Capstone Project: Car accident severity

By: Vimmi Seth

Contents

[I. Introduction | Business Understanding 2](#_Toc52713908)

[1. Background 2](#_Toc52713909)

[2. Problem Description 2](#_Toc52713910)

[3. Target Audience 2](#_Toc52713911)

[II. Data 2](#_Toc52713912)

[1. Data Understanding 2](#_Toc52713913)

[2. Data Preparation and Cleaning 3](#_Toc52713914)

[3. Exploratory Data Analysis 8](#_Toc52713915)

[4. Predictive Modeling 8](#_Toc52713916)

[III. Result 12](#_Toc52713917)

[IV. Discussion 12](#_Toc52713918)

[V. Conclusion 12](#_Toc52713919)

|  |  |
| --- | --- |
| **Title Collisions** | All Years |
| **Abstract** | All collisions provided by SPD and recorded by Traffic Records. |
| **Description** | This includes all types of collisions. Collisions will display at the intersection or mid-block of a segment. Timeframe: 2004 to Present |

## Introduction | Business Understanding

### Background

Road traffic are becoming one of the leading causes of death across all age groups globally. Analyzing a significant range of factors, including weather conditions, special events, roadworks, traffic jams among others, an accurate prediction of the severity of the accidents can be performed.

These insights will allow law enforcement bodies to allocate their resources more effectively in advance of potential accidents, preventing when and where a severe accidents can occur as well as saving both, time and money. In addition, this knowledge of a severe accident situation can be warned to drivers so that they would drive more carefully or even change their route if it is possible or to hospital which could have set everything ready for a severe intervention in advance.

### Problem Description

In an effort to reduce the frequency of car collisions in a community, an algorithm must be developed to predict the severity of an accident given the current weather, road and visibility conditions. When conditions are bad, this model will alert drivers to remind them to be more careful.

### Target Audience

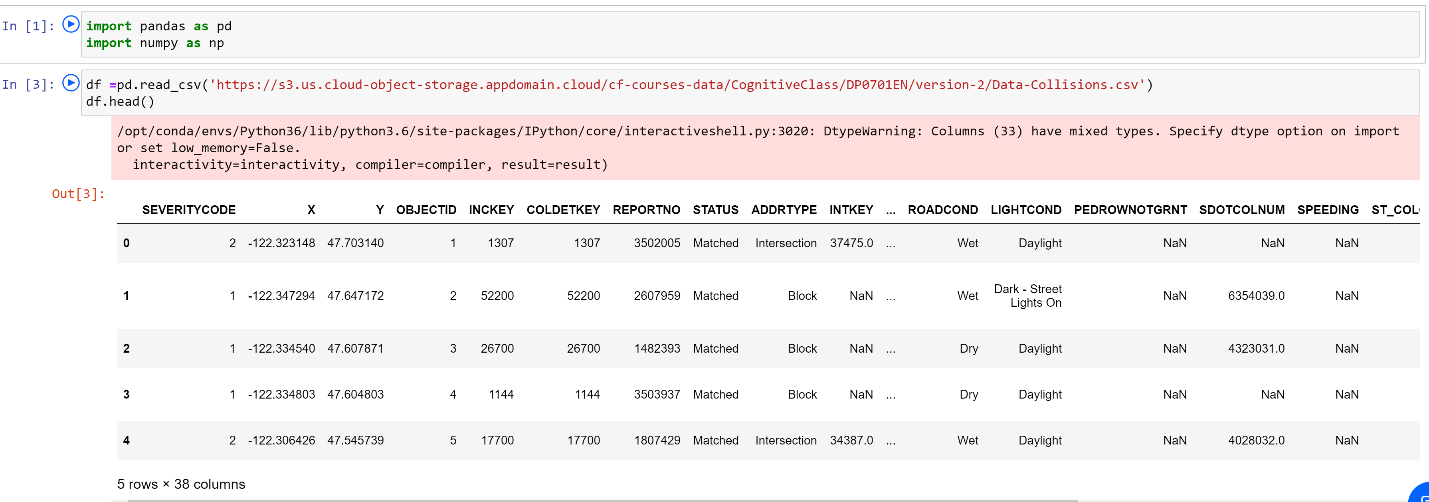
Governments should be highly interested in accurate predictions of the severity of an accident, in order to reduce the time of arrival and thus save a significant amount of people each year.

The public especially the travelers to a city would also be interested in accurate predictions of the severity of an accident under current conditions.

## Data

### Data Understanding

For this project I’ll use data collected by “SDOT Traffic Management Division, Traffic Records Group”, in “Data\_Collision.csv”. The data contains all types of collisions provided by SPD and recorded by Traffic Records from Timeframe: 2004 to Present.



This dataset contains 194674 rows and 38 columns with most of the information that will be needed for the project such as current weather, road and visibility conditions.



Our predictor or target variable will be 'SEVERITYCODE' because it is used measure the severity of an accident in 5 levels [0,1,2,2b,3] within the dataset.

