**STUDENT JOB PREDICTION**

**A PROJECT REPORT**

*Submitted by*

**AADITYA RAJ [Reg No: RA1511008010713]**

**KUMAR OMAN BURMAN [Reg No: RA1511008010666]**

*Under the Guidance of*

**MS. PUGHAZOLI S.**

(Associate Professor, Department of Information Technology)

*In partial fulfillment of the Requirements for the Degree*

*Of*

**BACHELOR OF TECHNOLOGY**



**DEPARTMENT OF INFORMATION TECHNOLOGY**

**FACULTY OF ENGINEERING AND TECHNOLOGY**

**SRMINSTITUTE OF SCIENCE AND TECHNOLOGY**

**KATTANKULATHUR – 603203**

**NOVEMBER 2019**

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

KATTANKULATHUR-603203

BONAFIDE CERTIFICATE

Certified that this project report titled “**STUDENT JOB PREDICTION”** is the bonafide work of **“AADITYA RAJ[Reg No: RA1511008010713], KUMAR OMAN BURMAN [Reg No: RA1511008010696]”** who carried out the project work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion for this or any other candidate.

MRS PUGHAZOLI S. Dr. G. VADIVU

**GUIDE HEAD OF THE DEPARTMENT**

Assistant Professor (O.G) Dept. of Information Technology

Dept. of Information Technology

Signature of Internal Examiner Signature of External Examiner

**ACKNOWLEDGEMENT**

We would like to express our gratitude and appreciation to all those who gave us the possibilityto complete this project and report. A special thanks to our final year project coordinator Dr. KayalvizhiJayavel whose help, stimulating suggestions and encouragement helped us incoordinating the project and writing the report and our guide Ms. PUGHAZOLI whopropelled the idea of this project and his full effort in guiding the team in achieving the goal aswell as his encouragement to maintain our progress on track.We would also like to acknowledge with much appreciation the crucial role of our panelDr. A. Razia Sulthana and Dr. S. Suresh who reviewed our project with much assistance andadvocacy along the journey.

**AADITYA RAJ**

**KUMAR OMAN BURMAN**

**TABLE OF CONTENTS**

CHAPTER NO. TITLE PAGE NO.

ABSTRACT

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

1. INTRODUCTION
   1. OVERVIEW OF STUDENT JOB ROLE PREDICTION
   2. PROBLEM STATEMENT
   3. SOLUTION TO PROBLEM
   4. CONCEPTS INVOLVED
      1. MACHINE LEARNING
      2. WEB SERVICES AND API

1.4.3 CLASSIFICATION OF ML AND WEB TECHNOLOGIES

1.5 FLOW OF PROCESS

1.5.1 COLLECTION OF THE DATA

1.5.2 DATA PRE-PROCESSING

1.5.3 ANALYSIS AND RESULTS

1.5.4 CONFIGURING AND POSTING DATA

1.5.5

1.6 LIFE CYCLE OF MACHINE LEARNING

1. LITERATURE REVIEW

2.1 RESEARCH ON LITERATURE PAPERS

2.2 CONCERNS IN PREVIOUS SYSTEM

3. SYSTEM ANALYSIS

3.1 PROBLEM DEFINITION

3.2 ADVANTAGES

3.3 DISADVANTAGES

3.4 REQUIREMENTS

3.5 ISSUES FOUND IN EXISTING SYSTEM

4. SYSTEM DESIGN

4.1 ALGORITHMS

4.1.1 SVM

4.1.2 RANDOM FOREST

4.1.3 XG BOOST

4.2 DESIGN DIAGRAMS

4.3.1 ARCHITECTURE DIAGRAM

4.3.2 PROPOSED SYSTEM ARCHITECTURE

4.3.3 USE CASE DIAGRAM

4.4.4 USE CASE SYNTAX

4.4.5 CLASS DIAGRAM

4.4.6 SEQUENCE DIAGRAM

4.4.7 ACTIVITY DIAGRAM

4.4.8 DATASETS

5. MODULE IMPLEMENTATION

5.1 GATHERING INFORMATION FROM PRIOR SOURCE

5.2 CALCULATION FOR PLOTTING THE VALUES

5.3 APPLICATION OF ALGORITHMS

5.4 COMPARISON OF ALGORITHMS

5.5 BUILDING FLASK API

6. SOFTWARE TESTING

6.1 OBJECTIVES

6.2 TYPES OF TESTING

6.2.1 UNIT TESTING

6.2.2 INTEGRATION TESTING

6.2.3 FUNCTIONAL TESTING

6.2.4 SYSTEM TESTING

6.2.5 WHITE BOX TESTING

6.2.6 BLACK BOX TESTING

6.2.7 ACCEPTANCE TESTING

6.3 POSTMAN API TESTING

7. CONCLUSION

8. REFERENCES

9. FUTURE ENHANCEMENTS

10. APPENDIX

11. PLAGARISM REPORT

**ABSTRACT**

The latest computer technology that supports leaning technology and tools, that helps students to learn and gain knowledges through smart teaching systems. This new technology system also helps to collect and analyze data of students from various tutoring systems, simulation tech, games through various data analyzing designs and patterns.

