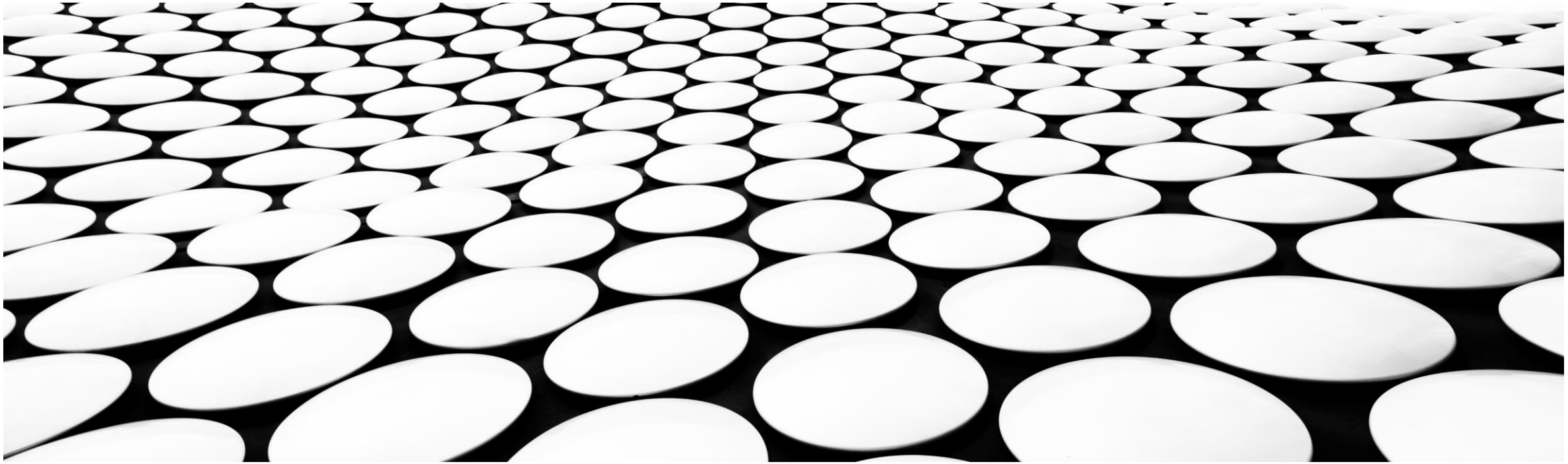
PREDICTING SEVERITY OF CAR ACCIDENT FROM WEATHER, ROAD AND LIGHT CONDITION

DAVID OH



PREDICTING ACCIDENT SEVERITY HELPS TO REDUCE POSSIBILITY OF ACCIDENT AND BE MORE RESPONSIVE WITH GOOD PREPARATION

- If there is something in place that could warn driver, given the weather and the road conditions about the possibility of you getting into a car accident and how severe it would be, so that you would drive more carefully or even change your travel if you are able to. Subsequently, It will help to reduce possibility of accident.
- Weather, Road and Light condition in Collision data set might help to understand relationship with Severity of accident. By learning what type of sub-condition in respective condition, how much particular or combined conditions related with Severity, the Model will be able to predict Severity of car accident

- Predicting severity of accident will be even more preventive if some actions could be made on frequently accident occurring location depending on impact of certain conditions, like notifying head-up to police station or hospital near to those location so that they will be able to prepare well in response to accident.

DATA ACQUISITION AND CLEANING

- Collisions-All year data set provided by SPD and recorded by Traffic Records which includes all types of collisions displayed at the intersection or mid-block of a segment with timeframe: 2004 to Present.
- Example Dataset by Clicking here. Metadata by Clicking here
- The Severity code is the labeled data , which describes the fatality of an accident and it is unbalanced labels. To avoid biased ML model, it needs balance the data
- Replace null value in Weather, Road and Light condition column with ‘Unknown’ because it could be considered as ‘Unknown’ as self-explanatory by meaning
- Select count attribute (Personcount, Vehcount) for visualize data exploration

