**SMARTBRIDGE SUMMER INTERNSHIP**

**PROGRAM-2019**

**TEAM NAME : INNOVATIVE GEEKS**

TEAM MEMBERS:

1)**MARLA Ranya**

2)**BAIRI SAGARIKA**

3)**MACHERLA VAISHNAVI**

4)**AMBATI INDRAJA**

**GRADUATE ADMISSION PREDICTION**

**1.1 INTRODUCTION:**

The project aims to help students choose the right Universities by predicting whether a student will be admitted to a specific University. Graduation is a dream for every under graduate students. But they are unaware of how to maintain a good striking score for easily cracking the admission process. This model can be used by Graduate Admission Committee to filter very low scoring or very high scoring applicants.

Graduate Admissions can be thought as a mapping problem between Students and Universities where each end strives for the best they can get. Based on all factors and the university rating that the individual secured during his/her educational career scores like CGPA,GRE,LOR and SOP ratings. With each student applying to multiple Universities, the total application cost may lie in 100’s of million USD per year. Admission committees have to spend a lot time evaluating tons of applications.

Our objective is to help either side by providing accurate recommendations to students based on their profile. This might help students cut costs on Admission counselling and applying to a smaller set of Universities. On the other side, lesser number of application will save a lot of time to evaluate them. Graduate Admissions process is generally carried out online via Universities own interface or third party websites like apply yourself.

Using this application, the entrance seat allotment becomes easier and efficient GRADE uses historical admissions data to predict how likely the committee is to admit each new applicant. It reports each prediction as a score similar to those used by human reviewers, and accompanies each by an explanation of what applicant features most inﬂuenced its prediction.

**1.2 OBJECTIVES OF RESEARCH:**

* College admission predictor is a boon to many students. This helps the student not only to help in filling out the application forms but also give the students an idea about their future college by calculating their cut off.
* Our objective is to help both admission committee and students by providing accurate recommendations to students based on their profile.
* Project aim is to help students choose the right Universities by predicting whether a student will be admitted to a specific University.

**1.3 PROBLEM STATEMENT:**

As there will be huge competition for graduate applications, it will be difficult for selection process. So we used prediction analysis.

**2. REVIEW OF LITERATURE:**

There has been various studies dealing with admission process. But there are very few papers which deal with assisting the decision making process using Machine Learning approach.

* Data Science has been used to partially automate the Graduate Admission process.
* Data Science models are being used to eliminate extremely good and bad application from the pool.
* This technique saves upto 74% of time for Graduate Admission Committee members.

**3. DATA COLLECTION:**

Data collection is a process of gathering information from different universities.

For prediction of graduate admission process, we have taken a sample dataset from Kaggle website. Dataset includes few parameters and predict the chances of admission based on the independent parameter values.

The predicted output is dependent on the individual independent parameters.

**4. METHODOLOGY:**

After data pre-processing is done we used regression algorithms for predicting admissions of graduates as it comes under regression part.

These algorithms comes under the regression algorithms:

1)Linear Regression

2)Multi Linear Regression

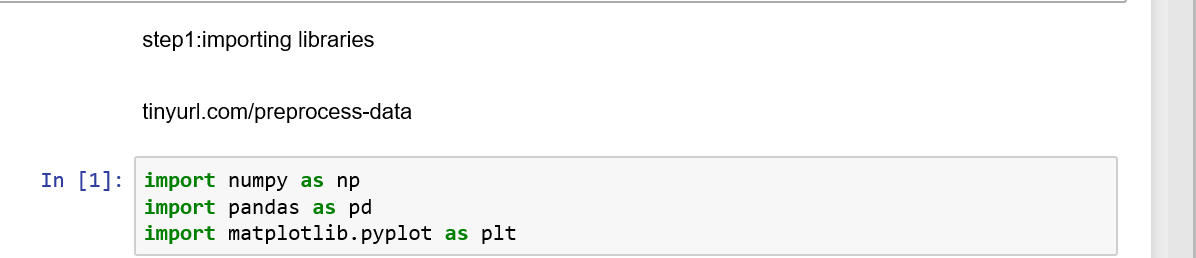
3)Polynomial Regression

4)Decision tree Regression

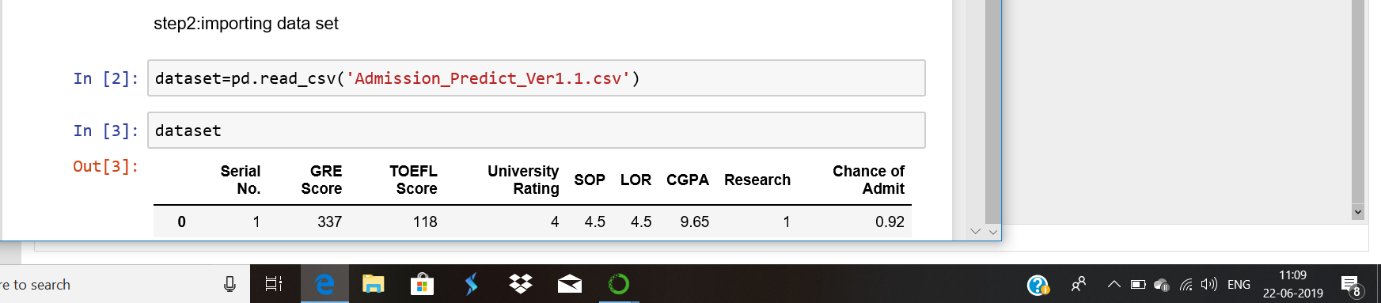
5)Random Forest Regression

* Linear Regression is not applicable to our project as we have multiple dimensions in the dataset.
* Polynomial Regression is not applicable as our dataset is continuous.
* Multi-linear Regression, Decision tree and Random Forest Regression can be applicable for predicting the result.
* Steps involved in Regression algorithms are:
* Step1: Data pre processing.

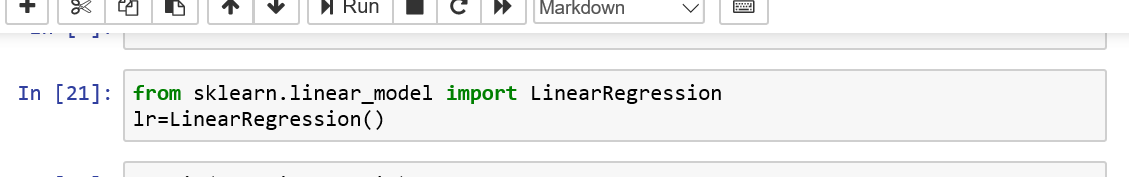
a)Importing libraries:- A library is essentially a collection of modules that can be called and used.



b)Importing dataset:- We will need to locate the directory of the CSV file and read it using a method called read\_csv which can be found in the library called pandas.



c)Encoding the categorical data:- Sometimes you may find some data are missing in the dataset. The library that we are going to use for the task is called [Scikit Learn](http://scikit-learn.org/stable/index.html" \t "_blank) pre-processing.



d)Avoiding the dummy variable:- Sometimes our data is in qualitative form, that is we have texts as our data. We can find categories in text form. Now it gets complicated for machines to understand texts and process them, rather than numbers, since the models are based on mathematical equations and calculations. Therefore, we have to encode the categorical data. There’s a class in the library called LabelEncoder which we will use for the task.

As we do not have any strings in our dataset. So we do not use LabelEncoder.

from sklearn.preprocessing import LabelEncoder

 the next step is usually to create an object of that class. We will call our object labelencoder\_X.

labelencoder\_X = LabelEncoder()

To do our task, there’s a method in the LabelEncoder class called fit\_transformwhich is what we will use. Once again, just like how we did it before, we will pass two parameters of X — row selection and column selection.

X[:,0] = labelencoder\_X.fit\_transform(X[:,0])

To accomplish the task, we will import yet another library called OneHotEncoder.

from sklearn.preprocessing import LabelEncoder, OneHotEncoder

Next we will create an object of that class, as usual, and assign it to onehotencoder. OneHotEncoder takes an important parameter called categorical\_features which takes the value of the index of the column of categories.