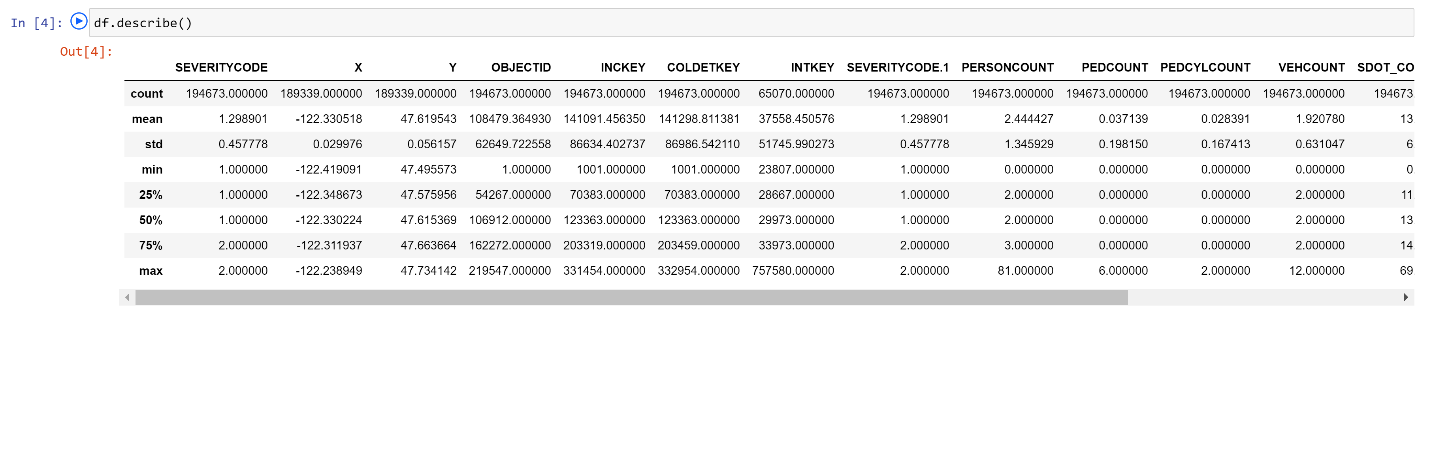
Severity codes are as follows:

|  |  |
| --- | --- |
| **Severity Code** | **Severity Description** |
| 3 | fatality |
| 2b | serious injury |
| 2 | injury |
| 1 | prop damage |
| 0 | unknown |

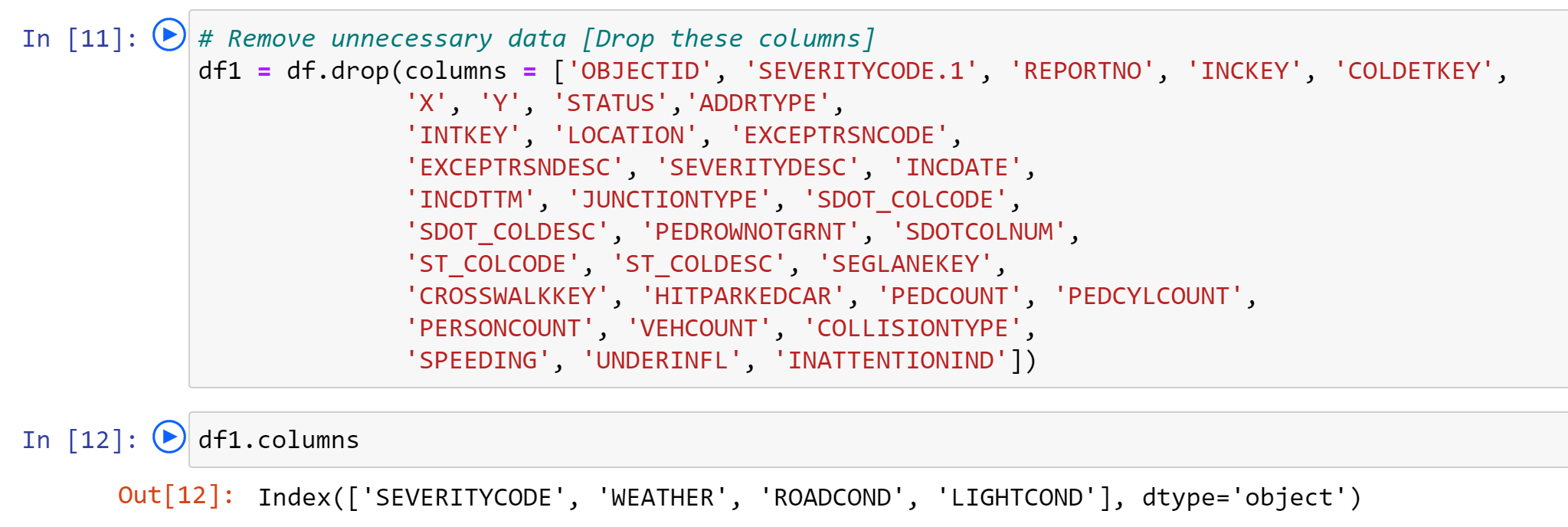
Attributes used to weigh the severity of an accident are ‘'WEATHER', 'ROADCOND' and 'LIGHTCOND'.