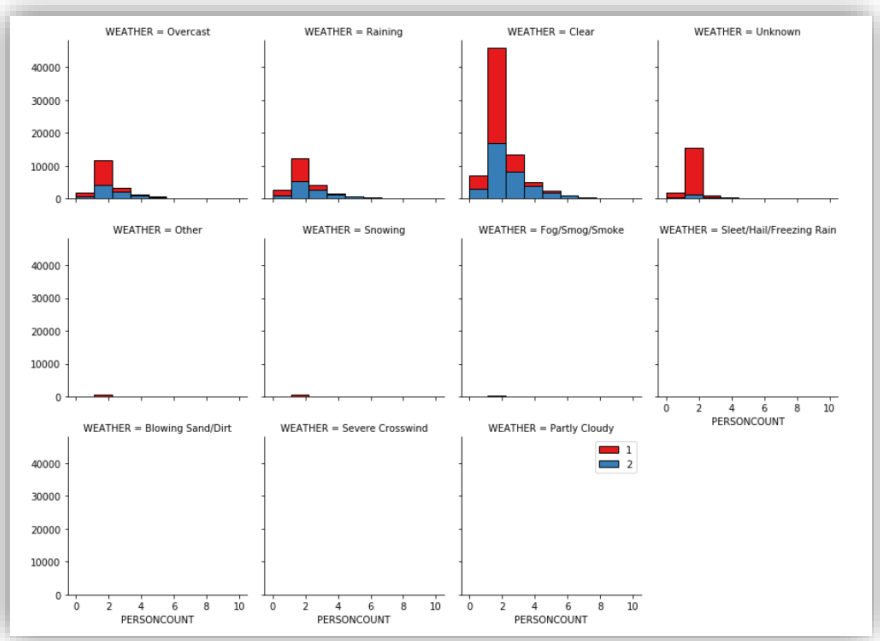
FEATURE SELECTION

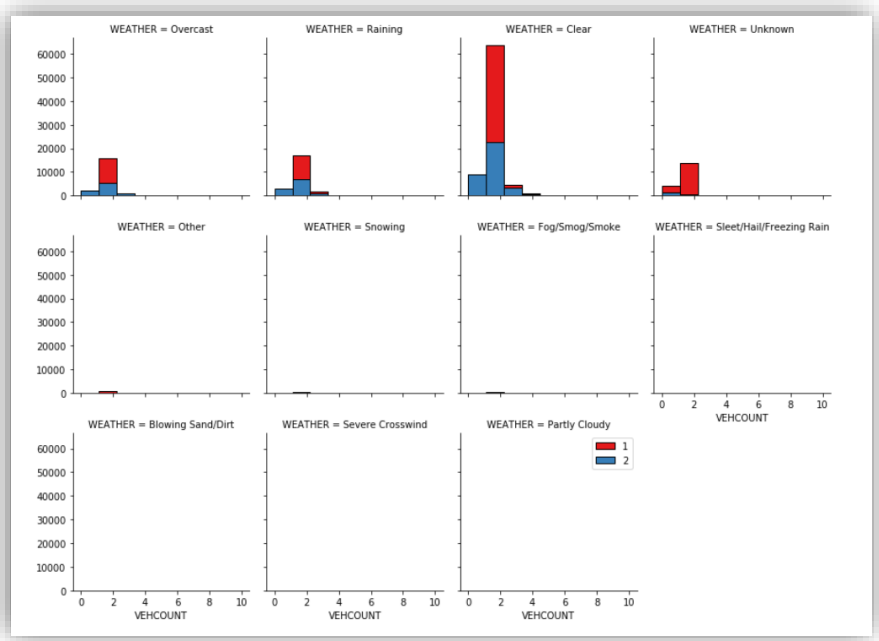
- Upon examining 29 features from condition attributes (11 features in Weather, 9 features in Road condition, 9 features in Light condition), the meaning of each feature and proportion of value within in feature.