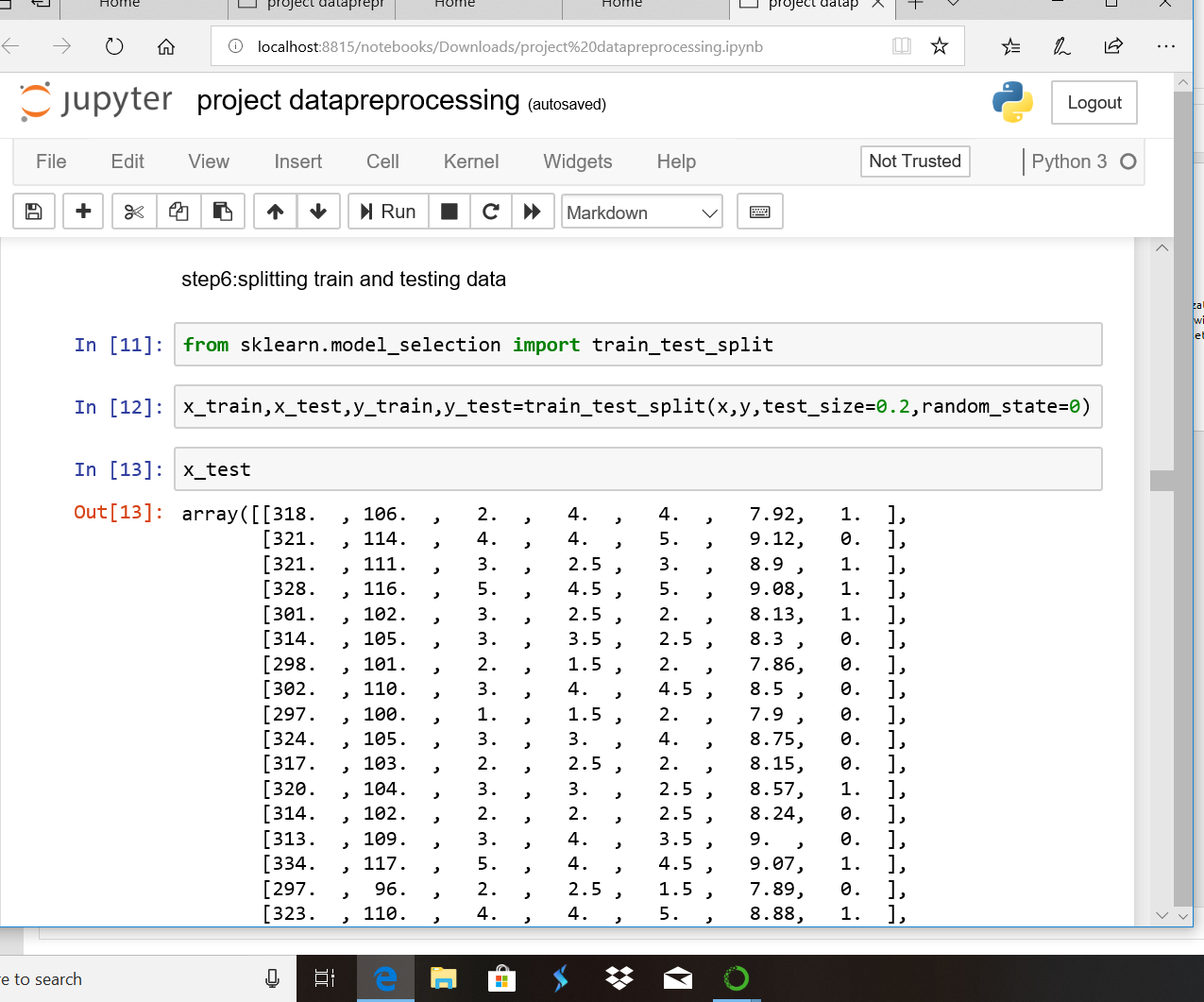
onehotencoder = OneHotEncoder(categorical\_features =[0])

Just as we used fit\_transform for LabelEncoder, we will use it for OneHotEncoder as well but also have to additionally include toarray().

