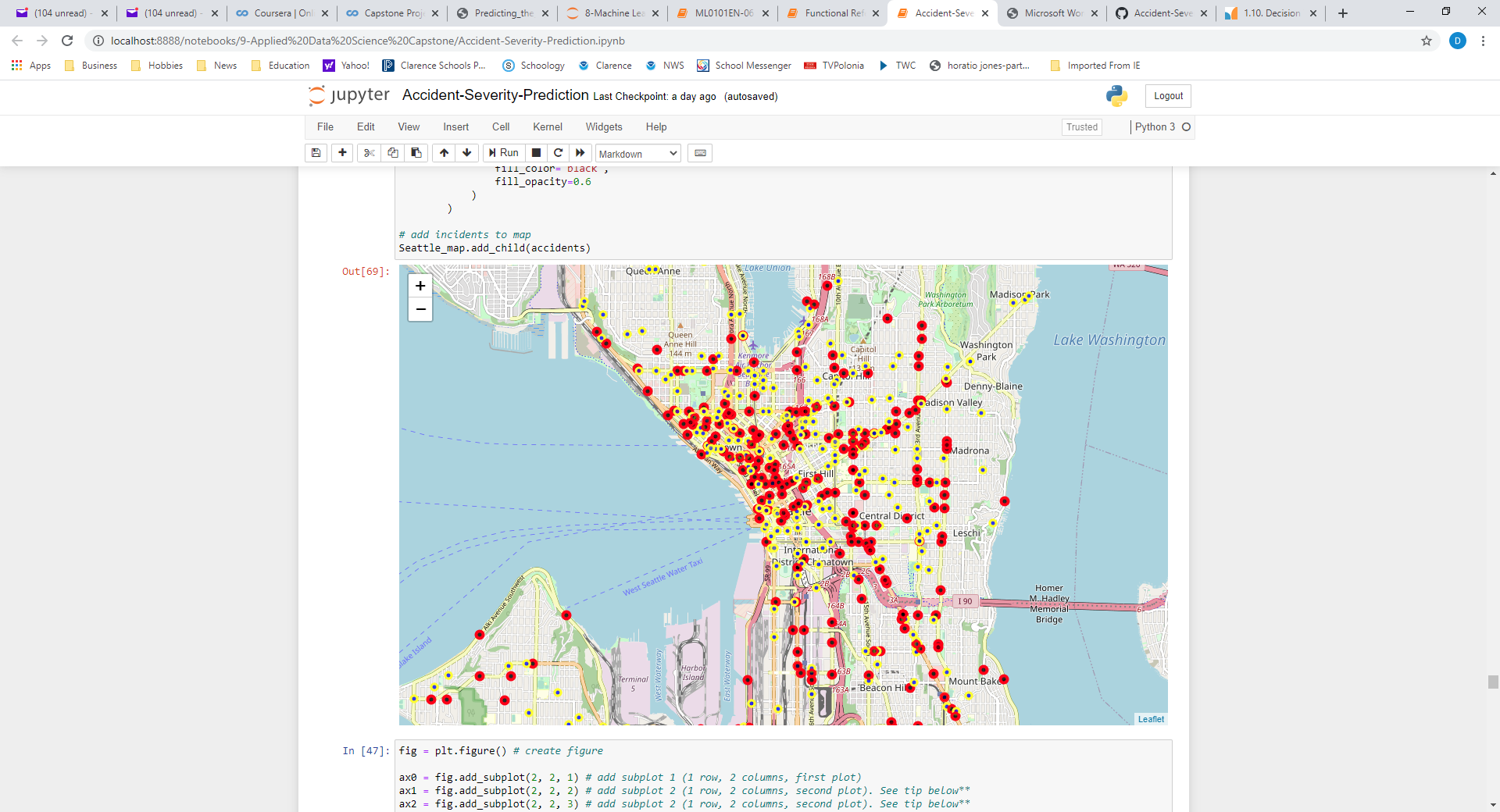
**Predicting the Severity of Automobile Accidents**

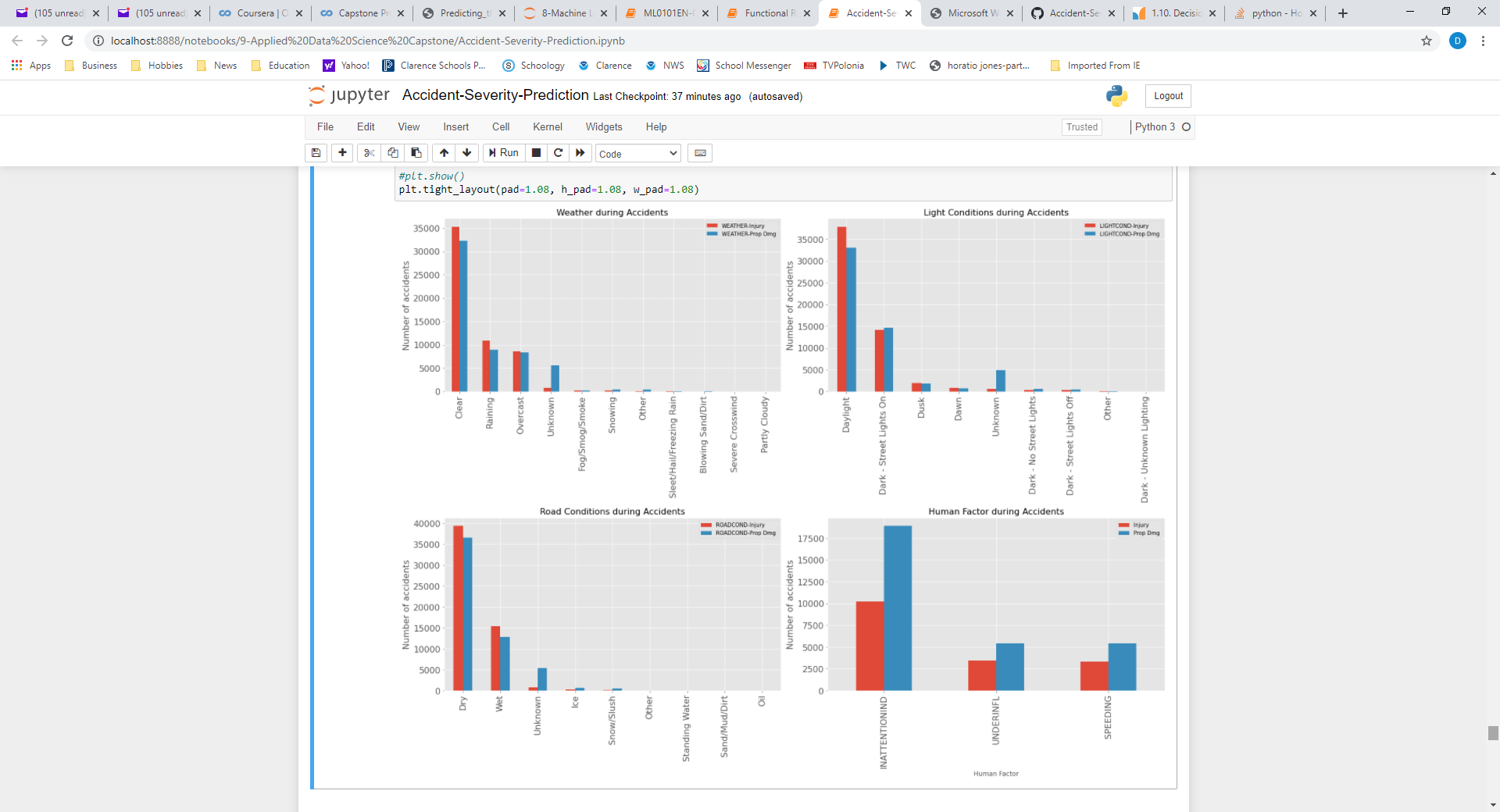
Daniel T. Pawlak

September 20, 2020

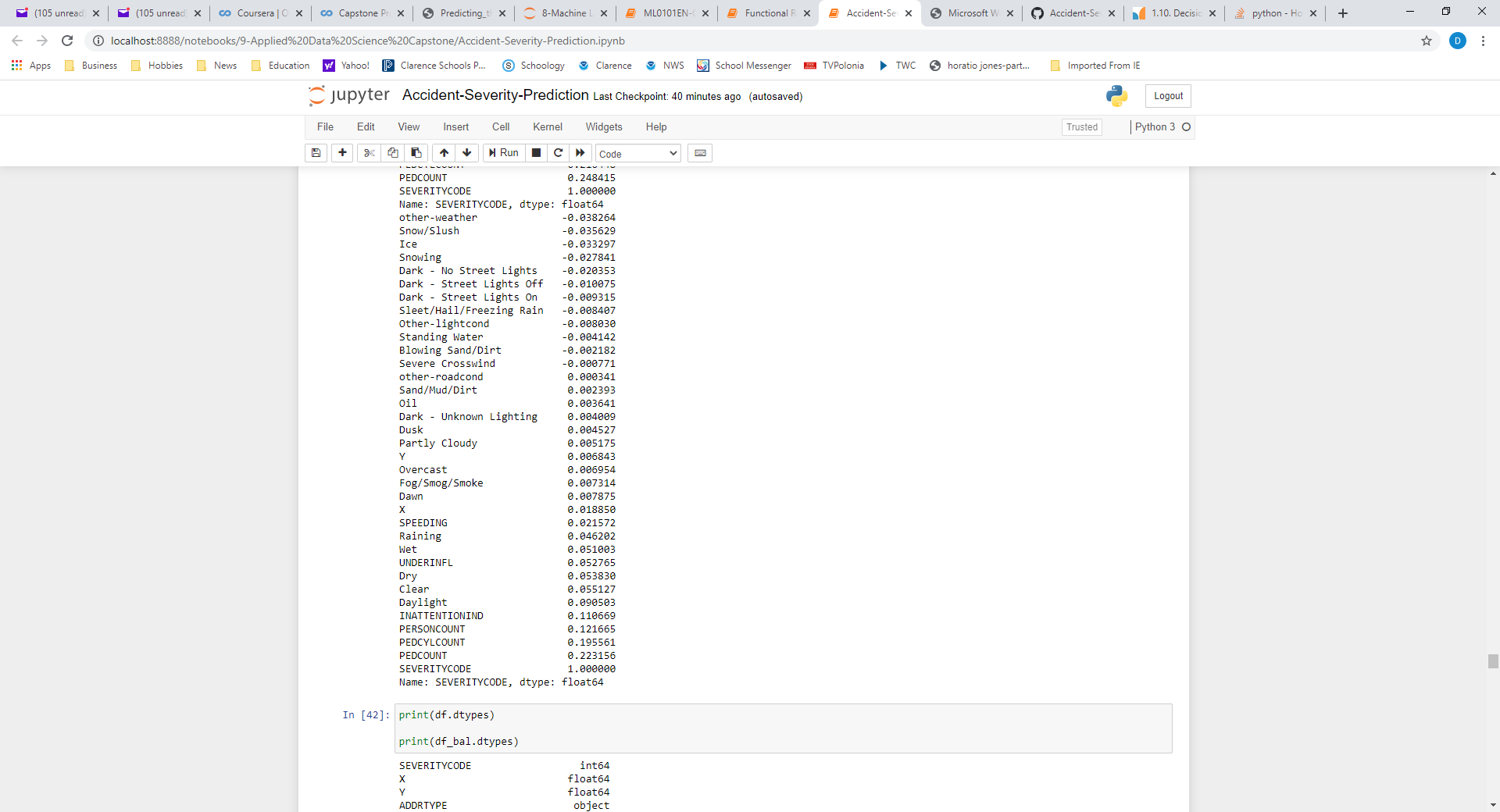
1. **Introduction:**
   1. **The Business Problem:** The focus of this project is the development of a model that can predict the severity of traffic accidents, given weather and road conditions. For this project, severity is one of two categories: either property damage or injury.
   2. **Key Stakeholders**: The immediate stakeholders of the project are public safety officials hoping to warn drivers of the potential severity of an accident to get them to drive more safely or change their travel plans.
2. **Data:**
   1. **The data that was used to solve the problem:** 
      1. The dataset is rich, with more than 150k observations and many attributes. Beyond the weather and road conditions, there are other variables which seem likely to contribute to the severity of an accident such as light conditions, whether the driver was speeding, distracted, or under-the-influence of a substance.
      2. Key to the development of the model is the fact that the severity label indicates the severity of the observed accidents as either fatal or property damage.