In the new system, we focus on collecting data through various analytical techniques and then the collected data goes through the various advanced algorithm to draw patterns and predict a well-suited career for computer science undergraduates based on his skills, experiences, interests, and opportunities. The data are collected from various systems contains many variables and also there's some data which is not applicable (related) for our systems. These data are needed to normalize so that we can train our systems. The methods and data patterns applied from statistics, data mining and various machine learning algorithm to train and aggregate data gathered from various sources. As students for them, it is very important to assay their interests and their capabilities which helps them to identify which career area to direct them in. This also helps to enhance and advance their capability and also motivates them towards the designated career.

While recruiters to recruiting candidates this also helps them to clinch which job role suits best for the candidate based on his/her performance from various collected resources. This exploration mainly focuses on job role prediction of computer science candidates.

**LIST OF TABLES**

Sl. No. TABLE NAME PAGE NO.

3.1 Software Requirements

3.2 Hardware Requirements

**LIST OF FIGURES**

Figure No. FIGURE NAME PAGE NO.

1.1 Machine learning life cycle …………………………………

1.2 ML and Traditional programming ………………………….

4.1 SVM ………………………………………………………...

4.2 XG BOOST …………………………………………………

4.3 System architecture…………………………………………..

4.4 Proposed architecture…………………………………………

4.5 Use case diagram……………………………………………..

4.6 Class diagram…………………………………………………

4.7 Sequence diagram…………………………………………….

4.8 Activity diagram …………………………………………….

4.9 Data sets …………………………………………………….

5.1 Collection of Data …………………………………………..

5.2 Creating dummies variable ………………………………….

5.3 Encoding for the categorical variables ……………………..

5.4 Normalization ……………………………………………….

5.5 Correlation heat map ………………………………………..

5.6 Principal component analysis ……………………………….

5.7 Feature importance ………………………………………….

5.8 Visualization of feature importance ………………………...

5.9 Random forest regressor ……………………………………

6.1 Unit testing …………………………………………………..

6.2 Black Box Testing ………………………………………….

5.10 Random Forest Algorithm …………………………………..

5.11 Prediction by SVM ………………………………………….

8.1 Comparison of different algorithms …………………………

**ABBREVIATIONS**

**ML** Machine Learning

**SVM** Support-Vector Machine

**API**Application Programming Interface

**RFR** Random forest regressor

**CHAPTER 1**

**INTRODUCTION**

**1.1 OVERVIEW OF STUDENT JOB PREDICTION**

In today's society, the competitions are increasing day by day. Every third day a new technology comes to the market, a new and improved version of applications arrives every day in the market. When it comes to competing with this new generation and to reach any goal, the students needed to plan and organize from initial stages of their education that is when starting to choose new skills to learn or focusing to gather new technical skills and interests. This is very important to constantly assay his/her performance and also keeps track of how close they are from their goal and assess that they are the right path toward their target. This not only helps in improving self but also helps in motivating self and keeps on the track throughout the goal.

While recruiters recruiting people, they not only consider candidates for one parameter, they search for candidates to be able to handle more than one specification. They search for the candidate who knows more than one skills. The recruiters also try to find, if got selected that this job role is the best fit for their career. There are many job roles like Developer, Testing Manager, Network Analyst, Data scientist, Database Administrator, Debugger, Data Analyst and so on. And for all these job roles there are some pre-requisites knowledge needed to be known. So recruiters also check this pre-requisites knowledge while hiring a candidate. With the help of our system model, these kinds of predictions make very easy for recruiters and also for the students to grasp the pre-requisite knowledge that helps them best-suited job role with their learning interests. Basically, the system model recommends the best-suited job role on the basis of user input. The inputs are given by what users or candidates have gathered and learned throughout the career. These prediction model also helps recruiters to select the best candidate suited for the selected job role. Our belief system will rotate around setting up an opportunity for a superior tomorrow, with all the pre-imperative a hopeful must be fit for the activity profile they prefer to have.

**1.2 PROBLEM STATEMENT**

Nowadays, Already there are numerous job prediction systems available such as LinkedIn Jobs, Monster Jobs, Co-Cubes, Naukri.com, and various others. These systems take input factors like technical abilities and psychometry of student's studies and suggest job roles suited for their performance.

**1.3 SOLUTION TO PROBLEM**

As denotes that as it may, here several factors consolidating capacities of stand-ins in games, scholastics, and their such interests, aptitude and information, academics performance also comes under consideration. By acknowledging all these information parameters and extracurricular we have an entirety of 39 such inputs to consider under our model that is used to train the data to get the well-fitted job that is fixed to 15 numbers.

The input parameters that are accepted to consider are tremendous in number and running of the regular programming and traditional algorithm or ordinary calculation such as Decision tree cannot give ideal output. Therefore we impelled AI calculations like SVM, OneHot Encoding, XG Boost, Random Forest to train the information and get the desired output.

**1.4 CONCEPTS INVOLVED**

**1.4.1 MACHINE LEARNING**

Machine learning is the sub part of Artificial intelligence. It is the field of study where computers can learn from any example without any human support. Machine learning basically develop the computer learning process according to their experiences without being programmed openly.

