Predicting Accident Severity analysis

1. **Introduction**

**1.1 Background**

In most cases major uncontrollable factors for accidents are weather, light and road conditions. Accidents can be prevented by revealing hidden patterns in the data and giving this invaluable data as a warning to the local government, traffic police and the respective drivers of the vehicles on the identified roads.  Also this data can also be shared with Road Design teams who would ensure that future designs are made considering these aspects in the design & construction phase itself.

**1.2 Problem Statement**

Data that contributes to determining severity of accident include addresstype, junctiontype, weather condition, Road condition and Light condition. This project aims to predict Severity of accident based on these metrics.

**1.3 Interest**

The target audience of the project is local government, police, rescue groups, road design teams, car insurance companies and last but not the least the drivers themselves. The model and its results are going to provide some key insights for the target audience to make impactful decisions in reducing the number of accidents and injuries in their localities.

* **Data acquisition and cleaning**

**2.1 Data source**

Data provided by the Seattle Department of Transportation (SDOT) on vehicle accidents along with its severity is used to derive insights and patterns on how and when these accidents have taken place with the environmental factors like weather, road conditions etc. The dataset consists of 40 columns  having different kinds of data like, collision severity, road conditions, number of people involved, location of collision, weather etc.

Update Frequency: This dataset is updated weekly.

Attribute information:  
<https://www.coursera.org/learn/applied-data-science-capstone/supplement/Nh5uS/downloading-example-dataset>

Meta-data of the dataset can be viewed

**2.2 Data cleaning**

Dropped Irrelevant columns based on problem statement and data understanding

**Handling Null/Missing values:** There are some null values for ADDRTYPE, JUNCTIONTYPE, WEATHER, ROADCOND and LIGHTCOND attributes which are replaced with value ‘others’.

**2.3 Feature selection**

Based on the problem statement, our predictor or target variable will be ‘SEVERITYCODE’ because it is used to measure the severity of an accident from 0 to 5 within the dataset. Attributes used to weigh the severity of an accident are ‘ADDRTYPE’, JUNCTIONTYPE’, ‘WEATHER’, ‘ROADCOND’ and ‘LIGHTCOND’. And for exploratory data analysis I wanted to choose some of numerical variables such as (PERSONCOUNT, VEHCOUNT, PEDCOUNT, PEDCYLCOUNT)

 Therefore considering below relevant features by dropping others.

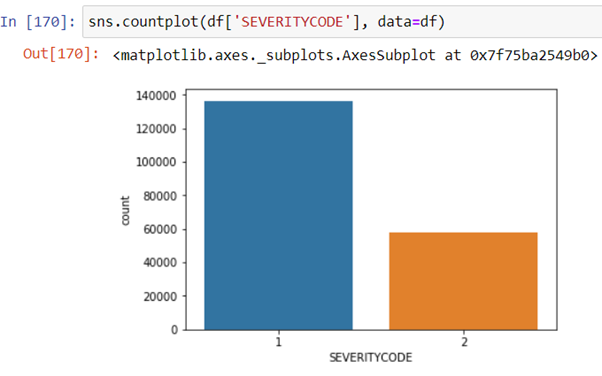
|  |  |
| --- | --- |
| Kept features | Dropped features |
| ‘SEVERITYVODE’,‘ADDRTYPE’, JUNCTIONTYPE’,‘WEATHER’, ‘ROADCOND’,‘LIGHTCOND’ ‘PERSONCOUNT’,‘PEDCOUNT’ ‘PEDCYLCOUNT’,’VEHCOUNT’ | ‘X’, ‘Y’, ‘COLDETKEY’, ‘REPORTNO’, ‘INTKEY’, ‘LOCATION’, ‘EXCEPTRSNCODE’,       ‘EXCEPTRSNDESC’, ‘SEVERITYCODE.1’, ‘SEVERITYDESC’,        ‘INCDATE’,’INCDTTM’, ‘, ‘SDOT\_COLCODE’, ‘SDOT\_COLDESC’,     ‘INATTENTIONIND’, ‘UNDERINFL’, ‘PEDROWNOTGRNT’, SDOTCOLNUM’, ‘SPEEDING’, ‘ST\_COLCODE’, ‘ST\_COLDESC’,’SEGLANEKEY’, ‘CROSSWALKKEY’, ‘HITPARKEDCAR’, ‘PERSONCOUNT’, ‘PEDCOUNT’, ‘PEDCYLCOUNT’, ‘VEHCOUNT’, ‘OBJECTID’, ‘COLLISIONTYPE’,’STATUS’ |

