**LOAN PREDICTION**

**Submitted by**

Kolisetty Pushpak AM.EN.U4CSE17308

Tungala Yashwanth AM.EN.U4CSE17340

Yerramsetti Raghuram AM.EN.U4CSE17359

**1. PROBLEM STATEMENT**

Dream Housing Finance company deals in all home loans. They have presence across all urban, semi urban and rural areas. Customers first apply for a home loan after that company validates the customer eligibility for the loan.

The company wants to automate the loan eligibility process (real time) based on customer detail provided while filling an online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, they have given a problem to identify the customers segments, those are eligible for loan amounts so that they can specifically target these customers. Here they have provided a partial data set.

**Tungala Yashwanth - AM.EN.U4CSE17340**

**2. Dataset 1**

Dream Housing Finance company deals in all home loans. They have presence across all urban, semi urban and rural areas. Customers first apply for a home loan after that company validates the customer eligibility for the loan. Company wants to automate the loan eligibility process (real time) based on customer detail provided while filling an online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, they have given a problem to identify the customers segments, those are eligible for loan amounts so that they can specifically target these customers.

<https://www.kaggle.com/ninzaami/loan-predication>

**3. PREPARE DATA**

PRE-PROCESSING

* Find the null values
* Fill the missing values for numerical terms - mean
* Fill the missing values for numerical terms - mode
* Changing data from string to int

SUMMARIZATION

* Found dimensions and statistical summary
* Brokedown the data by class using group by

