Analyzing Severity of Accidents

30.08.2020

Subhadra Jamkar

1.1 INTRODUCTION

1.1 Background

According to the World Health Organisation (WHO), road traffic crashes result in the deaths of approximately 1.35 million people around the world each year and leave between 20 and 50 million people with non-fatal injuries. More than half of all road traffic deaths and injuries involve vulnerable road users, such as pedestrians, cyclists and motorcyclists and their passengers.

The young are particularly vulnerable to the world's roads and road traffic injuries are one of the leading causes of death for children and young adults aged 5-29. Young males under 25 years are more likely to be involved in road traffic crashes than females, with 73% of all road traffic deaths occurring among young males in that age. So these were the facts. But my major interest is going to be knowing and accessing the factors that impact accident severity.

1.2 Problem and Interest

In this project, I am going to study some of the **parameters that affect the severity of accidents.** By doing so, I will be able to highlight the major conditions in which most of the accidents occur.

Data from official websites will be used, like in my case Accident data from the UK government's website will be analyzed. This will, in turn, help the concerned authorities take effective steps to prevent as many numbers of accidents as possible. On a hopeful note, the findings from this data analysis may even save several lives. The purpose of this project is to highlight impactful variables while operating a vehicle to improve accident prevention. **Using machine learning algorithms we can create models that can point out the impacting variables which are directly related to the severity of the accidents.**

This is the first time I am trying my hands at such a project, so the report will be quite simple and straightforward with not many complicated terms. Hope it works.

2. DATA

I collected the dataset from the UK government's official website; the government authorities amassed the traffic information based on numerous reports from the police.
The dataset can be accessed by clicking the following link-
https://data.gov.uk/dataset/6efe5505-941f-45bf-b576-4c1e09b579a1/road-traffic-accidents

For the sake of simplicity and convenience, I rather chose to study the data for the years 2014 to 2016. Also, this being my first project of the kind, I preferred to go for a less complicated data set.
The data of the accidents have been recorded based on these following fields:-

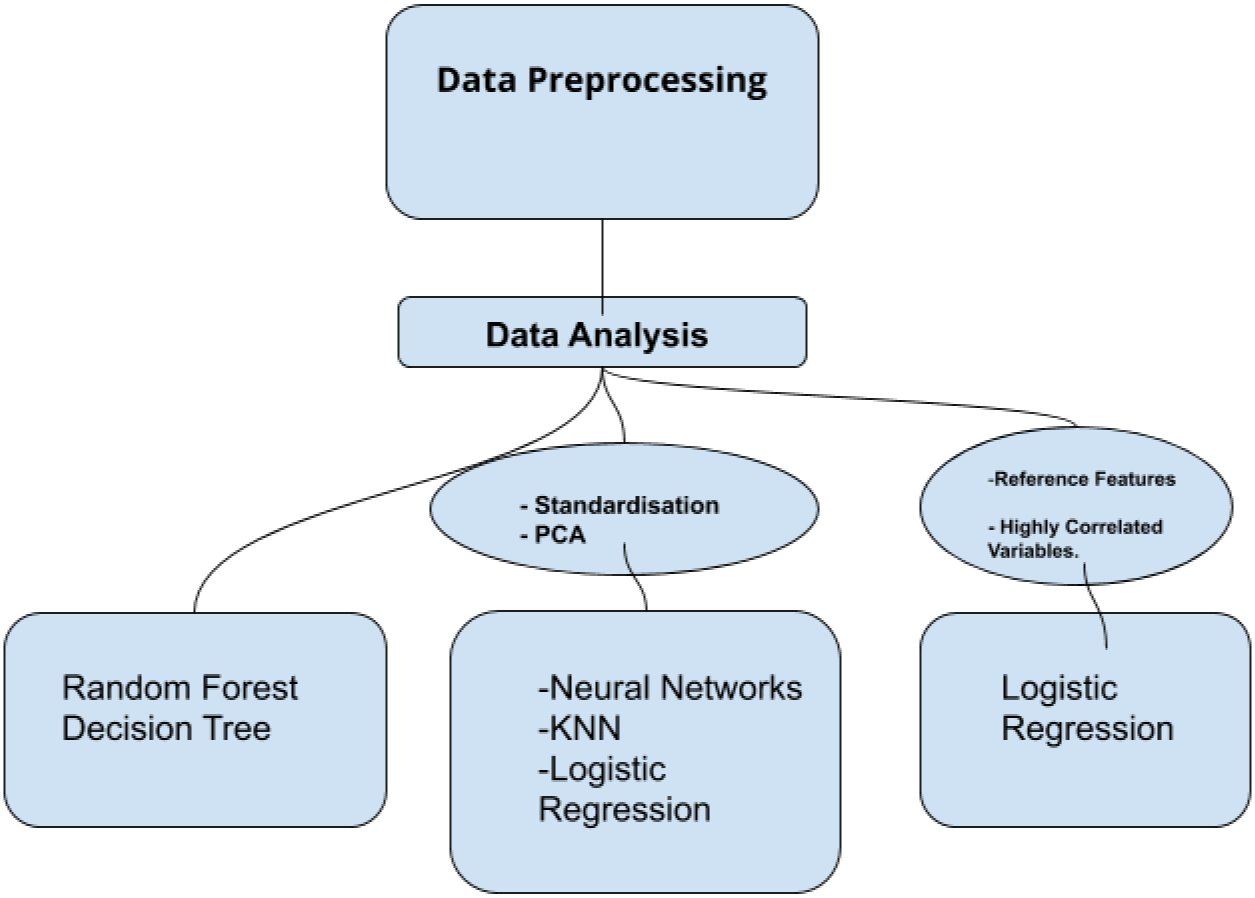
- Reference Number
- Grid Ref: Easting
- Grid Ref: Northing
- Expr1
- Severity
- Day of the week
- Time (24hr)
- 1st Road Class
- Road surface