Kept Features	Dropped Features	Reason for dropping
Overcast, Raining, Clear, Snowing, Fog/Smog/Smoke in Weather condition	Unknown, Others Sleet/Hail/Freezing Rain, Blowing Sand/Dirt, Severe Crosswind, Partly Cloudy	Less meaningful information Very small number of cases over total cases
Wet, Dry, Snow/Slush, Ice in Road condition	Unknown, Others Sand/Mud/Dirt, Standing Water, Oil	Less meaningful information Very small number of cases over total cases
Daylight, Dark-Street Light On, Dark-No Street Lights, Dusk, Dawn, Dark-Street Lights Off in Light Condition	Unknown, Other Dark-Unknown Lighting	Less meaningful information Very small number of cases over total cases over total cases

RELATIONSHIP BETWEEN SEVERITY AND WEATHER

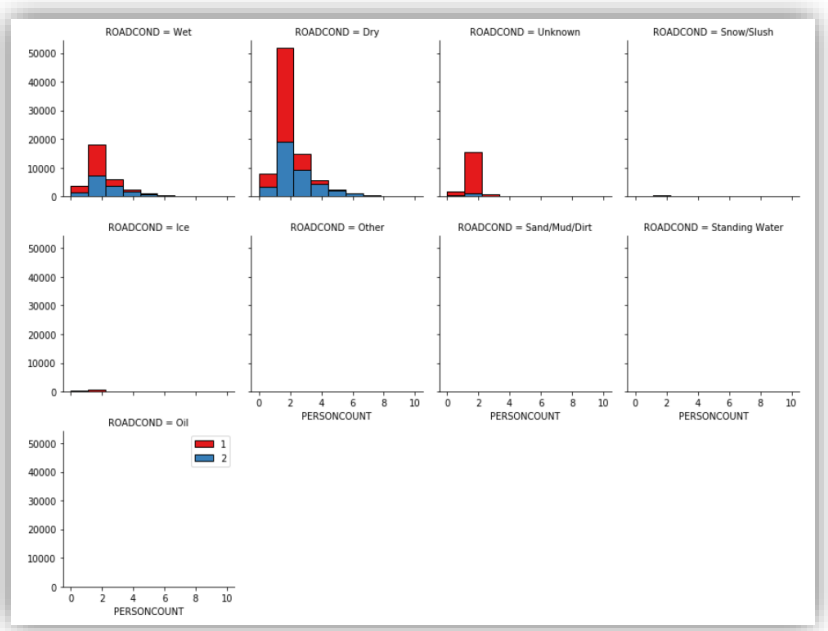
- Majority cases were happened in ‘Clear’, ‘Raining’ and ‘Overcast’, around 70% Severity code 1, 30% Severity code 2.
- Due to ambiguity of interpretation, ‘Unknown’ and ‘Other’ were excluded from feature selection for modeling
- Values occurred in very low volume were excluded from feature selection to avoid biased model training

