Predicting Car Accident Severity

Nicholas (Yangshuai) Liu

Background

- 1. Car accident is ranked 8th (7th by 2030 by prediction)leading cause of death and 1.35 millions people died from car accidents globally every year;
- 2. Car accidents cost most countries 3% of their gross domestic product (GDP); and it was reported that the cost of medical care and productivity losses associated with motor vehicle crash injuries was over 99 billion in the United States in 2010
- 3. Car accident prediction models provides important information for emergency personnel to assess the severity of accidents, assess the potential impact of the accident, and implement effective accident management procedures

IBM Cognitive Class
IBM
Capstone Project
Applied Data Science
Powered By
coursera
IBM Digital Credential

Data cleaning

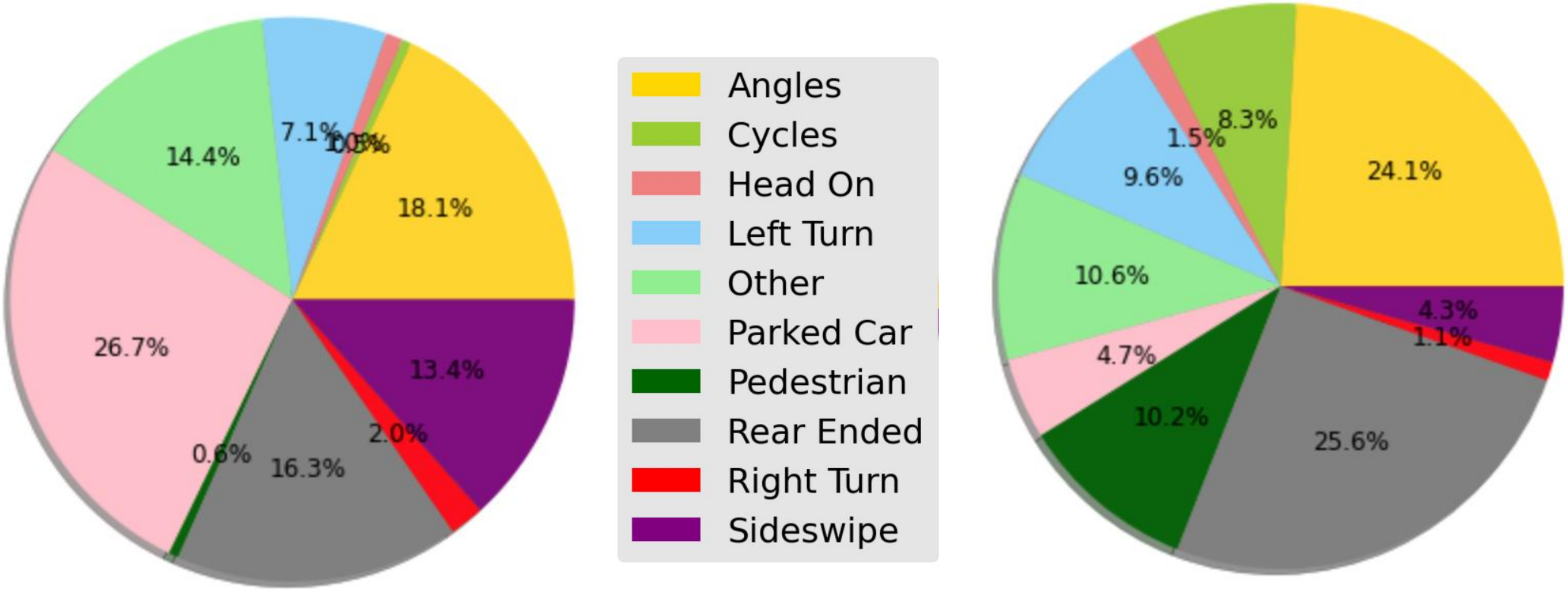
- Car accident data was acquired directly from the link provided by the course
- Original data has 194673 rows and 38 features, among which only correlated attributes were selected; especially, the date format was transformed to “workday” and “weekend” for further investigation
- All the missing and “Unknown” values of the selected attributes were dropped
- 167980 rows with 9 columns of dataset were obtained after data cleaning