### Data Preparation and Cleaning

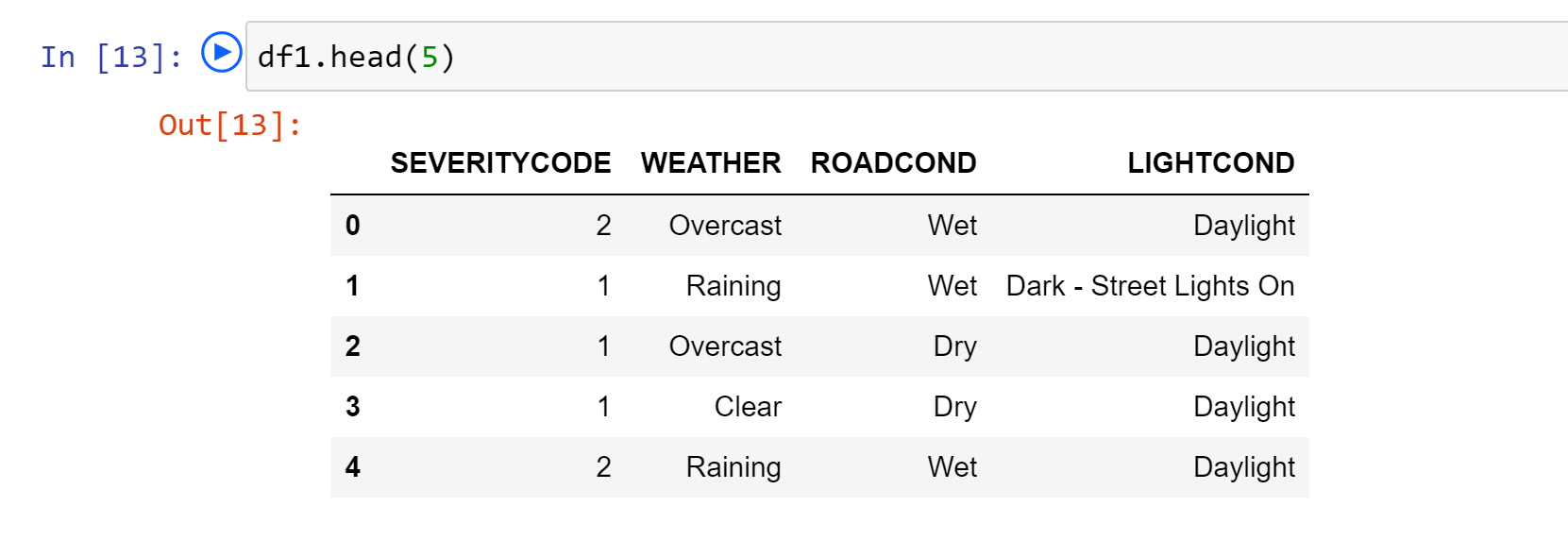
However, this dataset contains mixed datatypes and needs extensive cleaning before it can be used for the project.



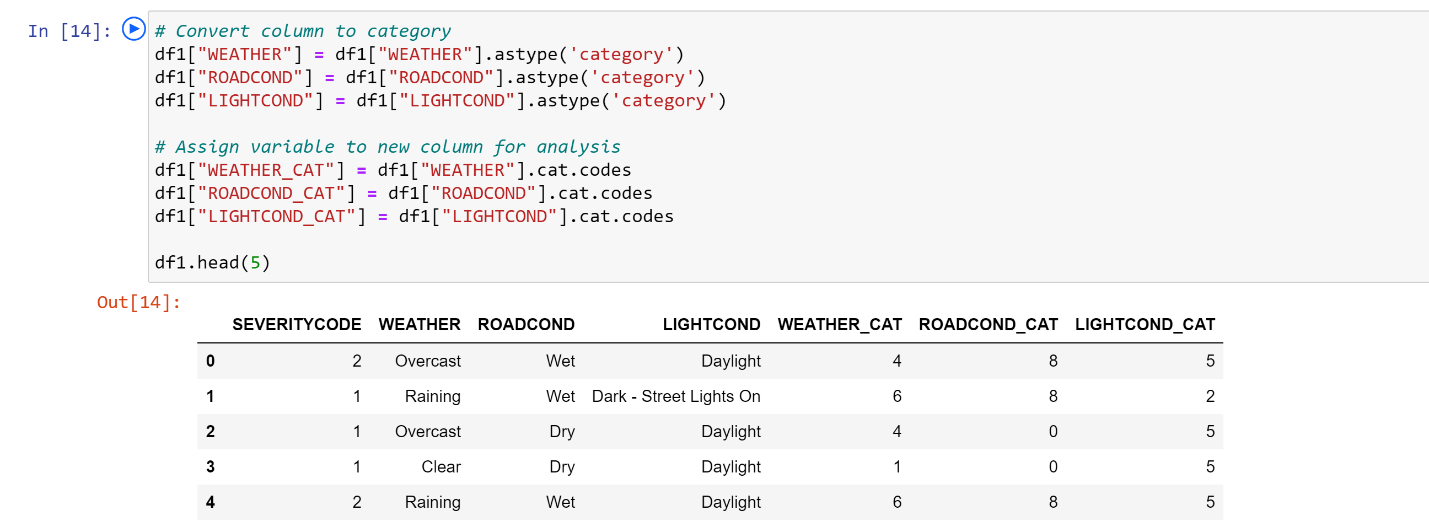
* Select data that is related to our problem
* Remove unnecessary data [Drop these columns]
* Remove the rows with null values.
* Do data profiling to validate correctness of data.
* Convert columns with text field to numerical data like categorizing the weather conditions
* Eliminated rows that greatly contributed to the imbalance of the data and were not significant in volume
* Replaced null values with fillers
* Remove unnecessary data [Drop these columns]