DATA VISUALIZATION

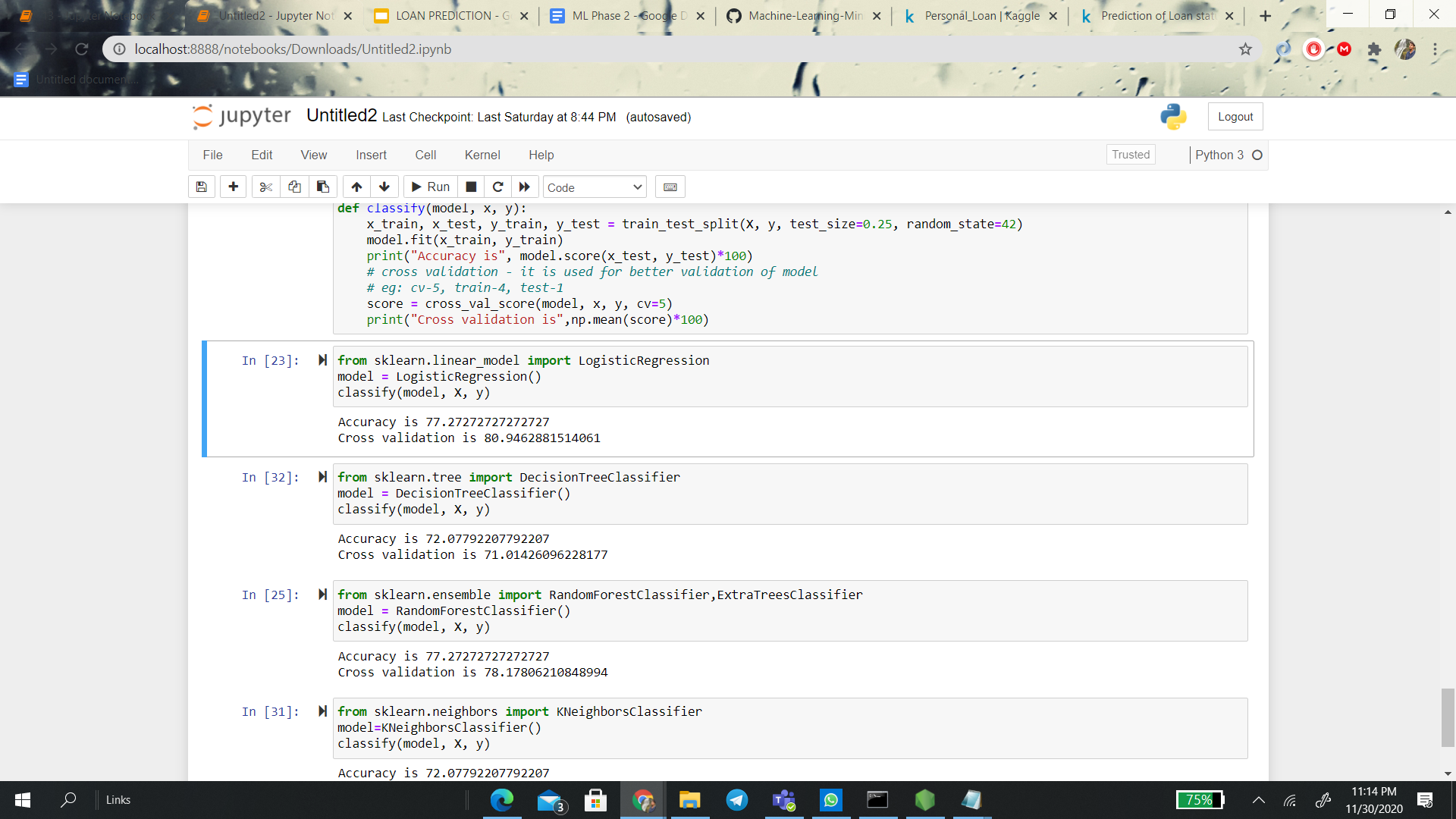
* Heat map
* Bar plot

**4. PYTHON PACKAGES**

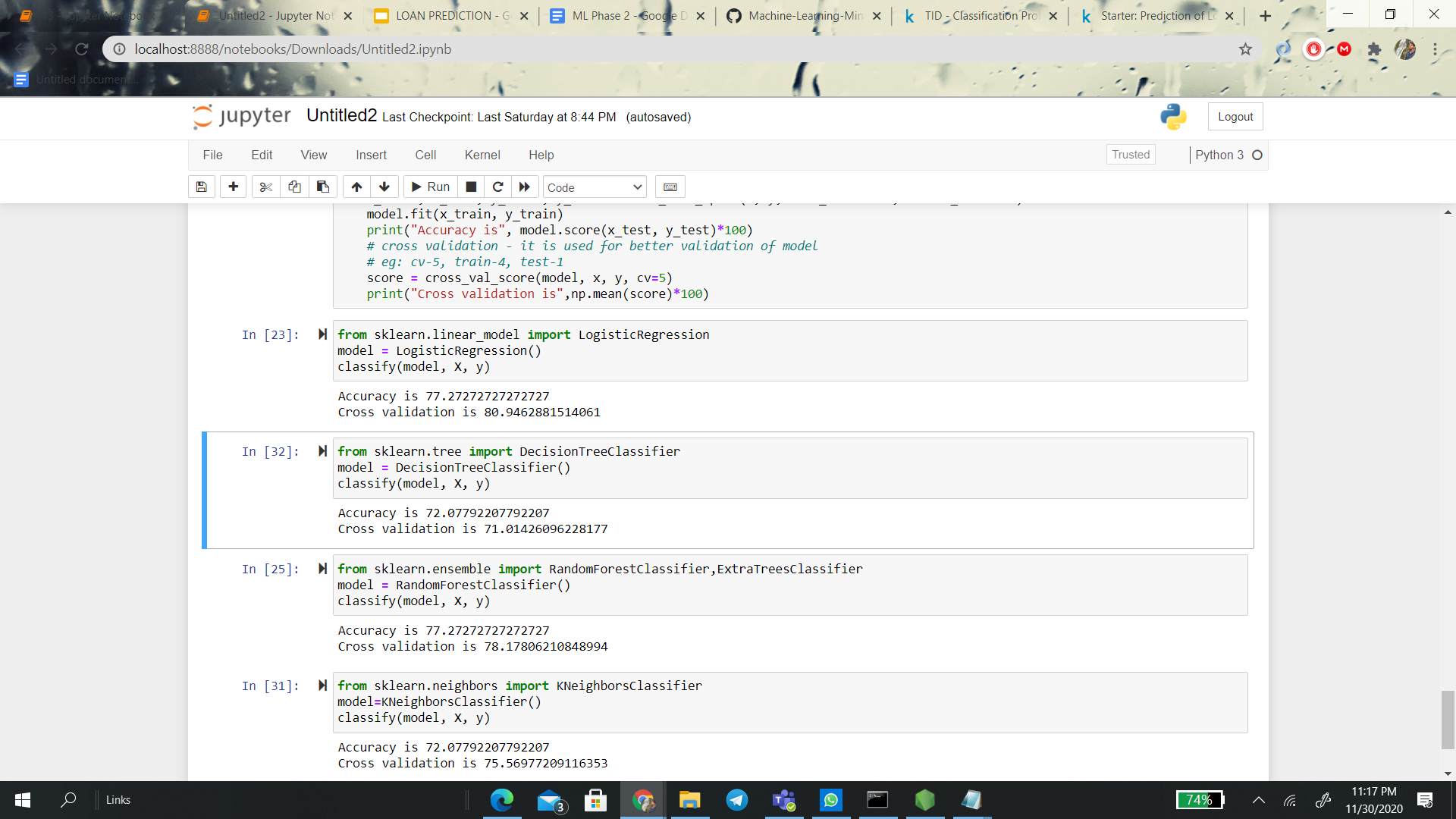
* pandas
* numpy
* seaborn
* matplotlib
* pyplot
* warnings
* sklearn for label encoding, model training, hyperparameter tuning, confusion matrix, logistic regression, decision tree, KNN