Data Visualization

-Collision type vs. car accident severity

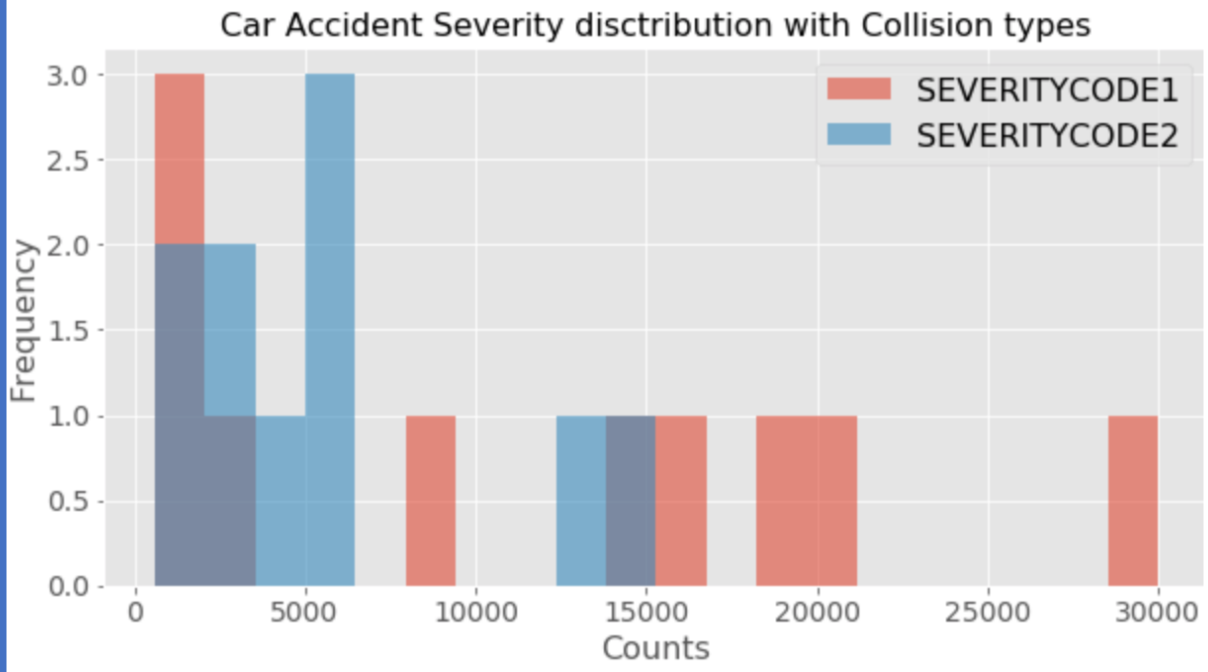
SEVERITYCODE1

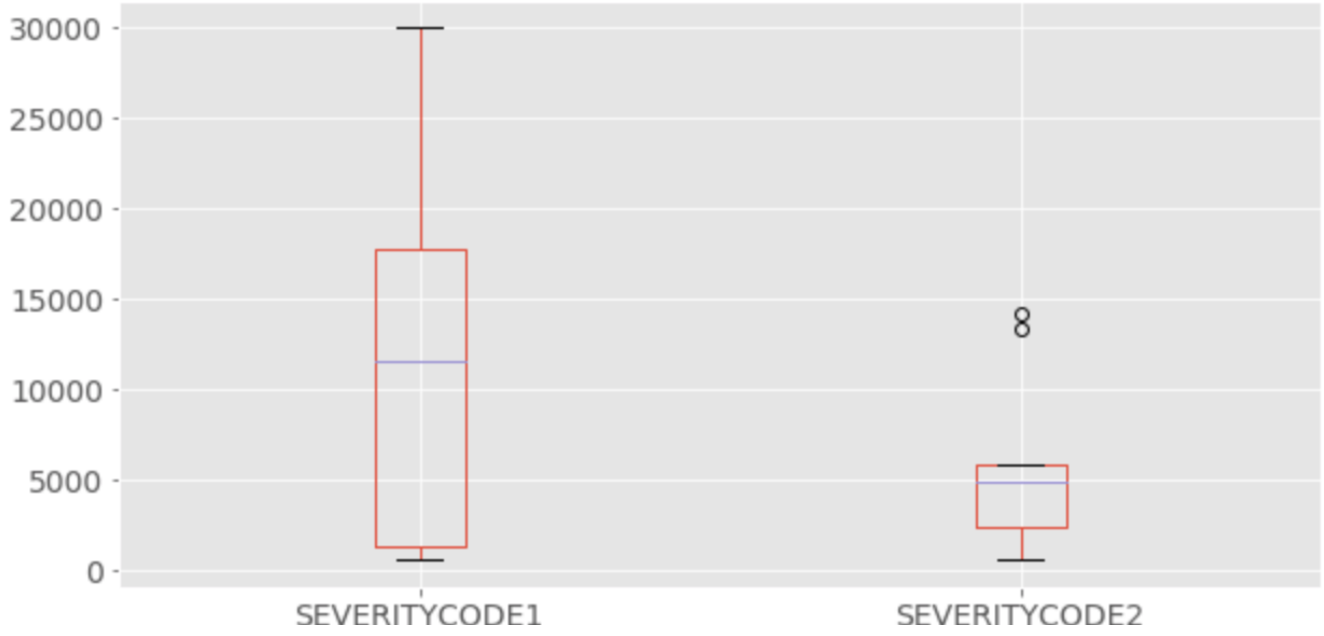
SEVERITYCODE2



- 26.7% of the accidents associated with SEVERITYCODE1 occurred between parked cars, while 24.1% of those with SEVERITYCODE2 occurred between rear ended cars, which means accidents caused by parked cars are less likely to lead to severe injuries compared to those caused by rear ended cars
- The percentage of car accidents caused by different collision types vary widely in SEVERITYCODE1 and SEVERITYCODE 2 categories

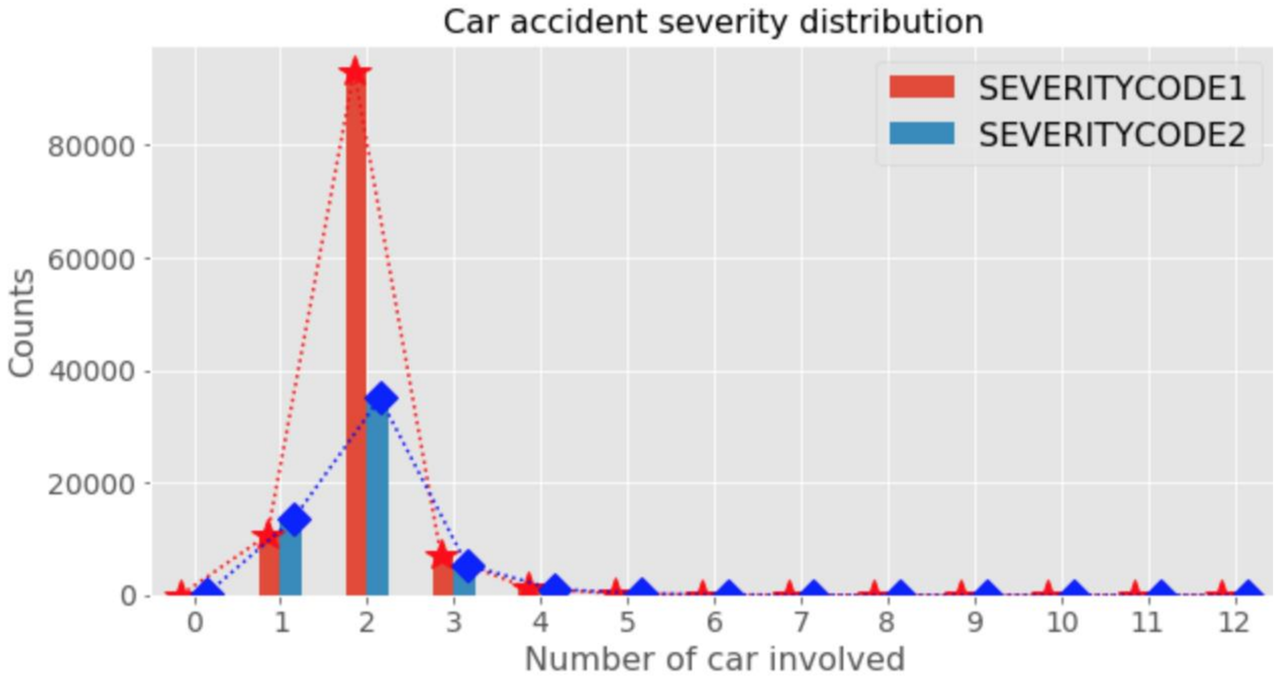
Data Visualization
-Collision type vs. car accident severity