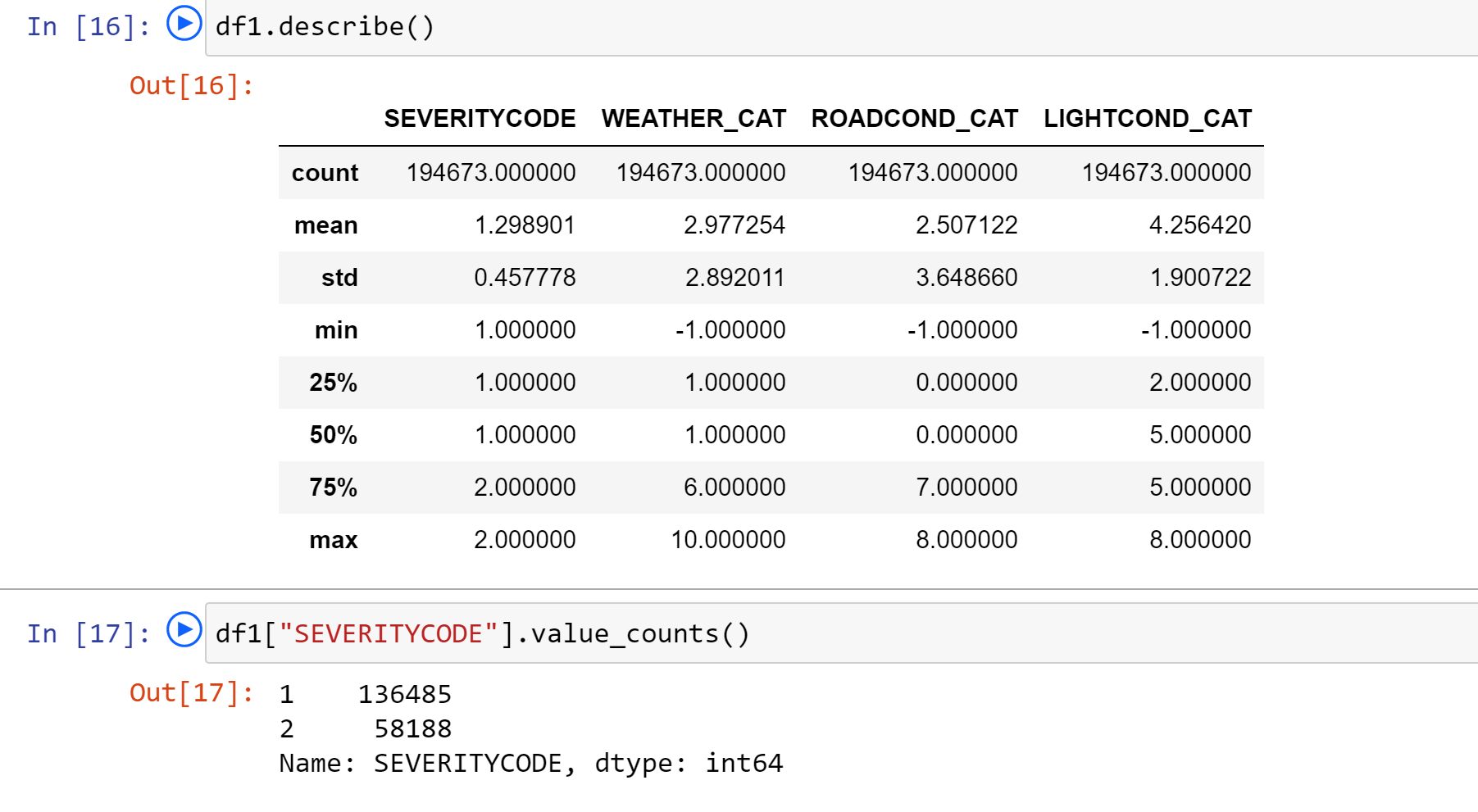
* Select data that is related to our problem