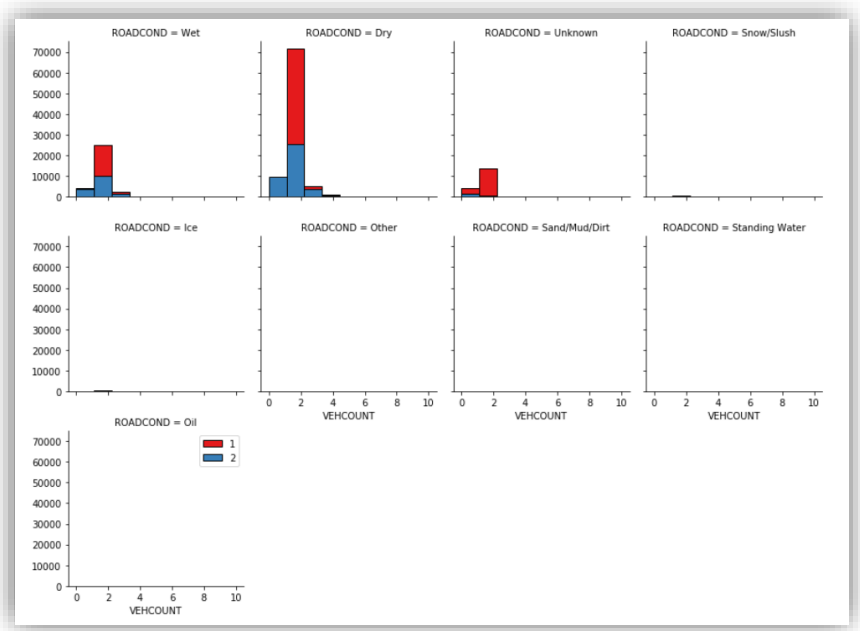




RELATIONSHIP BETWEEN SEVERITY AND ROAD CONDITION

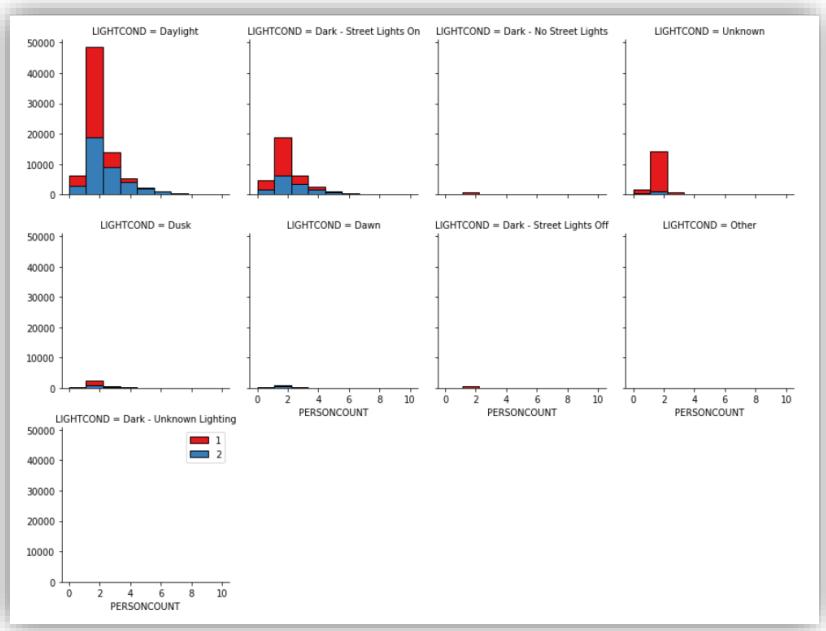
- Majority cases were happened in ‘Dry’ and ‘Wet’, around 70% Severity code 1, 30% Severity code 2.
- Due to ambiguity of interpretation, ‘Unknown’ and ‘Other’ were excluded from feature selection for modeling
- Values occurred in very low volume were excluded from feature selection to avoid biased model training

