**This project is to predict the severity code of an accident in the city of Seattle.**

**Business Understanding**

The Seattle government is going to prevent avoidable car accidents by employing methods that alert drivers, health system, and police to remind them to be more careful in critical situations.

In most cases, not paying enough attention during driving, abusing drugs and alcohol or driving at very high speed are the main causes of occurring accidents that can be prevented by enacting harsher regulations. Besides the aforementioned reasons, weather, visibility, or road conditions are the major uncontrollable factors that can be prevented by revealing hidden patterns in the data and announcing warning to the local government, police and drivers on the targeted roads.

The target audience of the project is local Seattle government, police, rescue groups, and last but not least, car insurance institutes. The model and its results are going to provide some advice for the target audience to make insightful decisions for reducing the number of accidents and injuries for the city.

# Data

The data was collected by the Seattle Police Department and Accident Traffic Records Department from 2004 to May 2020.

The data consists of 37 independent variables and 194,673 rows. The dependent variable, “SEVERITYCODE”, contains numbers that correspond to different levels of severity caused by an accident from 0 to 4.

Severity codes are as follows:

0: Little to no Probability (Clear Conditions)

1: Very Low Probability — Chance or Property Damage

2: Low Probability — Chance of Injury

3: Mild Probability — Chance of Serious Injury

4: High Probability — Chance of Fatality

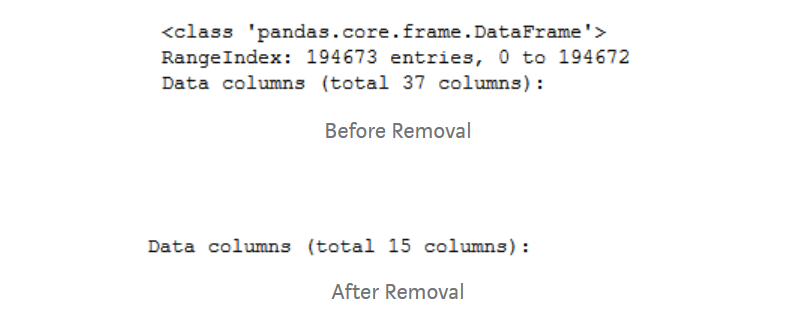
Furthermore, because of the existence of null values in some records, the data needs to be preprocessed before any further processing.

# Data Preprocessing

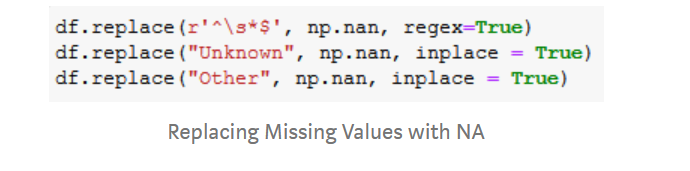
The dataset in the original form is not ready for data analysis. In order to prepare the data, first, we need to drop the non-relevant columns. In addition, most of the features are of object data types that need to be converted into numerical data types.

After analyzing the data set, I have decided to focus on only four features, severity, weather conditions, road conditions, and light conditions, among others to reduce the complexity and dimensionality of the data set

After performing this step, the dimensionality dropped from 37 to 15.

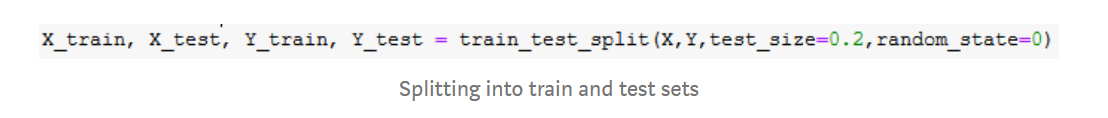


Empty boxes, ‘Unknown’ and ‘Other’ were values considered as missing values. These were replaced with NA to make the dataset uniform.



The dataset is split into two datasets, X and Y, where Y contains the target feature (SEVERITYCODE) and X contains all the independent features/variables.

The datasets X and Y are split into X\_train, Y\_train, X\_test,and Y\_test. The first two will be used for training purposes and the last two will be used for testing purposes. The split ratio is 0.8, 80% of data is used for training and 20% of is used for testing.



Feature scaling of data is done to normalize the data in a dataset to a specific range. It also helps improve the performance of the ML algorithms. Standard Scaler metric is used to scale/normalize all the numerical data for both, the X\_train and X\_test datasets. This completes the pre-processing stage.