      3. Data preparation required the handling of missing data, as well as re-structuring certain attributes such as weather conditions. As an example, weather exists in the dataset as a single attribute with several categories. This was re-structured using indicator variables to correlate each separate weather condition with the severity of an accident. The same was done with other categorical variables. The original dataset is also unbalanced with regards to the labeled observation. That is, there are far more observed occurrences of accidents with property damage than there are with injuries. The dataset was balanced to remove a source of bias in the predictive model.
   2. **The source of the data:** 
      1. The data used for the testing and development of the predictive model was the shared dataset compiled by the Seattle Department of Transportation, Traffic Management Division, Traffic Records Group.
      2. All collisions provided by SPD and recorded by Traffic Records. Timeframe: 2004 to 05/2020
      3. A more detailed description of the data may be found at: <https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf>
3. **Methodology section:**
   1. **Exploratory data analysis.** This was done to better understand the dataset and how the data could be used to develop a predictive model.
      1. Data shape. The original dataset has 194,673 rows with 38 columns.
      2. Attributes. Of the 38 columns, some are explicitly mentioned in the project, such as severity code, weather, road condition, and light condition. Other factors likely to be predictive of injury or not include human conditions, such as inattention (distracted driver), under-the-influence, or speeding. Also, attributes that specify the number of people involved in an accident, number of pedestrians, and number of cyclists seem likely to be correlated with likelihood of injury. Other attributes are merely descriptive and therefore unlikely to add any predictive value.
      3. Data Types. Although there are some floats and int variables, many of the attributes are of type object. Of the attributes to be used for the development of the predictive model, the object variables were generally converted to int, since they were mainly 1’s and 0’s.
      4. Missing Values. There were many missing values indicated by nan. Some of these were eliminated by deleting an entire row, such as if one of the key variables were missing. Other nan’s were replaced by appropriate values. This was especially necessary when using indicator variables to convert from categorical attributes.
      5. Unique Values. Identifying the unique values of the attributes helped to understand which variables were categorical and which were numerical.
      6. Spatial plot of sample by severity. A sample plot of the accidents was created to determine whether any obvious patterns could be detected visually.



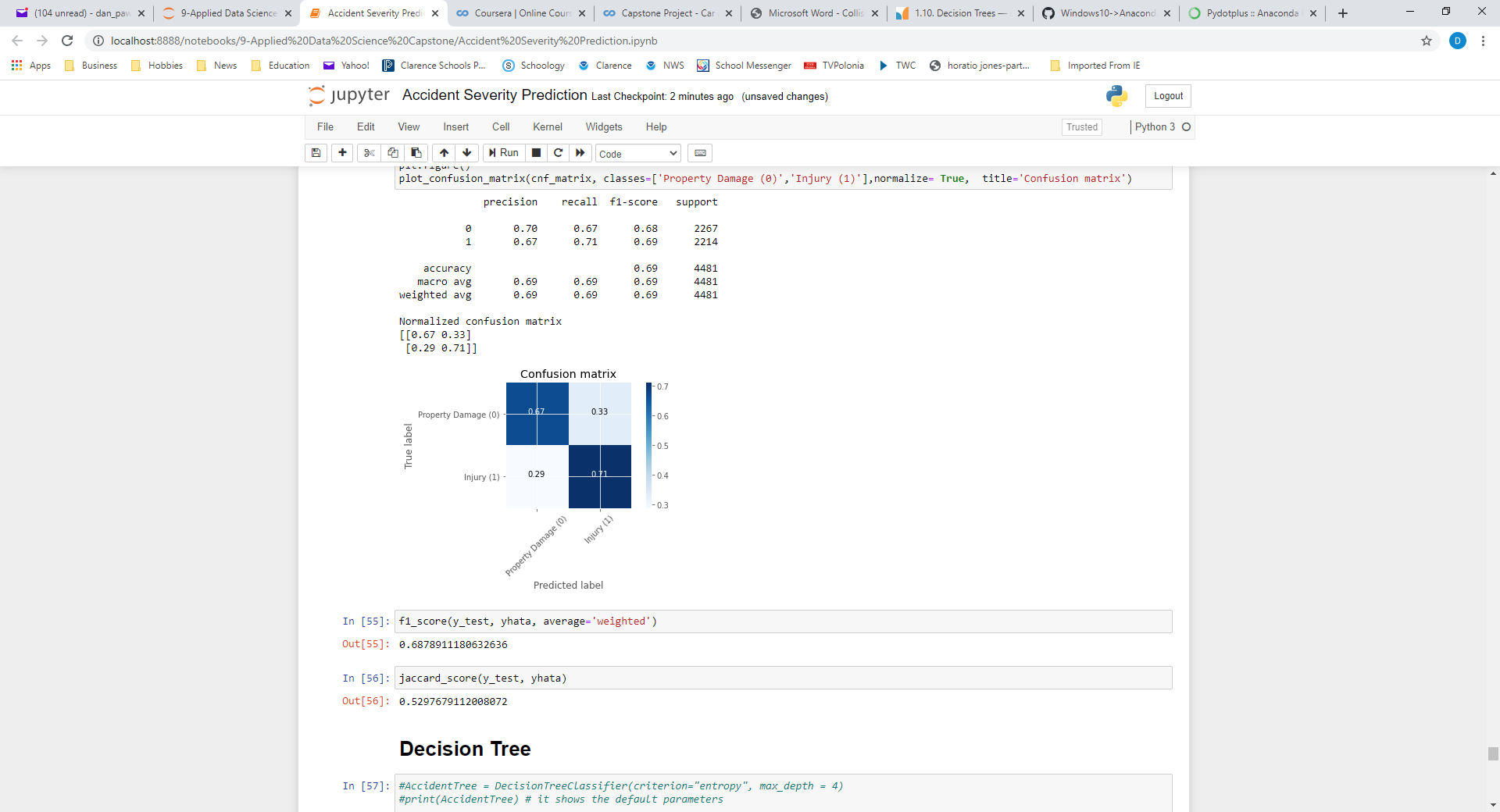
* + 1. Bar graphs of key independent variables were created to see the number of positive occurrences of the attributes grouped by severity code.