**5. SUPERVISED LEARNING ALGORITHMS**

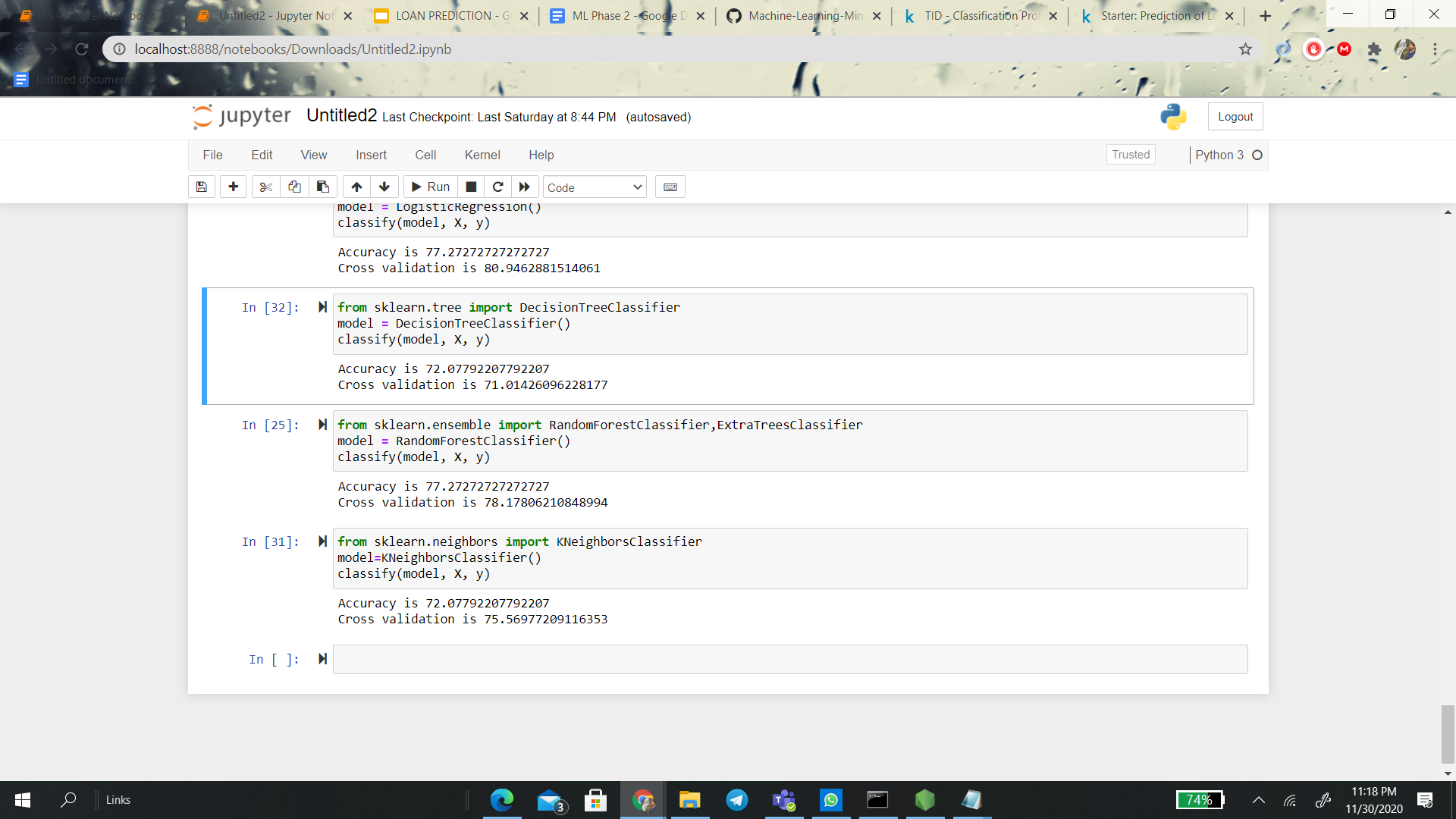
Logistic Regression



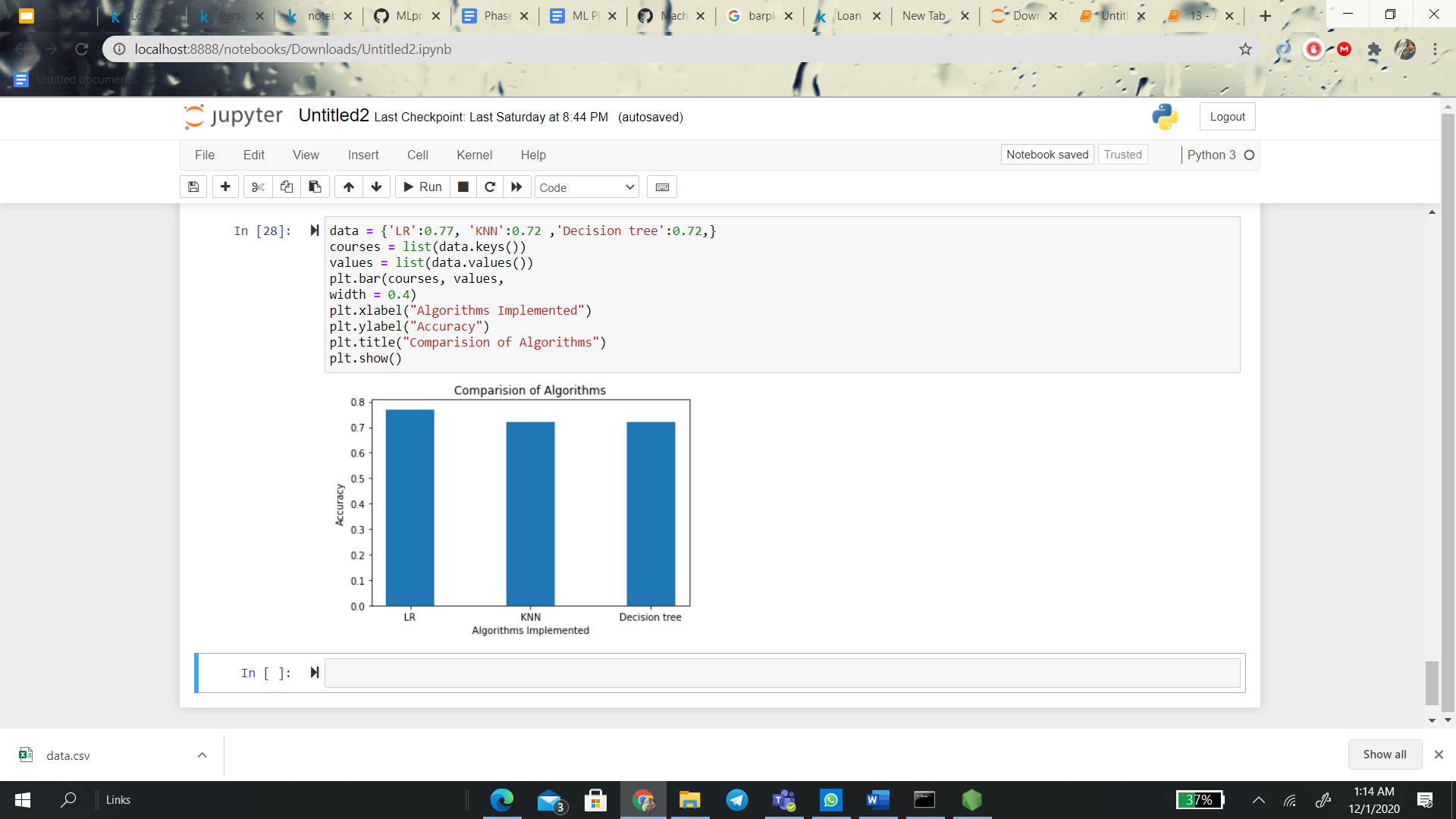
Decision Tree



KNN



**ALGORITHM COMPARISON**



After comparing the accuracies of all three algorithms, Logistic Regression is more accurate among three algorithms

**Kolisetty Pushpak - AM.EN.U4CSE17308**

**2. Dataset 2**

Among all industries, the insurance domain has the largest use of analytics & data science methods. This data set would provide you enough taste of working on data sets from insurance companies, what challenges are faced, what strategies are used, which variables influence the outcome etc. This is a classification problem. The data has 347 rows and 10 columns.