**Table1. Simple feature selection during data cleaning**

**3. Exploratory Data Analysis**

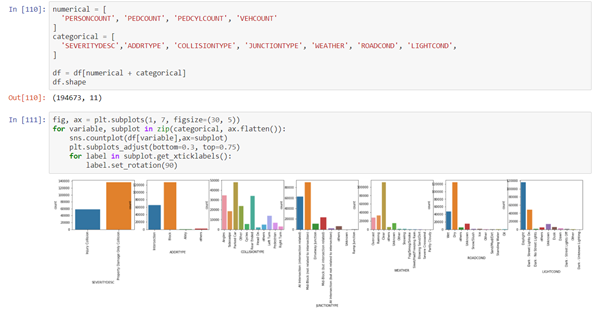
**3.1 Understanding of Target variable**

Verified count of accidents by SeverityCode using bar chart, based on the given data set, there are only two types of accidents listed that are severity 1(Prop damage) and severity 2(Injury).



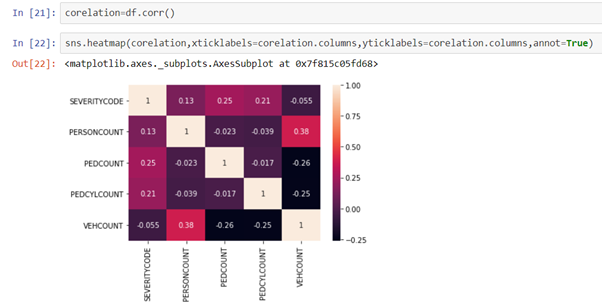
**3.2 Categorical Variables analysis**

Based is the categorical variable analysis, based on which we can see that more accidents are reported with Collisiontype-parked cars, Junctiontype-Mildblock (not related to intersection),Weather-Clear,RoadCondition-Dry,LihgtCondition-Daylight,Addresstype-Block



**3.3 Corelation Matrix of numerical variables**

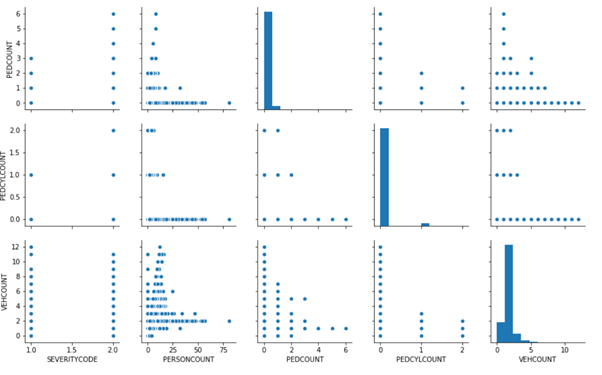
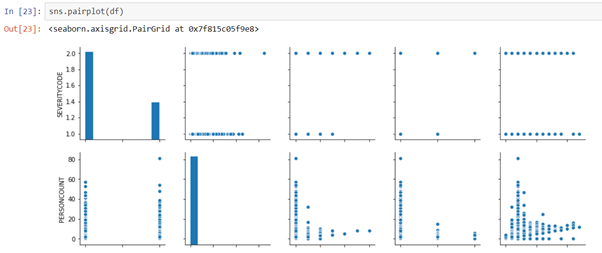
With the below correlation matrix, on numerical variables we can say that there is no much correlation between any of these variables



**3.3 PairPlot on numerical variables (VEHCOUNT, PERSONCOUNT, PEDCOUNT, PEDCYLCOUNT)**

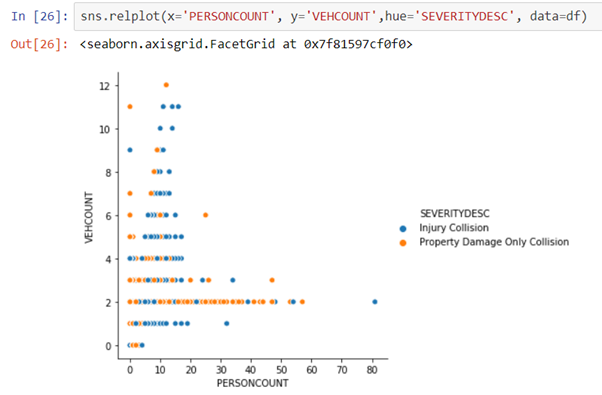
Based on below Pairplot analysis, we can say that there might be strong relationship between personcount,pedcount and personcount,vehcount but other relationships make no sense.