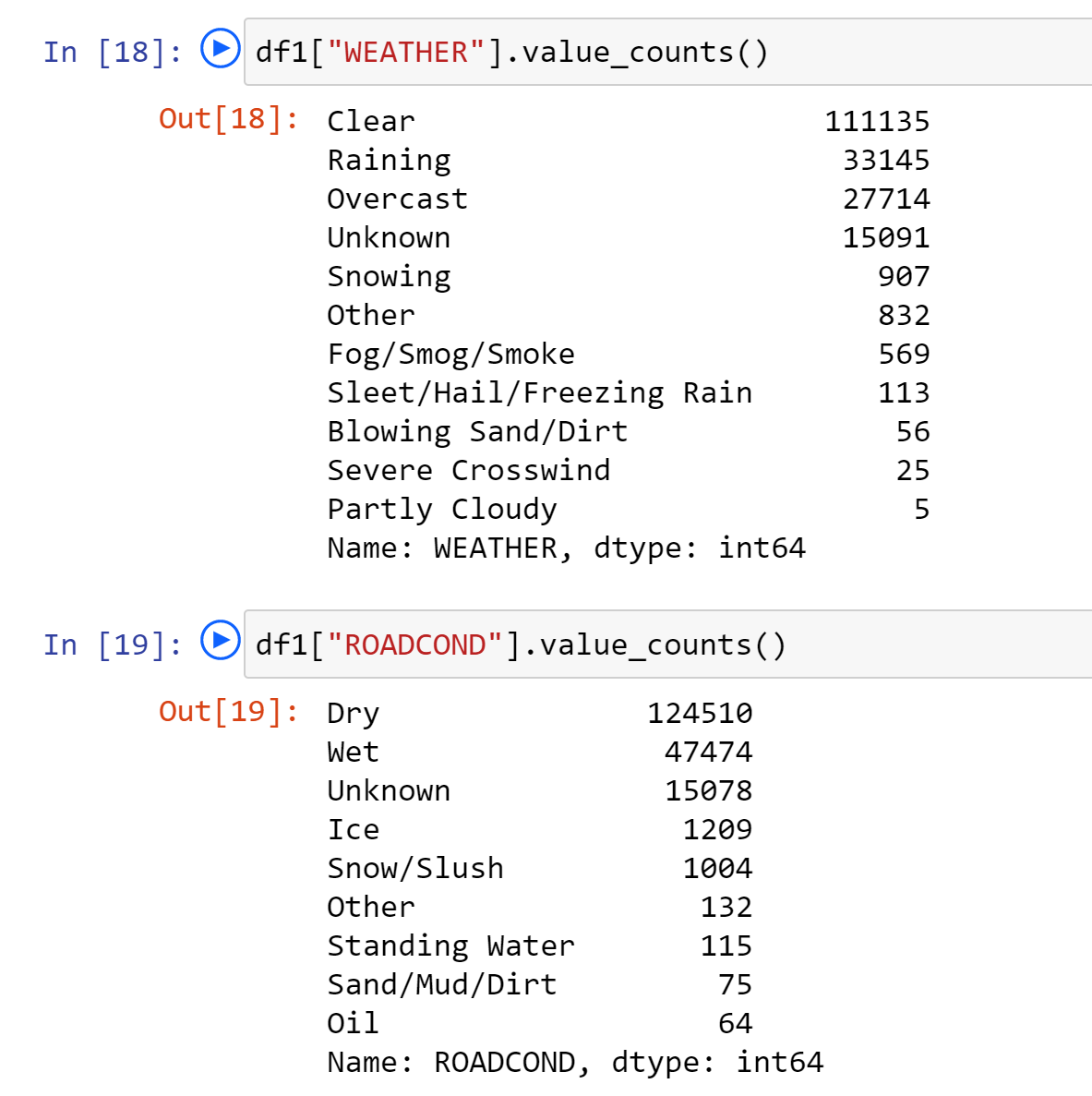


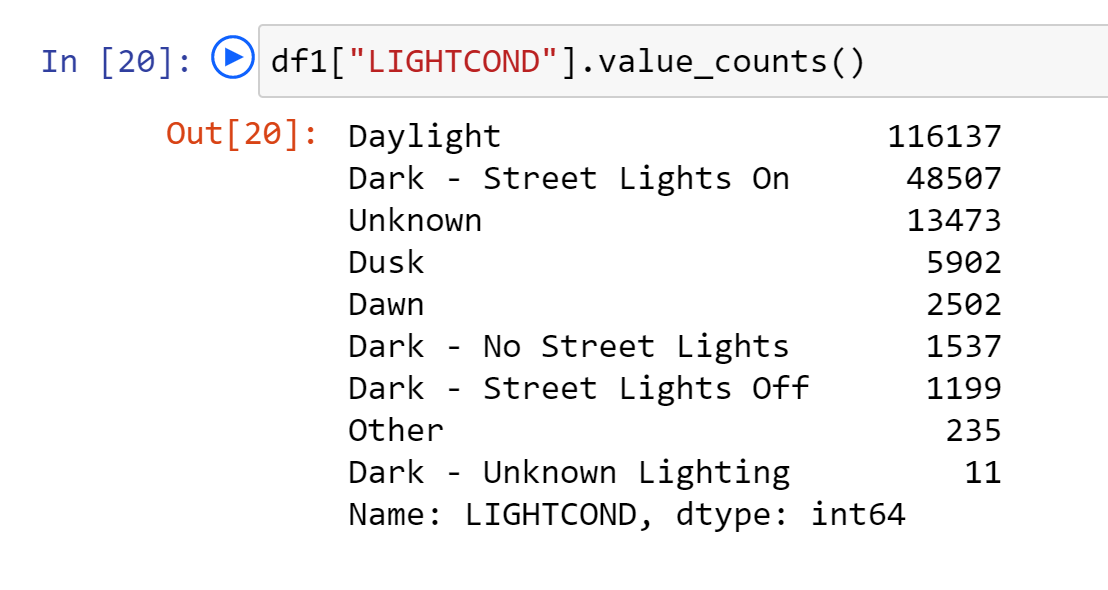
* Convert columns with text field to numerical data like categorizing the weather conditions



* Eliminated rows that greatly contributed to the imbalance of the data and were not significant in volume







Our target variable SEVERITYCODE is not balanced. SEVERITYCODE in class 1 is nearly three times the size of class 2.