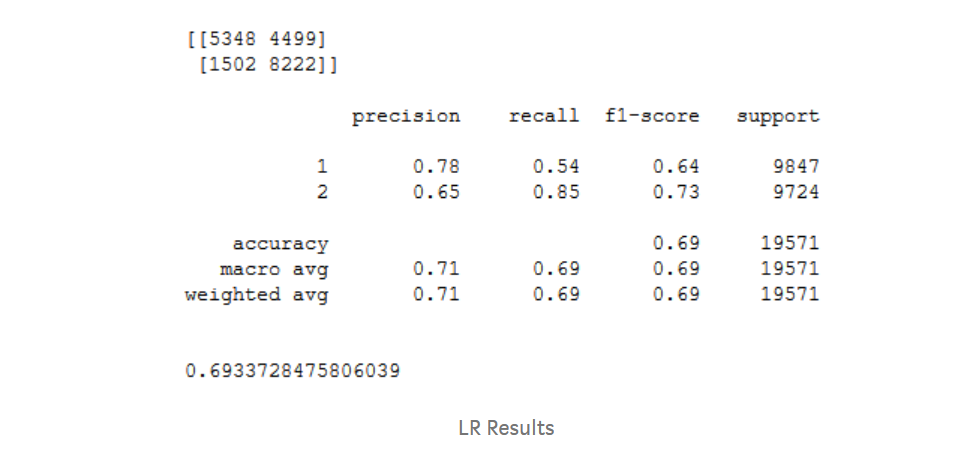
# Machine Learning Algorithms

A total of six ML algorithms were trained on the pre-processed dataset and their accuracies were compared. A brief explanation on how each of them works along with their results in shown below.

**1)Logistic Regression Classifier**

Logistic Regression is a classifier that estimates discrete values (binary values like 0/1, yes/no, true/false) based on a given set of an independent variables. It basically predicts the probability of occurrence of an event by fitting data to a logistic function. Hence it is also known as logistic regression. The values obtained would always lie within 0 and 1 since it predicts the probability.

The chosen dataset has only two target categories in terms of the accident severity code assigned; hence it was possible to apply this model to the same. The results, confusion matrix, classification report and accuracy, are: Image for post Image for post LR Results

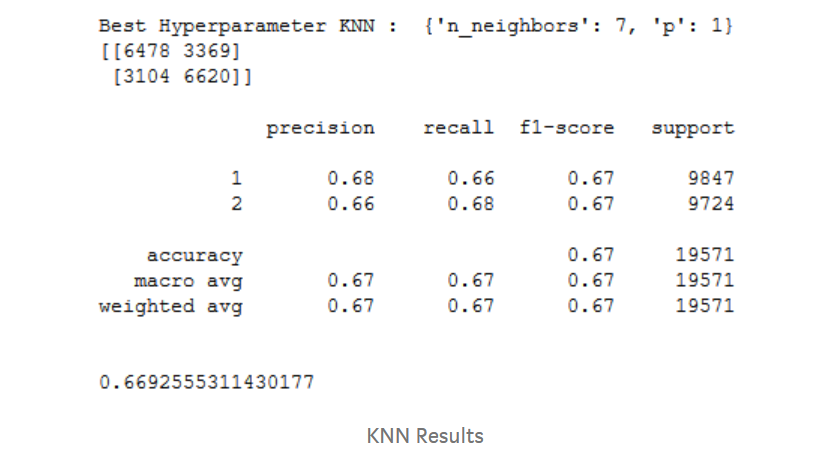


**2)K Nearest Neighbours Classifier**

K nearest neighbours algorithm used for both classification and regression problems. It basically stores all available cases to classify the new cases by a majority vote of its k neighbours. The case assigned to the class is most common amongst its K nearest neighbours measured by a distance function (Euclidean, Manhattan, Minkowski, and Hamming).

In order to arrive at the optimum values for nearest neighbours (k) and the distance metric (Euclidean and Manhattan), a hyper parameter KNN was used. The best accuracy was obtained for 7 nearest neighbours with Euclidean being the distance metric when applied for the problem in question.

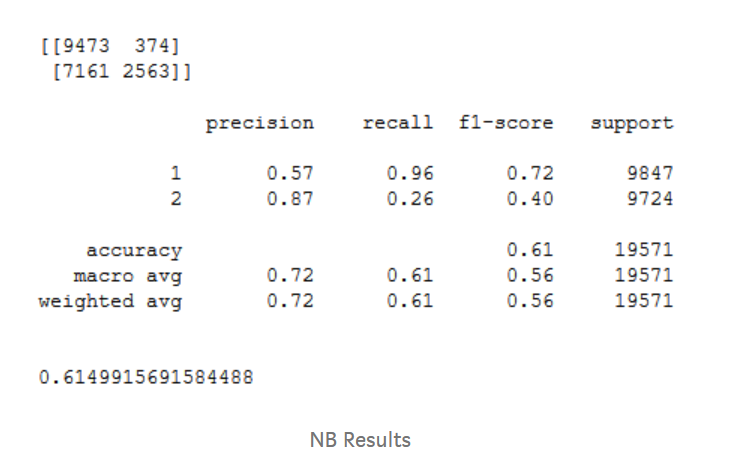
The results, confusion matrix, classification report and accuracy, are: Image for post Image for post KNN Results



**3)Naïve Bayes Classifier**

Naive Bayes classifies objects based on Bayes’ Theorem with an assumption that the predictors (features) are independent of each other. Bayes theorem is a way to calculate posterior probability P(c|x) from the P(c), P(x), P(x|c). Naive Bayes is naive because it assumes the presence of a particular feature is completely unrelated to the presence of another, and each of them contributes to the posterior probability independently.