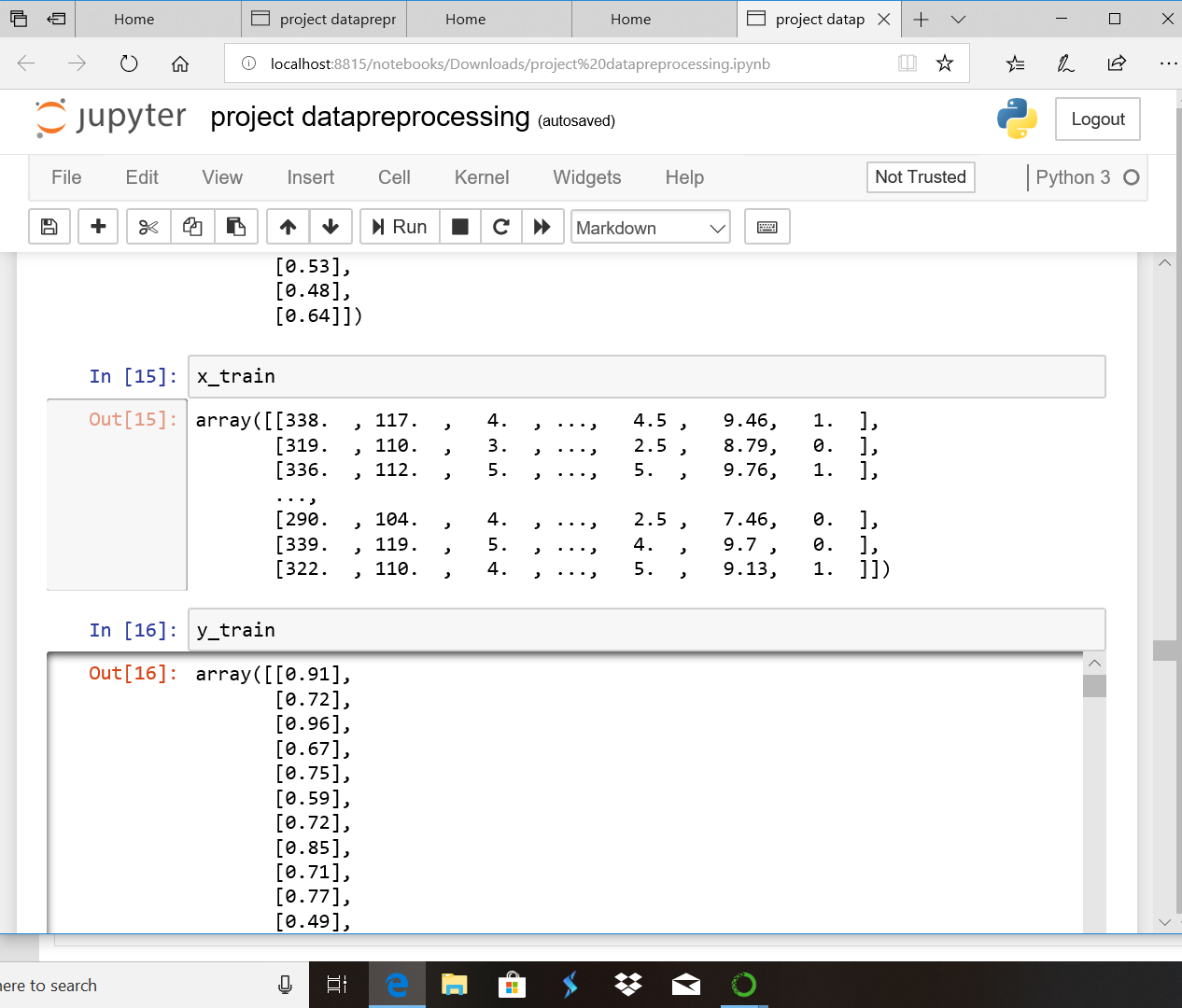
X = onehotencoder.fit\_transform(X).toarray()

If you check your dataset now, all your categories will have been encoded to 0s and 1s.

e)Splitting the dataset into training set and test set:- Now we need to split our dataset into two sets — a Training set and a Test set. We will train our machine learning models on our training set, i.e our machine learning models will try to understand any correlations in our training set and then we will test the models on our test set to check how accurately it can predict. A general rule of the thumb is to allocate 80% of the dataset to training set and the remaining 20% to test set. For this task, we will import test\_train\_split from model\_selection library of scikit.



