

**RATING PREDICTION PROJECT**

**Submitted By:**

**Lakshmi Rajendra Thute**

**ACKNOWLEDGEMENT**

I am very much Thankful to FlipRobo Technologies for giving me the opportunity to work with them and to work on this project and also, I am very grateful to Data Trained Education Team for their support and help to understand each and every concept of machine learning which helped me a lot while working on this project. I thought, I am fortunate to become a part of FlipRobo Technology.

References:

Google website

Stack overflow

Analytics Vidya

Medium

Data trained notes

**INTRODUCTION**

We have a client who has a website where people write different reviews for technical products. Now they are adding a new feature to their website i.e. The reviewer will have to add stars(rating) as well with the review. The rating is out 5 stars, and it only has 5 options available 1 star, 2 stars, 3 stars, 4 stars, 5 stars. Now they want to predict ratings for the reviews which were written in the past and they do not have a rating. So, we must build an application which can predict the rating by seeing the review.

We must scrape at least 20000 rows of data. We can scrape more data as well, more the data better the model in this section we need to scrape the reviews of different laptops, Phones, Headphones, smart watches, Professional Cameras, Printers, Monitors, Home theatre, Router from different ecommerce websites. Basically, we need these columns1) reviews of the product. 2) rating of the product. We can fetch other data as well.

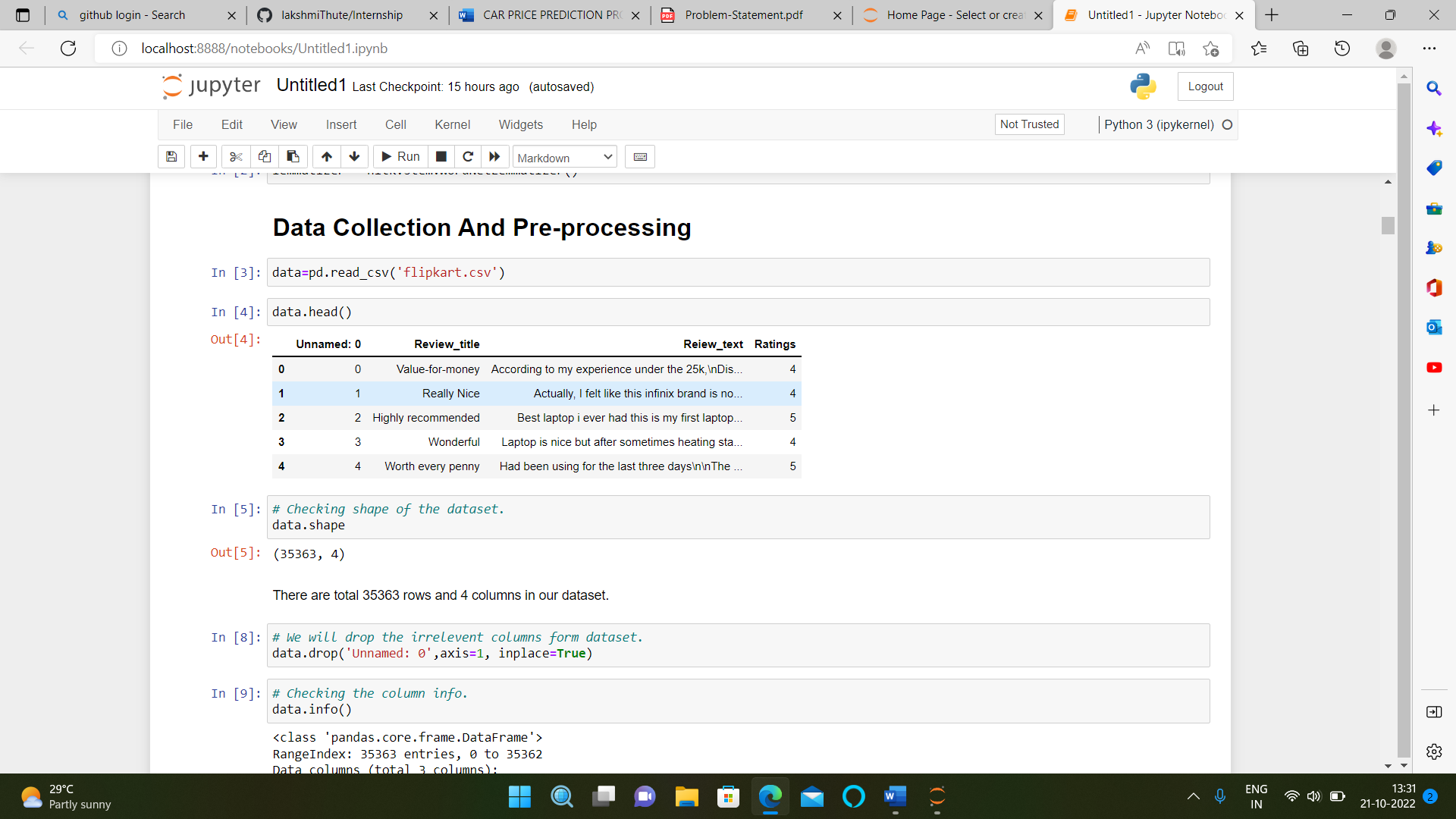
**ANALYTICAL PROBLEM FRAMING**

**• Mathematical/ Analytical Modelling of the Problem**

Here I have scraped the ratings data from different e-commerce websites then did the Data Pre-processing using NLP, Exploratory Data Analysis, then Encoding and lastly model Building and Evaluation.