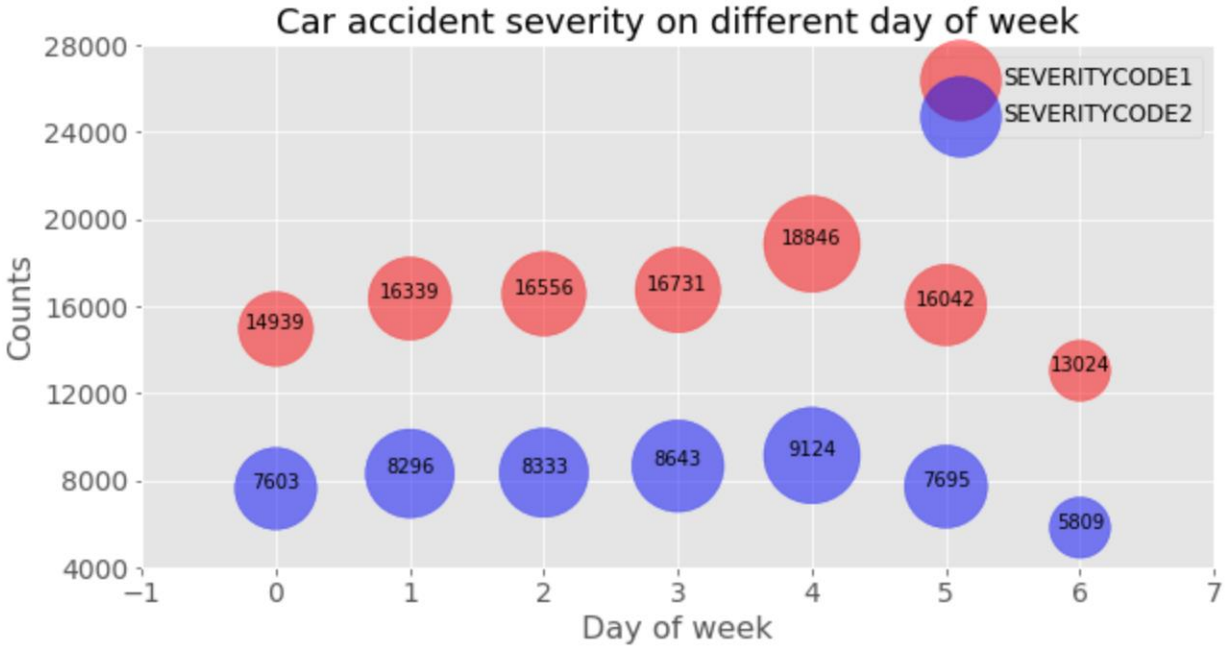

Car Accident Severity disctribution with Collision types



- SEVERITYCODE1 and SEVERITYCODE2 demonstrated different frequency distribution pattern with different collision types as shown in the histogram.
- The distribution of SEVERITYCODE1 is more continuous from 0 to 30000 while there are two outliners for SEVERITYCODE2 and most of SEVERITYCODE2 drop below 10000. The median counts number for SEVERITYCODE 1 and 2 are ~12000 and ~5000, respectively, which can also be effected by the unbalanced dataset as shown in the box plot.

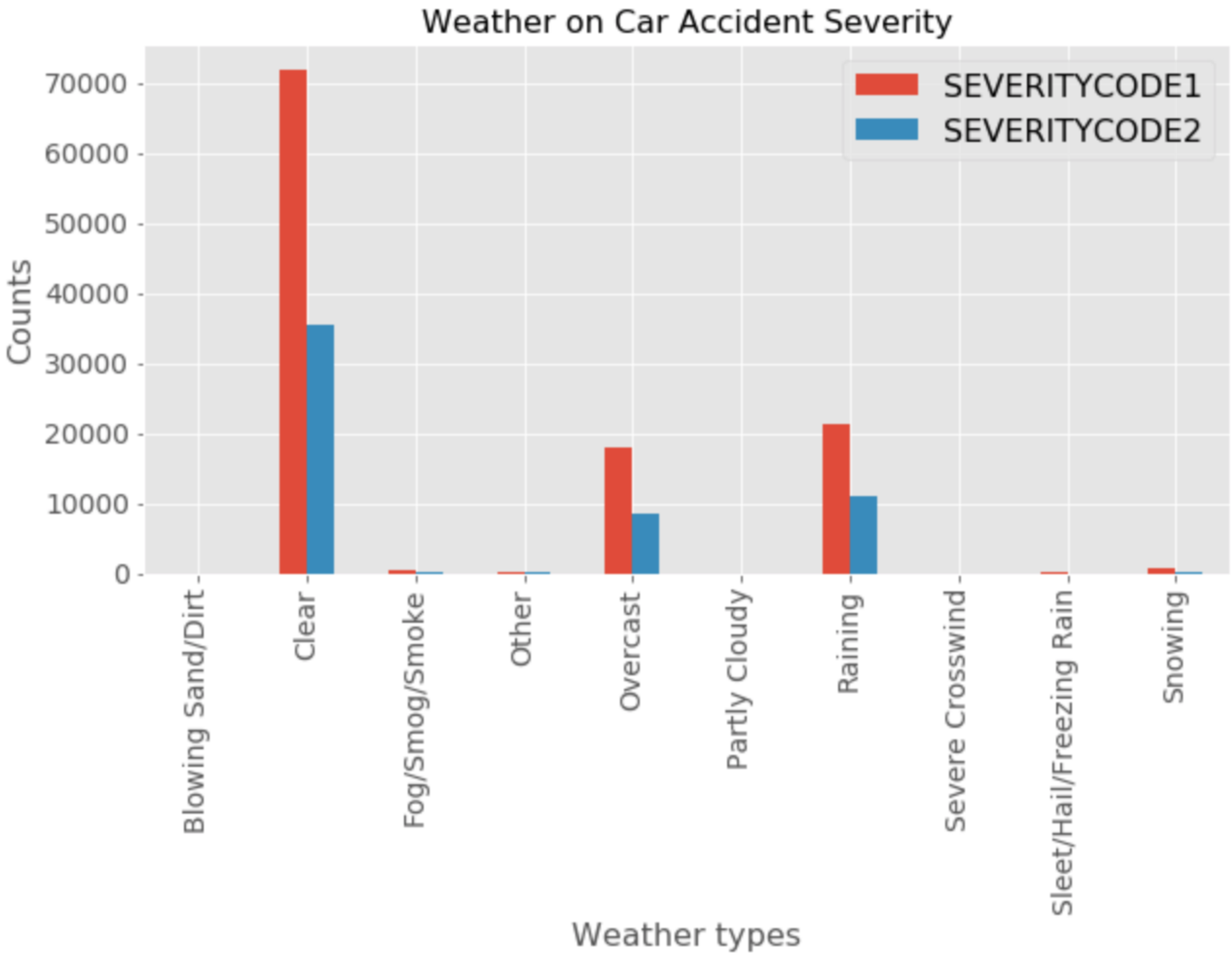
Data Visualization
-Number of cars/day of week vs. car accident severity