[**https://www.kaggle.com/ritikmeena/prediction-of-loan-status**](https://www.kaggle.com/ritikmeena/prediction-of-loan-status)

**3. PREPARE DATA**

PRE-PROCESSING

# Convert Categorical features to numerical values

* Checked null value

SUMMARIZATION

* Found dimensions and statistical summary
* Brokedown the data by class using group by

DATA VISUALIZATION

* Bar plot

**4. PYTHON PACKAGES**

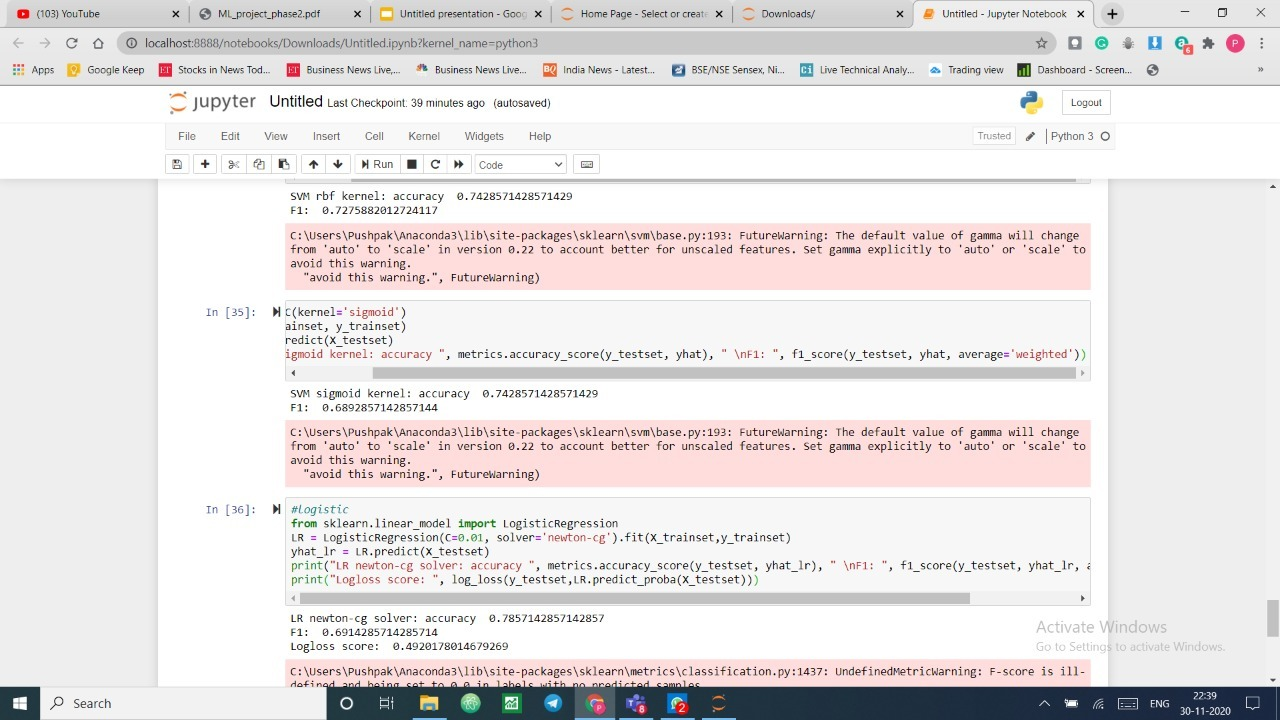
* Itertools
* numpy
* matplotlib.pyplot and matplotlib.ticker
* NullFormatter
* pandas
* seaborn
* sklearn for one hot encoding, feature selection, normalize data, model training, SVM, logistic regression, decision tree.