Now to build our training and test sets, we will create 4 sets— X\_train (training part of the matrix of features), X\_test (test part of the matrix of features), Y\_train (training part of the dependent variables associated with the X train sets, and therefore also the same indices) , Y\_test (test part of the dependent variables associated with the X test sets, and therefore also the same indices). We will assign to them the test\_train\_split, which takes the parameters — arrays (X and Y), test\_size (if we give it the value 0.5, meaning 50%, it would split the dataset into half. Since an ideal choice is to allocate 20% of the dataset to test set, it is usually assigned as 0.2. 0.25 would mean 25%, just saying).



To accomplish the job, we will import the class StandardScaler from the sckit preprocessing library and as usual create an object of that class.

from sklearn.preprocessing import StandardScaler

sc\_X = StandardScaler()

Now we will fit and transform our X\_train set (It is important to note that when applying the Standard Scalar object on our training and test sets, we can simply transform our test set but for our training set we have to at first fit it and then transform the set). That will transform all the data to a same standardized scale.

X\_train = sc\_X.fit\_transform(X\_train)

X\_test = sc\_X.transform(X\_test)

* Step2: Filling the regression algorithms to the required training set.
* Step3: Predicting the test set results.

**4.1 EXPLORATORY DATA ANALYSIS:**

GRE : **GRE** scores are **used** to assess the suitability of applicants for **graduate**-level **study** across many different subjects. The Subject Tests assess knowledge in a particular field, while the General Test assesses verbal reasoning, quantitative reasoning and analytical writing.

TOFEL: The report found that **TOEFL** scores were even more **important** to **admissions** officers than student qualifications and grades, and even their college application essays. So, it's safe to say it's pretty **important** to meet the required score.

UNIVERSITY RATING: Usually, **university rankings matter** a lot with degrees such as law, business, liberal arts, and other degrees that do not teach highly marketable technical skills and are more or less, considered “soft subjects”, subjects that require personality and people skills in order to be successful in them.

Rankings have been proven to significantly help maintain and build institutional position and reputation. Having your institution rank as highly as possible only improves the chances of falling into a prospective students shortlisting process.

University rankings can influence national and international partnerships and collaborations. This ties in with the positive branding that manifests from rankings, but for institutions with an advantageous rank, it will improve willingness of others to partner with them or [support their membership of academic or professional associations.](http://www.oecd.org/education/imhe/39802910.pdf)

SOP: Statement of Purpose is a reflection of your personality that you portray in front of the admission committee (AdCom) for getting enrolled in universities abroad. It is your chance of impressing the committee to present your profile strongly by describing your personality traits that are not covered in your academic documents.

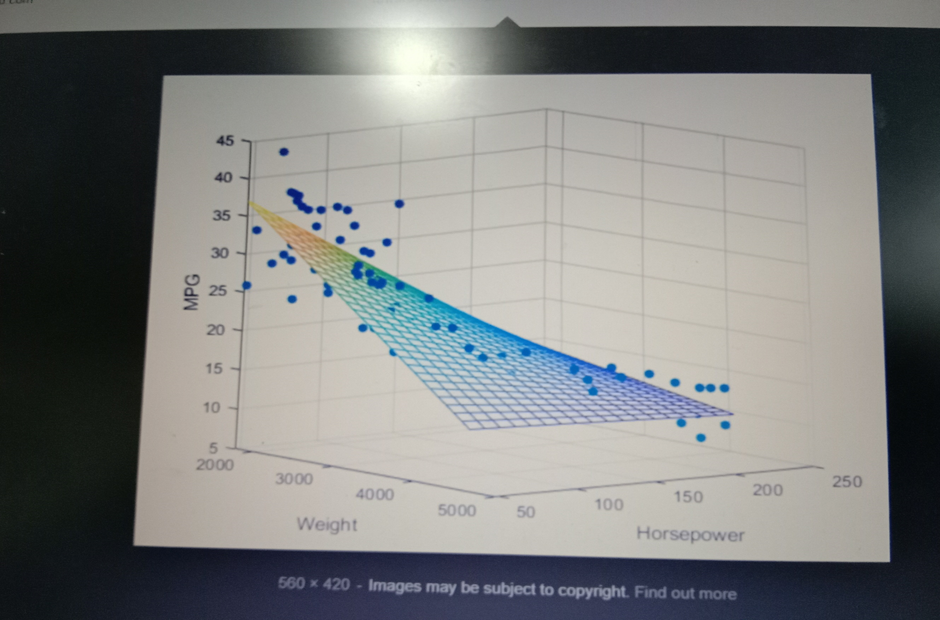
LOR: The **LOR** is a document that provides the admission officers with a comprehensive insight into your suitable candidate for admission into the concerned University. The full form of **LOR** is a Letter of Recommendation.