Car accident severity distribution


Car accident severity on different day of week

- Most of the accidents associated with SEVERITYCODE1 and 2 involve two cars while there are more SEVERITYCODE 2 accidents involving one car
- Both SEVERITYCODE1 and 2 accident counts slightly increase from Monday to Friday then sharply drop through the weekend

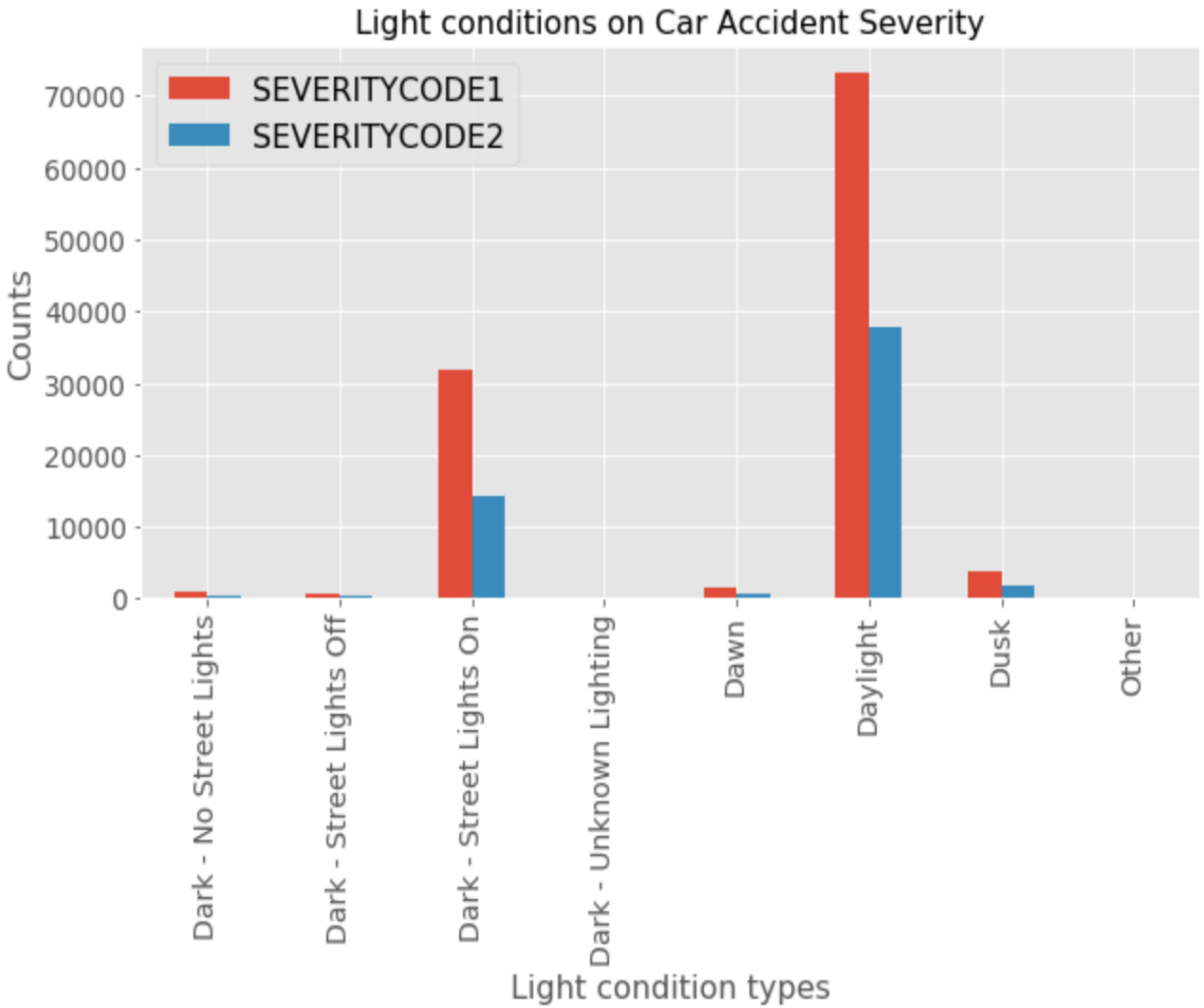
Data Visualization
-Weather type vs. car accident severity