**5. SUPERVISED LEARNING ALGORITHMS**

SVM



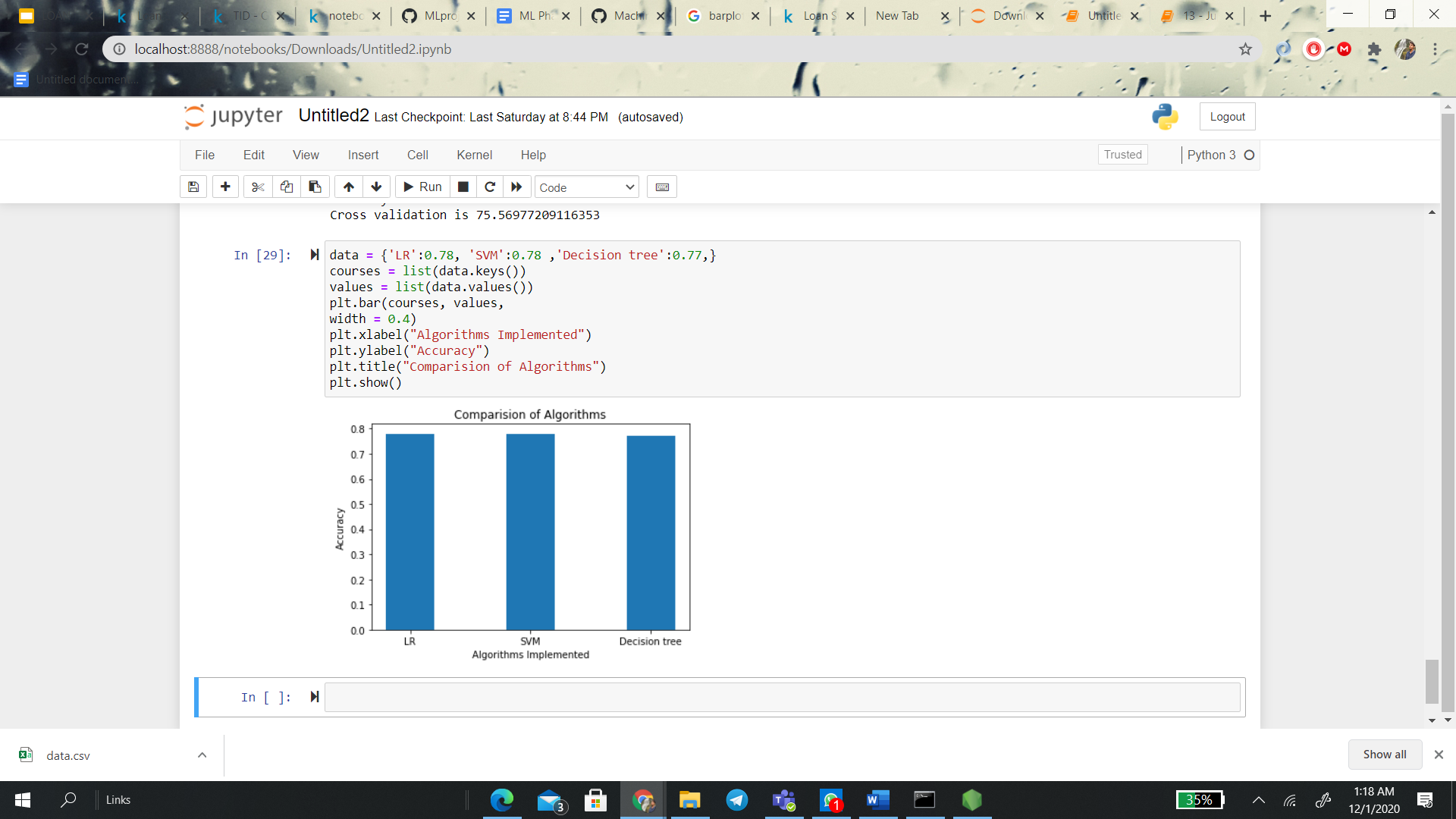
Logistic Regression



Decision Tree



**ALGORITHM COMPARISON**



After comparing the accuracies of all three algorithms the Logistic Regression is more accurate among three algorithms.

**Yerramsetti Raghuram - AM.EN.U4CSE17359**

**1. Dataset 3**

This case is about a bank (Thera Bank) which has a growing customer base. Majority of these customers are liability customers (depositors) with varying sizes of deposits. The number of customers who are also borrowers (asset customers) is quite small, and the bank is interested in expanding this base rapidly to bring in more loan business and in the process, earn more through the interest on loans. In particular, the management wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors). A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9% success. This has encouraged the retail marketing department to devise campaigns with better target marketing to increase the success ratio with minimal budget.

The department wants to build a model that will help them identify the potential customers who have a higher probability of purchasing the loan. This will increase the success ratio while at the same time reduce the cost of the campaign.

The file given below contains data on 5000 customers. The data include customer demographic information (age, income, etc.), the customer's relationship with the bank (mortgage, securities account, etc.), and the customer response to the last personal loan campaign (Personal Loan). Among these 5000 customers, only 480 (= 9.6%) accepted the personal loan that was offered to them in the earlier campaign.

<https://www.kaggle.com/itsmesunil/bank-loan-modelling>

**2. PREPARE DATA**

Pre-processing:

* Checked for null values
* Analysing data and finding wrongly entered data
* Checked the normality of the columns

Summarization:

* Found dimensions and statistical summary.
* Brokedown the data by class using group by.

