Capstone project



**Author**

Jegath Prabhu

6th September 2020

**Predict Road Accident Severity**

Contents

[Problem Statement 1](#_Toc50320343)

[Data Section 1](#_Toc50320344)

# Problem Statement

Anyone would rejoice over the fact of taking a drive to a new city. Little is known about the hazards involved in taking up an unknown driveway. A driver approaching the road is unaware of the traffic, weather conditions, visual distance, etc. Thereby, the chances of him/her meeting up with an accident is high and the severity of the accident occurrence in a particular road is not known. Given a scenario, if the driver is provided with a prediction to estimate the severity of accident in for a given weather, light conditions, etc. he/she will be well equipped to take a uniform decision and drive carefully.

This stakeholders for this model would be

1. Drivers who are about to get onto a motorway
2. Emergency response team who should be on alert based on traffic and weather conditions
3. Pedestrians or cyclists who are about to get on to the road for quick getaway

# Data Section

To build this machine learning algorithm for predicting the severity of a accident; the dataset to be used to train the algorithm is as provided by SDOT Traffic Management Division, Traffic Records Group. This includes all types of collisions. Collisions will display at the intersection or mid-block of a segment. Timeframe: 2004 to Present.

The input variables or feature set would be the below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Sl.** | **Attribute** | **Data type, length** | **Description** |
| 1 | JUNCTIONTYPE | Text, 300 | Category of junction at which collision took place |
| 2 | ROADCOND | Text, 300 | The condition of the road during the collision. |
| 3 | LIGHTCOND | Text, 300 | The light conditions during the collision. |
| 4 | WEATHER | Text, 300 | A description of the weather conditions during the time of the collision. |
| 5 | PERSONCOUNT | Double | The number of pedestrians involved in the collision. |
| 6 | VEHCOUNT | Double | The number of vehicles involved in the collision |
| 7 | SPEEDING | Text, 1 | Whether or not speeding was a factor in the collision. (Y/N) |

The labeled data would be SEVERITYCODE (Text,100) which is a code that corresponds to the severity of the collision:

• 3—fatality

• 2b—serious injury

• 2—injury

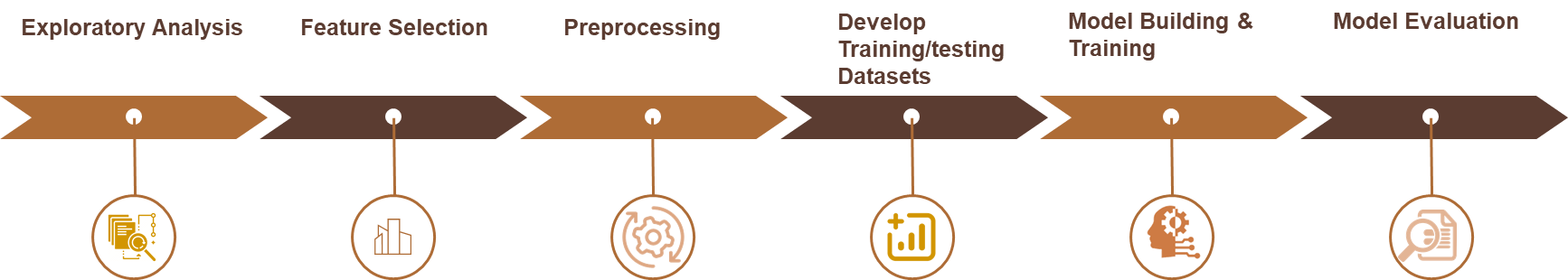
• 1—prop damage

• 0—unknown

Examples of feature set is given a scenario, if the driver is provided with a prediction to estimate the severity of accident in for a given weather, light conditions, etc. he/she will be well equipped to take a uniform decision and drive carefully. An emergency response team would be on alert during bad weather on a specific junction of road. Traffic inspector can monitor highly vulnerable road junctions during bad weather and heavy traffic.

# Approach & Methodology

With the aim to build a model for predicting the severity of road accident using supervised machine learning technique; we would approach the project in a phased approach as shown below:



**Exploratory Analysis**

The dataset we chose is the downloaded CSV file collated with data on accident records from the Seattle Department of Transportation. There are total of 194,673 records. Initially, we explore the data to identify the key input variables that could possibly provide insights in predicting the severity of accident. The input variables such as Weather, Road Condition, Light Condition, Speeding and Junction Type are checked for any interrelations.