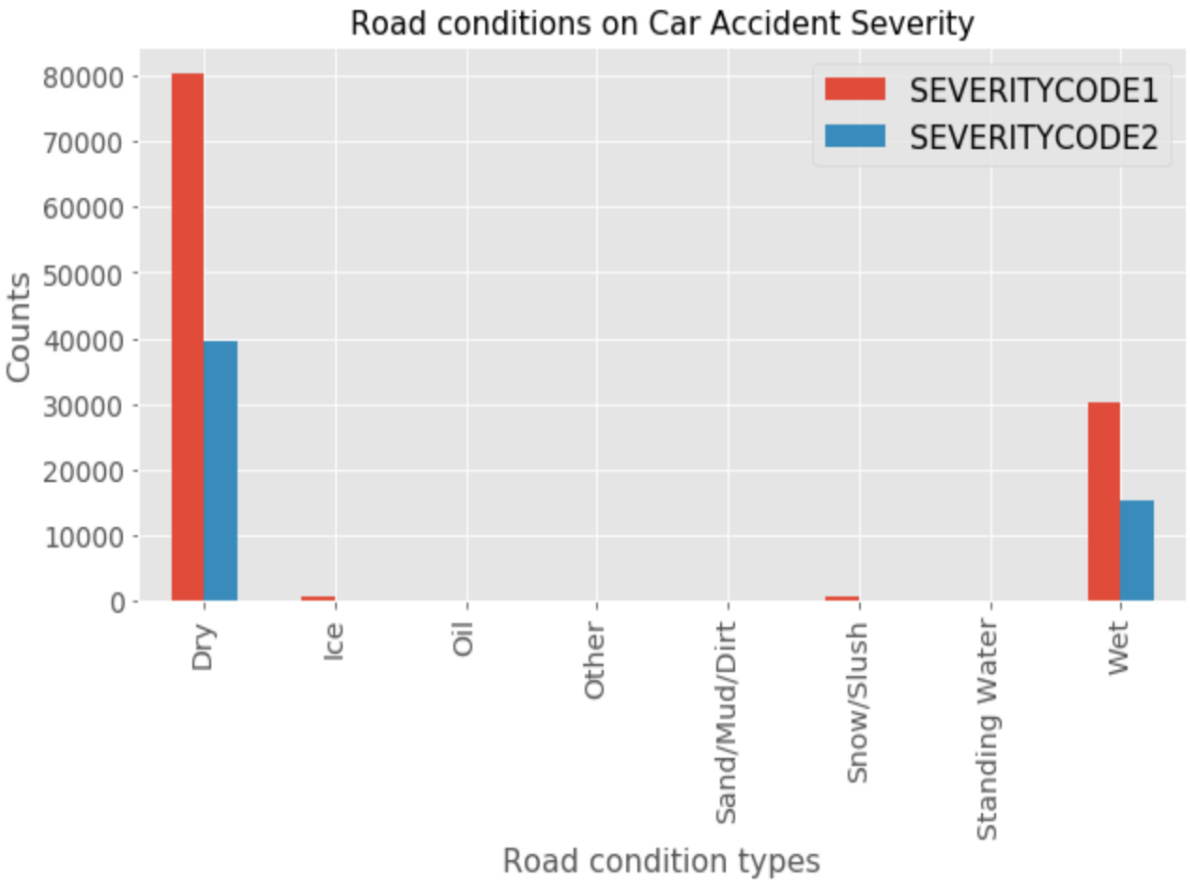


- Most car accidents happen on clear days, followed by raining days and overcast
- The counts of both severitycode1 and severitycode2 show the similar pattern versus different weather types
- The data quantity of severitycode1 is twice of that of severitycode2