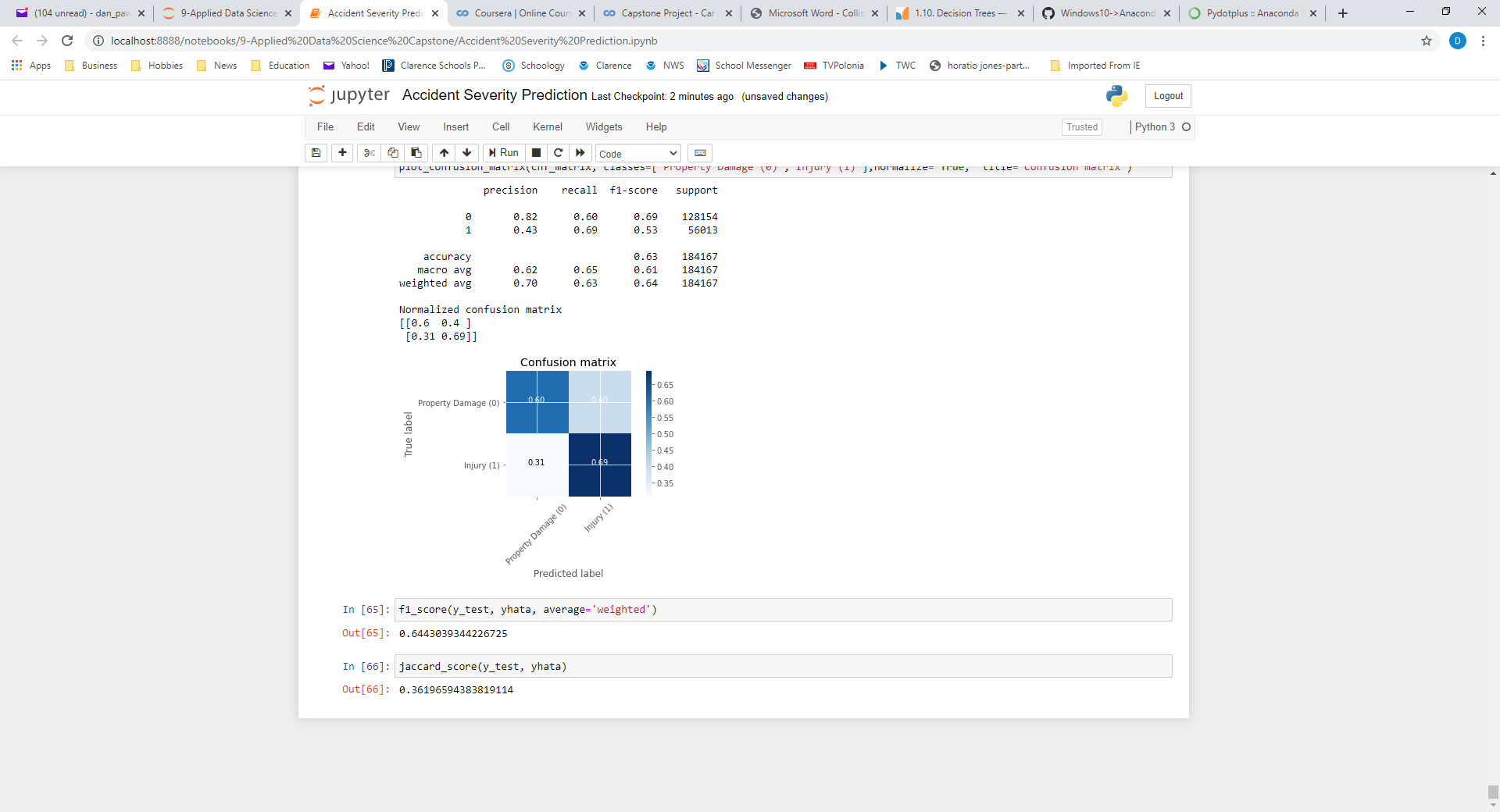


* 1. **Inferential statistical testing**
     1. Pearson correlation (df.corr) was used to determine which variables had the greatest likelihood of predicting injury vs property damage. When the dataset was unbalanced all the correlations were quite low (most less than 5%). Even with the dataset balanced, most attributes were only weakly correlated, but the attributes most strongly correlated were the counts of people involved (approx. 20%), Inattention (nearly 10%), Under-the-Influence, Wet, and Raining (near 5%). Other variables with high correlation with Injury include daylight, dry, and clear weather conditions (5-10%), but that’s because those are the most common weather conditions in which most accidents of all kinds occur.

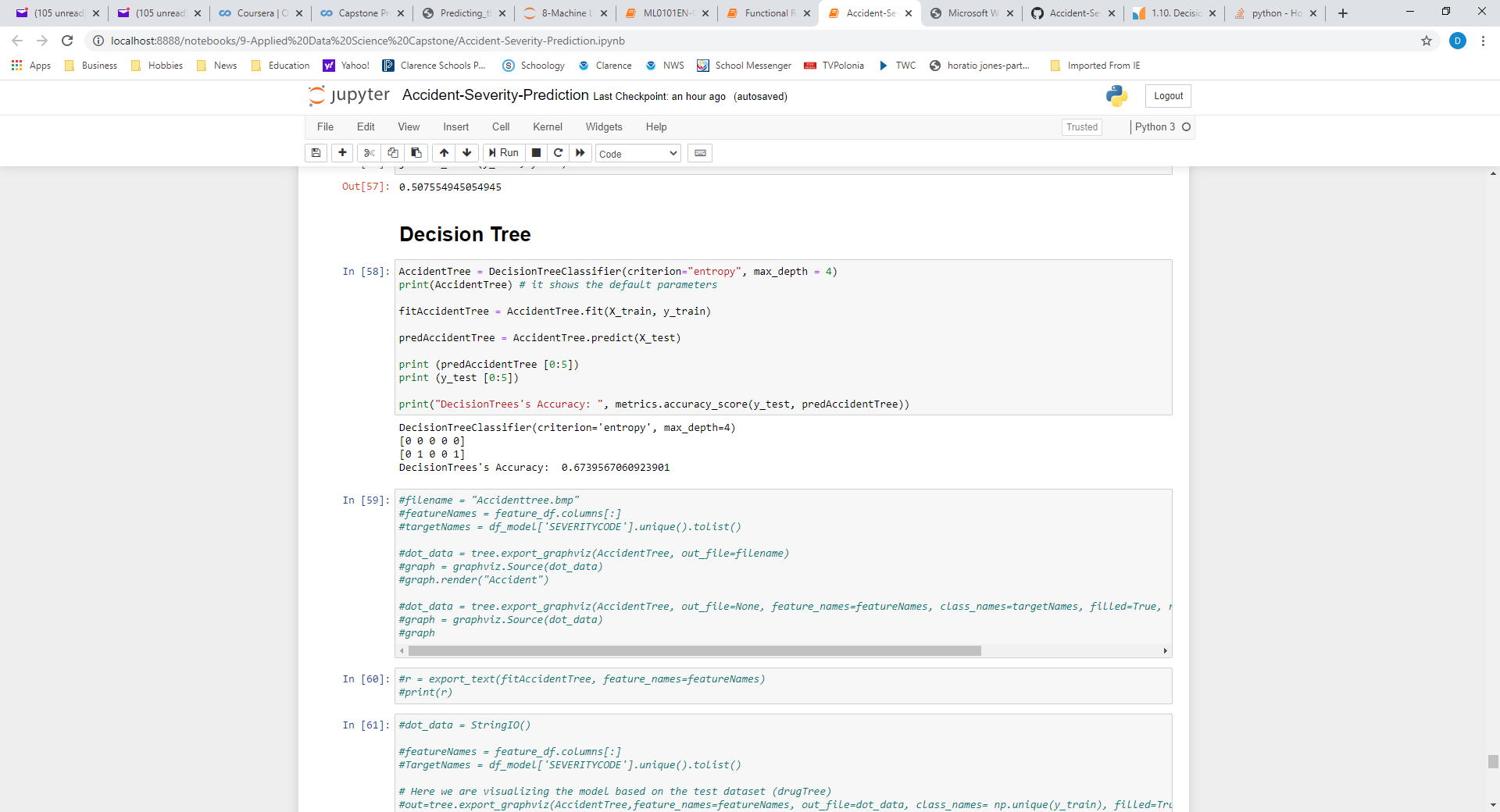


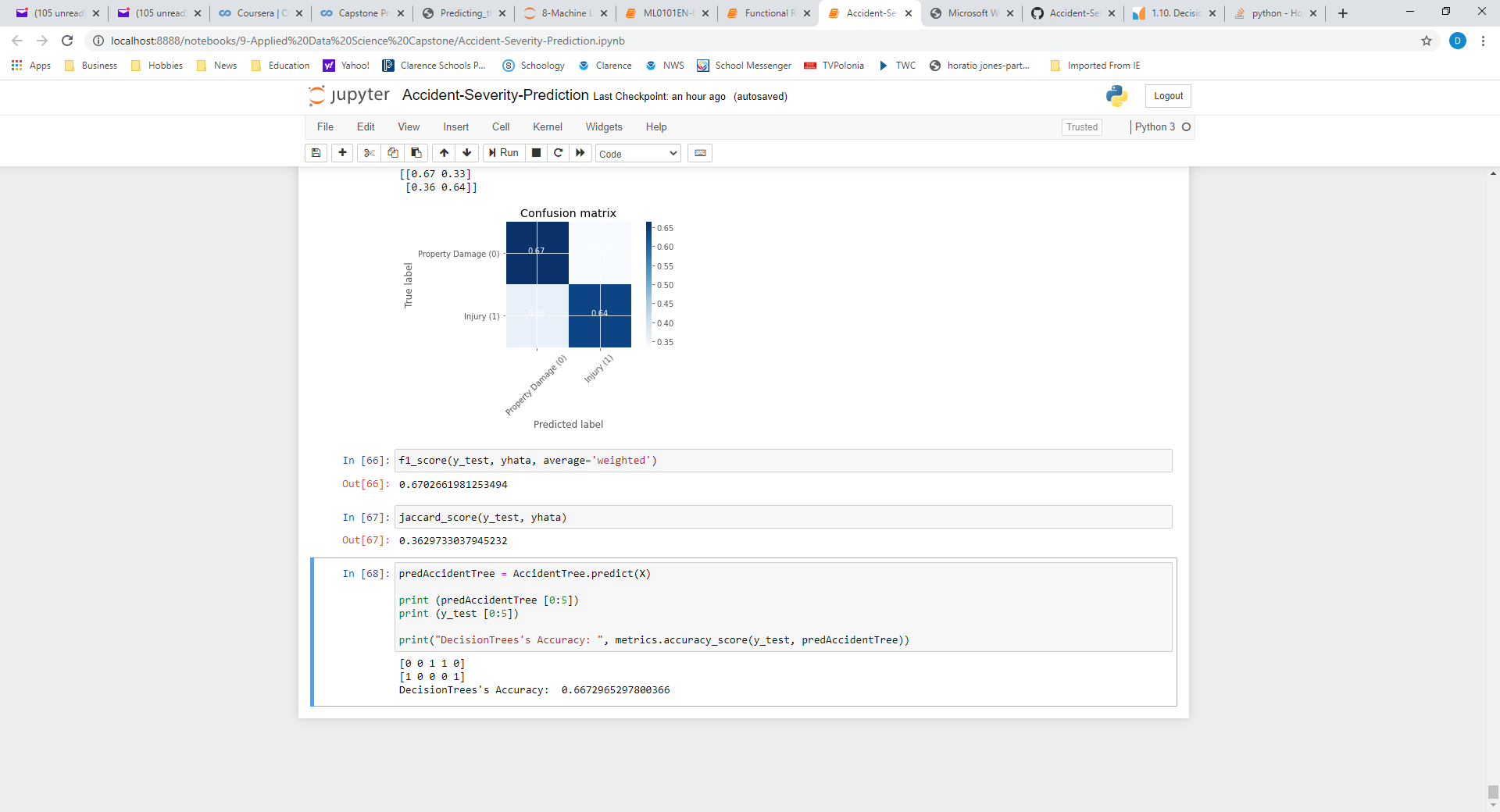
1. **Machine learning algorithms**
   1. SVM was used because SVM works by mapping data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. That way characteristics of new data can be used to predict the group to which a new record should belong. The SVM model was also run with four different kernels (linear, rbf, poly, and sigmoid) to determine which provided the most accurate results.
   2. Decision Tree was also used because the goal of a decision tree is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. If a given situation is observable in a model, the explanation for the condition is easily explained by Boolean logic.