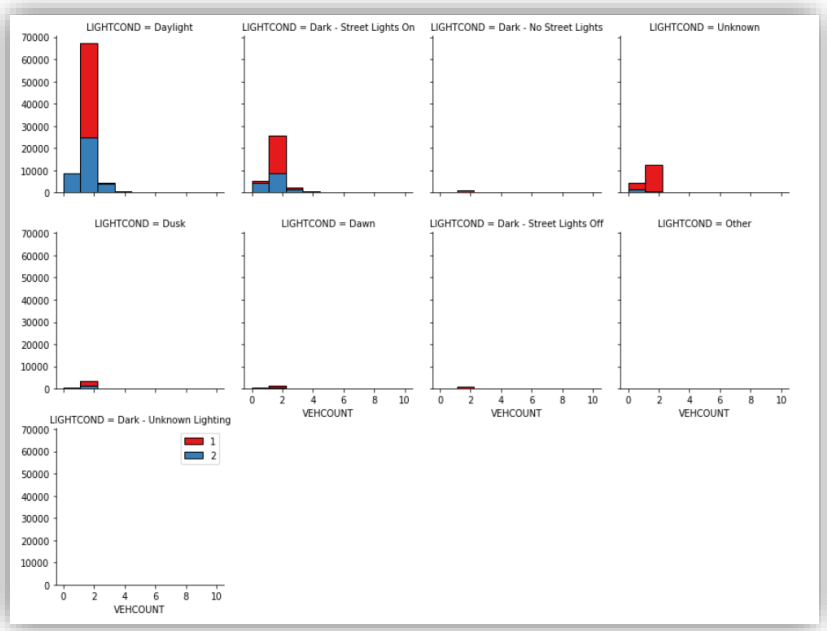




RELATIONSHIP BETWEEN SEVERITY AND LIGHT CONDITION

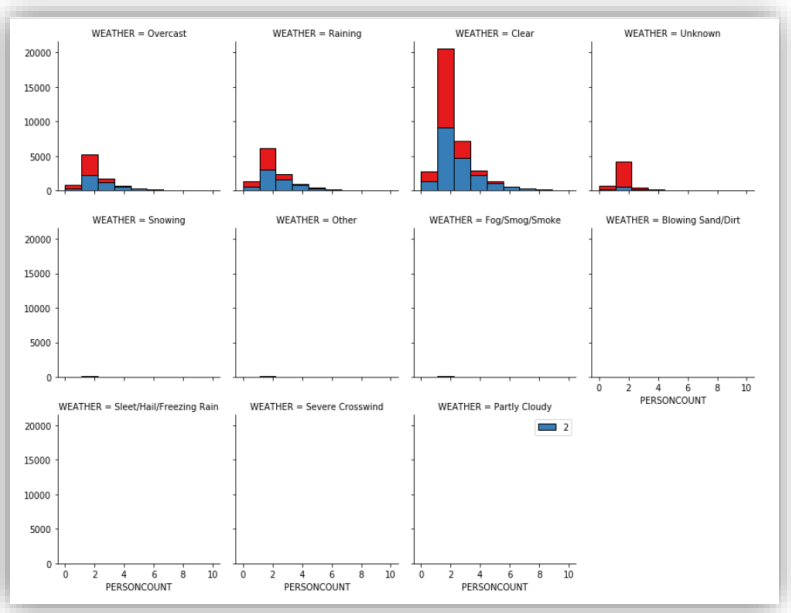
- Majority cases were happened in ‘Daylights’ and ‘Dark-Street Lights On’, around 70% Severity code 1, 30% Severity code 2.
- Due to ambiguity of interpretation, ‘Unknown’ and ‘Other’ were excluded from feature selection for modeling
- Values occurred in very low volume were excluded from feature selection to avoid biased model training