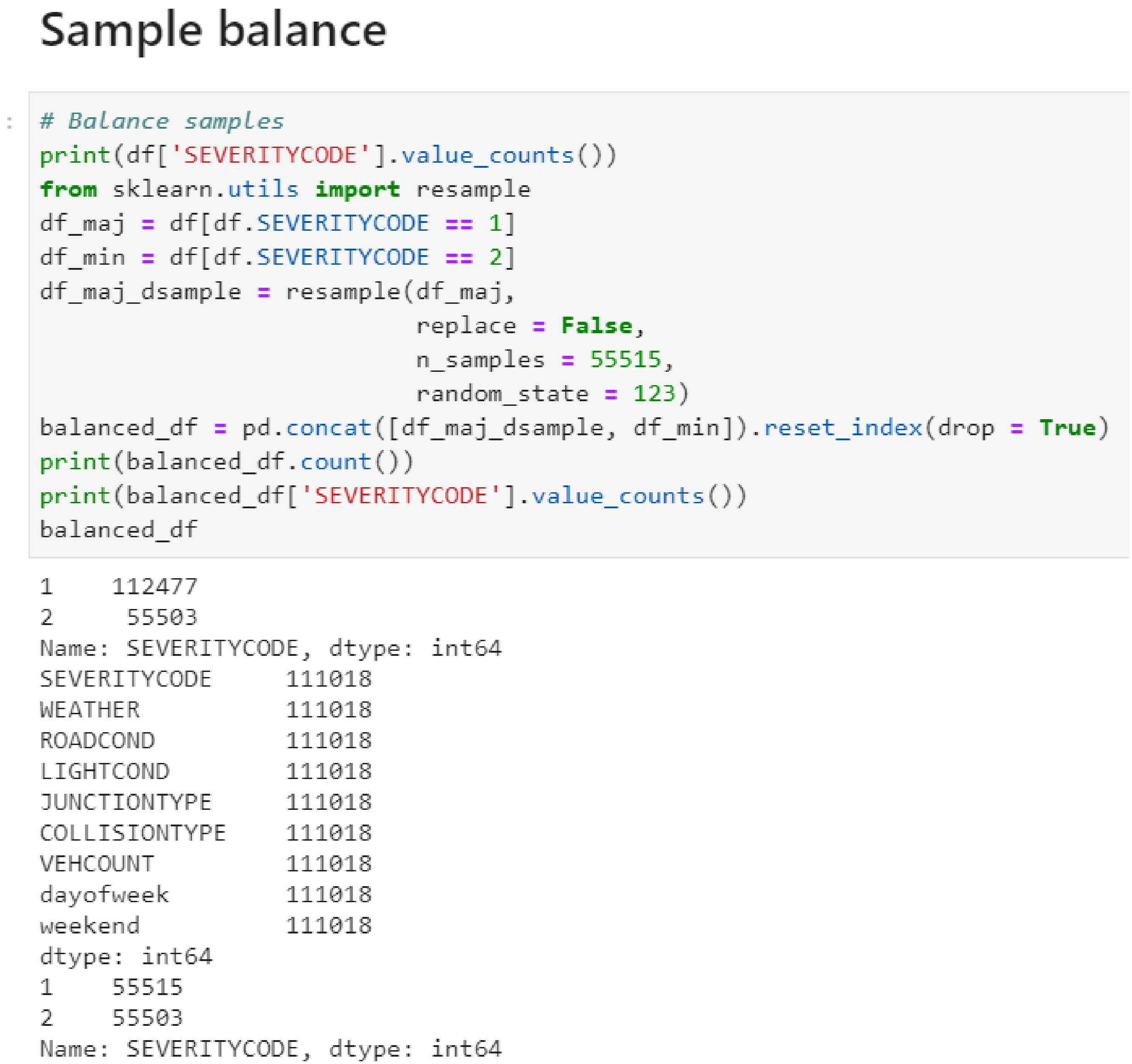
Data Visualization

-Light/road conditions vs. car accident severity





Data Balance



Sample balance

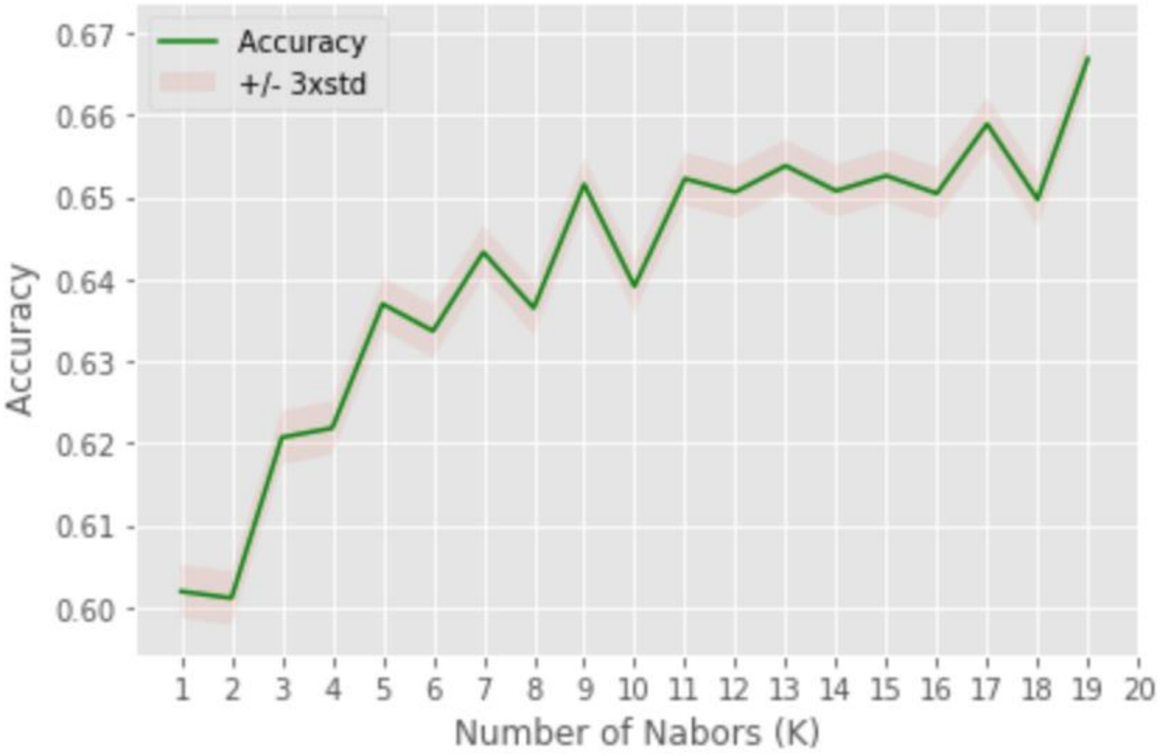
```
# Balance samples
print(df['SEVERITYCODE'].value_counts())
from sklearn.utils import resample
df_maj = df[df.SEVERITYCODE == 1]
df_min = df[df.SEVERITYCODE == 2]
df_maj_dsample = resample(df_maj,
                          replace = False,
                          n_samples = 55515,
                          random_state = 123)
balanced_df = pd.concat([df_maj_dsample, df_min]).reset_index(drop = True)
print(balanced_df.count())
print(balanced_df['SEVERITYCODE'].value_counts())
balanced_df
```

```
1    112477
2     55503
Name: SEVERITYCODE, dtype: int64
SEVERITYCODE      111018
WEATHER           111018
ROADCOND          111018
LIGHTCOND         111018
JUNCTIONTYPE      111018
COLLISIONTYPE     111018
VEHCOUNT          111018
dayofweek         111018
weekend           111018
dtype: int64
1    55515
2    55503
Name: SEVERITYCODE, dtype: int64
```