   3. For each model algorithm the model was trained with a balanced dataset with 56,013 observations (That’s how many Severity = Injury observations there were after processing), and tested with the full unbalanced dataset.
      1. Of the full balanced dataset, 20% (22,405 observations) was used for train/test to avoid overtraining the model.
      2. Of the train/test dataset 20% (4481 observations) was used for test.
      3. For final model evaluation, the full unbalanced dataset (184,167 observations) was used as the test dataset.
2. **Results section:**
   1. For the SVM, the kernel that provided the most accurate results (determined by highest sum of f1 and Jaccard scores) overall was the rbf with an f1 score of .688 for the balanced test sample and .644 for the full unbalanced set. The Jaccard score was .530 for the balanced test sample and .362 for the full unbalanced set. The rbf kernel also predicted the best combination of highest true positives and lowest false negatives in the confusion matrix.





* 1. The Decision Tree had an accuracy of .674 for the balanced sample dataset and .667 for the full unbalanced dataset.





1. **Discussion section:** 
   1. I was surprised that the bad weather variables didn’t have as strong a correlation with injury as I imagined, especially snow, ice, and weather limiting visibility. The correlation might be strengthened by combining some attributes that reduce visibility (such as fog and darkness) or create slippery conditions (such as rain, ice, snow, and slush).
   2. This model development did not include accident location as predictive variables, although accidents occurring at intersections may have a greater likelihood of injury. This is an opportunity for further testing and potentially increased model accuracy.
   3. This model also did not consider the time of day or day of week of an accident, although those things may also add predictive value. More accidents may occur at morning or evening rush hour when the roads are nearer to capacity, or may occur later at night when people are more likely to be tired. Weekends versus weekdays may also be important. These considerations are an opportunity for possible model improvement.
   4. Although the Decision Tree model was run successfully, technical difficulties prevented the visualization of the tree. The difficulties were associated with StringIO() and graphviz, and the model developer was not able to overcome the problems at the time of this report.
2. **Conclusion section:**
   1. The model developed for this project can predict the severity of traffic accidents, given weather and road conditions and other attributes. The model does have predictive capability sufficient to provide useful information to stakeholders that may be passed on to drivers to get them to drive more safely or change their travel plans.
   2. Although the model has predictive capability, this report identifies additional areas to continue model improvement.