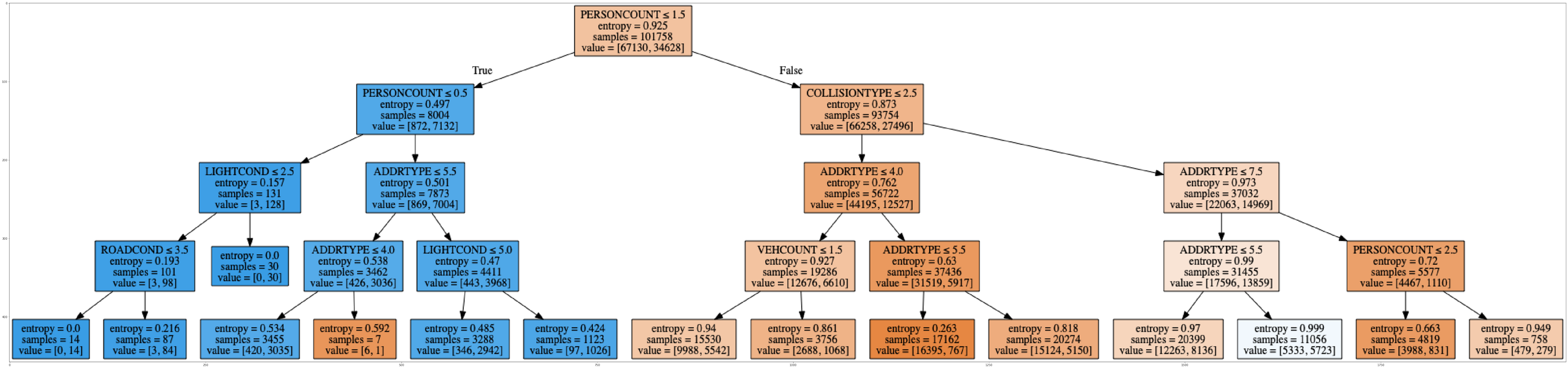
- Determining the K using accuracy computation