The whole process get started with feeding the data and then our machine(computer) is trained by building various models using the different algorithms like SVM,XG Boost, Random forest, etc and the data we have. According to our type of the data and the task we are trying to computerize.

**Practical example of machine learning:**

When the students go for the exam, they usually understand the concept rather than cramming that part. Before going for the examination, students use the different type of data(resources like books, you tube,wikis,etc) to feed their machine(brain). But in reality they train their brain with the input(various resources of study) and the output(the logic or concept to use for solving the problem). Every time they solve the practice papers and find their performance i.e (accuracy/score).Slowly the student increase his performance with the approach or way he followed to prepare for the test. Similarly the model of the machine learning is built, both inputs and the outputs are given to the model to train the computer or system. The algorithms train with the input data and get the model score by comparing the actual output which was not fed while the training to the answer we get from the system.

**DATA (INPUT)**

**Traditional programming**

**OUTPUT**

**PROGRAM**

**DATA (INPUT)**

**Machine Leanining**

**PROGRAM**

**OUTPUT**

**Figure 1.1:** ML AND TRADITIONAL PROGRAMMING

In the traditional programming we give the data as our input and the logic behind the program to run it.

While training we give the data along with the output class ,then the machine creates its own program(logic),which is used while testing for the evaluation**.**

**1.4.2 CLASSIFICATION**

Classification is the supervised learning method which is used to predict the category of the class that it will belong.Classification in machine learning specifies the process of the prediction of the class for the given data sets. Classes are actually the target(main output) or labels or the various categories. It maps the function fn(x) from the input variables to the output variables.

Example:- Suppose we have a data set of the student consisting of the various details like name, reg no, age, dob, grades in all the subjects in each level of exam, the projects he has done etc and the output class will be whether the student will pass the final exam or not. It will be a parallel order issue where the yield will be yes or no. A classifier will utilize some of the training data to understand how the input variables are linked and mapped to the output variables. When the classifier we are applying is trained accurately, it can be easily predict whether the student will pass or not.

There are the various classification algorithms in Machine learning:-

1. Linear Classifiers: Logistic Regression, Naive Bayes Classifier
2. Support Vector Machines
3. Decision Trees
4. Boosted Trees
5. Random Forest(Regressor and classifiers)
6. Neural Networks
7. Decision Tree

**1.5 THE KNOWLEDGE OF THE PROCESS**

**1.5.1 DATA COLLECTION**

Information accumulation is the most significant errands of any AI ventures. Since we need to encourage the information as a contribution to sustain the calculations. Subsequently the calculation productivity and exactness relies on the nature of the information we have gathered.

For understudy work expectation, numerous parameters are required like scholastic scores in the different subjects , every one of the specializations the understudy has done, programming and logical abilities, all the individual subtleties like interests, sports, hackathons, memory, rivalries and different workshops identified with the distinctive fields.

A portion of the information is gathered from representatives working in various associations, some measure of information is gathered through LinkedIn programming interface, some measure of information is arbitrarily produced and other from school graduated class database.Inthis dataset we collected, there are 20000 rows and 39 columns where the target variable isSuggested Job Role

**1.5.2 PRE-PROCESSING**

Data collection is one assignment and making that information to be helpful is another important task. When we collect information from different methods, it will be in a disorderly arrangement. There may be a lot of unwanted data and may be null values.

Steps in data pre-processing:-

1. To replace the missing values in the data set.
2. Check the categorical data and change to the those data sets through encoding.
3. Divide the data set in training and testing set
4. Feature scaling and extraction

All the cases of the data set must be verified or replaced in the case of null values.

**1.5.3 RESULTS AND ANALYSIS**

The experimental setup consists of two components. The first one is collecting data and analyzing it. We collect data from CSV File. The second component being the learning models. We first applied the feature extraction and selection. Among the three types of filtering method(Embedded method, Filter method and Wrapper method),we used Filter method(Univariate method selection). After doing this, we train our model with different algorithms later on the algorithm having higher accuracy is taken and the results are predicted based on that.

**1.6 MACHINE LEARNING LIFE CYCLE**

The machine learning life cycle is the cyclic process that highlights those steps which are essential for an organization to be benefitted by machine learning and artificial intelligence to derive practical value. It requires all the projects from initiation to completion and provides broad level outlook of how the data science is designed in real sense.

**Model scoring**

**Data Ingestion**

**Discovery of data**

**Feature engineering**

**Model training**

**Model development**

LIFECYCLE MANAGEMENT

**Figure 1.2: Life Cycle Of ML**

**CHAPTER 2**

**2.1 LITERATURE REVIEW**

**Title: Prediction of Student Performance by Using MapReduce**

**Author:** Dr. N. Tajunisha , M. Anjali

Expanding digitization of understudy records implies prescient investigation is required to change instructing and become a key device in getting familiar with understudies. Prescient investigation is a procedure in which information gathered about the understudy, ordinarily participation, subjects taken, evaluation is utilized to comprehend learning designs, recognize ability holes, anticipate execution and distinguish learning openings. The viable element choice technique is required to investigate the effective characterization calculation. Taking care of enormous information, a novel segment system in MapReduce is additionally required. The examinations demonstrate that the understudy information grouping calculation can improve information area perceptibly to contrast and default scheduler and it additionally can improve other scheduler's territory. Besides, compute the information mining procedure's normal culmination time which depends on the understudy information level. It improves the accuracy of order's residual time assessment. At last the outcome demonstrates that the proposed framework has higher order exactness in huge information and furthermore decreases the time unpredictability.[1]