But with all observations of all pairplots, we can see extreme values/outliers, which can be removed



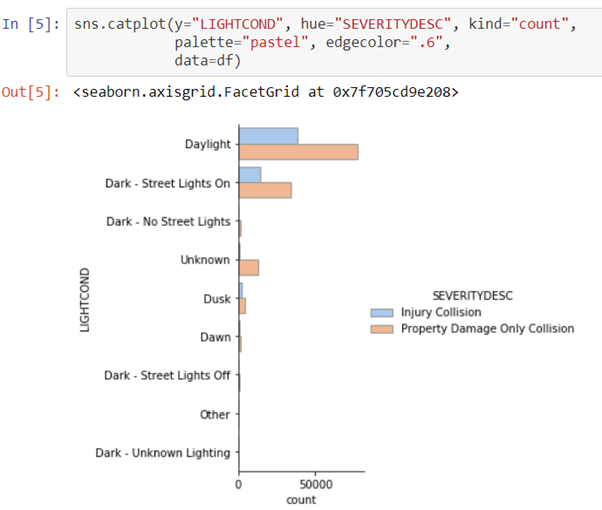
**3.4 Relationship between VEHCOUNT, PERSONCOUNT based on SEVERITY**

Based on below relationship plot, we can say that personcount and vehiclecount has strong relation



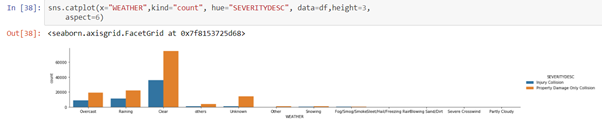
**3.5 Relationship between LIGHTCOND by SEVERITY of Accidents**

Based on below categorical plot between LIghtcondition by Severity of accidents, we can say that more number of accidents related to property damage are reported during Daylight



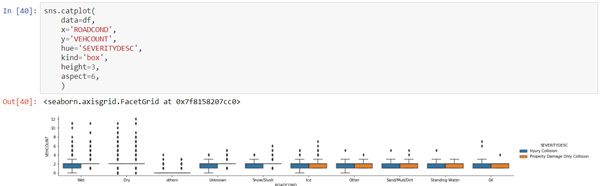
**3.6 Relationship between WEATHER by SEVERITY of accidents**

Based on below categorical plot between Weather by Severity of accidents, we can say that more number of accidents related to property damage are reported during Clear weather



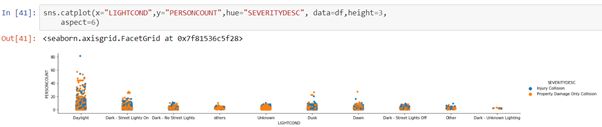
**3.7 Relationship between ROADCOND, VEHCOUNT by SEVRITY of accidents**

Based on below box plot analysis between **ROADCOND, VEHCOUNT** by Severity of accidents, we can see some outliers of VEHCOUNT when weather is Wet, Dry and Oil



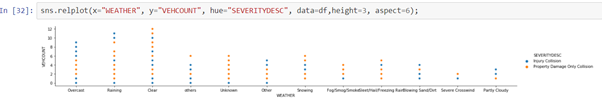
**3.6 Relationship between LIGHTCOND, PERSONCOUNT by SEVERITY of accidents**

Based on below categorical plot between LIGHTCOND, PERSONCOUNT by Severity of accidents, we can say that more number of people involved in accidents related to property damage and are reported during Daylight



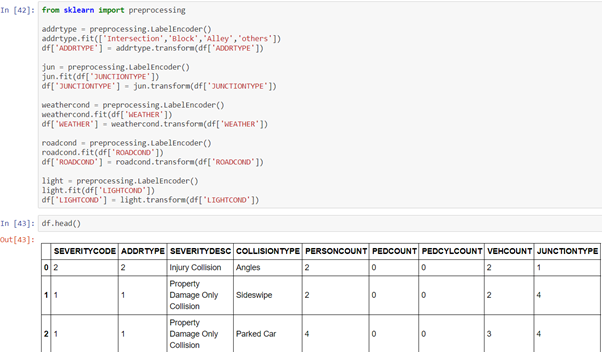
**3.6 Relationship between WEATHER, VEHCOUNT by SEVERITY of accidents**