Analysis

Decision Tree

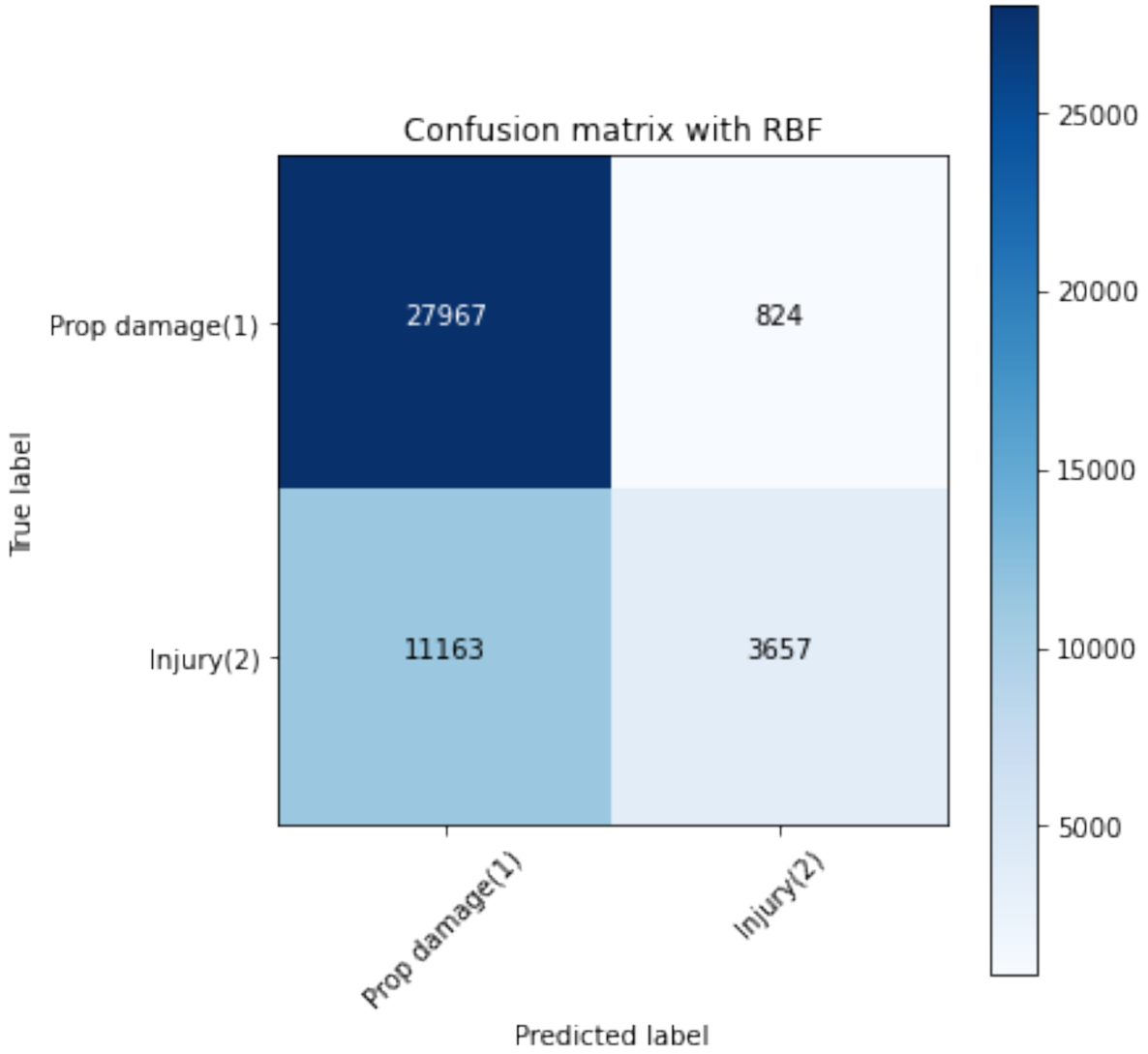
- With max depth=4



Analysis

Support vector machine

- With Radial Basis Function



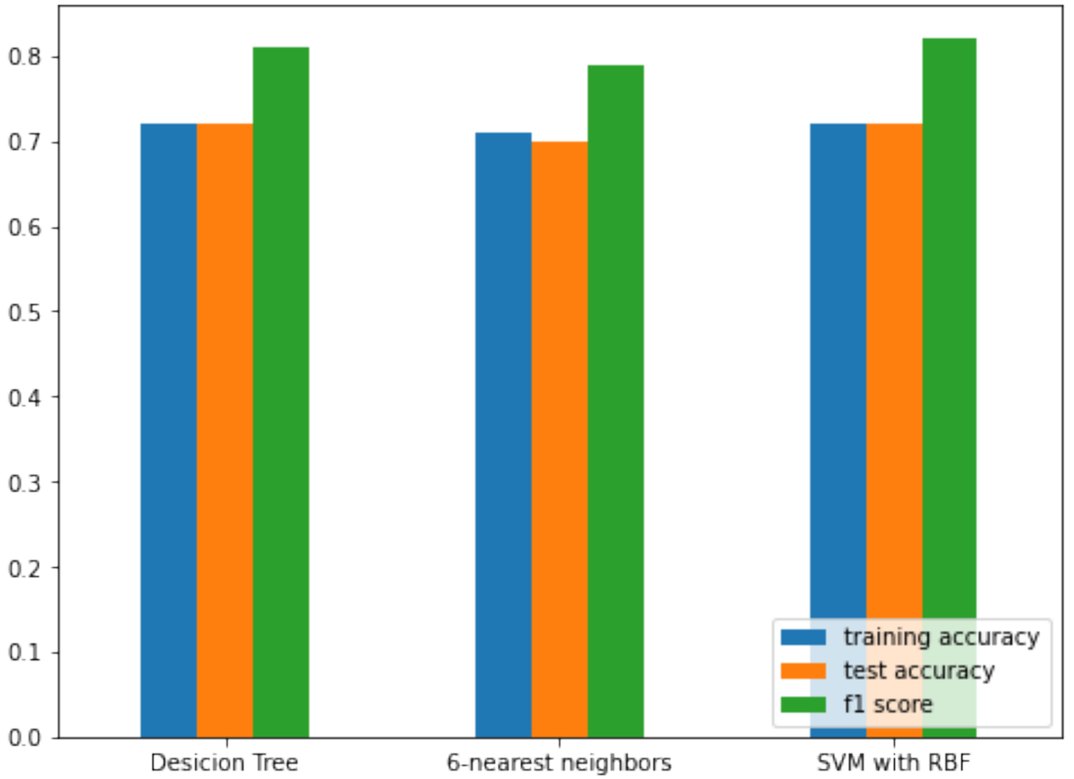
Result and Discussion

Accuracy comparison

- Training accuracy
- Test accuracy
- F1 score

  The above parameters in three different

  method are computed and compared



Conclusion

Best prediction model

- Support vector machine was revealed as the best prediction model with F1 score 0.82
- Using this method, the severity (either property damage or injury) can be predicted.
- This result can benefit the stakeholders.