**Title: Prediction of Math Performance from Raw Large-Scale Educational Assessments**

**Data: A Machine Learning Approach**

**Author:** MirkaSaarela, Bulent Yener, Mohammed J. Zaki, TommiKarkk

They characterized distinguishing the understudies that are probably going to succeed or fall flat math things of certain trouble as a forecast issue. The objective was to prepare a managed learning calculation that predicts achievement or disappointment from the information. Hotheyver there are a few issues with distinguishing the names essential for this methodology. To start with, the conceivable qualities can't be utilized, since that would be much the same as designing a definitely known recipe. Second, the understudies theyre managed distinctive intellectual tests and the single things in the tests differ in their trouble (OECD, 2014), which is the reason they can't just utilize the all out total of right things for every understudy as their name. The crude scored intellectual information has a high level of missing information and no accumulated test scores and no thing troubles are accessible. Other than the PVs, the main accessible data about the genuine execution of every understudy in the subjective test is the reality whether the person was regulated a thing and in the event that the thing was controlled the score the understudy acquired for it. The score esteems can be either 0 (come up short), 1 or 2. To have the capacity to work with the accessible information, they structured a calculation to remove names from crude information and utilize these names to prepare a prescient model. For each extraordinary test/booklet, they summed up the all out scores of the included math things. At that point, they relegated every math thing that was incorporated into the test—a synopsis of the bunch of various things of the fundamental tests was given in Table 1—to a canister which they signify as trouble level so that every trouble level is of same size (i.e., incorporates a similar number of things). They picked the quantity of trouble levels for our name framework Λ to be seven, on the grounds that the OECD characterized seven math capability levels in the PISA 2012 specialized report by the OECD (2014)). Thusly, it is expected that the majority of the diverse booklets are steady as to their normal trouble, which is bolstered by the way that each test ought to be reasonable and feasible inside.[2]

**Title:**Prediction of Secondary Students Performance by mining educational data

**Author:** A. Dinesh Kumar, V. Radhika

A few information mining arrangements have been displayed for instructive information mining. Choice tree characterization got critical consideration in the region of foreseeing the understudy execution. In this area, a schematic review is given of highlight determination, Use full preparing set which is utilized for proposed calculation. It is having just concentration with the pertinent qualities through element determination strategy utilizing Ranker Search.In this investigation, those factors whose likelihood esteems were more noteworthy than 0.60 were considered and very impacting factors with high likelihood esteems have been appeared table 2. These highlights were utilized for forecast mode development. For both variable determination and expectation model development were actualized in seven days. The need of forecast over understudy execution is to support instructors and guardians to concentrating their understudies and kids to ad lib their execution just as specialist to choose among the choice tree classifier calculation to locate the best classifier for foreseeing the understudy execution.[3]

**4. Title: Educational Data Mining and Research Trends Survey**

**Author:** Rajni Jindal, Malaya Dutta Borah

Instructive Data Mining (EDM) is a rising field investigating information in instructive setting by applying distinctive Data Mining (DM) strategies/apparatuses. EDM acquires properties from regions like Learning Analytics, Psychometrics, Artificial Intelligence, Information Technology, Machine learning, Statics, Database Management System, Computing and Data Mining. It tends to be considered as interdisciplinary research field which gives characteristic information of instructing and learning process for successful training.[4]

**2.2 CONCERNS IN PREVIOUS SYSTEM**

There were a number of issues that were found in the existing system that is meant to be cleared by the proposed project such as the accuracy of the prediction of the jobs gets decreased, when dataset is smaller, data mining techniques has to be extended for collecting information for the students skills. When the data set is smaller, the algorithms can not map the target variable to the different attributes of the data set properly.

To achieve more valid results there is a need for larger amounts of data that is, larger datasets and the usage of different classifiers for predicting various jobs for the candidate.

**CHAPTER 3**

**SYSTEM ANALYSIS**

**3.1 PROBLEM DEFINITION:**

In this paper we focus upon helping students finding jobs based upon their area of interest. The goal they have targeted and dreamt of their amazing journey of learning.

The focus is to provide a job which a candidate must love to do, and to keep them updated with the latest tech world knowledge, we are providing them with all the latest courses which they can easily understand and which will eventually be more beneficial for their career perhaps.

There are several levels in building any platform and bringing it to the market in the real world. Our focus is mainly upon the easing the work of developers as well as for the clients as well. Now, in future if any student, willing to look around for a job based on our platform, they can easily view jobs related to their area of interest. On our platform the ideology which is applied is as the following- for an instance say, a student comes for searching a job in the field of web designing, which is not so well payed in the IT industry nowadays according to the research done recently, so as the student is quite interested in the Web Designing field, he/she might opt for Web Developer too! Which will eventually pay them more in the IT industry and the love for web still lies in the student target career option.

