**Driving Conditions and Accident Severity Machine Learning Data Analysis in Seattle**

**I. Introduction**

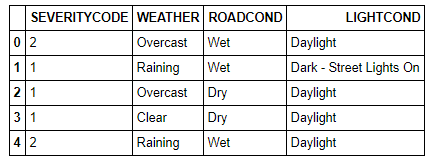
Seattle is a city that is infamously known for low sunlight and rough road conditions. There are about 637 vehicles per 1000 Seattle residents--a higher per-capita rate that Los Angeles. [1] This means that the majority of residents are commuting to work and the roads are often occupied with vehicles. According to 2017 WSDOT data, a car accident occurs every 4 minutes in Washington State and a person dies due to a car crash every 20 hours. [2] If we are given only certain road conditions, weather, and light conditions, can we predict the severity of an accident on any given day? The model created in this article aims to do this by using three different machine learning algorithms.

**II. Business Understanding**

Using Machine learning techniques, it may be possible to predict a severe accident event by creating a model that factors in data from past accidents. In this model, we are given uncontrollable accident factors and the result of the model will be the severity of the accident based on the different values of the factors. The main stakeholders in this project are emergency first responders and local police because of their potential increased ability to efficiently assign a certain number of personnel at a time when there may be a large number of severe accidents. Also, they can warn drivers of a high probability of a severe accident in attempt to avoid them. Another large stakeholder would be insurance companies because they would be able to predict severe accident events in the future and use this data to manipulate insurance rates. The main goal of this model is to prevent severe accidents from happening by controlling the factors that may contribute to them.

**III. Data**

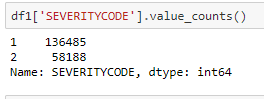
The data used in this model is from collisions provided by SPD and recorded by Traffic Records in Seattle from 2004 to present. The data contains past accidents and records each accident on a scale of 0 to 3 based on the severity of the injury. After inspecting the contents in the SEVERITYCODE column, it is apparent that there are only severity codes 1 and 2 present in the data. Code 1 refers to property damage and Code 2 refers to an injury. In order to focus on the main factors that contribute to severe accidents, we can narrow down the columns to only vital information that we can then manipulate. The independent variables in this model are the weather, road conditions, and the light conditions. The dependent variable is the severity code (1 or 2). The expected outcome of the model is to predict a code 2 severity when given certain conditions. The table below shows the first five lines of the dataset after unnecessary or redundant columns were removed.



**IV. Methodology**

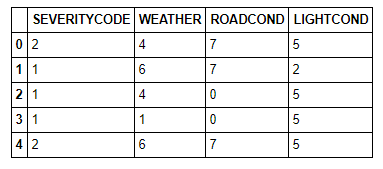
**Pre-Processing**

The first step in creating a predictive data model is to pre-process the data and shape it so that it can be fed into the model and create accurate results. The figure below shows that there are significantly more Severity 1 entries in the dataset than Severity 2 entries. Having an unbalanced dataset can skew the results of the model by creating a bias toward one outcome. Balance can be created by downsampling the number of entries with Code 1 values.



Next all rows that have missing or unknown values in them are removed so that the only data remaining is complete data. This can be done by marking unknown data as null values and removing the row using the dropna function.

Since machine learning models can only analyze numerical values and not categorical values, all of the values in the independent variable columns must be converted to numerical values using Label Encoders. Below is the new table with numerical values corresponding to each category for every independent variable.



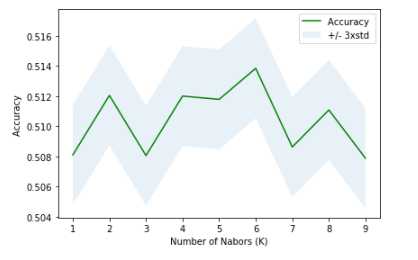
Now that the data has been preprocessed, it can be fed into three different supervised learning algorithms.

**K-Nearest Neighbours Algorithm**

KNN is one of the simplest algorithms used for regression and classification problems that uses data to classify new data points based on similarity measures. Classification is done by a majority vote to its neighbors and assigns the data to the class which has the nearest neighbors.

Using the sklearn library, the dataset can be split into a training set and a test set. The training set in this case is 80% of the data and the test set is made up of the remaining 20%. The training set is fed into the model and the predicted values from the model are then compared to the test values.