Based on the observations below, we see that most accidents have occurred during daylight, dry road and clear weather but in Mid-Block(not related to intersection). This gives us an insight that all variables, if modelled together should give an indication of the severity of the accident. Standalone variables indicate the most ideal scenario out which we are not able to pull out any insights. Also, based on the data provided the Severity Codes are only ‘1’ and ‘2’; there is no data for ‘2b’, ‘3’ and ‘4’. Hence, the **limitation of our prediction model is that we will only be able predict for “Property Damage” or “Injury**”.

Figure 1: Seattle Map with Accident Location Plotted

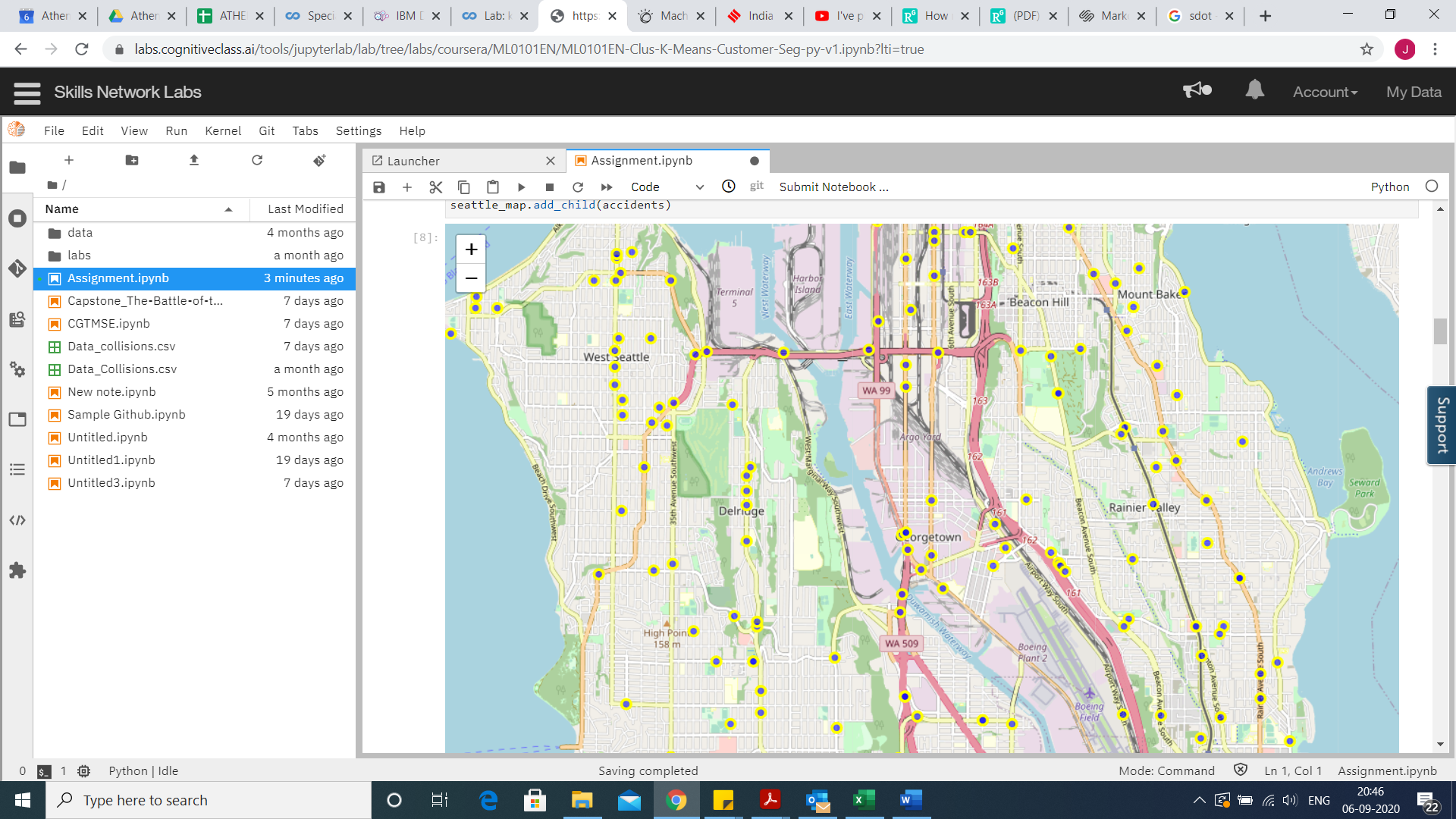


Figure 2: # Accidents in Junctions

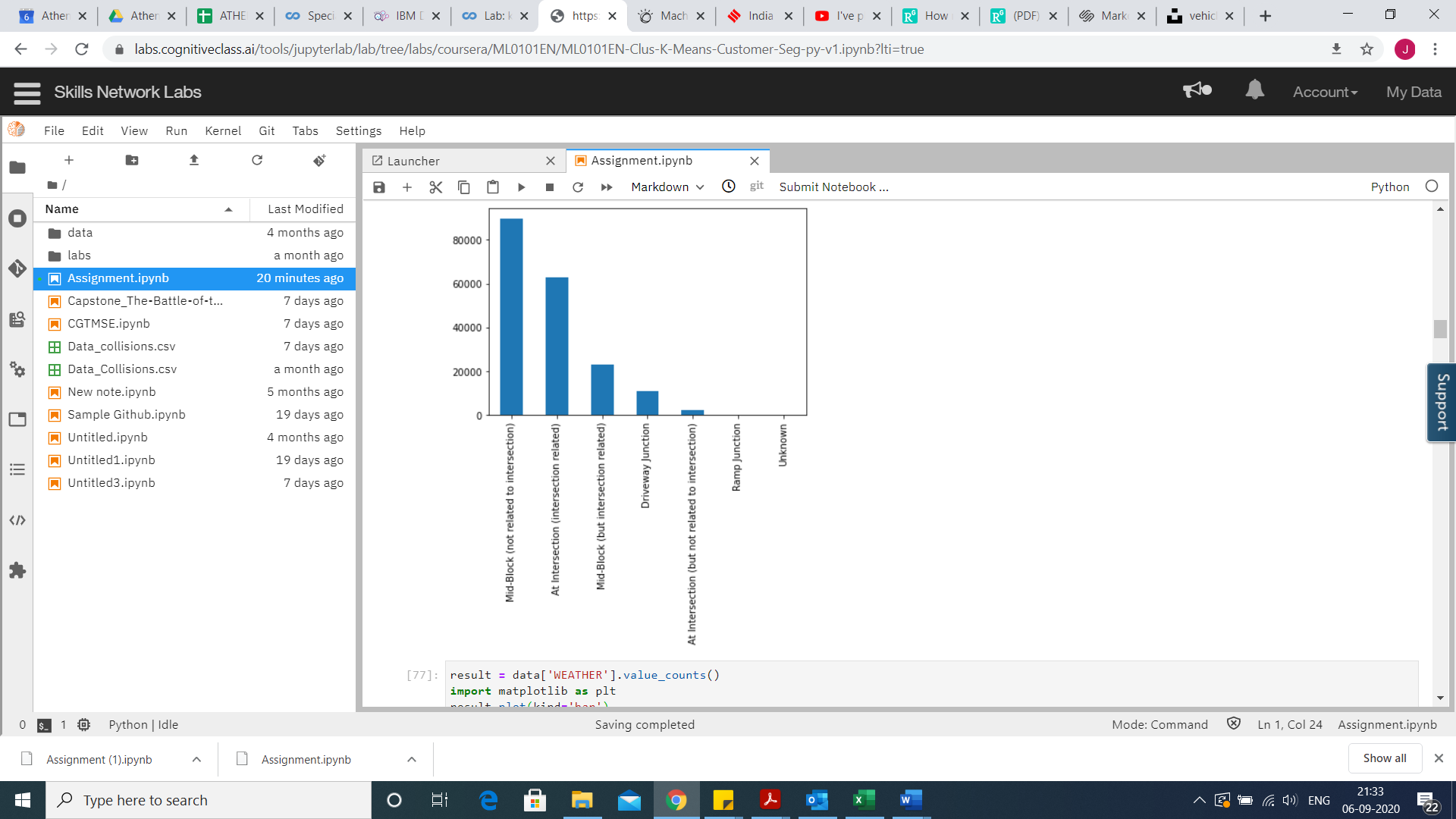