- Accident date
- Weather condition
- Lighting conditions
- Number of vehicles
- Casualty class
- Sex of casualty
- Age of casualty
- Type of vehicle

I will try to scrape out the necessary parameters like the road conditions, casualty type, etc, from this data to put up a proper understanding of the accident severity. More facts regarding the handling of the dataset would be dealt with in the Methodology section.

3. Methodology

Flow chart for the approach-



I. Data Preprocessing

Before diving into analysis, I had to make a few changes in the acquired dataset.

- **Merging datasets** - The data for three different years i.e. 2014-16 was in three different files. So I had to merge them for better handling.
- **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 labelled 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** (it's better to have certain data in int or float datatype) 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

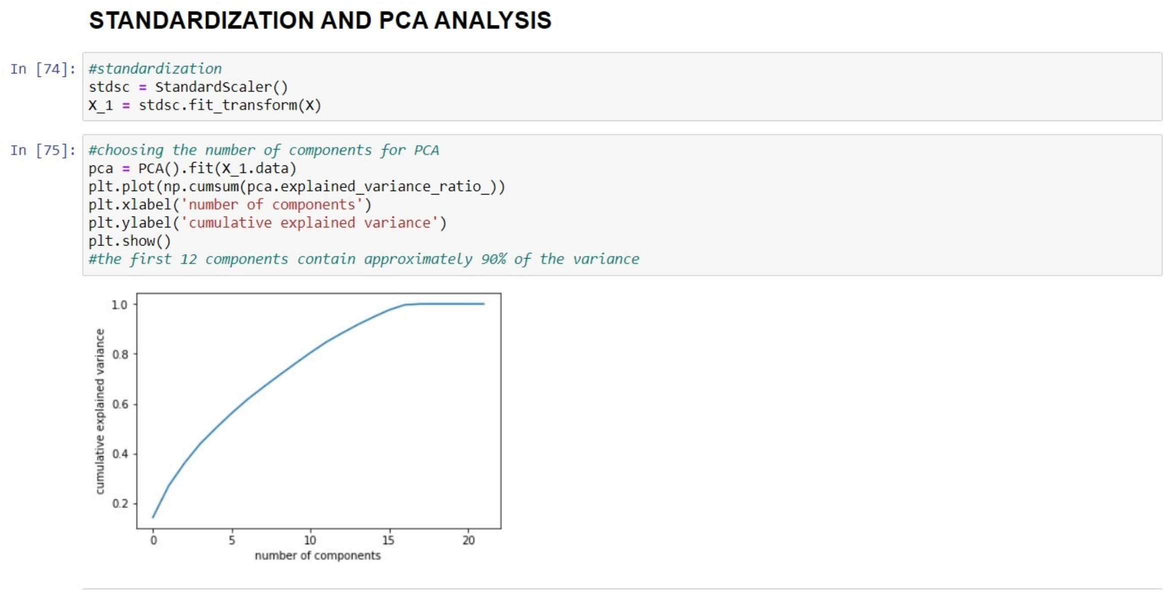
Undersampling from slight to serious- Slight: 957, Serious: 957, Fatal: 48