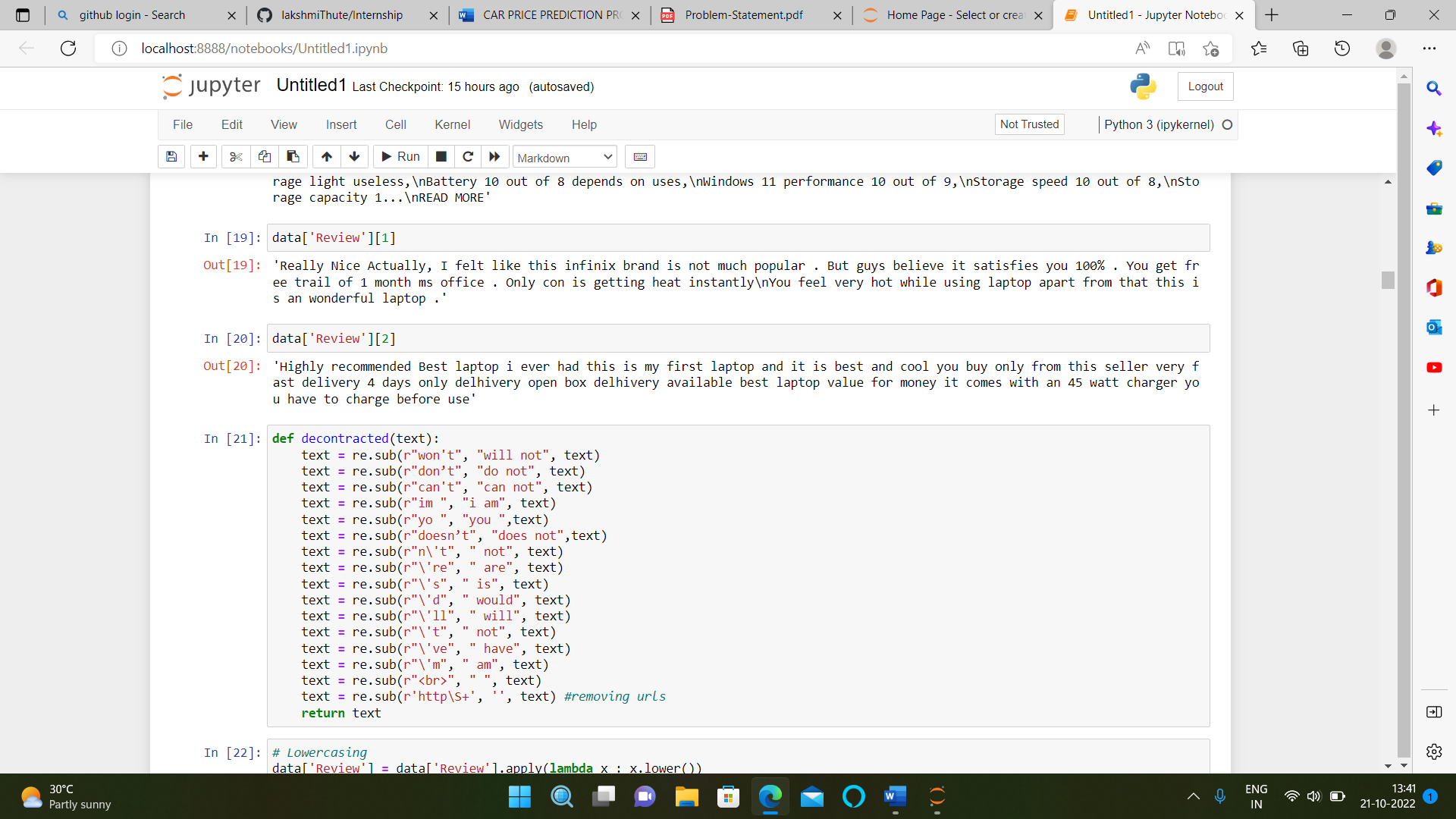
**• Data Sources and their formats**

I got the dataset in CSV format, and I read the data in Jupyter Notebook using pandas data frame.



**• Data Pre-processing Done**

The dataset contains 35363 rows and 4 columns, The pre-processing is done using NLP.