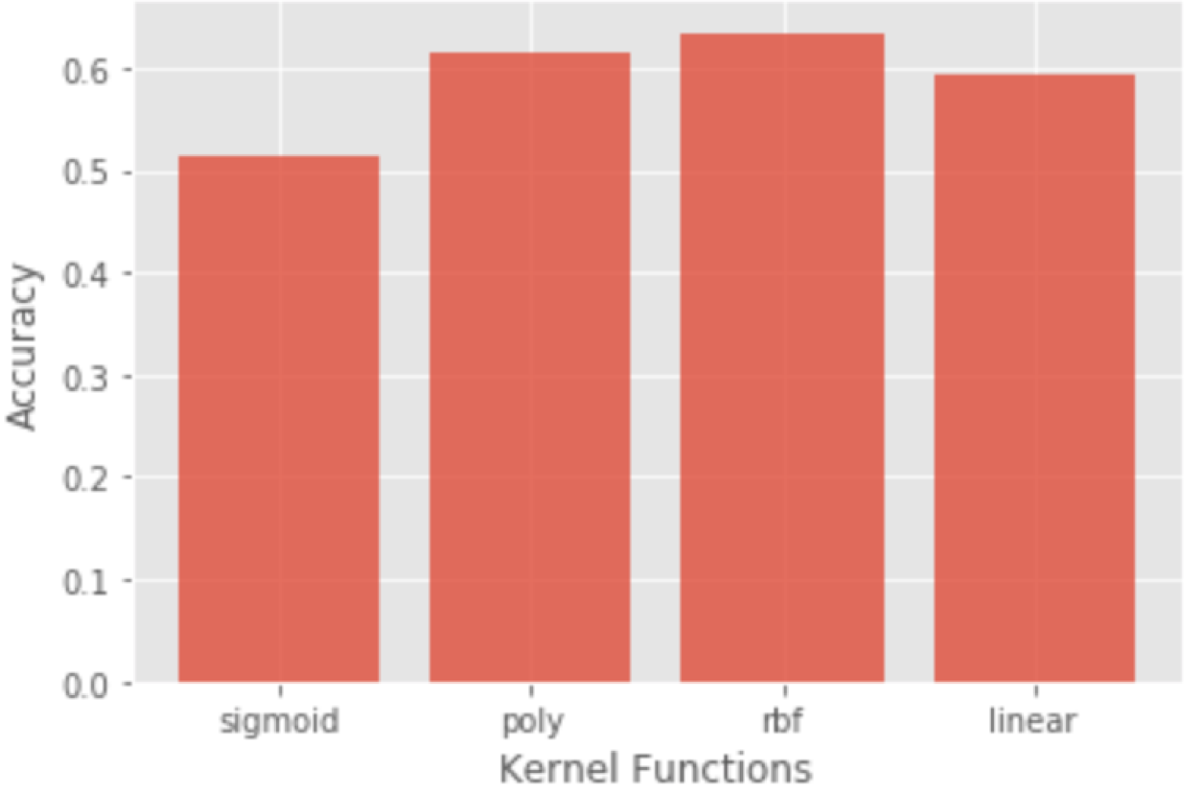
- Before data balance, the data quantities with independent variable SEVERITYCODE1(11247 rows) is double of that of SEVERITYCODE2 (55503 rows), which will lead to a biased model for further prediction
- After data balance, the data quantities with independent variable SEVERITYCODE1(55515 rows) is almost equal to that of SEVERITYCODE2(55503 rows)

Data modeling(machine learning)

KNN model accuracy score with different K values



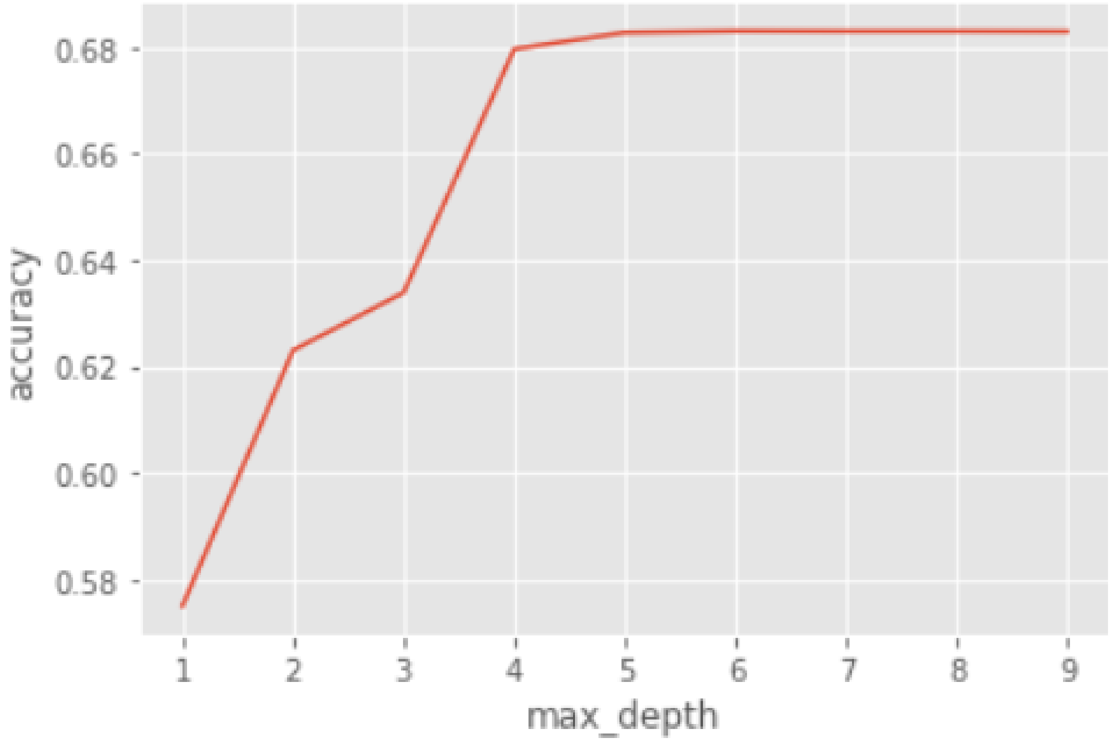
SVM model accuracy score with different Kernal functions