Figure 3: # Accidents in Weather Conditions

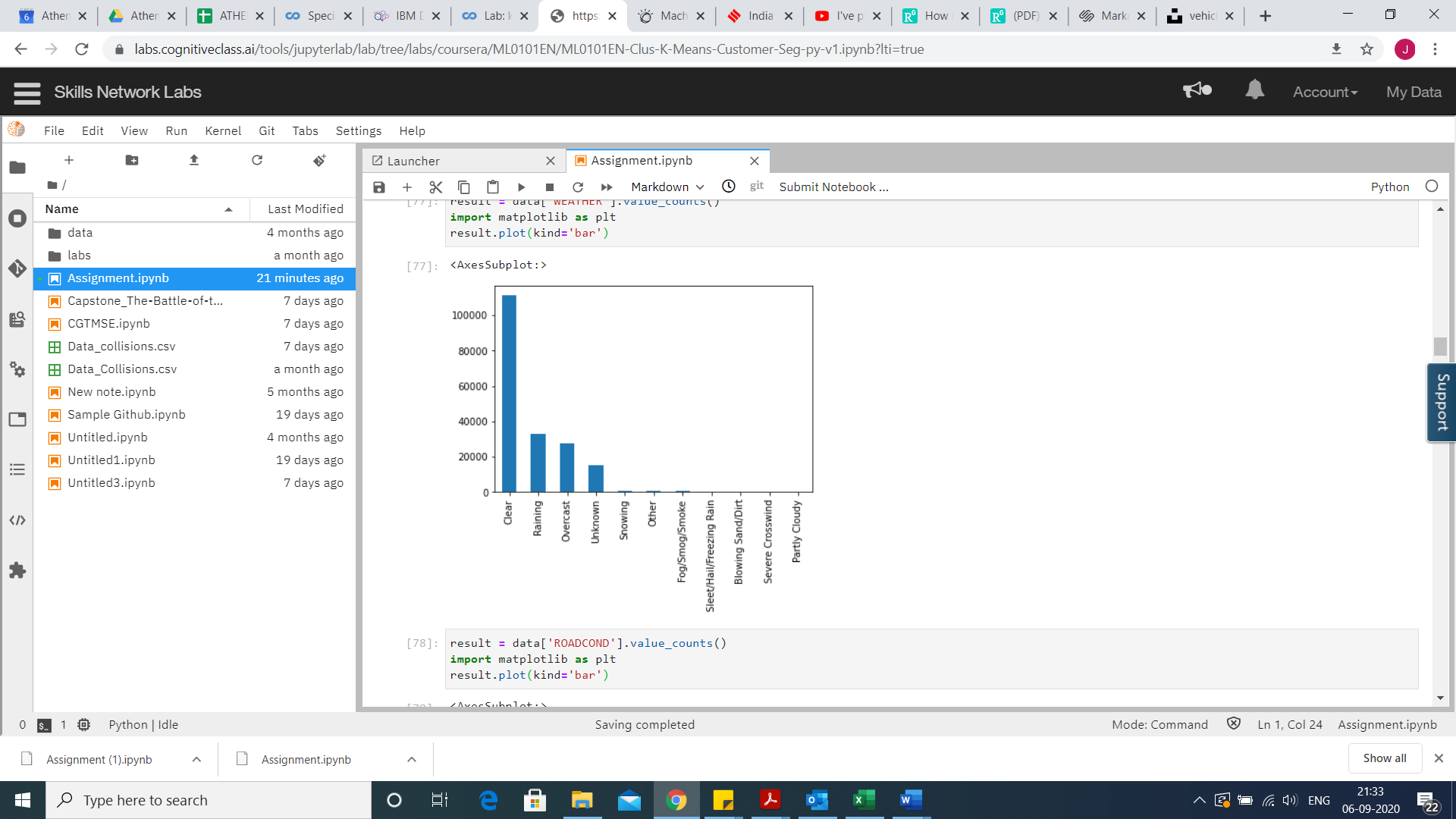


Figure 4: Accidents in Road Condition

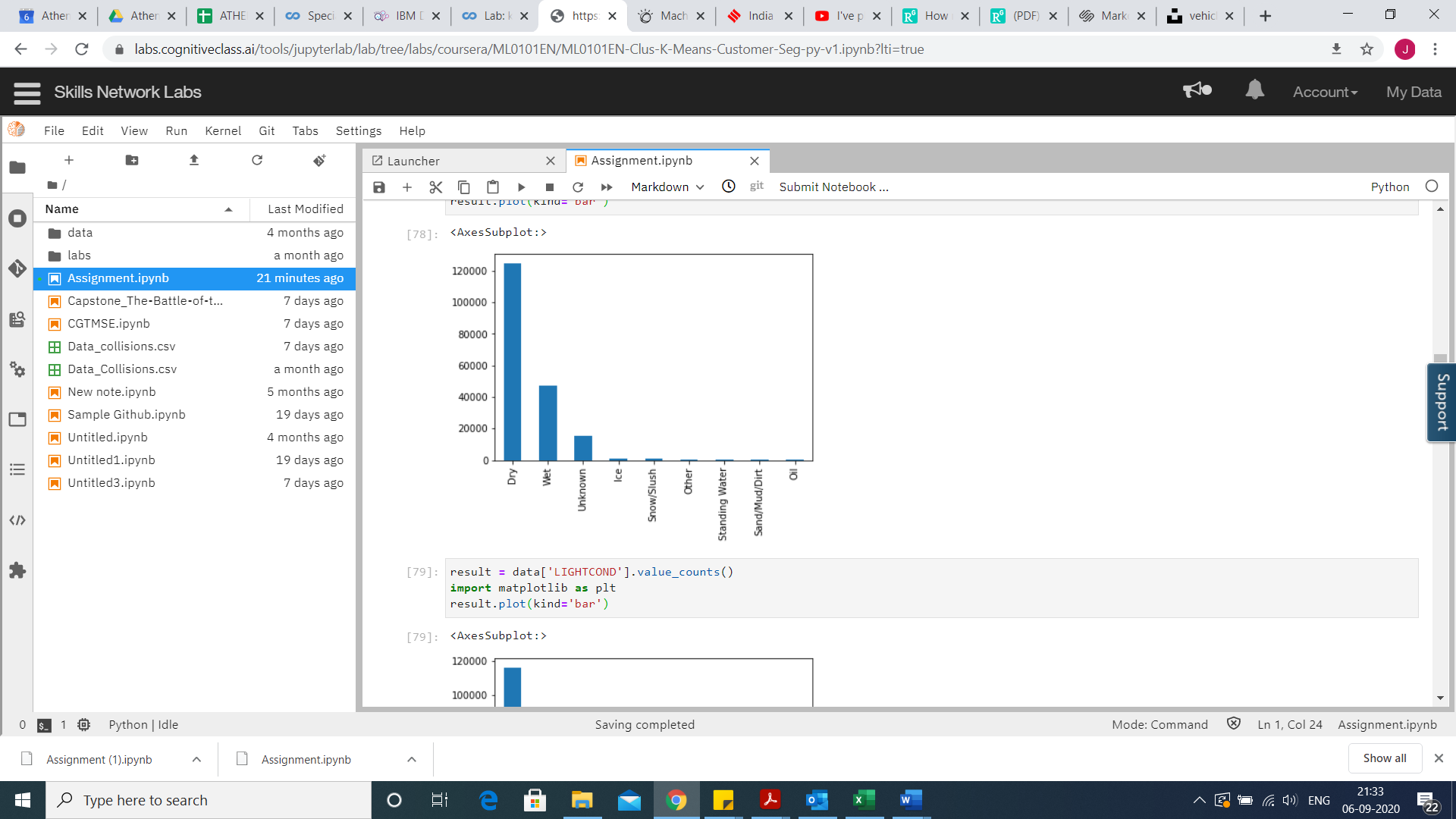
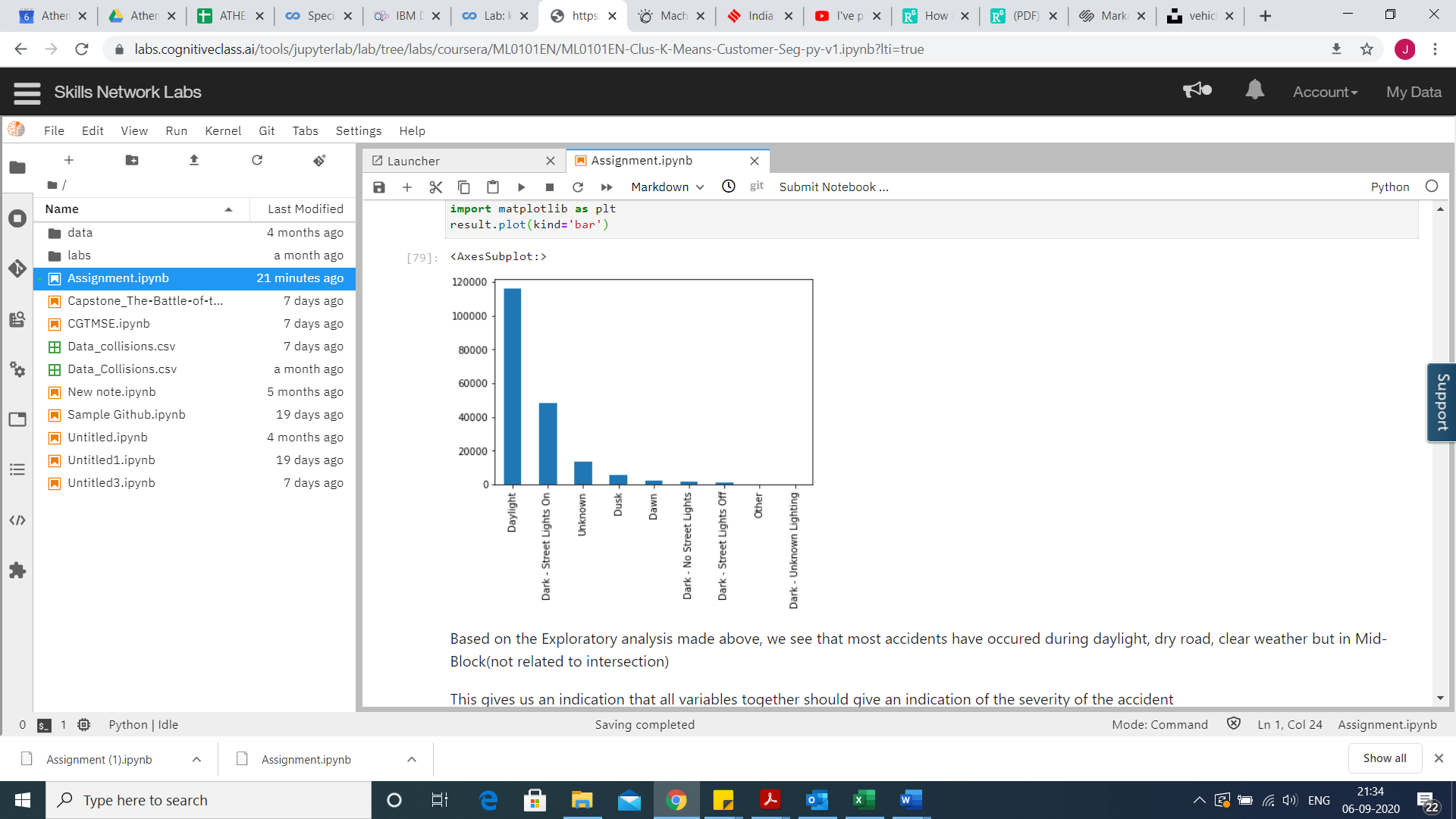