**• Hardware and Software Requirements and Tools Used**

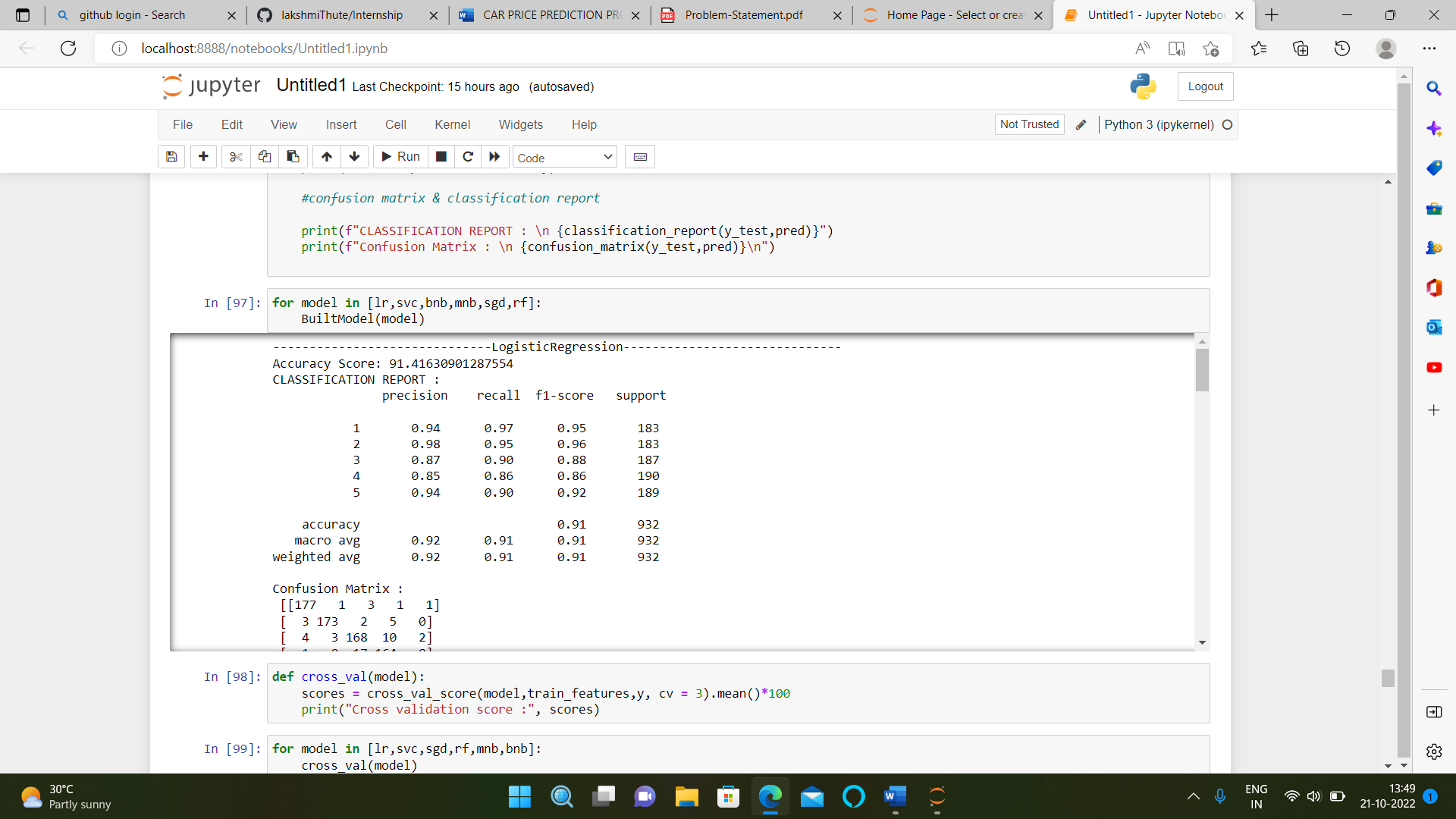
Here for this project, I used Jupyter notebook and tools used pandas and NumPy for mathematical operations, matplotlib and seaborn for various type of data visualizations.