Data Visualization:

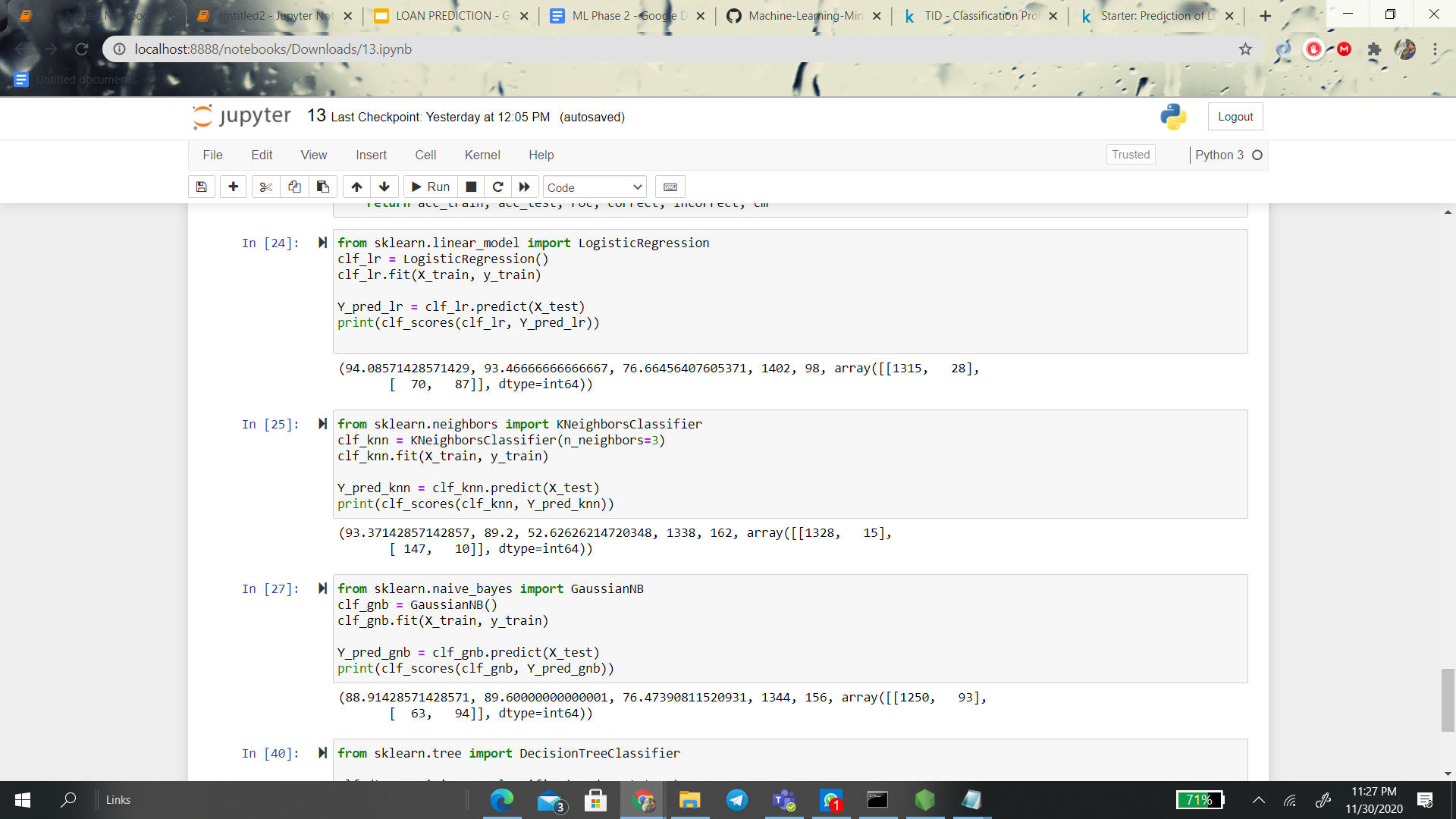
* Heatmap
* barplot

**3. PYTHON PACKAGES**

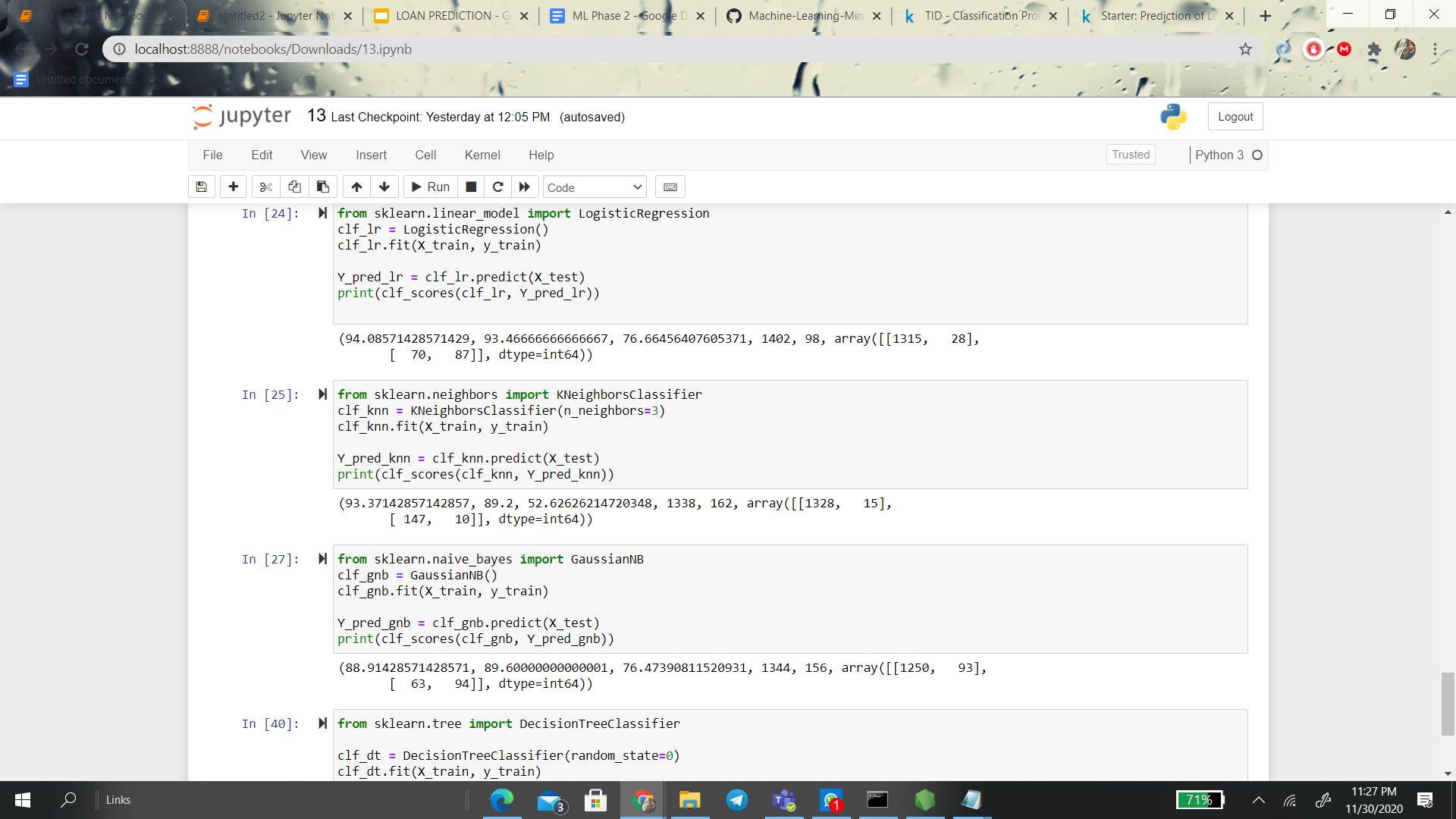
* Numpy and Pandas
* Sklearn for, confusion matrix,KNN, logistic regression, Naive Bayes , roc\_aoc\_score