Based on below categorical plot between WEATHER, VEHCOUNT by Severity of accidents, we can say that more number of vehicles are involved in accidents related to property damage, that are reported during Clear and Rain weather



**Encoding data:- Convert categorical variables to numeric values**

Features in this dataset are categorical ADDRTYPE, JUNCTIONTYPE, WEATHER, ROADCOND, LIGHTCOND, Sklearn KNN, Decision Trees, Logistic models do not handle categorical variables. But still we can convert these features to numerical values.



**4. Predictive Modeling**

For implementing the ML-Predictive modeling, I have used Github as a repository and IBM Watson studio for running Jupyter Notebook to preprocess data and build Machine Learning models. Regarding coding, I have used Python and its popular packages such as Pandas, NumPy and Sklearn.

Once I have load data into Pandas Dataframe, used ‘*dtypes’*attribute to check the feature names, counts and their data types. Based on exploratory data analysis, I can see that the categorical variables have key impact in predicting severity in accidents hence I have selected these most important features to predict the severity of accidents in Seattle. Among all the features, the following features have the most influence in the accuracy of the predictions:

* “WEATHER”,
* “ROADCOND”,
* “LIGHTCOND”
* “ADDRTYPE”
* “JUNCTIONTYPE”

Also, as I mentioned earlier, “SEVERITYCODE” is the target variable.

After standardizing the input feature, the data has been ready for building machine learning models.

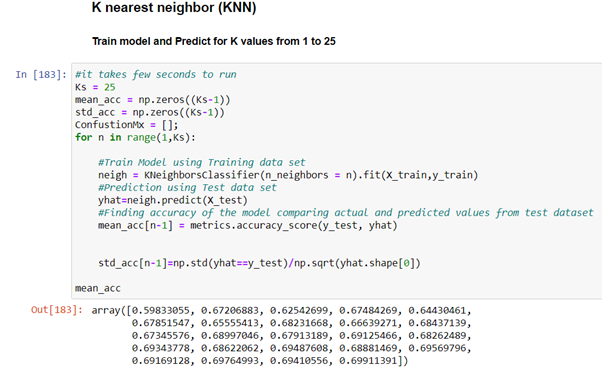
I have employed three machine learning models:

* K Nearest Neighbour (KNN)
* Decision Tree
* Logistic Regression

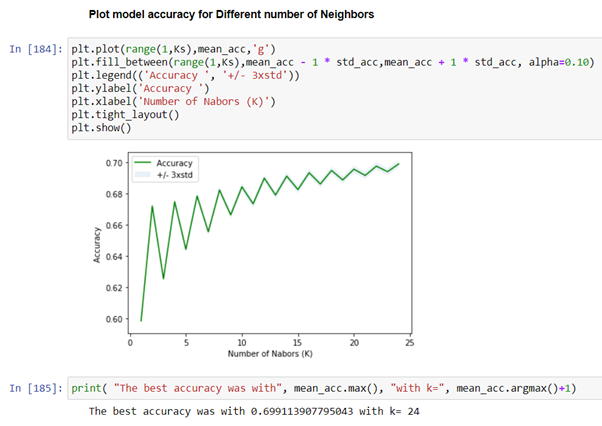
After importing necessary packages and splitting pre-processed data into test and train sets, for each machine learning model, I have built and evaluated the model and shown the results as follow:

