Prediction of Automotive Accident Severity

MOTIVATION:

The motivation behind our research is the understanding of specific conditions that affect the severity of an automotive accident. The purpose of this project is to highlight impactful variables while operating a vehicle in order to improve accident prevention.

GOAL:

We executed a classification based on U.K. road accidents ranging from 2014 to 2016 using the methodologies covered in class (Logistic Regression, Neural network, KNN, Decision tree). Our classification specifies the impact of certain features on car wreckage.

DATA SOURCE:

The data collected comes from the U.K. government who amassed traffic data based on police reports. The analysis of data executed here is composed of the U.K. road accidents from 2014 to 2016.

Accidents are recorded according to these features:

- Reference Number
- Grid Ref: Easting
- Grid Ref: Northing
- Expr1
- Severity
- Day of the week
- Time (24hr)
- 1st Road Class
- Road surface
- Accident date
- Weather condition
- Lighting conditions
- Number of vehicles
- Casualty class

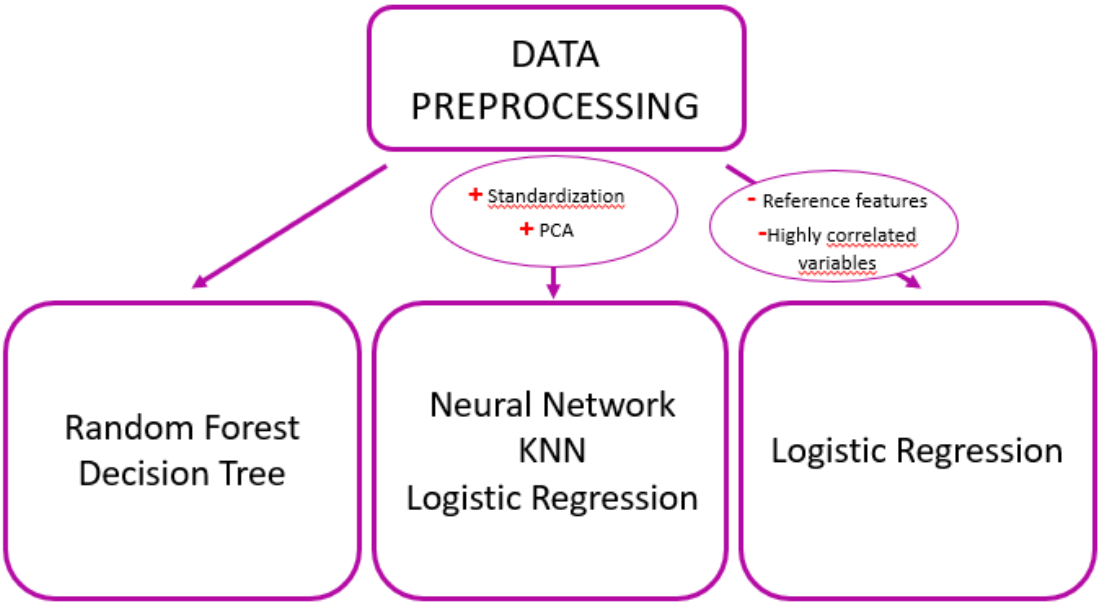
- Sex of casualty
- Age of casualty
- Type of vehicle

DATASET SOURCE:

https://data.gov.uk/dataset/6efe5505-941f-45bf-b576-4c1e09b579a1/road-traffic-accidents

METHODOLOGY:

I. Data preprocessing:

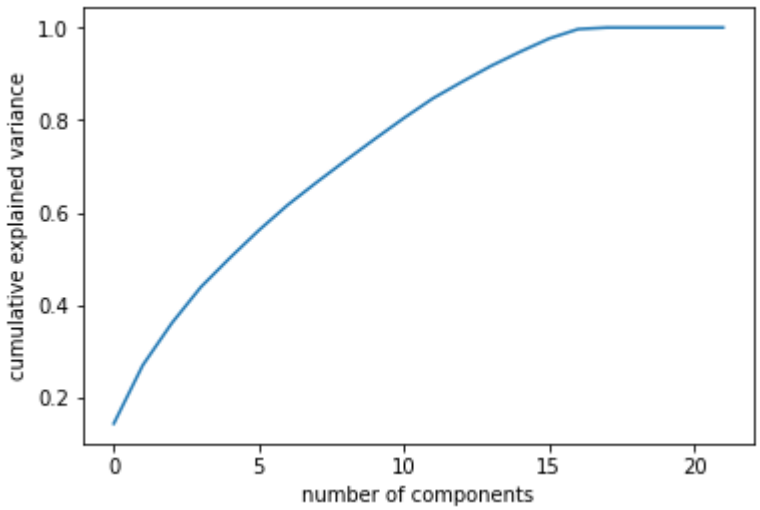


- Merging datasets
- Dropping columns containing references (Reference number, Grid Ref: Easting, Grid Ref: Northing) and correlated variables (Lighting conditions, Accident Date).
- Dealing with missing data by deleting observations that are labeled with NaNs.
- Listing variables:
    - Time (24hr): Day-time, Night-time
    - Weather conditions: Fine, Snowing, Raining, Fog, Other
    - Type of Vehicle: Car, Bus, Goods vehicles, Motorcycle, Other
    - Day: Weekday, Weekend
    - Casualty class: Passenger, Pedestrian, Driver

- Creating dummies out of categorical variables and dropping variables containing the same information (Sex of casualty_Female, Day_Weekday, Time (24hr)_Day-time)
- Resampling unbalanced data
    - Slight: 6739, Serious: 957, Fatal: 48
        - Undersampling from slight to serious
    - Slight: 957, Serious: 957, Fatal: 48
        - Oversampling from fatal to serious
    - Slight: 957, Serious: 957, Fatal: 957

II. Standardization and PCA:

- Standardization
- PCA



We utilize the first 12 components as they make up approximately 90% of the variance.