CGPA: It stands for Cumulative Grade Point Average. It refers to average of overall marks. It is a grade pointing system.

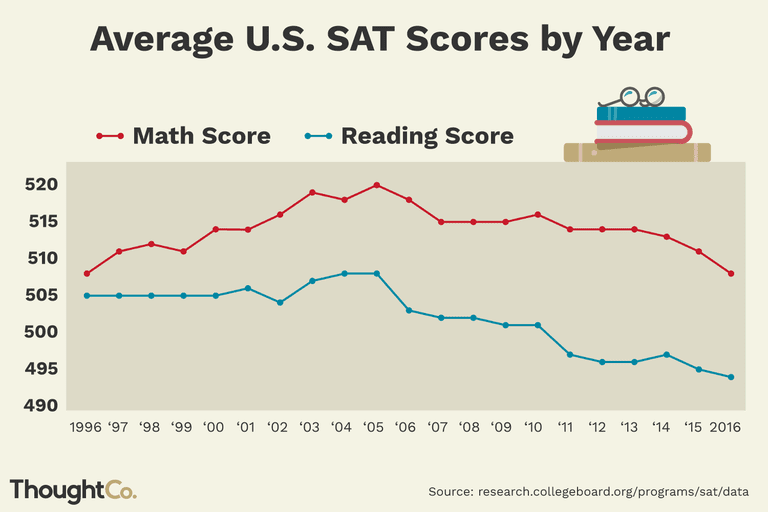
RESEARCH: It is a systematic investigation in order to establish the new facts and new conclusions for getting admission into good universities.

CHANCE OF ADMIT: It is the probability that the student is having a chance of getting admission in top rated universities.

**FIGURES AND TABLES:**

****

Multi linear regression graph

****

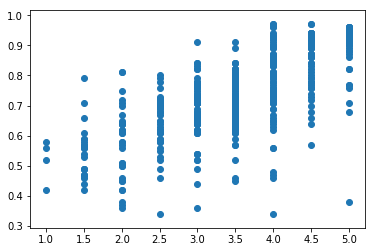
University scores every year

**4.2 STATISTICAL TECHNIQUES AND DATA VISUALIZATION:**

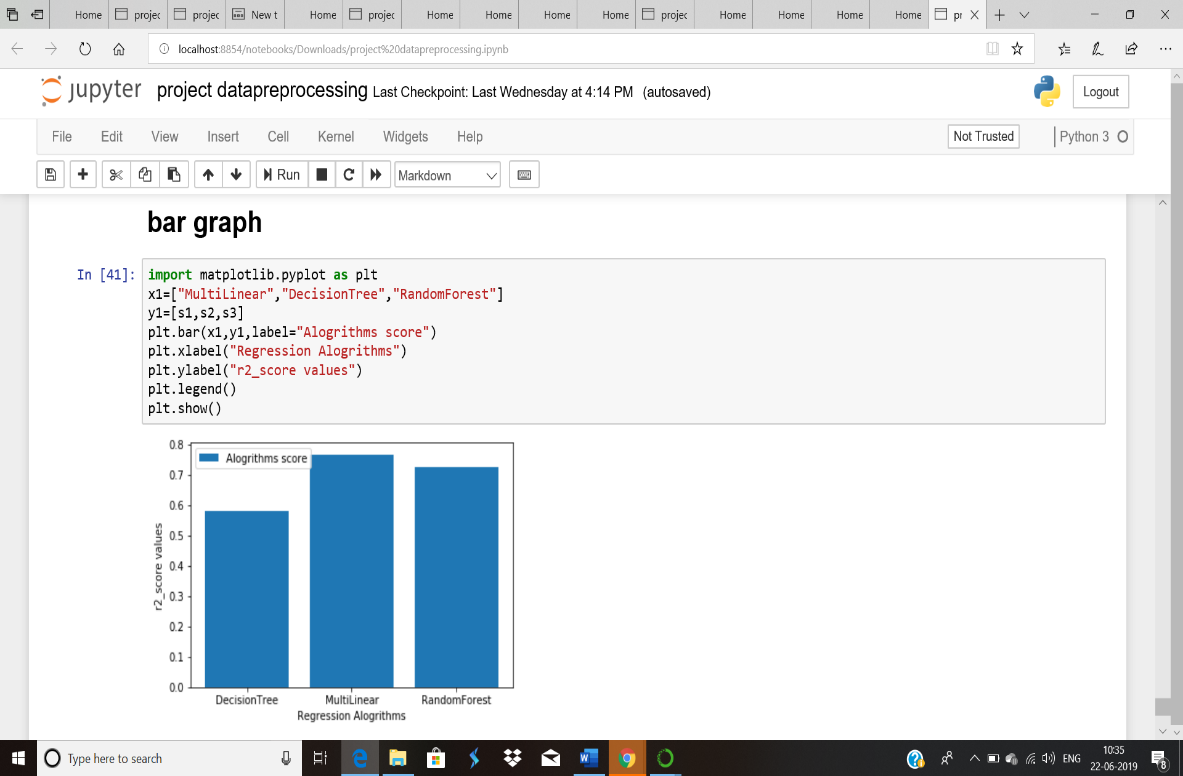
* For plotting the graph of multi-linear regression