**KNN**

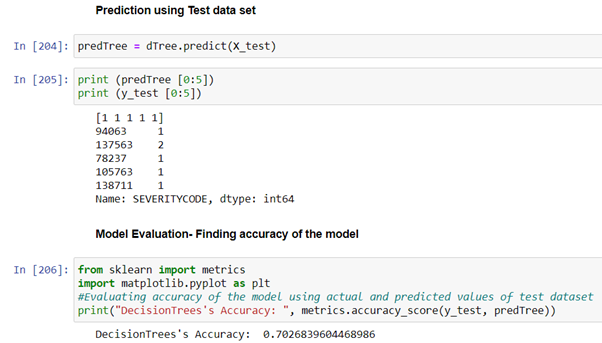
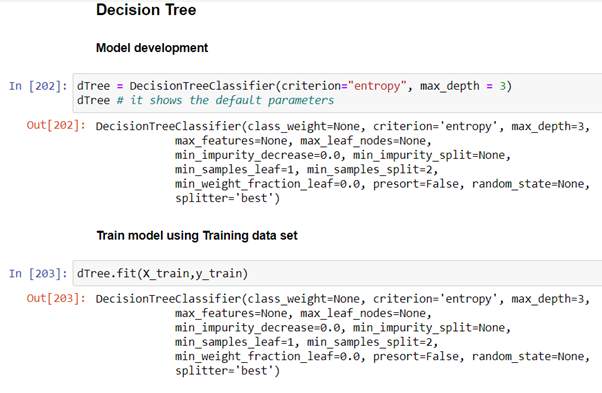
Here while implementing KNN algorithm, I have taken it for set of 25 values  the model will be trained on training set of data, and then predicting the values based on test data set, Atlast the model accuracy will be calculated by comparing actual values with predicted values of test data set.



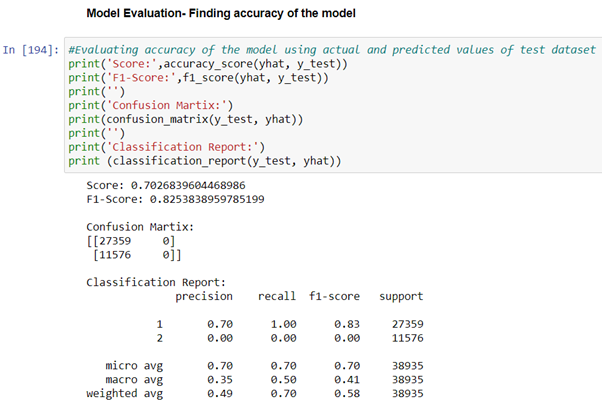
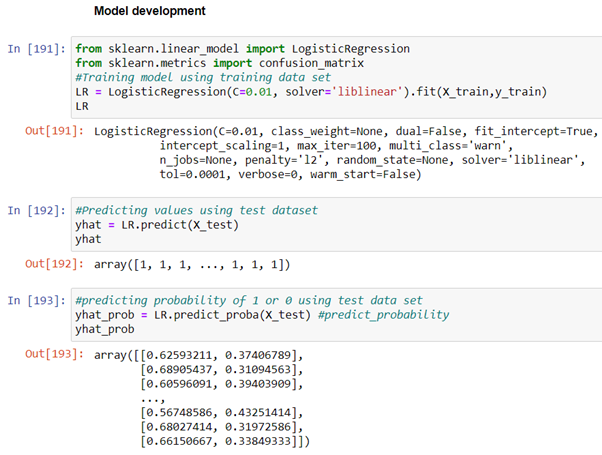
While evaluating model, plotting all 25 accuracy values, so the highest value of is the best accuracy KNN Model. In this case it is for k=24 we have got maximum accuracy of 0.69



**Decision Tree**



**Logistic Regression**



**5. Results and Evaluation**

Once we analysed and cleaned the data, it was then fed through three ML models; K-Nearest Neighbour, Decision Tree and Logistic Regression. Although the first two are ideal for this project, logistic regression made most sense because of its binary nature.

Evaluation metrics used to test the accuracy of our models were jaccard index, f-1 score and logloss for logistic regression. Choosing different k values helped to improve our accuracy to be the best possible.

The final results of the model evaluations are summarized as below

|  |  |
| --- | --- |
| MODEL | Accuracy |
| KNN | 0.69 |
| Decision Tree | 0.70 |

|  |  |  |  |
| --- | --- | --- | --- |
| MODEL | Jaccard Score | F1 Score | Logloss |
| Logistic Regression | 0.70 | 0.82 | 0.63 |

**6. Conclusion**

I have got a decent accuracy value for all classification algorithms. So the best classifier of this problem are Decision Tree, Logistic and KNN based on their accuracy value.

By revealing hidden patterns in predicting severity in accidents based on the features Weather, Road and Light conditions, addresstype, junctiontype have significant impact on whether to travel or not which often result in injury and property damage.