Figure 5: Accidents in Different Light Conditions



**Feature Selection**

Based on Exploratory Analysis choose the Input Variables as Junction Type, Weather Condition, Road Condition, Light Condition, Speeding, No. of persons and no. of Vehicles involved

**Preprocessing**

As required, the Feature dataset is cleaned, normalized and standardized for further processing

**Develop Training/testing Datasets**

The feature dataset is split into 30,000 records for developing training set and training the different algorithms. Since certain algorithms would take hours to process all ~200,000 records, we have taken a smaller dataset for first round. We have kept 4000 records separately for model evaluation purpose and this will not be used to train the model.

**Model Building & Training**

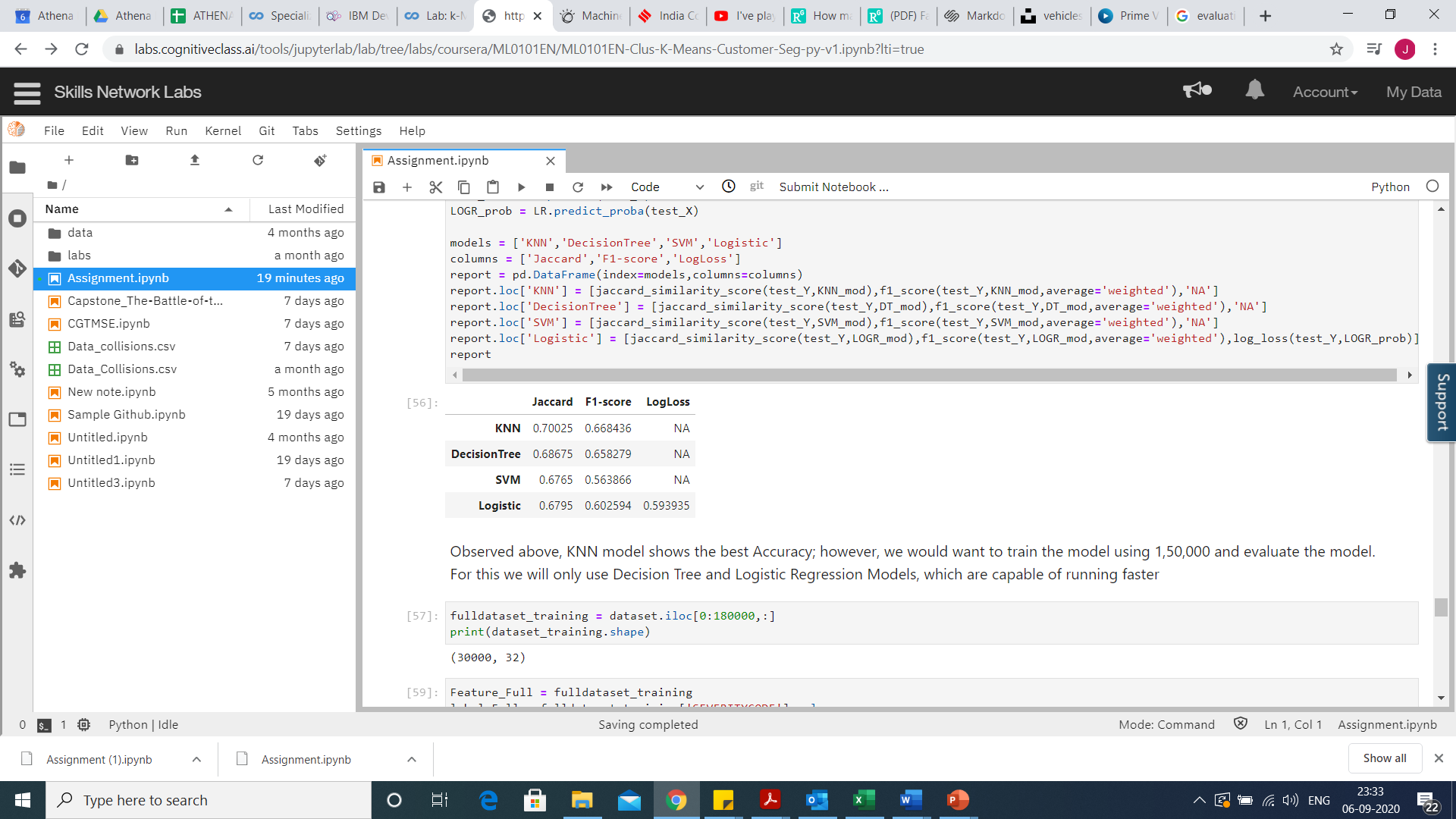
Using Training Dataset, we built

* K Nearest Neighbor Classification model (KNN)
* Decision Tree Classification
* Support Vector Machine Classification (SVM)
* Logistic Regression