Oversampling from fatal to serious- Slight: 957, Serious: 957, Fatal: 957

II. Data Analysis

- **Standardization**
- **PCA** -Principal Component Analysis (PCA) is an unsupervised, non-parametric statistical technique primarily used for dimensionality reduction in machine learning.

From the graph shown below, it can be understood that one must utilize the first 12 components, as they approximately make up 90% of the variance in the data.



STANDARDIZATION AND PCA ANALYSIS

```
In [74]: #standardization
         stdsc = StandardScaler()
         X_1 = stdsc.fit_transform(X)
```

```
In [75]: #choosing the number of components for PCA
         pca = PCA().fit(X_1.data)
         plt.plot(np.cumsum(pca.explained_variance_ratio_))
         plt.xlabel('number of components')
         plt.ylabel('cumulative explained variance')
         plt.show()
         #the first 12 components contain approximately 90% of the variance
```

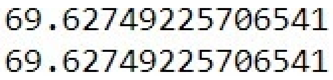
- **Prediction**

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

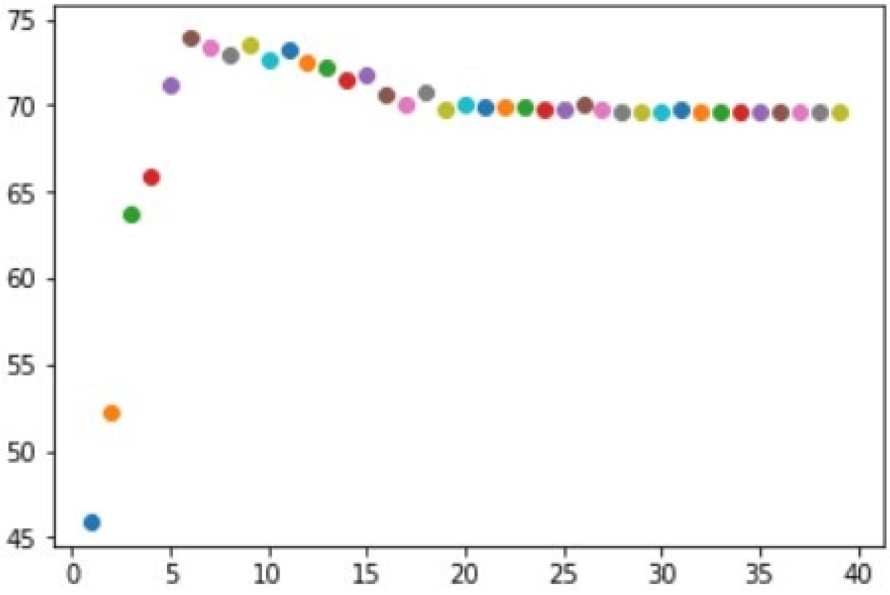
b. To avoid overfitting, we used **K-fold cross-validation method** with ten splits.

⟶ **DECISION TREE** - Upon running the decision tree model, several points became clear.