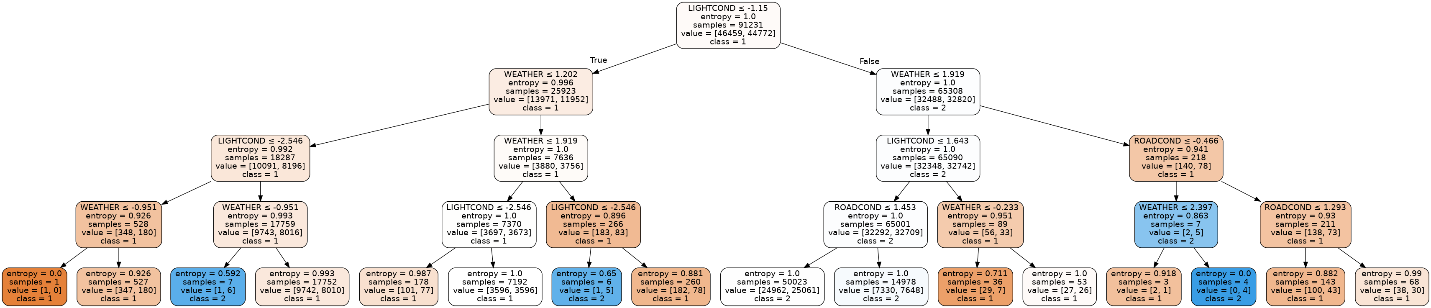
The k value in a KNN algorithm is the number of nearest neighbors. Below is a graph that describes the different accuracies for different k values. A k value of 6 produces the highest accuracy.



**Decision Tree Algorithm**

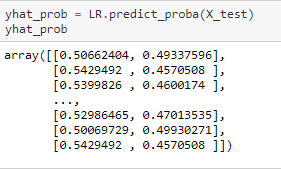
A Decision Tree algorithm is also a supervised learning algorithm that can be used for regression and classification problems. Decision Trees work by creating a training model to predict class or value of target variables by learning decision rules inferred from prior data.

Using the previous training and test sets, the data is fed into the Decision Tree algorithm and transformed in to a prediction tree. The prediction tree can be used to compare predicted values to the test set values. Entropy is the measure of randomness or uncertainty of a random variable. By calculating the entropy measure of each attribute, we can find the information gain at each node. Information gain calculates the expected reduction in entropy due to sorting on the attribute. Below is the full Decision Tree with entropy values at each node.



**Logistic Regression Algorithm**

Logistic Regression is the most appropriate algorithm to use for this problem because it is mainly used when the dependent variable is binary. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and multiple independent variables. [3]

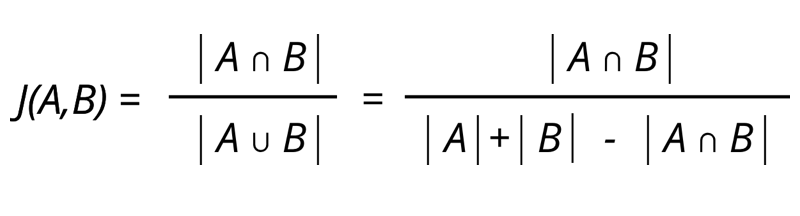


As shown in the figure above, the training and test sets are fed into the Logistic Regression model and an array is created to describe the probability of predicting the test set value.

**V. Results**

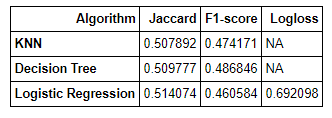
In order to evaluate the results of the different model, a Jaccard, F1-score, and Logloss analysis are done to test for accuracy.

Jaccard Similarity measures the similarities between sets and is defined as the size of the intersection divided by the size of the union of two sets. It is computed using the following formula:



F1-score is the measure of a model's accuracy on a dataset. It is used to evaluate binary classification models such as the one given in this problem and is defined as the harmonic mean of the model's precision and recall.

Lastly, Logloss is used only to evaluate Logistical Regression algorithms and is also referred to as binary cross-entropy loss. It is defined as the negative log-likelihood of a logistic model that returns true predicted values for its training data.



The table above shows the Jaccard, F1-score, and Logloss values for the three different algorithms.

**VI. Discussion**

The three algorithms used to fit the dataset recorded similar values for Jaccard and F1-score. This means that all three algorithms were equally accurate in predicting test results and can all be used to evaluate binary problems. The closer the Jaccard score is to 1, the more similar the two sets of data are. Similarly, an F1-score close to 1 indicates high precision and recall. The fact that the Jaccard value is around 50 illustrates that the independent and dependent variables share some similarities but are not significantly similar. Also, having an F1-score of around 48 means that the results are somewhat precise but not significantly precise.

From this data we can conclude that road conditions, light conditions, and weather all have a relation to severe accidents but cannot be used to accurately predict a severe accident event.

**VII. Conclusion**

Although the driving conditions that were analyzed in this model had a relationship to severe accidents, they can not be used as the only factors that contribute to a severe accident. This means that larger factors must be accounted for in the model to create more accurate results. Future machine learning models should look to use more than these three driving conditions to predict a severe accident.

**VIII. References**

* [1] [Curbed Seattle](https://seattle.curbed.com/2017/9/21/16346824/seattle-commute-data-bus-drive)
* [2] [Washington State Car Accident Rate](https://www.injurytriallawyer.com/library/car-accident-statistics-seattle-washington-state.cfm#:~:text=Fatal%20Car%20Accidents%20In%20Washington,the%20death%20of%20555%20people.)
* [3] [Logistic Regression](https://www.statisticssolutions.com/what-is-logistic-regression/)