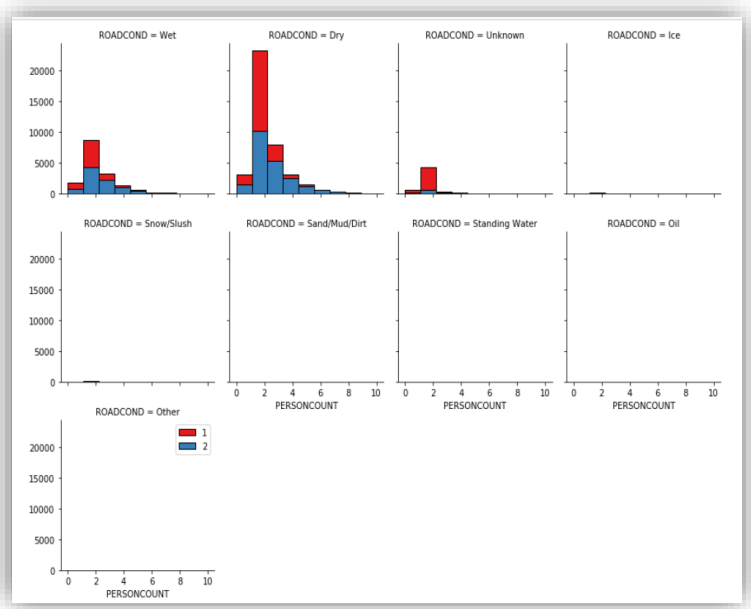


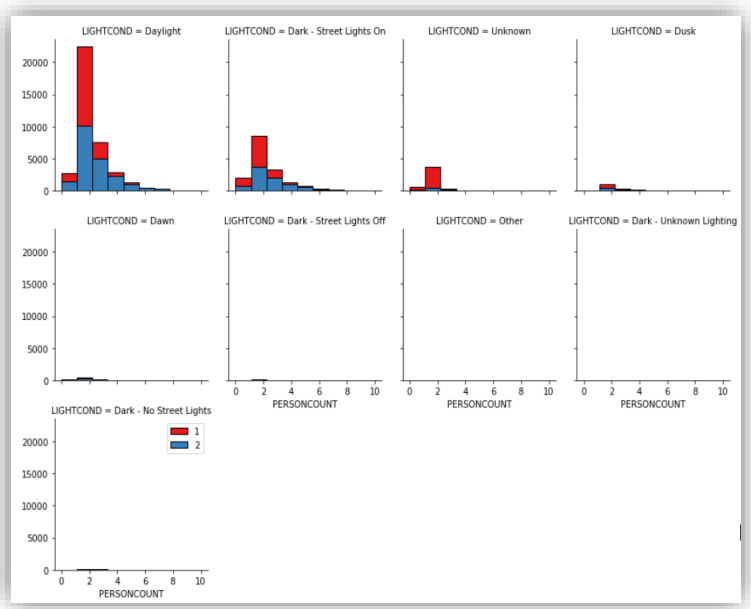


RELATIONSHIP WITH FREQUENT ACCIDENT OCCURRING LOCATION

- To see any different dependency in highly frequent accident occurring location, Dataset was filtered with the number of accidents in the location > 20 (mean 3, 75% 8) and used for visualization.
- It turned out that no significant difference from no-filtered case in weather, road and light condition



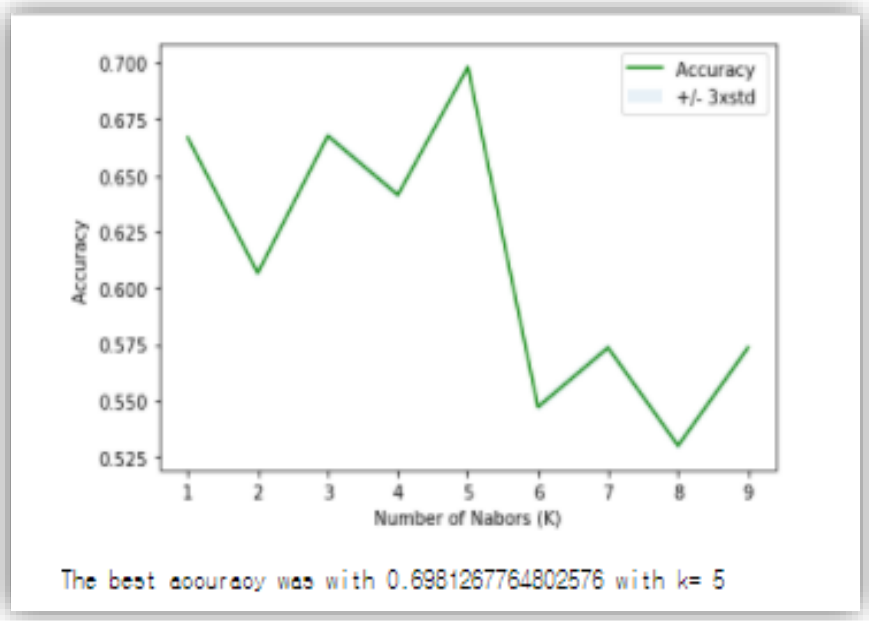




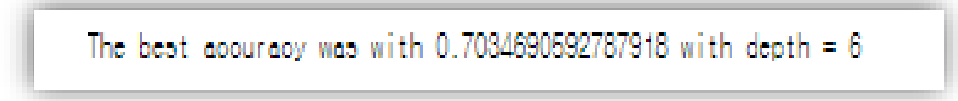
BEST PARAMETER OF KNN AND DECISION TREE ALGORITHM