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



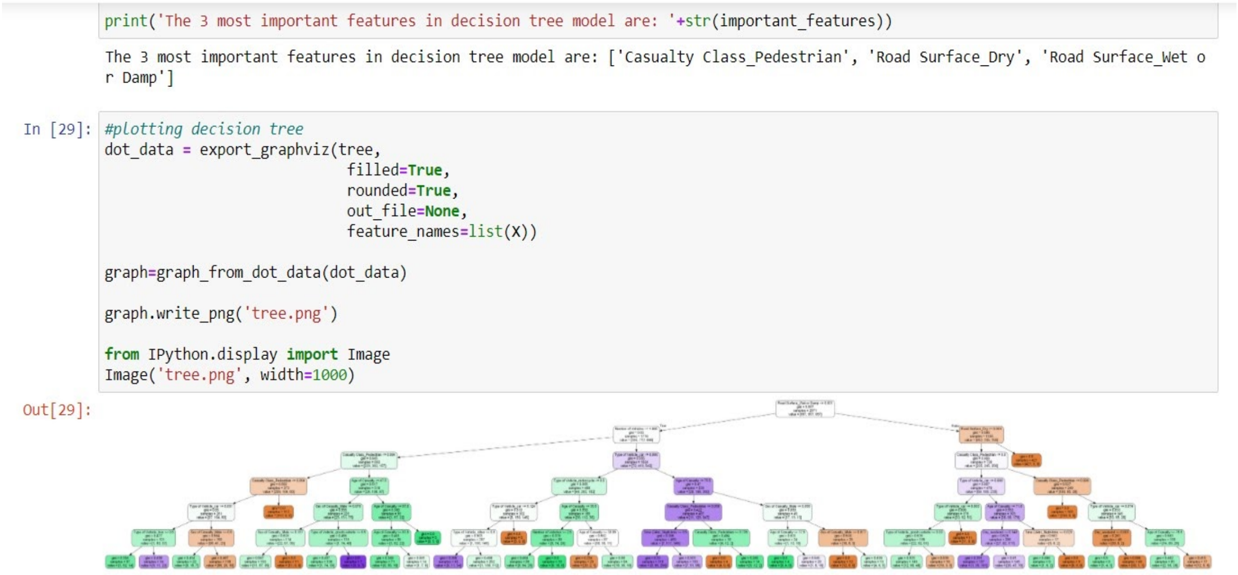
```
69.62749225706541
69.62749225706541
```



From the graph, we get to know the depth of the decision tree, which must be 6.

K-fold best mean accuracy is 73.95% (standard deviation 2.64%) for a decision tree depth equal to six.

Let us take a look at how the decision tree will look-



```
print('The 3 most important features in decision tree model are: '+str(important_features))
```

```
The 3 most important features in decision tree model are: ['Casualty Class_Pedestrian', 'Road Surface_Dry', 'Road Surface_Wet o
r Damp']
```

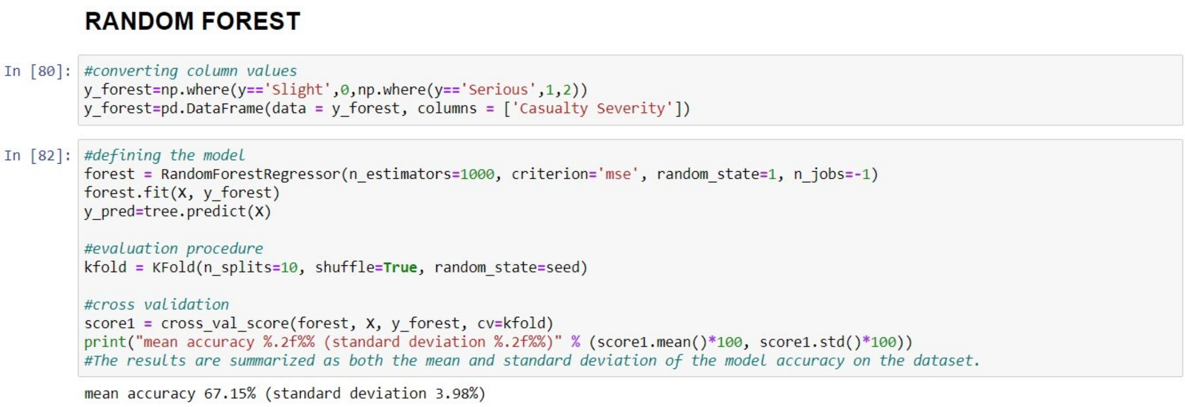
```
In [29]: #plotting decision tree
         dot_data = export_graphviz(tree,
                                    filled=True,
                                    rounded=True,
                                    out_file=None,
                                    feature_names=list(X))

         graph=graph_from_dot_data(dot_data)

         graph.write_png('tree.png')

         from IPython.display import Image
         Image('tree.png', width=1000)
```

```
Out[29]:
```