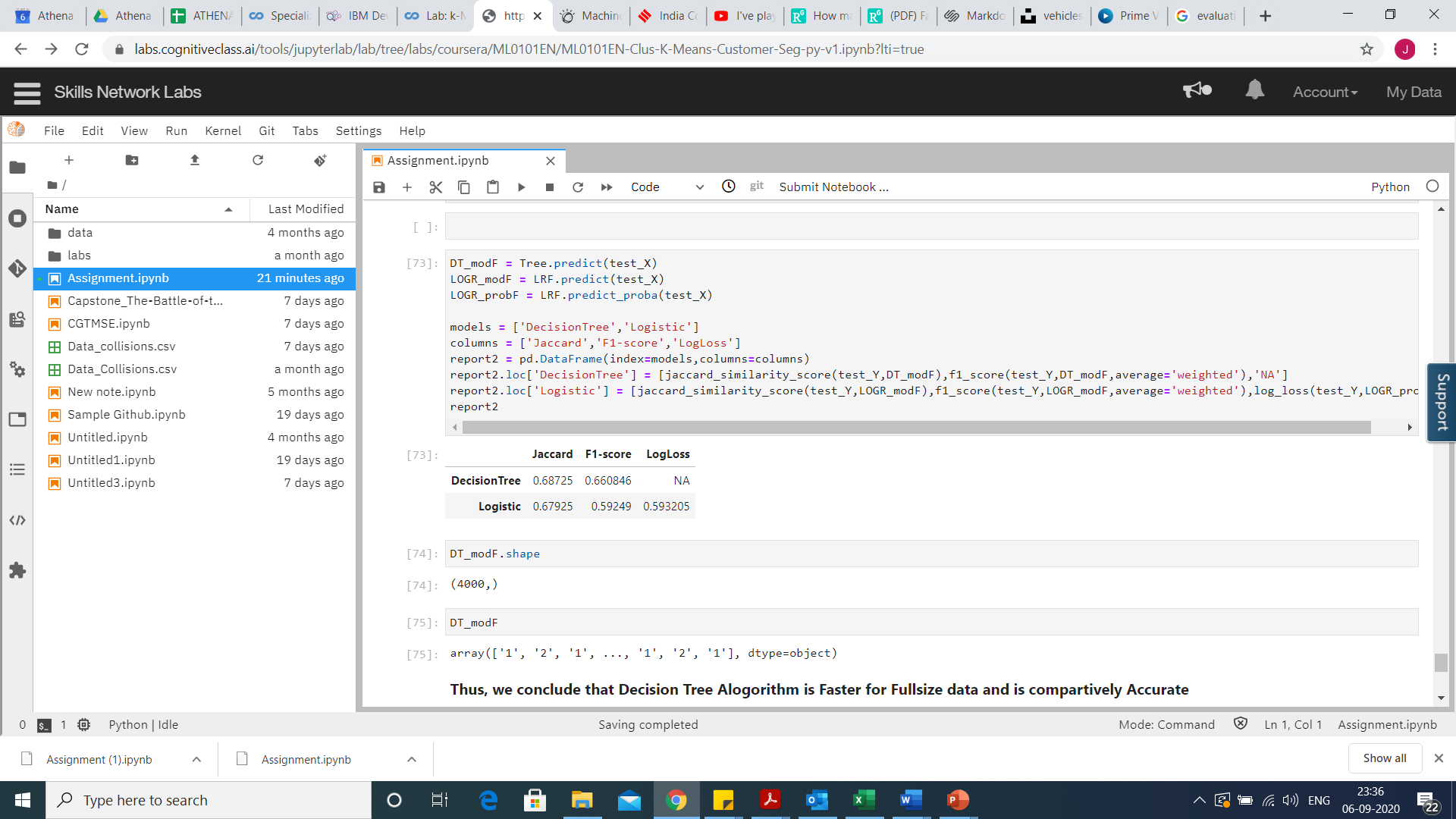
It is clearly evident that KNN and SVM models are quite slower compared to Decision Tree and Logistic Regression.

**Model Evaluation**

In the model evaluation step, we will test for Model Accuracy using Jaccard, F1-Score, log loss.



As shown above, KNN is good model with best accuracy figures; however, to further verify we choose Decision tree and Logistic regression and trained the model using 150,000 input datasets. The evaluation of the model is as below:



In the second iteration we see the Decision tree classifier has better accuracy.

# Conclusion

Thus, we conclude that **Decision Tree classification** model is best suited model for Predicting the Severity of Road Accidents; due to the best accuracy and faster processing time taken.