So, working upon such cases, which will be our future motive for the research, our platform is totally focused upon increasing the productivity of the work done by candidates, will give more profitable figures to the companies, and will be profitable to the developers as well!

**3.2 ADVANTAGES:**

1. The approaches are styled up in the several ways and these will be gone forward in an intuitive manner.
2. We consider the Larger datasets which is having an output of higher accuracy with the usage of recently classifiers that is not used by the previous research papers.

**3.3 DISADVANTAGE:**

1. When traditional classifier is used ,then the job prediction is unpredictable.

**3.4 REQUIREMENTS:**

**3.5.2 HARDWARE REQUIREMENTS:**

|  |  |
| --- | --- |
| Component | Description |
| * Processor * RAM * HardDisk * GPU | * Any intel processor above i3-5th Gen * DDR3/DDR4 abode 4GB * 120 GB or above * Any GPU above 2Gb with atleast 2MB Cache Memory |

**3.5.2 SOFTWARE REQUIREMENTS:**

|  |  |
| --- | --- |
| Software | Purpose |
| * Jupyter * Flask * PostMan * Xampp server * Atom * Anaconda | * Python Script * API building * API testing * MySql,Frontend * Python script building * Integration |

**3.7 NEW METHODOLOGY:**

1. To achieve more valid results there is a need for larger amounts of data i.e. large dataset.
2. Different classifier is used for predicting stock price movements.
3. Using and combining the three analyses will give more accurate result not only for the short term buying and selling of the stocks but also in the long run

The steps involved in the new methodology are:

1. Gathering the information from the existing system:
   1. Data Collection: Data is collected in Excel Sheet and extracted using pandas.
   2. Pre-Processing: The data is the classified into two sets - Training and Testing.
2. Apply Feature extraction and take with the best accuracy:
   * 1. Filter methods
     2. Embedded methods
     3. Wrapper methods
3. Calculating by using the following algorithms:
4. Decision tree(Comparison with this traditional approach)
5. Random Forest
6. SVM
7. XG Boost
8. Result with analysis is displayed using graphs that are plotted using mat plot.

**CHAPTER 4**

**SYSTEM DESIGN**

**4.1ALGORITHMS:**

1. Support Vector Machine Algorithm
2. Random Forest Algorithm
3. Logistic Regression
4. XG Boost

**4.2.1 SUPPORT VECTOR MACHINE ALGORITHM:**

It is a type of supervised machine learning algorithm which can be used in case of both regression and classification problem as well as for outlier detection. Here, a dot products of pairs of input data vectors is calculated. The computation is done by defining them in terms of kernel function k(x, y).

This is the type of a linear model which solve the problems related to linear and non-linear functions.It creates a line that acts as a hyperplane that is used to divide the data in the various classes.

Advantages of SVM:-

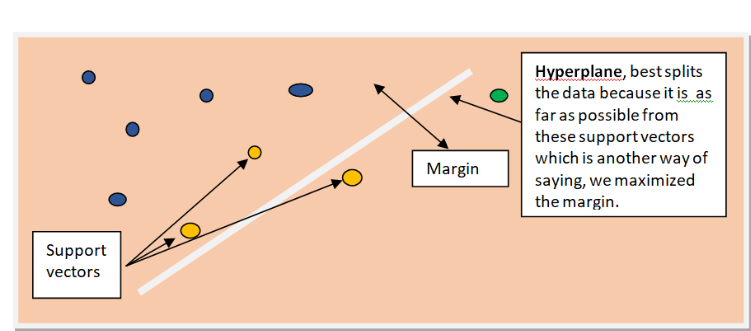
1-This algorithm works very well in high dimensional spaces.

2-If there are number of dimensions is bigger than the samples amount ,it is effective.

3-Memory efficiency is good.

Disadvantages of SVM:-

1-If the number of samples is much smaller than the number of features, it avoid over-fitting in choosing the functions of kernel.



**FIGURE 4.1:** SVM hyperplane

**4.2.2 RANDOM FOREST ALGORITHM**

Random forest as the name indicates is made up of trees which are randomly mixed. Decision Trees are the building blocks of random forest algorithm. Decision tree gives as all the possible solution present for a particular situation established on a decision. It starts from a node which is known as root and spreads into branches giving as multiple solutions to a single problem. The algorithm helps in finding the best solution possible to the problem rather than the most important solution. Hence the result is much better than other algorithms. For example, while planning for a vacation a person say X goes to different friends in order to suggest the places he should go. So when he went to his first friend, his friend asked question based on his previous vacations whether he liked it or not. Hence the friend gave X his opinion based on the answers. After this X goes and asks different friends. Then the most recommended place is chosen by X. Random Forest has a very great feature of calculating the importance of the different features at each level of the prediction since each level gives as the solution of the problem. This helps us to get the results easily since we can decide which solution to keep and which we can drop. Differences still can be seen between them. If we give different features to a dataset in a decision tree some set of rules will be generated which will give us a better prediction. For example, when an ad pops up on a person’s screen to predict whether the person will click on it or not is based on the person’s past history. Now if a decision tree is created by inputting the required features, we can get the prediction whether the person will see the ad or not. In this way a better prediction can be made. Another difference is in random forest smaller decision tree are created by taking the subsets of the required features for a particular problem.