The results, confusion matrix, classification report and accuracy, when Naïve Bayes was applied to the pre-processed accident severity dataset are: Image for post Image for post NB Results

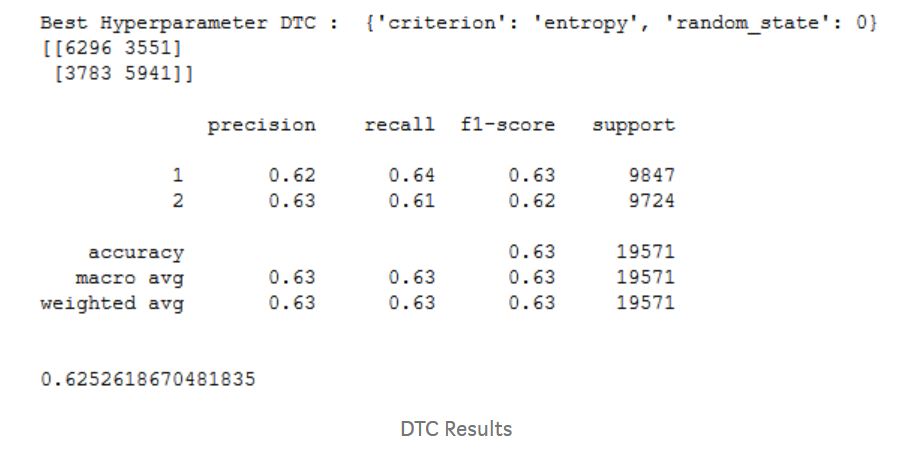


**4)Decision Tree Classifier**

Decision Tree makes decision with tree-like model. It splits the sample into two or more homogenous sets (leaves) based on the most significant differentiators in the input variables. To choose a differentiator (predictor), the algorithm considers all features and does a binary split on them (for categorical data, split by category; for continuous, pick a cut-off threshold). It will then choose the one with the least cost (i.e. highest accuracy), and repeats recursively, until it successfully splits the data in all leaves (or reaches the maximum depth).

Information gain for a decision tree classifier can be calculated either using the Gini Index measure or the Entropy measure, whichever gives a greater gain. A hyper parameter Decision Tree Classifier was used to decide which tree to use, DTC using entropy had greater information gain; hence it was used for this classification problem.

The results, confusion matrix, classification report and accuracy, are: Image for post Image for post DTC Results

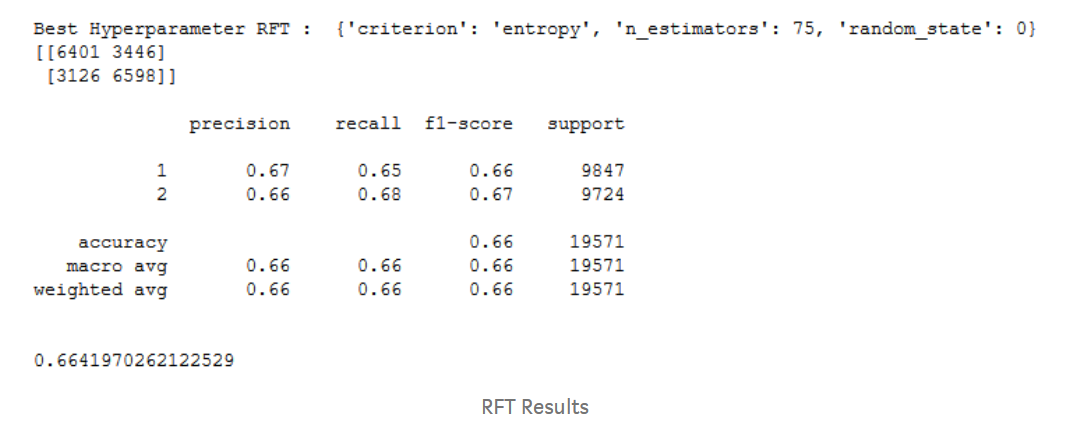


**5)Random Forest Tree Classifier**

Random Forest Classifier is an ensemble (algorithms which combines more than one algorithms of same or different kind for classifying objects) tree-based learning algorithm. RFC is a set of decision trees from randomly selected subset of training set. It aggregates the votes from different decision trees to decide the final class of the test object. Used for both classification and regression.