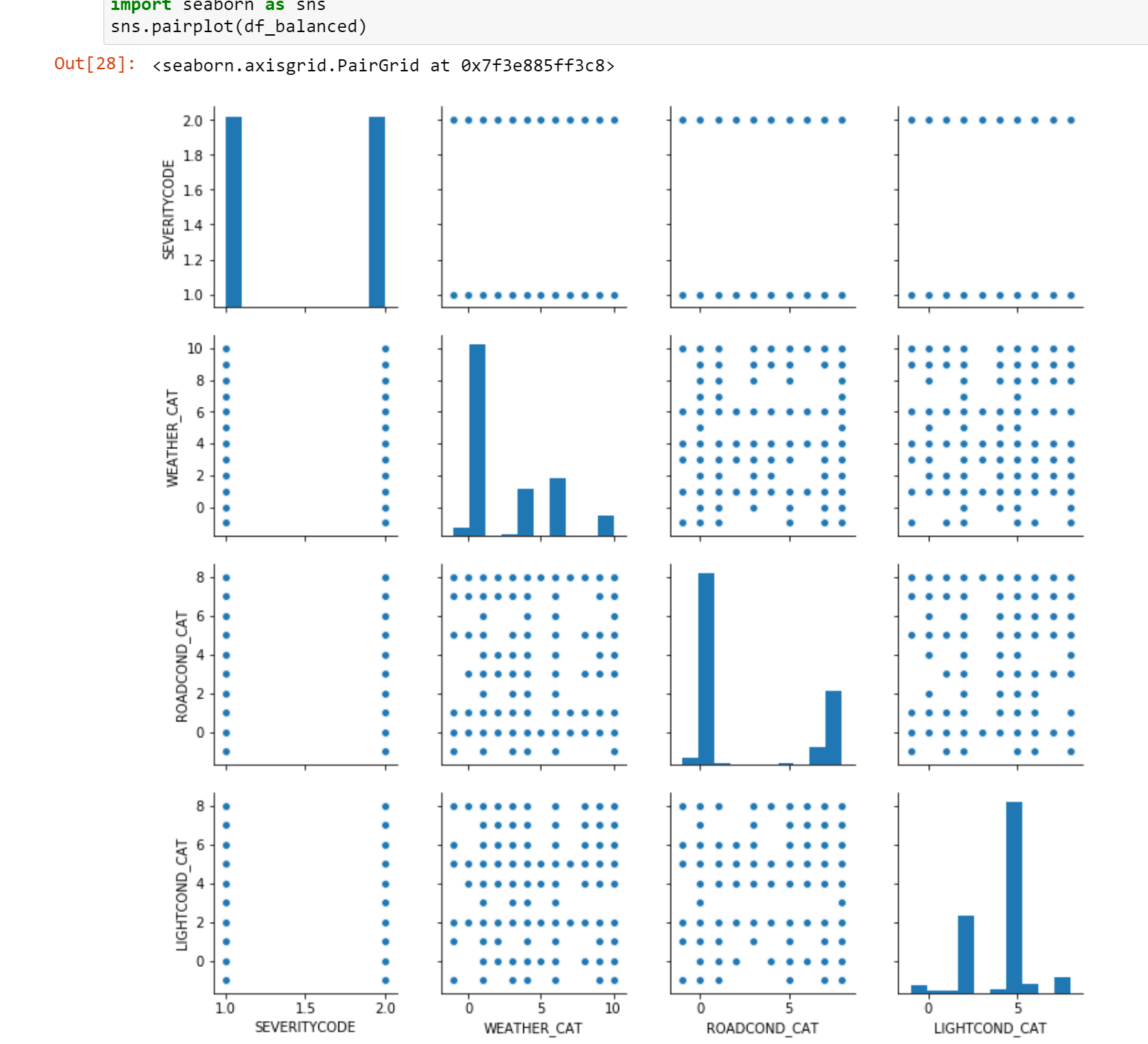
We can fix this by **downsampling** the majority class.

Down-sampling involves randomly removing observations from the majority class to prevent its signal from dominating the learning algorithm. Here are the steps:

* First, we'll separate observations from each class into different DataFrames.
* Next, we'll resample the majority class without replacement, setting the number of samples to match that of the minority class.
* Finally, we'll combine the down-sampled majority class DataFrame with the original minority class DataFrame.



### Exploratory Data Analysis



### Predictive Modeling

Different classification algorithms have been tuned and built for the prediction of the severity of the accident. These algorithms provided a supervised learning approach predicting with certain accuracy.

Our data is now ready to be fed into machine learning models. We will use the following models:

**K-Nearest Neighbor (KNN)**

KNN will help us predict the severity code of an outcome by finding the most similar to data point within k distance.

**Decision Tree, Random Forest**

A decision tree model gives us a layout of all possible outcomes so we can fully analyze the consequences of a decision. It context, the decision tree observes all possible outcomes of different weather conditions.

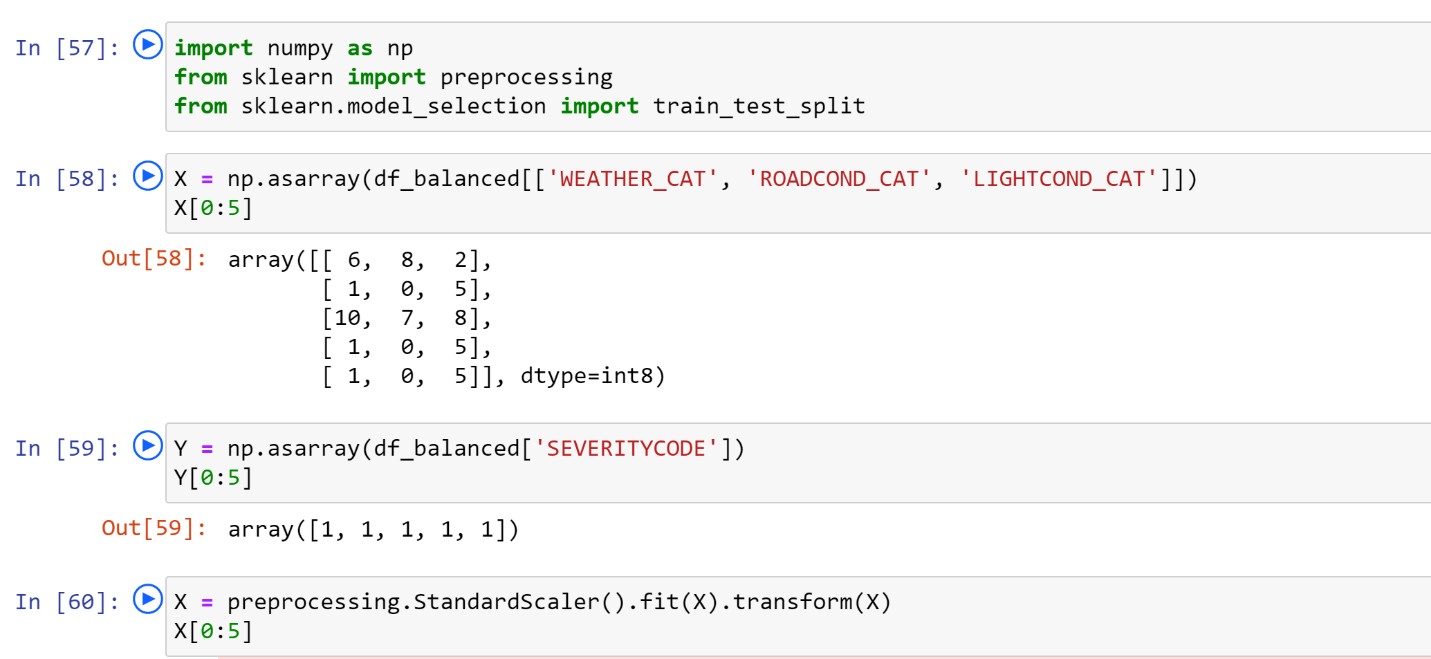
A random forest fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The number of decision trees is specified with the n\_estimators parameter.