**4.2.3XG Boost**

This calculation is like angle boosting system yet increasingly productive. It has both direct model solver and tree learning calculations. It's ability to do parallel calculation on a solitary machine makes it quick.

It underpins different target capacities including relapse, arrangement and positioning. It has extra highlights for doing cross approval and finding significant highlights.

This calculation permit to fit numerous feeble classifiers to reweighted renditions of the preparation information. It arranges last models by greater part casting a ballot.

While utilizing the boosting system all occasions in the informational collection are relegated a score that advises how hard to arrange they are. In every emphasis the calculation gives more consideration relegate greater loads to occurrences that were wrongly ordered already.

Pros:-

1-It is vigorous to the exceptions.

2-It ready to deal with the missing qualities.

3-It has computational dependability.

4-It can deal with blended predictors.(Qualitative and quantitative)

Cons:-

1-It can not extricate the straight mix of highlights.

2-Small prescient power as it has high difference.

**Weighted sample**

**Weighted sample**

**Weighted sample**

**Training sample**

···

**·**

**Mth week classifier**

**1st week classifier**

**2nd week classifier**

**3rd week classifier**

**Final classifier**

**Figure 4.2: XG boost algorithm**

**4.2.5 RESULT:**

After the data which is considered gives the required results the buyer can by any stock and the seller can sell its stocks. So that profit can be made. There is a process of comparing the values to previous and the next value. Therefore in this way the best prediction can me and results can be obtained.

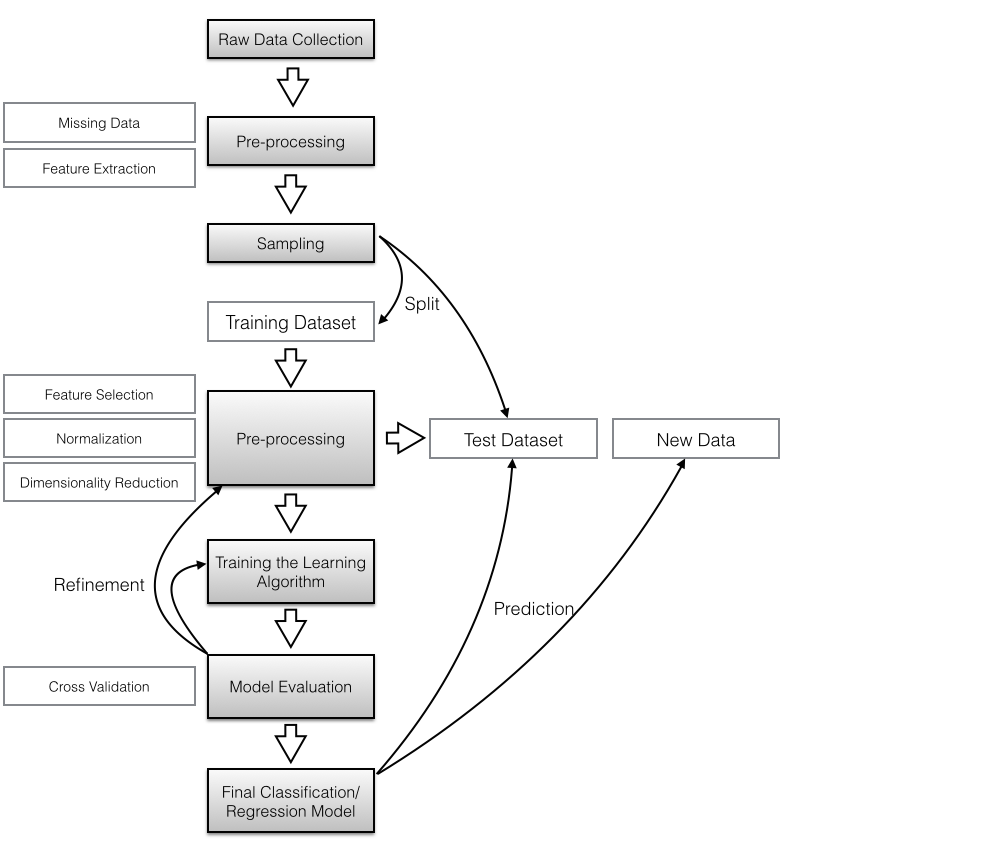
**4.2.6 REVIEW OF THE SYSTEM ANALYSIS:**

Based on all the reasons are considered it is clear that the input decides everything and plays a vital role in prediction of future. We have to implement the process on all the datasets which are considered. The results obtained in all the cases are accurate not fully but can be considered. Hence the final results are showed to the users.

And by using all the different algorithms we can say that more the number of features relates to particular problem more will the prediction will be accurate.