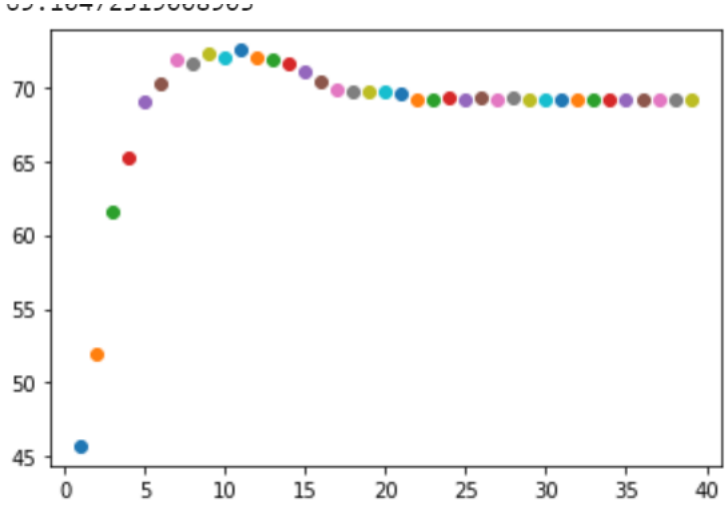
III. Prediction:

Accuracy of each of the following methods were examined to choose the best classifier for reaching our goal. To implement the methods mentioned below, scikit-learn and Keras were used.

To avoid overfitting, we used K-fold cross-validation method with ten splits.

Decision Tree

- The graph below shows the depth that returns the best accuracy based on the number of features that we have in the dataset.



- K-fold best mean accuracy is 70.29% (standard deviation 3.41%) for a decision tree depth equal to six.

- The three most important features in the decision tree model are: Casualty Class_Pedestrian, Road Surface_Dry, Road Surface_Wet or Damp.

Random Forest

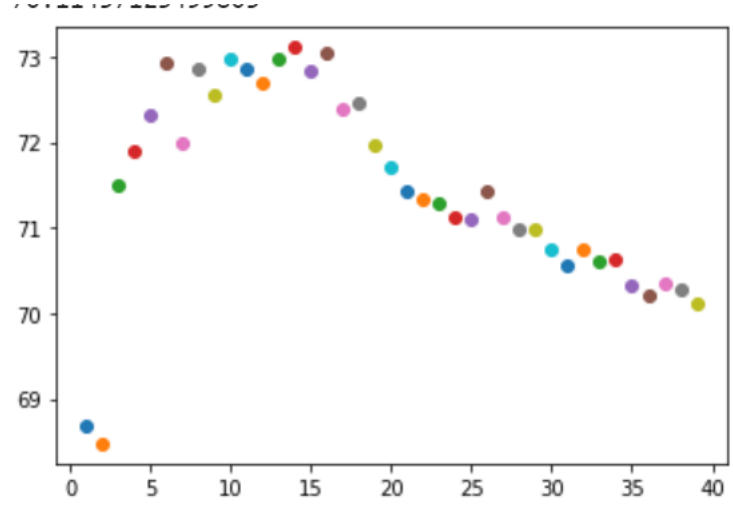
- The mean accuracy is equal to 67.15% (standard deviation 3.98%).

Neural Network

- Using preprocessed standarized data followed by PCA.
- Two hidden layers each containing 24 nodes.
- The mean accuracy is equal to 72.66% (standard deviation 2.95%).

KNN

- Using preprocessed standarized data followed by PCA.
- The graph below shows the number of neighbors that returns the best accuracy based on the number of features that we have in the dataset.



- K-fold best mean accuracy is 72.31% (standard deviation 2.77%) for number of neighbors equal to five.

Logistic Regression

using PCA

- The mean accuracy is equal to 57.58% (standard deviation 2.82)

without PCA

- Dropping reference variables.
- Dropping 'Weather Condition' variable due to its high correlation with 'Road Surface'.
- The mean accuracy is equal to 61.16% (standard deviation 3.02%).

Number of Vehicles	Age of Casualty	Day_weekend	Time (24hr)_Night-time	Road Surface_Flood (surface water over 3cm deep)	Road Surface_Frost or Ice	Road Surface_Snow	Road Surface_Wet or Damp	Casualty Class_Passenger	Casualty Class_Pedestrian	Sex of Casualty_Male	Type of Vehicle_bus
-1.30332	0.00917862	-0.494654	0.0387718	-0.0478659	-0.138922	-0.04959	-0.0581141	-1.01767	-1.44951	-0.0503227	-1.16631

SUMMARY

Algorithm	Mean Accuracy	Standard Deviation
Decision Tree	70.29	3.41
Random Forest	64.83	4.83
KNN	72.31	2.77
Logistic Regression with PCA	57.58	2.82
Logistic Regression without PCA	61.16	3.02

The best model is the decision tree with a mean accuracy of 70.29%.

We can conclude that the three most important features that affect the severity of an automotive accident are: Casualty Class_Pedestrian, Road Surface_Dry, Road Surface_Wet or Damp.