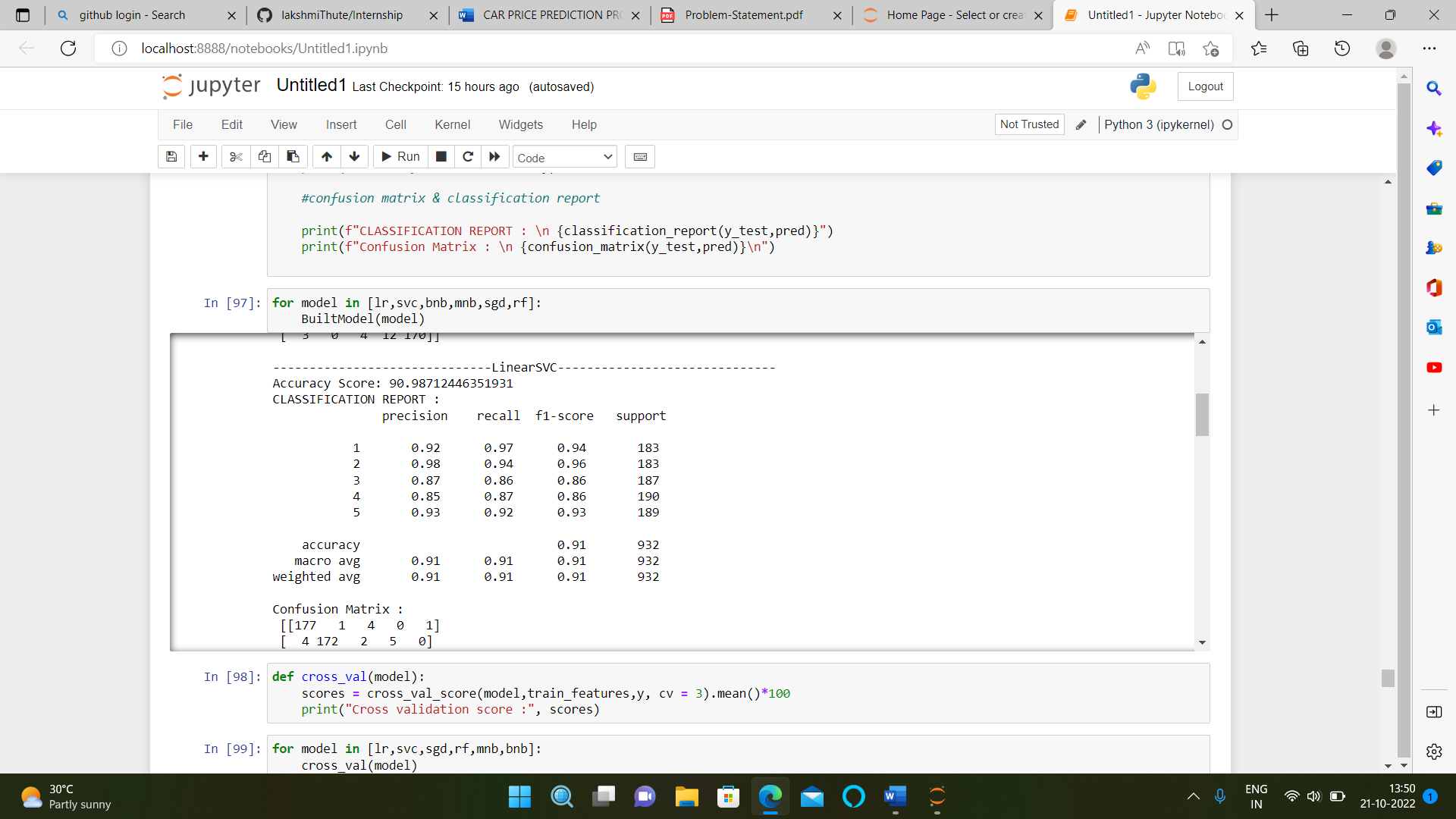
**• Identification of possible problem-solving approaches (methods)**