* 1. **DESIGN DIAGRAMS**

**4.2.1 ARCHITECTURE DIAGRAM**

****

**FIGURE 4.3:** Architecture of System

**4.2.2 PROPOSED SYSTEM ARCHITECTURE:**

**Data collection**

**Data pre processing**

**Feature selection and Model Selection**

**Front End-GUI**

**Flax API**

**Final model from outcome**

**Data base**

**FIGURE 4.4:** Proposed System Architecture

In the proposed architecture the data is fetched then processed. After the training of the

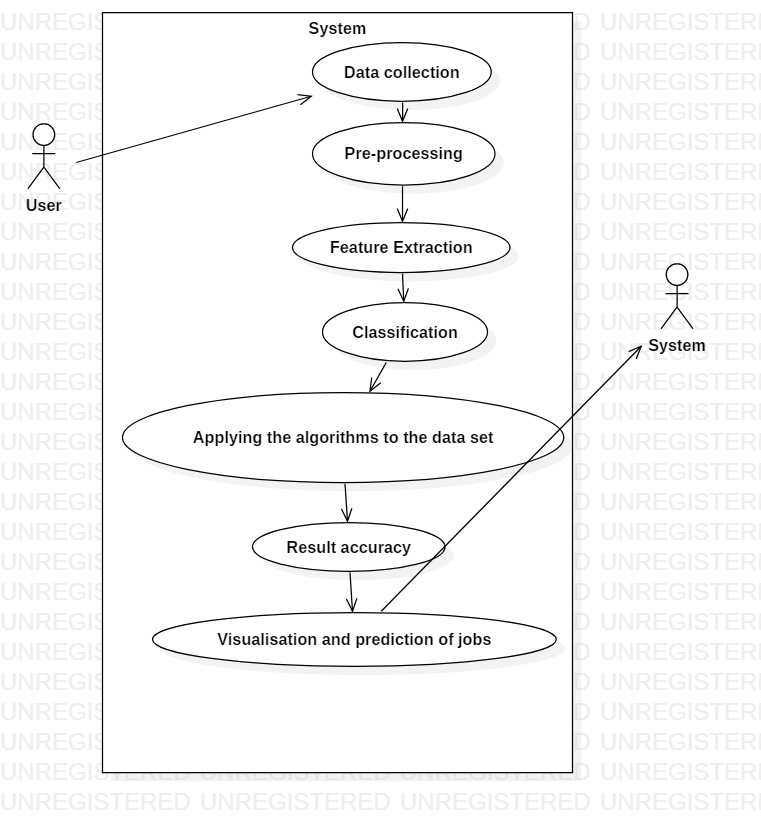
dataset the appropriate model with best accuracy is taken. Then the training part is done.

The model we will select ,we will use this model to train in the Flask API to get thE

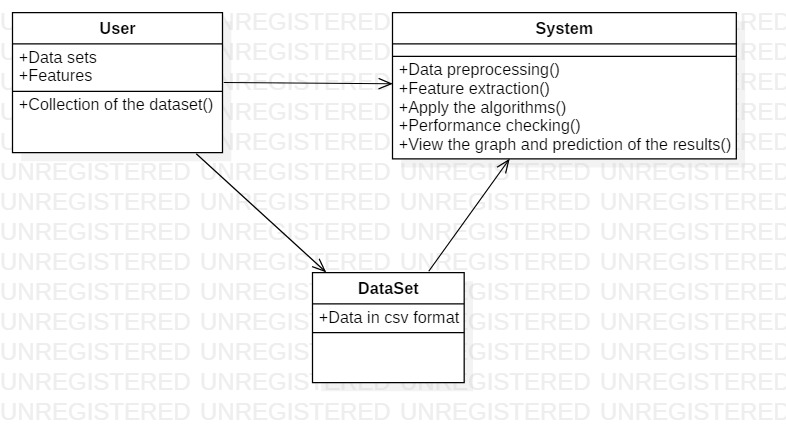
prediction done and then this will also act as a backend to the Front end gui.

The user will be able to view the results according to the job skills, he/she will have.