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

→ RANDOM FOREST -

Here is the algorithm-



RANDOM FOREST

```
In [80]: #converting column values
         y_forest=np.where(y=='Slight',0,np.where(y=='Serious',1,2))
         y_forest=pd.DataFrame(data = y_forest, columns = ['Casualty Severity'])
```

```
In [82]: #defining the model
         forest = RandomForestRegressor(n_estimators=1000, criterion='mse', random_state=1, n_jobs=-1)
         forest.fit(X, y_forest)
         y_pred=tree.predict(X)

         #evaluation procedure
         kfold = KFold(n_splits=10, shuffle=True, random_state=seed)

         #cross validation
         score1 = cross_val_score(forest, X, y_forest, cv=kfold)
         print("mean accuracy %.2f%% (standard deviation %.2f%%)" % (score1.mean()*100, score1.std()*100))
         #The results are summarized as both the mean and standard deviation of the model accuracy on the dataset.
```

```
mean accuracy 67.15% (standard deviation 3.98%)
```

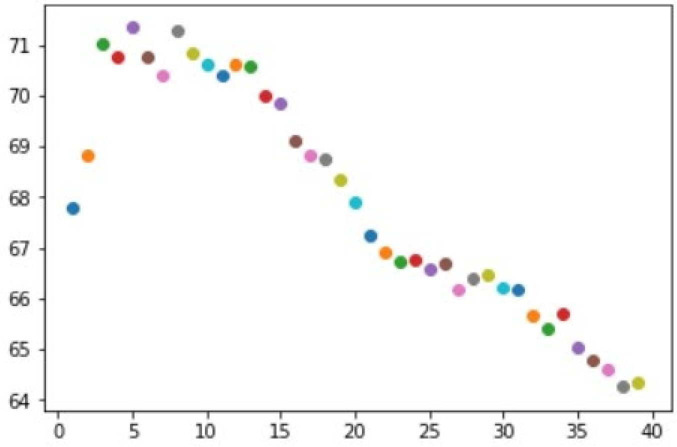
The outcome using this model was that the **mean accuracy equal to 67.15% (with a standard deviation of 3.98%).**

→ NEURAL NETWORK -

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

→ KNN -

- Used the preprocessed standardized data then applied the PCA technique.
- The following graph shows the number of neighbours that returns the best accuracy based on the number of features that we have in the dataset:-



- **The number of nearest neighbours is 5.**
- **K-fold best mean accuracy was 71.37% (with a standard deviation of 2.77%)** for the number of neighbours equal to five.

→ LOGISTIC REGRESSION

- Applied this model/algorithm using the PCA technique.
- **The mean accuracy was equal to 53.12% (standard deviation 2.32%)**