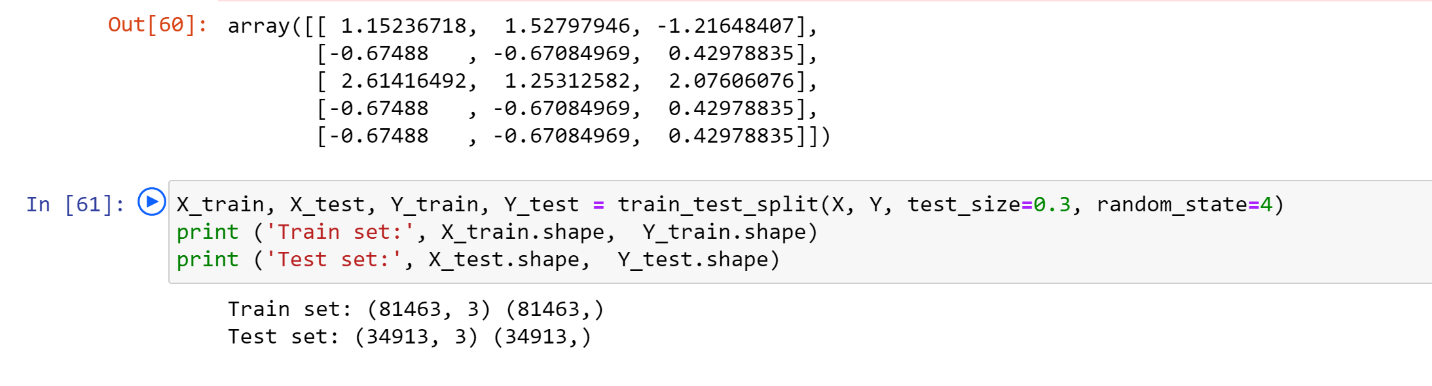
**Logistic Regression**

Because our dataset only provides us with two severity code outcomes, our model will only predict one of those two classes. This makes our data binary, which is perfect to use with logistic regression.