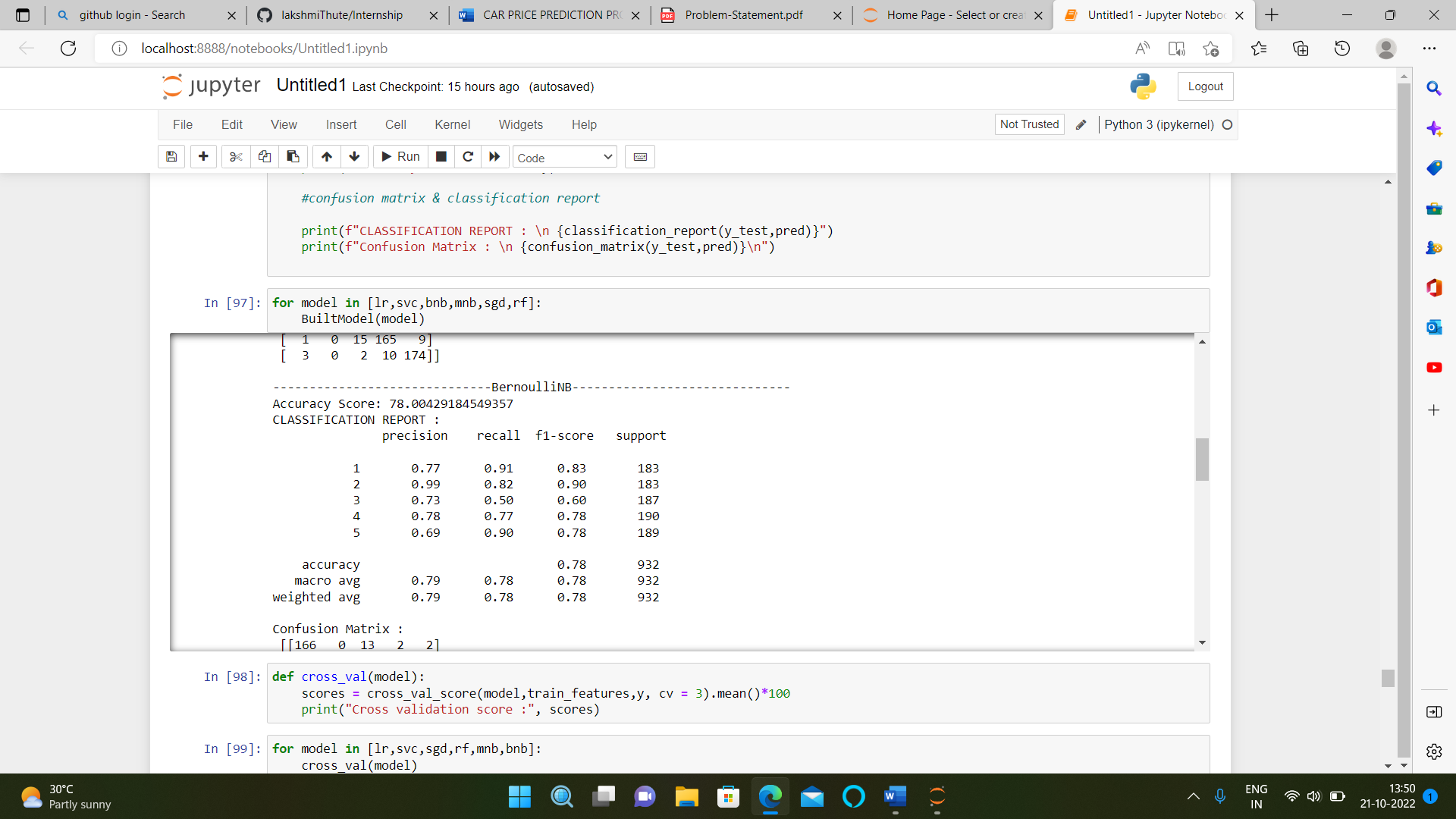
The statistical summary shows the total count of 35363 rows then mean, min value, max value, standard deviation, and quartiles shows up and down values that means the data contains outliers.

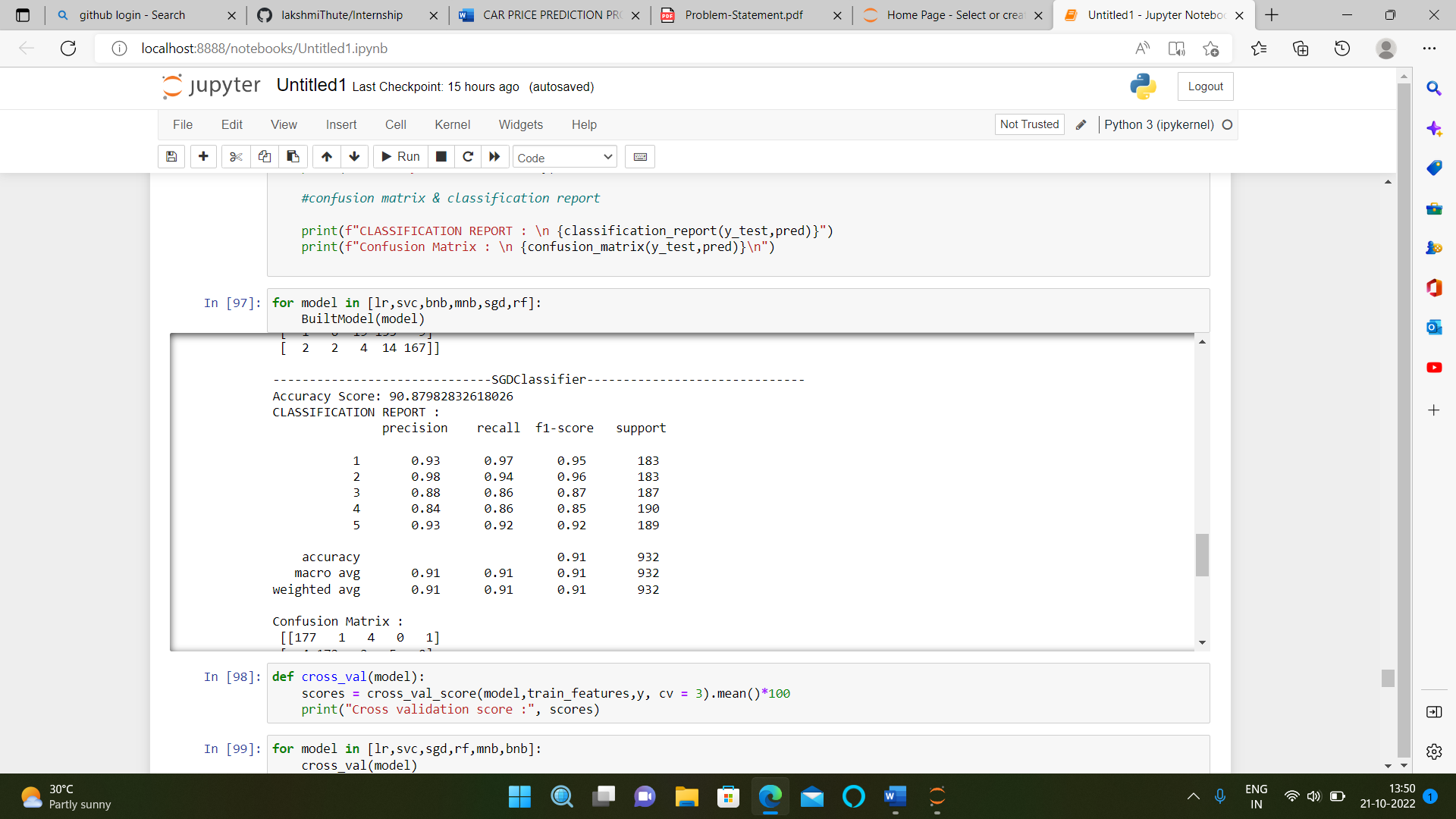
**• Testing of Identified Approaches (Algorithms)**