- The following correlation matrix was obtained-

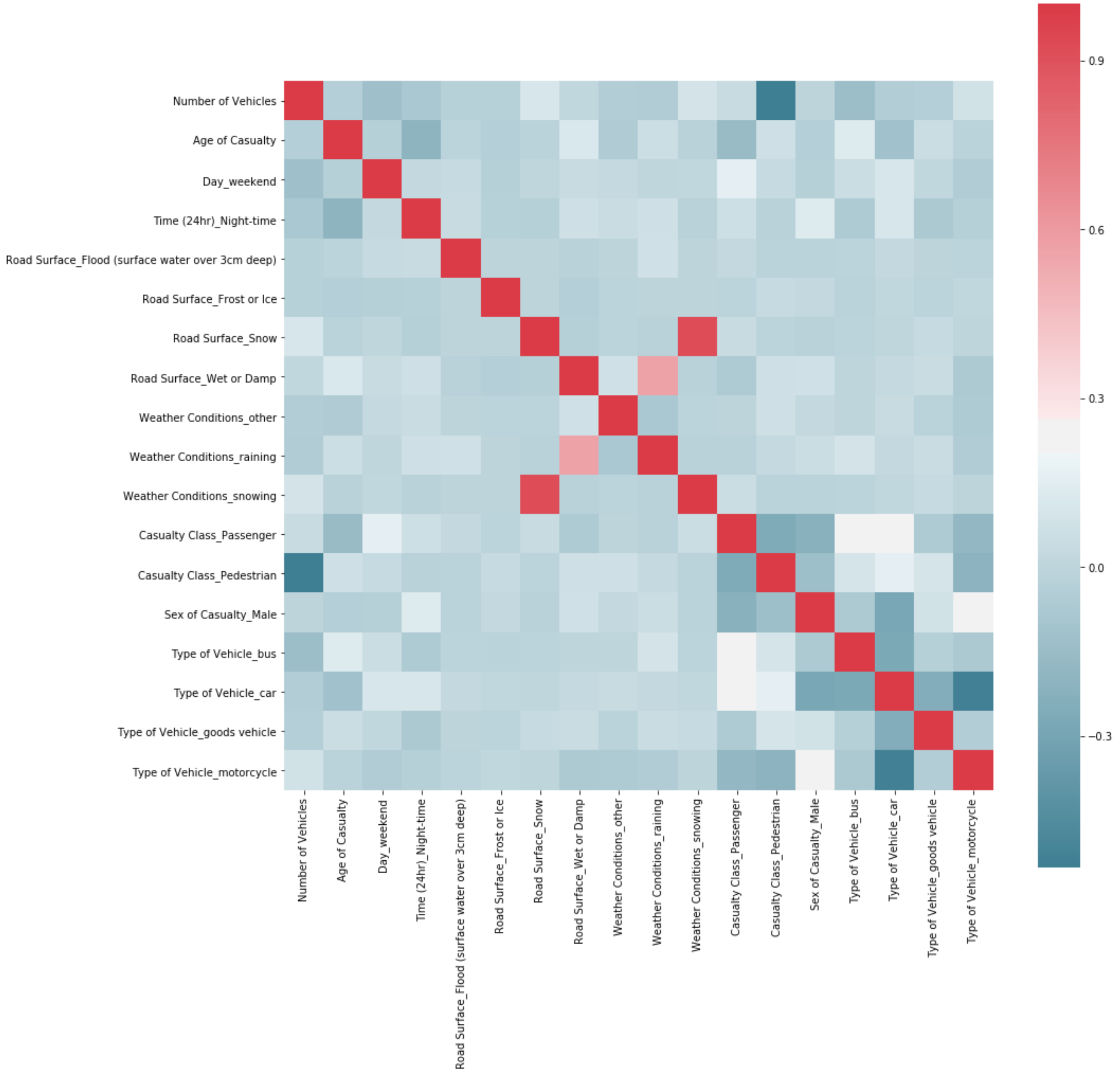


FIG: Correlation Matrix 1

- Then applied the logical regression without PCA.
- Certain changes had to be done to the dataset like dropping reference variables and Dropping 'Weather Condition' variable due to its high correlation with 'Road Surface'.
- **The mean accuracy was equal to 54.09% (with a standard deviation of 3.03%).**
- The following correlation matrix was obtained-