Similar to DTC, RFT requires an input that specifies a measure that is to be used for classification, along with that a value for the number of estimators (number of decision trees) is required. A hyper parameter RFT was used to determine the best choices for the above mentioned parameters. RFT with 75 DT’s using entropy as the measure gave the best accuracy when trained and tested on pre-processed accident severity dataset.

The results, confusion matrix, classification report and accuracy, are: Image for post Image for post RFT Results

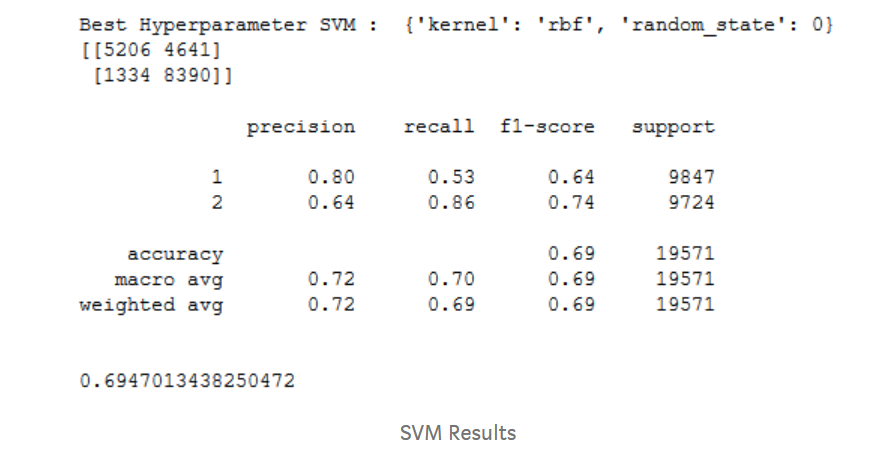


**6)Support Vector Machine Classifier**

Support Vector Machine is an algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, each data item is plotted as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, classification is performed by finding the hyper-plane that differentiates the two classes.

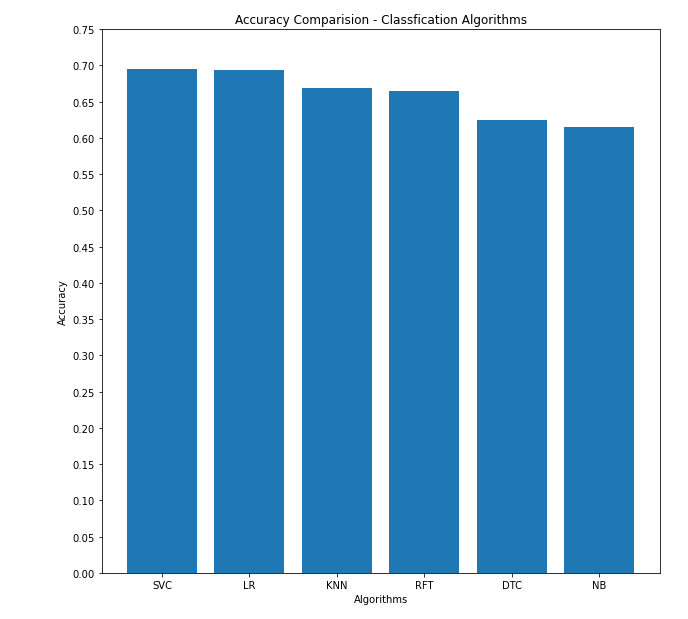
Hyper parameter SVC was used to choose between Linear SVC and a Kernel SVC and the latter arrived on top with a greater accuracy when applied on the dataset in question. It used the ‘radial basis function’ kernel for performing the classification.

The results, confusion matrix, classification report and accuracy, are: Image for post Image for post SVM Results



# Results

None of the algorithms implemented above gave an accuracy score equal to or greater than 0.7, they all ranged from 0.6 to 0.7. Meaning, these models can predict the severity code of an accident with an accuracy equalling 60–70%. A bar plot is plotted below with the bars representing the accuracy of each model in descending order respectively. Image for post



# Conclusion

The accuracy of the classifiers is not great, highest being 69%. This usually means that the model is under fitted i.e. it needs to be trained on more data. Though the dataset has a lot of variety in terms of scenarios, more volume of the data for such scenarios has to be collected.

Certain features with missing values were removed, this reduced the dimensionality of the dataset, these features could have been correlated to other important features but they had to be removed. A better effort has to be made to collect data to reduce the number of missing values.