* Random Forest Classifier
* Linear SVC
* Logistic Regression
* MultinomialNB
* BernoulliNB
* LGBM Classifier
* SGD Classifier
* **Run and evaluate selected model**





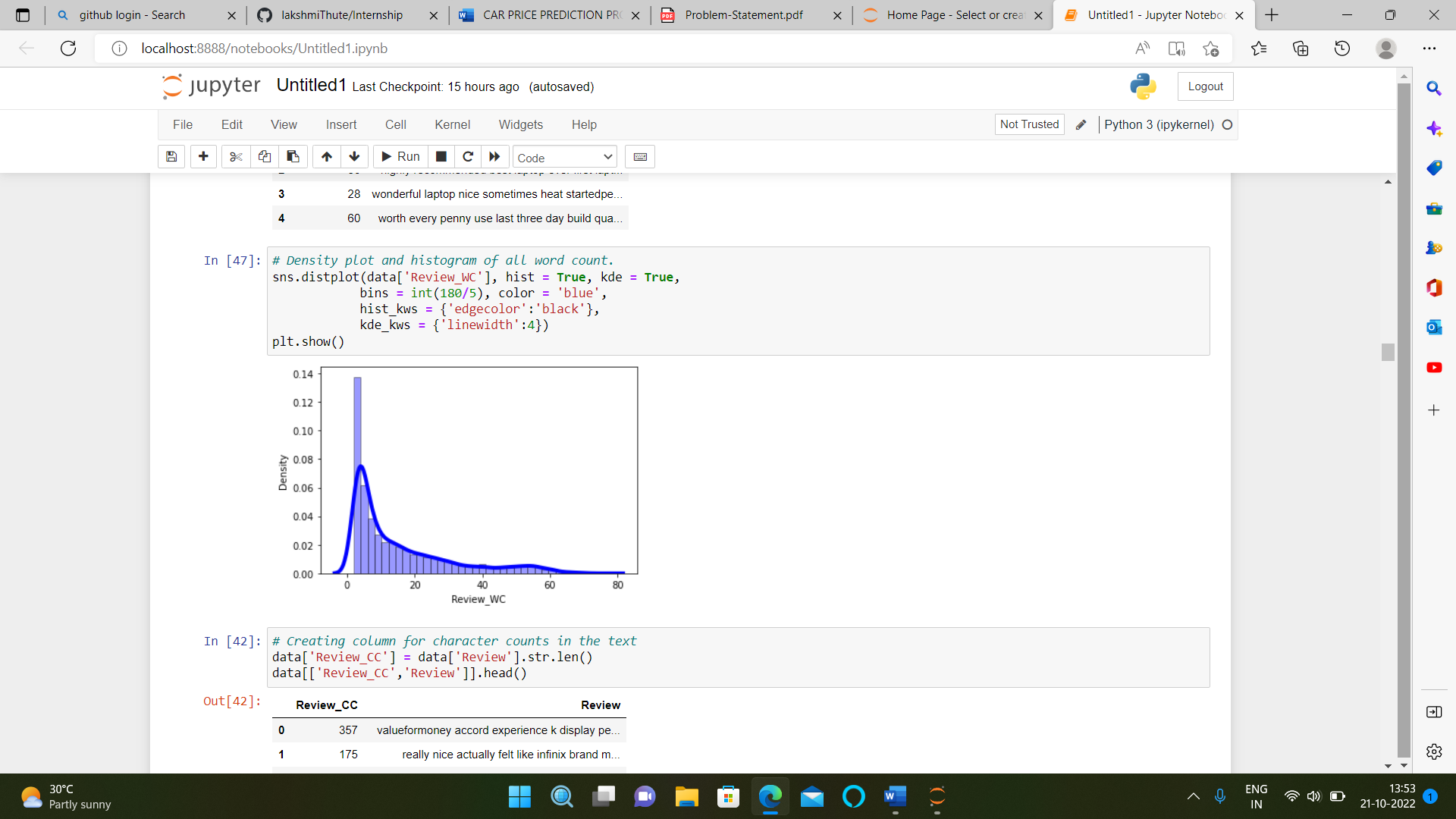


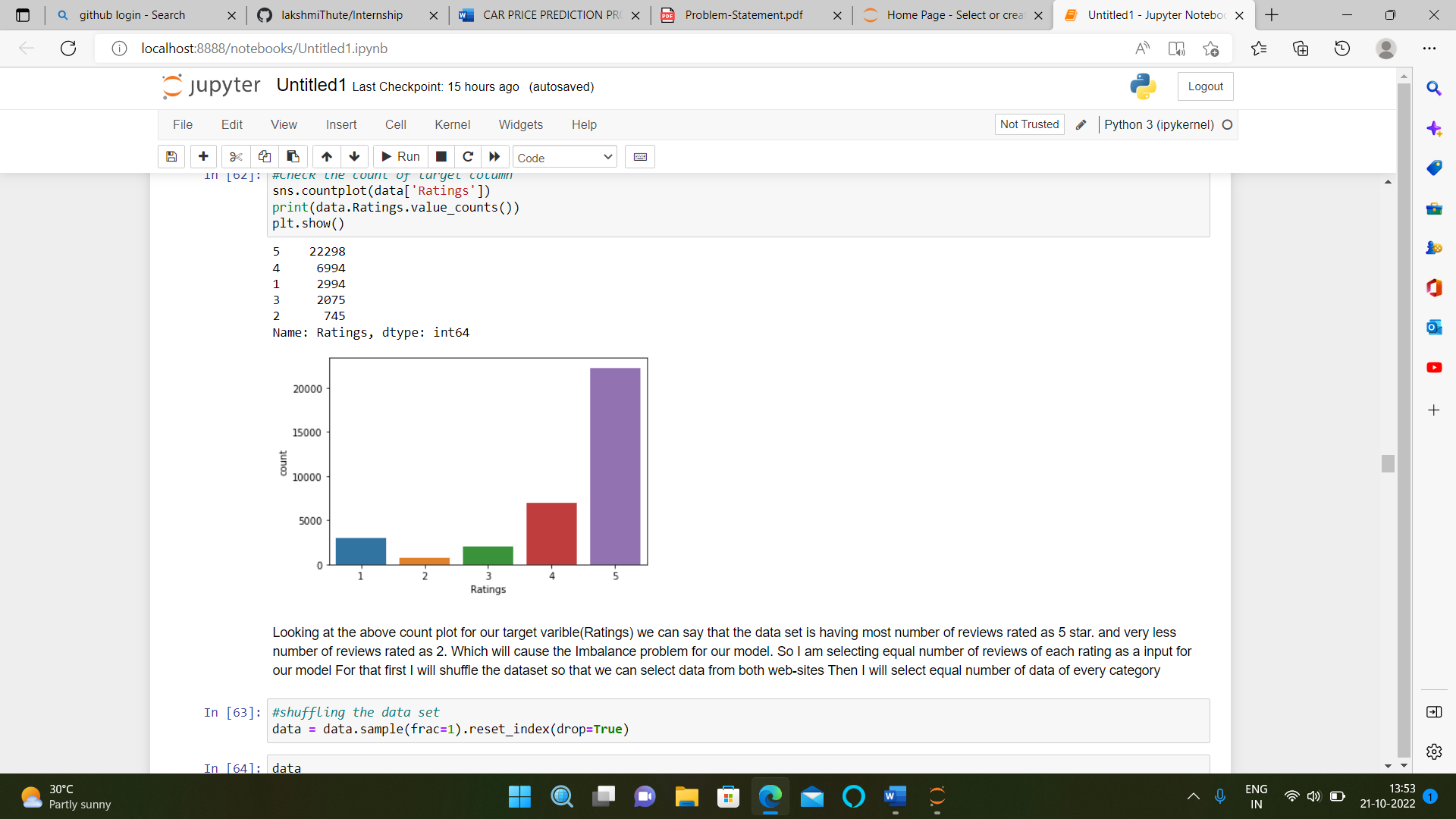


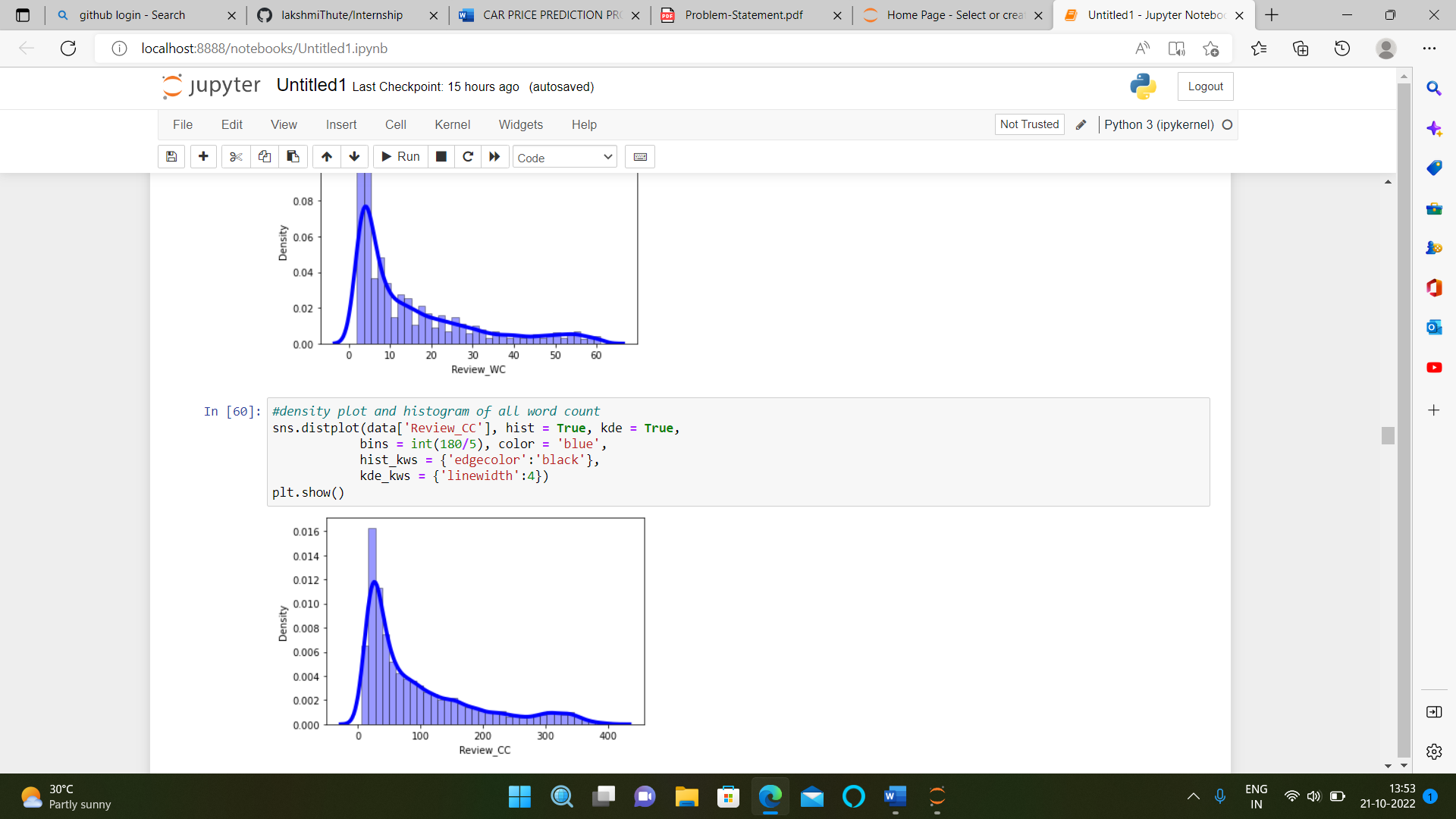
**• Key Metrics for success in solving problem under consideration**

A company's units can use these dashboards to create milestones and monitor their progress by tracking all the most relevant metrics in one location.

* **Visualizations**







**Observations:**

Looking at the above count plot for our target variable (Ratings) we can say that the data set is having the greatest number of reviews rated as 5 stars. and very a smaller number of reviews rated as 2. Which will cause the Imbalance problem for our model. So I am selecting equal number of reviews of each rating as a input for our model For that first I will shuffle the dataset so that we can select data from both web-sites Then I will select equal number of data of every category.

**• Interpretation of the Results**

Here after pre-processing we get the data encoded for framing the model and after visualization, we observe that the data does not contain skewness. After checking correlation, we observe that the dataset doesn’t contain multicollinearity. After data pre-processing and EDA we build 6 different algorithms for dataset and from them Linear SVC algorithm perform very well, Lastly we perform hyper parameter tunning to enhance the accuracy. Here we finalise Linear SVC as our best fit model.