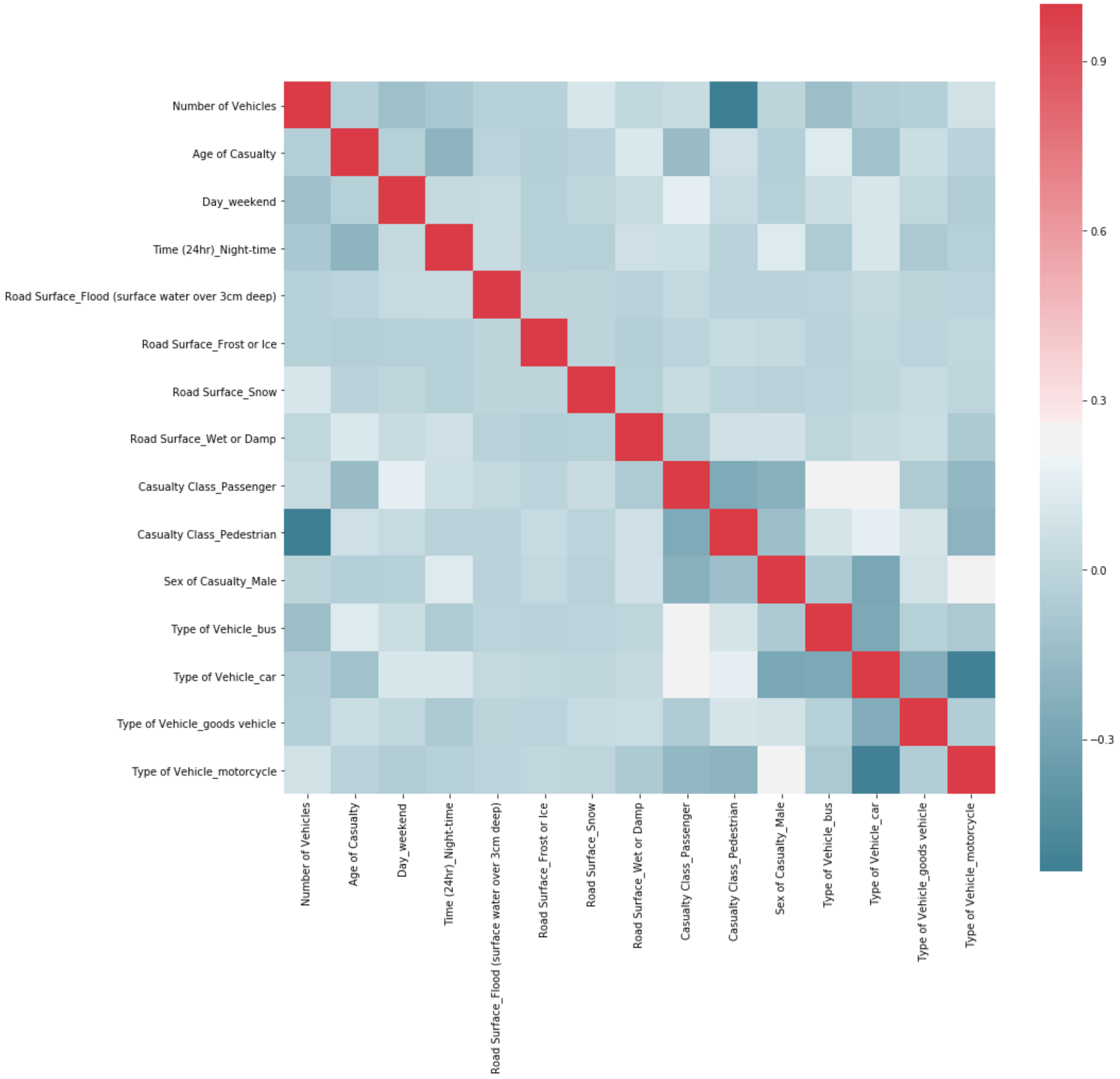
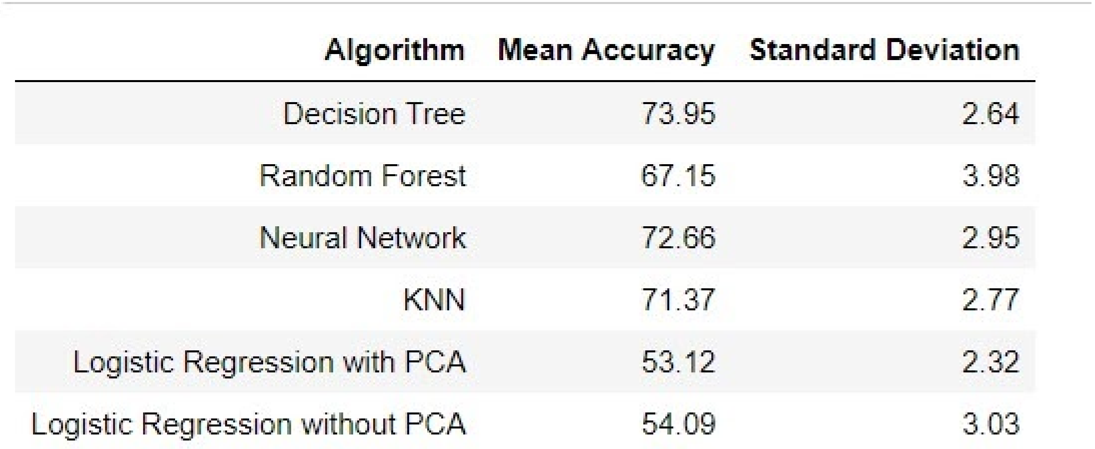