**4. SUPERVISED LEARNING ALGORITHMS**

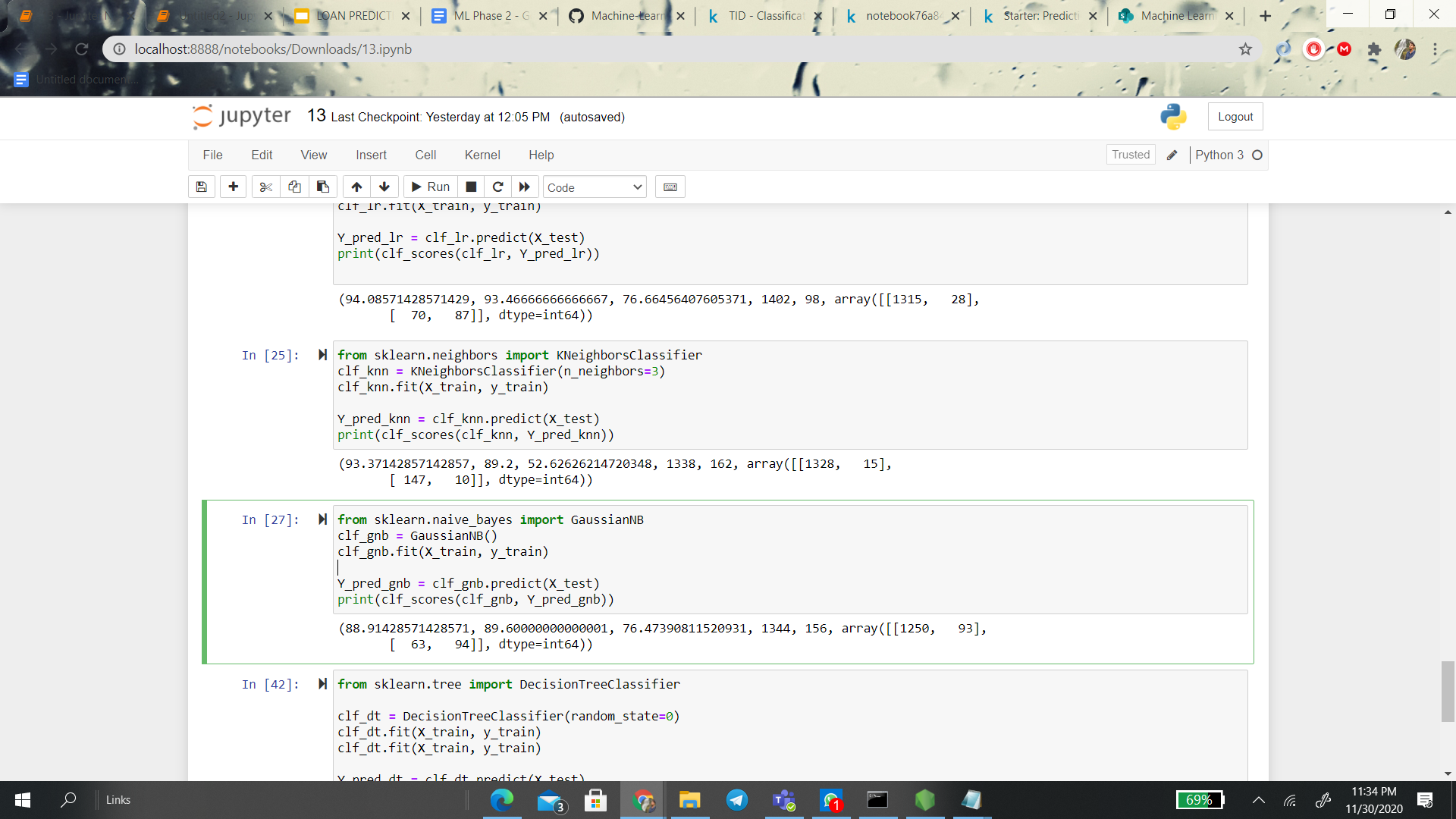
Logistic Regression



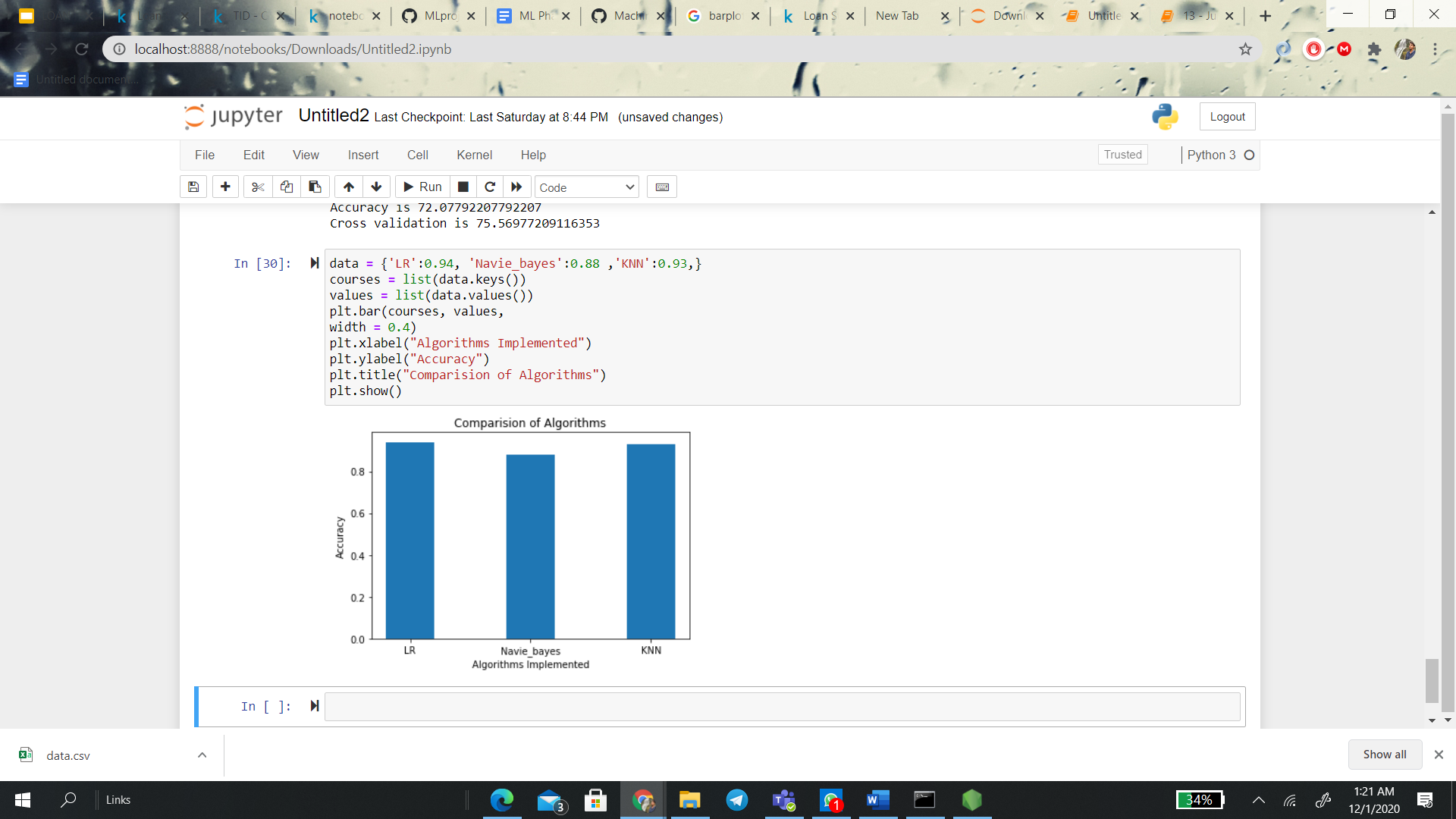
KNN



Naive Bayes



**ALGORITHM COMPARISON**



After comparing the accuracies of all three algorithms the Logistic Regression is more accurate among three algorithms.