Code: plt.scatter(x\_train[:,3],y\_train)

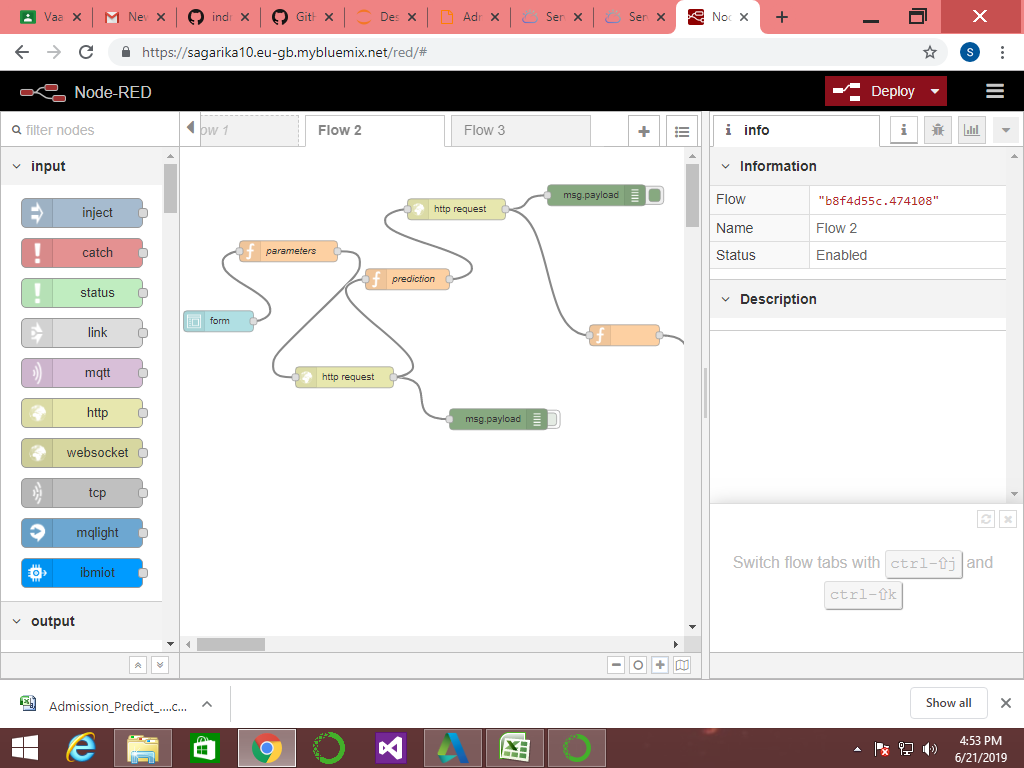
Output:

****

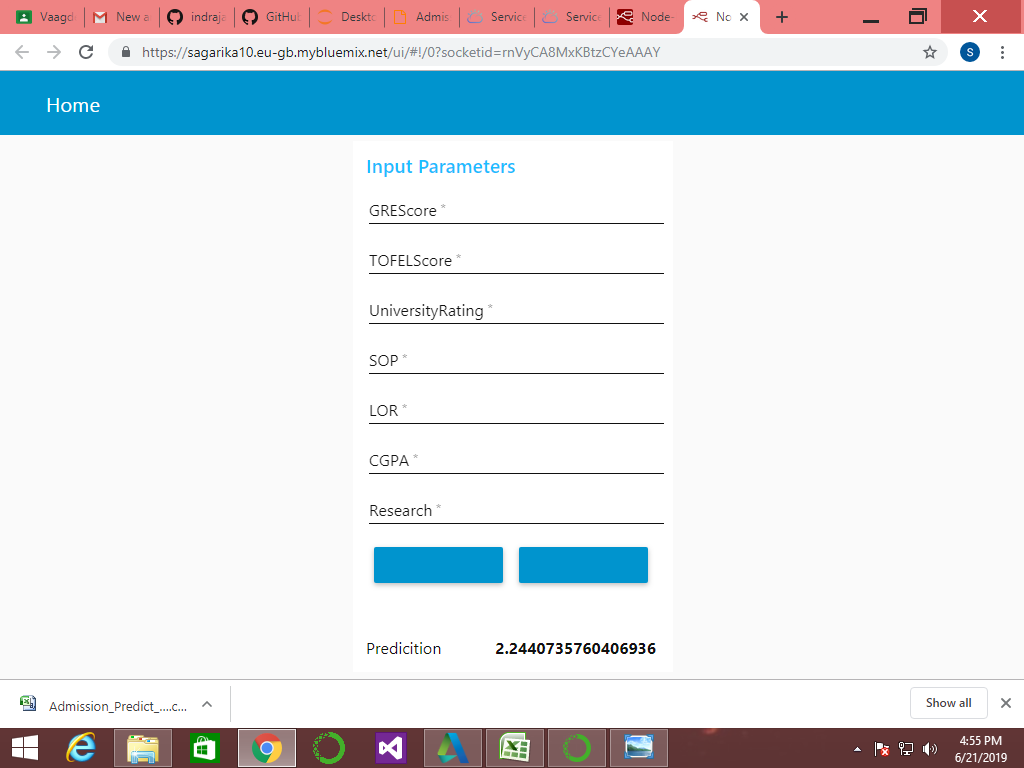
* Comparision of bar graph among the regression algorithms



* Node-Red deployment



* Final result of prediction



**4.3 DATA MODELING USING SUPERVISED ML TECHNIQUES:**

In statistical modelling, regression analysis is a set of statistical processes for estimating the relationships among the variables. It includes many techniques for modelling and analysing several variables, when the focus is on relationship between the dependent and independent variables.

Regression analysis helps one to understand how the typical value of the dependent variable changes when any of the independent variable is varied. It also estimates the conditional expectation of the dependent variable gives the independent variable. A function of independent variables is called as Regression function.

Regression analysis is widely used for prediction and forecasting. Regression analysis is also used to understand which among the independent variables are related to the dependent variable and to explore the forms of these relationships.

The performance of the regression analysis methods depends on the form of data generating process and how it relates to the regression approach. Regression analysis often depends on making assumptions about the process. These assumptions are sometimes testable if a sufficient quantity of data is available.

Regression model for prediction are often useful even when the assumptions are moderately violated. In many of the applications, with small effects or questions of causality based on observational data, regression methods can give the misleading results of prediction.

**5. FINDINGS AND SUGGESTIONS:**

* Best fit accuracy was found to be maximum in Multi-linear Regression among all the regression algorithms.
* As accuracy score is important for predicting the admission process of graduates. We calculated the accuracy scores in decision tree algorithm, random forest algorithm and multi-linear algorithm.
* By comparing the resulting accuracy score of the regression algorithms, we have decided that multi-linear model is resulting the best accuracy.

**6. CONCLUSION:**

We have used many real features to model admission process. First we tried to model all universities at once, but this seems to be quite difficult since it involves lot of variability. Each University has slightly different admission criteria so predicting results all universities at once doesn’t seem a good idea. Next, I tried to approach each university at a time.

This project was a great learning opportunity. I started right from website scrapping to data cleaning and finally modeling. It helped me to experience the real life challenges faced by a Data Scientist.