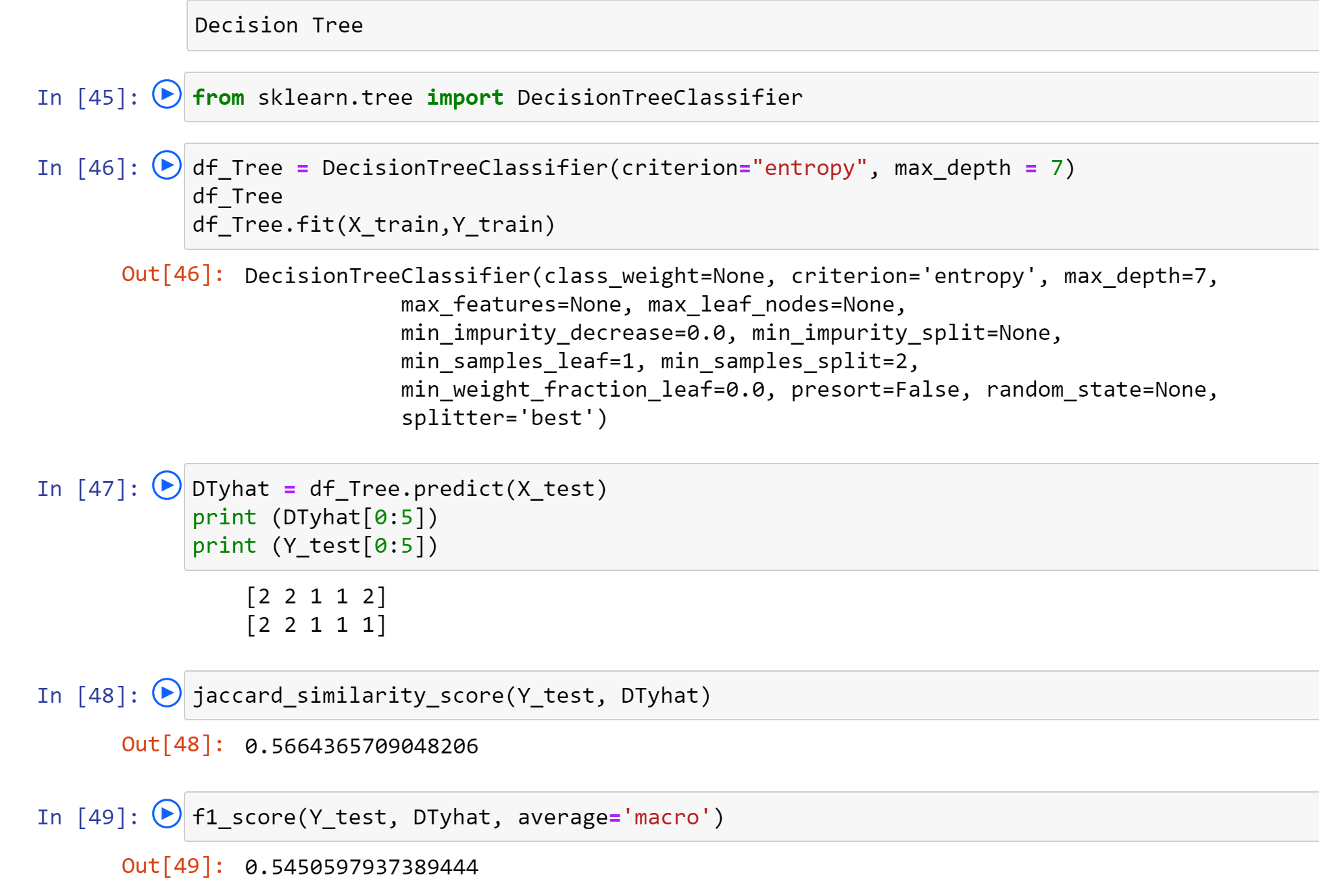
Initialization

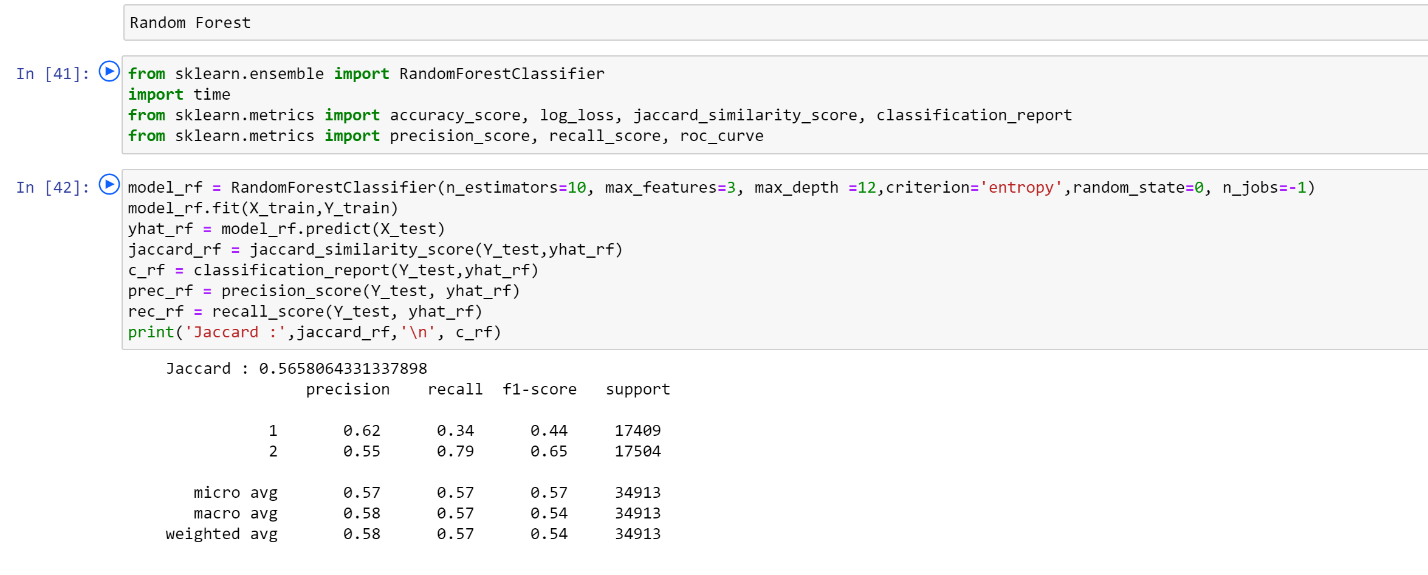
Define X and Y













## Result

| **Algorithm** | **Jaccard** | **f1-score** |
| --- | --- | --- |
| **K-Nearest Neighbor** | 0.564 | 0.540 |
| **Decision Tree** | 0.566 | 0.545 |
| **Random Forest** | 0.659 | 0.565 |
| **Logistic Regression** | 0.526 | 0.511 |

With no doubt the Random Forest is the best model.

## Discussion

* After uploading data from the provided data [csv file ] successfully, we determined the attributes (columns) that we will use to train your machine learning model.
* We cleaned the data by dropping the columns containing unrelated data.
* Data of type 'object' could be fed through an algorithm, so label encoding was used to created new classes that were of type int8; a numerical data type. We mapped the categorical data that was of type 'object' to int8.
* Next issue identified with data was - imbalanced data. Class 1 of SEVERITYCODE was nearly three times larger than class 2. The solution to this was downsampling the majority class. We used sklearn's resample tool to downsample the data to match the minority class exactly with 58188 values each.
* Once we analyzed and cleaned the data, it was then fed through four ML models; K-Nearest Neighbor, Decision Tree, Random Forest and Logistic Regression. Evaluation metrics used to test the accuracy of our models were jaccard index, f-1 score and logloss for logistic regression. Choosing different k, max depth and hyperamater C values helped to improve our accuracy to be the best possible.

## Conclusion

I analyzed the relationship between severity of an accident and other factors such as weather condition, road condition and light condition from data collected from Traffic Records. I used 4 different classification models to study & predict whether an accident would have a high or low severity ijn certain conditions.

Considering Jaccard value = 0.659 and f1-score = 0.565, Random Forest is the most suitable model for this scenario.

These models can have an important application in real life such as, predicting the Severity of an accident in real time, so that corrective measures can be taken in time to avoid them.