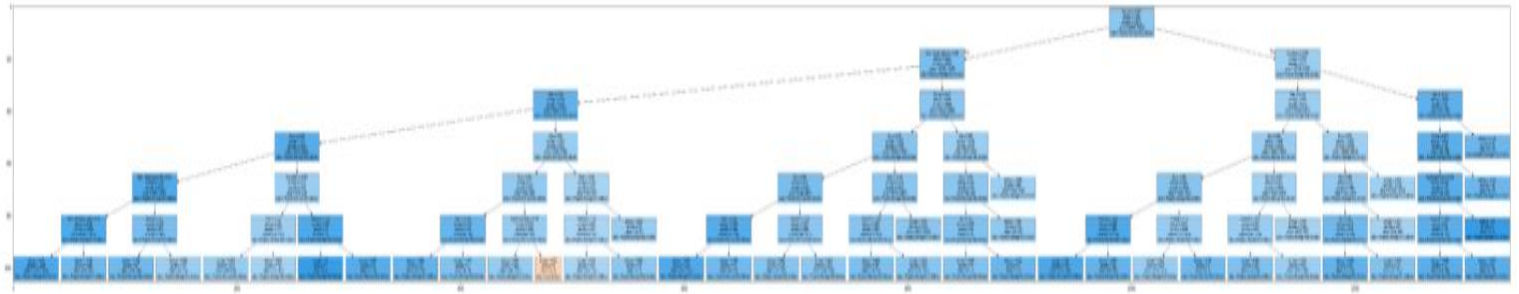
- Check the best option for classification model algorithm, K for K-Nearest Neighbor, Maximum depth for Decision Tree.

- K-Nearest Neighbor



- Decision Tree





BEST PARAMETER OF SVM AND LOGISTIC REGRESSION

- Check the best option for classification model algorithm, Kernel function for Support Vector Machine and Optimizer function for Logistic Regression.

- Support Vector Machine : It showed the best accuracy when kernel function was Linear but to apply for Non-linear problem, chose RBF

Kernel Function	Linear	Polynomial	Radial Basis Function	Sigmoid
Accuracy	0.70345194	0.70328071	0.70331495	0.63484127

- Logistic Regression : It showed the same accuracy regardless of optimization functions, so I chose liblinear which is widely used

Optimizer	Newton-cg	lbfgs	liblinear	sag	saga
Accuracy	0.70345194	0.70345194	0.70345194	0.70345194	0.70345194

EVALUATION OF CLASSIFICATION ALGORITHMS

- With the best option for each algorithm, evaluated the accuracy with metrics like Jaccard Index, F1-Score and Log Loss if applicable.
- Decision Tree and Logistic Regression showed the best accuracy with very small difference. I chose Logistic Regression with logarithmic loss, because it is more accurate in F1 Score and the result would probably be presented with probability for each class rather than just the most likely class.

Algorithm	Jaccard Index	F1-Score	Log Loss
K-Nearest Neighbor	0.698127	0.586849	NA
Decision Tree	0.703469	0.581031	NA
Support Vector Machine	0.703315	0.580988	NA
Logistic Regression	0.703452	0.580990	0.590949

CONCLUSION AND FUTURE DIRECTION

- Accident Severity could be predicted by certain condition of weather, road and light with 59% probability by class based on logistic regression model
- This model can be useful in identifying accident severity and helping police station and hospital preparing accordingly
- There will be more complex to apply in real situation for multiple classification (severity 1,2,2b,3 or else) and this model could be more improved on analyzing relationship among driver attention/inattention, geometry condition impact by capturing pattern of accident collision code
- Accident occurring pattern in time and date with environmental condition might contribute more insightful prediction to prevent by alerting to driver at certain period and location