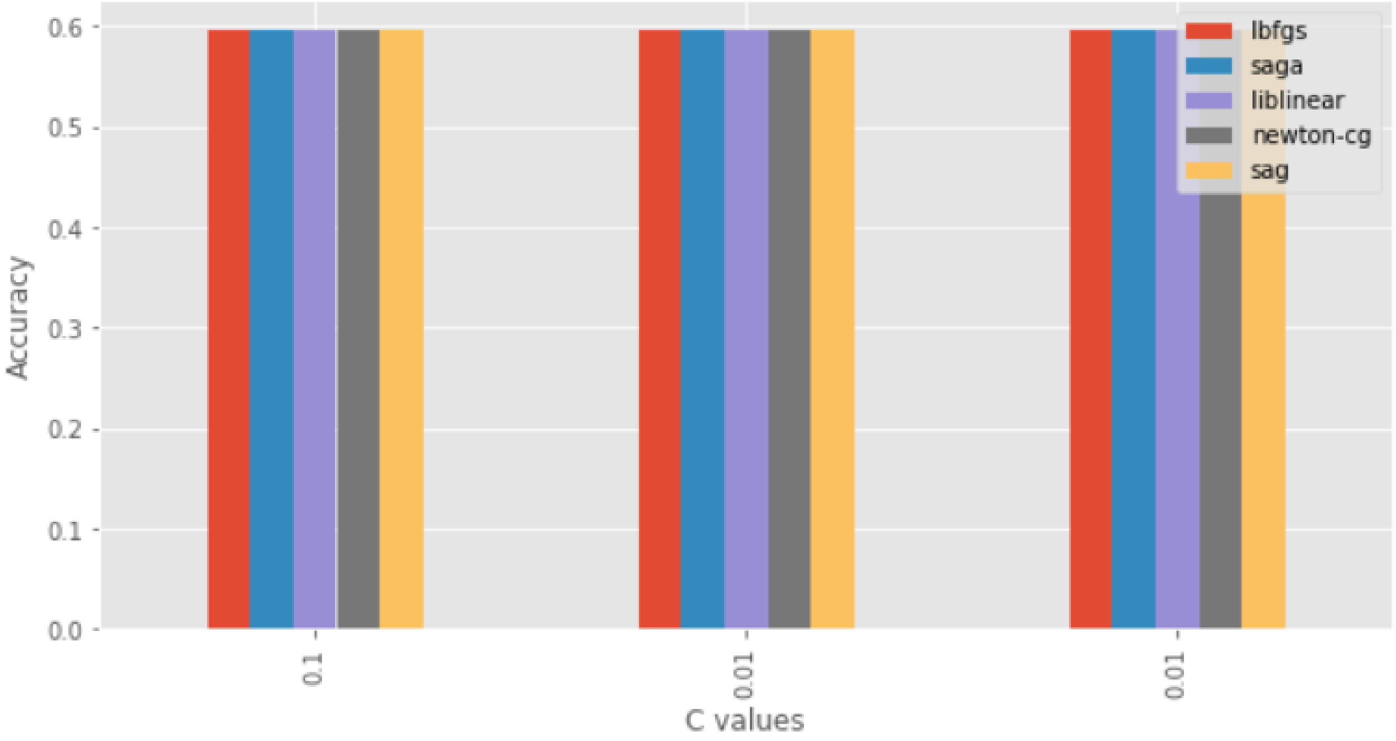
- 80% of the data was utilized as training data and the rest was used as test data
- KNN, SVM, LogiscticRegression, and Decision Tree models were established for prediction

Data modeling(machine learning)

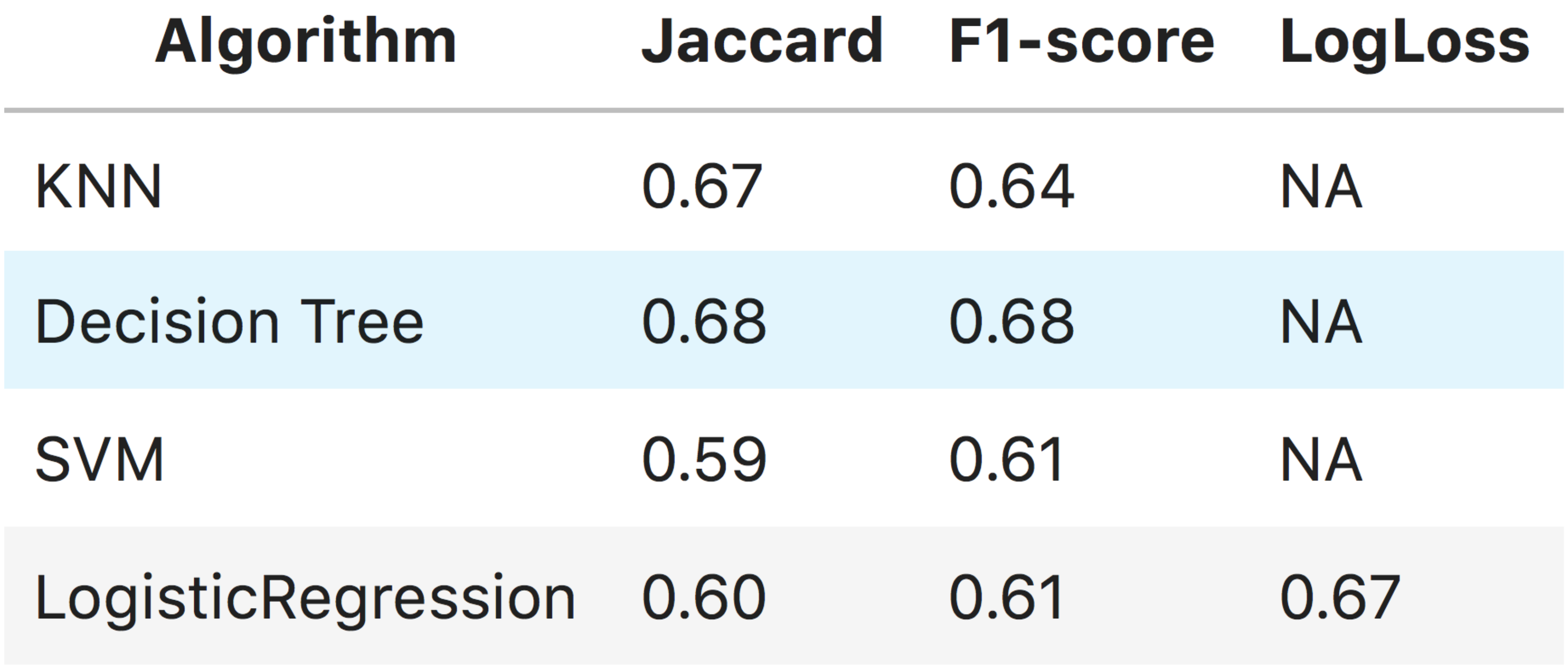
Decision tree models with different max_depth



LogisticRegression models with different C values & functions



Data evaluation



Algorithm	Jaccard	F1-score	LogLoss
KNN	0.67	0.64	NA
Decision Tree	0.68	0.68	NA
SVM	0.59	0.61	NA
LogisticRegression	0.60	0.61	0.67

- The best prediction model should be the Decision Tree with a max_depth of 5 with Jaccard similarity score of 0.68 and F1-score of 0.68