FIG: Correlation matrix 2

4. Result

After summarizing the accuracies from all the models and putting them into a tabular form like the following-

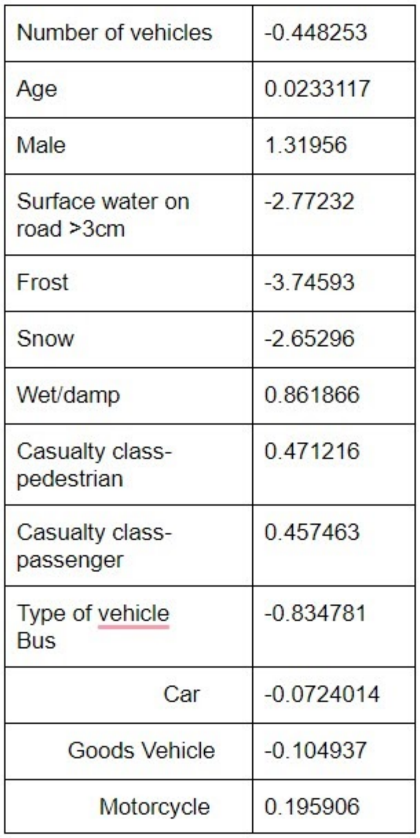


Algorithm	Mean Accuracy	Standard Deviation
Decision Tree	73.95	2.64
Random Forest	67.15	3.98
Neural Network	72.66	2.95
KNN	71.37	2.77
Logistic Regression with PCA	53.12	2.32
Logistic Regression without PCA	54.09	3.03

The best model predicting the severity is DECISION TREE with an accuracy of 73.95%.

- **Let us look into the logistic regression coefficients for various parameters.**
    - Let's say that the probability of success of some event is 0.8. Then the probability of failure is 1 – 0.8 =0 .2. **The odds of success are defined as the ratio of the probability of success over the probability of failure**. In our example, the odds of success are .8/.2 = 4. That is to say that the odds of success are 4 to 1. If the probability of success is .5, i.e., 50-50 per cent chance, then the odds of success is 1 to 1.
    - LR coefficients are the **log of odds of success of events.** In the above example, it will be log(4). Greater the probability, greater the odds of success and thereby greater log odds (LRs).
    - And to get the odds of success from LR coefficients, do their antilog.
    - **Lesser the LR, lesser are the odds of success, lesser the probability of that event taking place**

Here are the LR coefficients for various variables-



Number of vehicles	-0.448253
Age	0.0233117
Male	1.31956
Surface water on road >3cm	-2.77232
Frost	-3.74593
Snow	-2.65296
Wet/damp	0.861866
Casualty class-pedestrian	0.471216
Casualty class-passenger	0.457463
Type of vehicle Bus	-0.834781
Car	-0.0724014
Goods Vehicle	-0.104937
Motorcycle	0.195906

5. Discussion

The LR coefficients for Night-time is more, meaning the severity of some accident will be more at night. This even sounds logical. Same is for the road surface when it is damp or wet, the severity rises; the chances of vehicles skidding on the road increase. Pedestrians become victims of accidents and the severity is also high as seen from the LR coefficients, slightly higher than the passengers in the vehicle.

We cannot comment on other factors, like age, type of vehicle, whether a male or a female was involved, the number of vehicles involved, as they are highly correlated and vary from situation to situation.

6. Conclusion

We can conclude that some of the important features affecting the severity of an accident are-

⟶ If the road is wet or damp

⟶ If any pedestrians are involved in the accident

⟶ Whether the accident took place at night

This model explains that traffic authorities should increase pedestrian safety measures on roads. And if the road is too slippery, it shouldn't be open to traffic. Drivers must take extra precautions while driving during night hours.

Link to my Blogpost