**4.3.3 USE CASE DIAGRAM:**

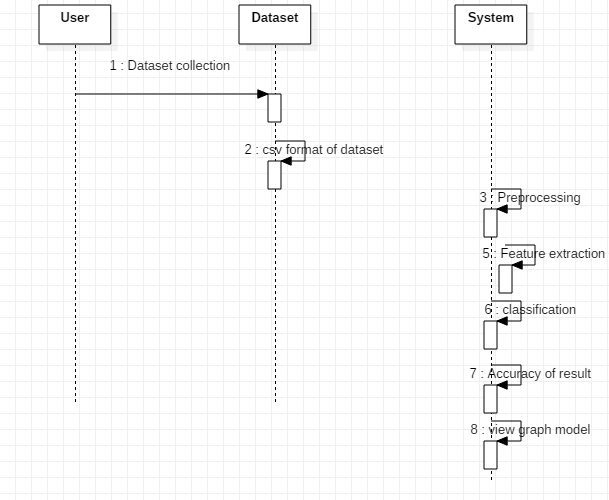
**FIGURE 4.5:** Use Case Diagram

**4.3.4 CLASS DIAGRAM:**

****

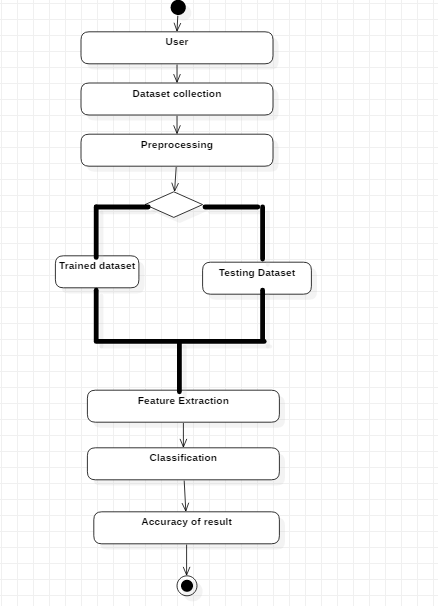
**FIGURE 4.6:** Class Diagram

**4.7 SEQUENCE DIAGRAM:**

****

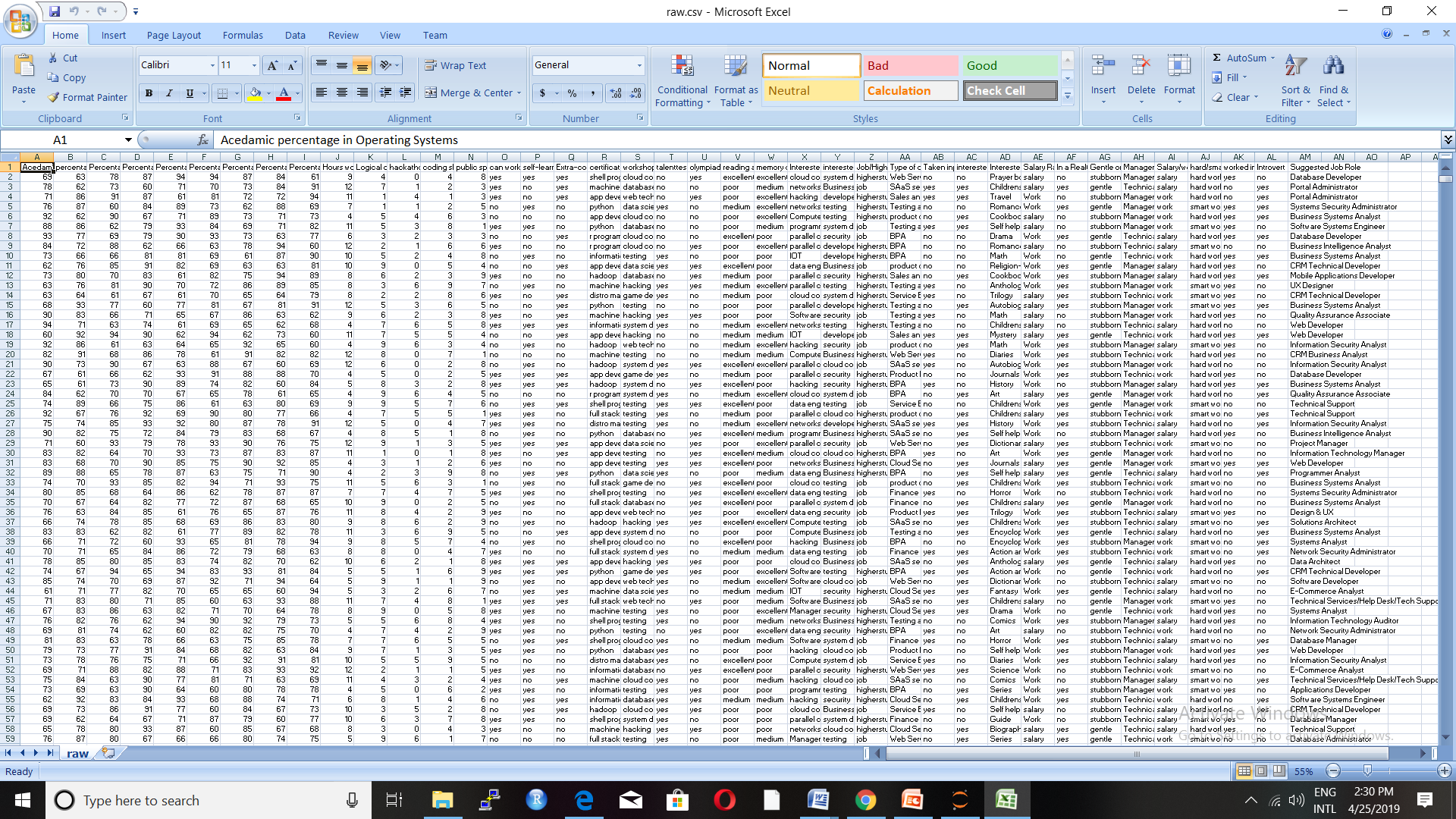
**FIGURE 4.7:** Sequences Diagram

**4.8 ACTIVITY DIAGRAM:**

****

**FIGURE 4.8:** Activity Diagram

**4.3.7 DATASETS:**

****

**FIGURE 4.9:** Datasets

The above data set consists of 20000 rows and 39 coloumns include the target variable “SUGGESTED JOB ROLE”.

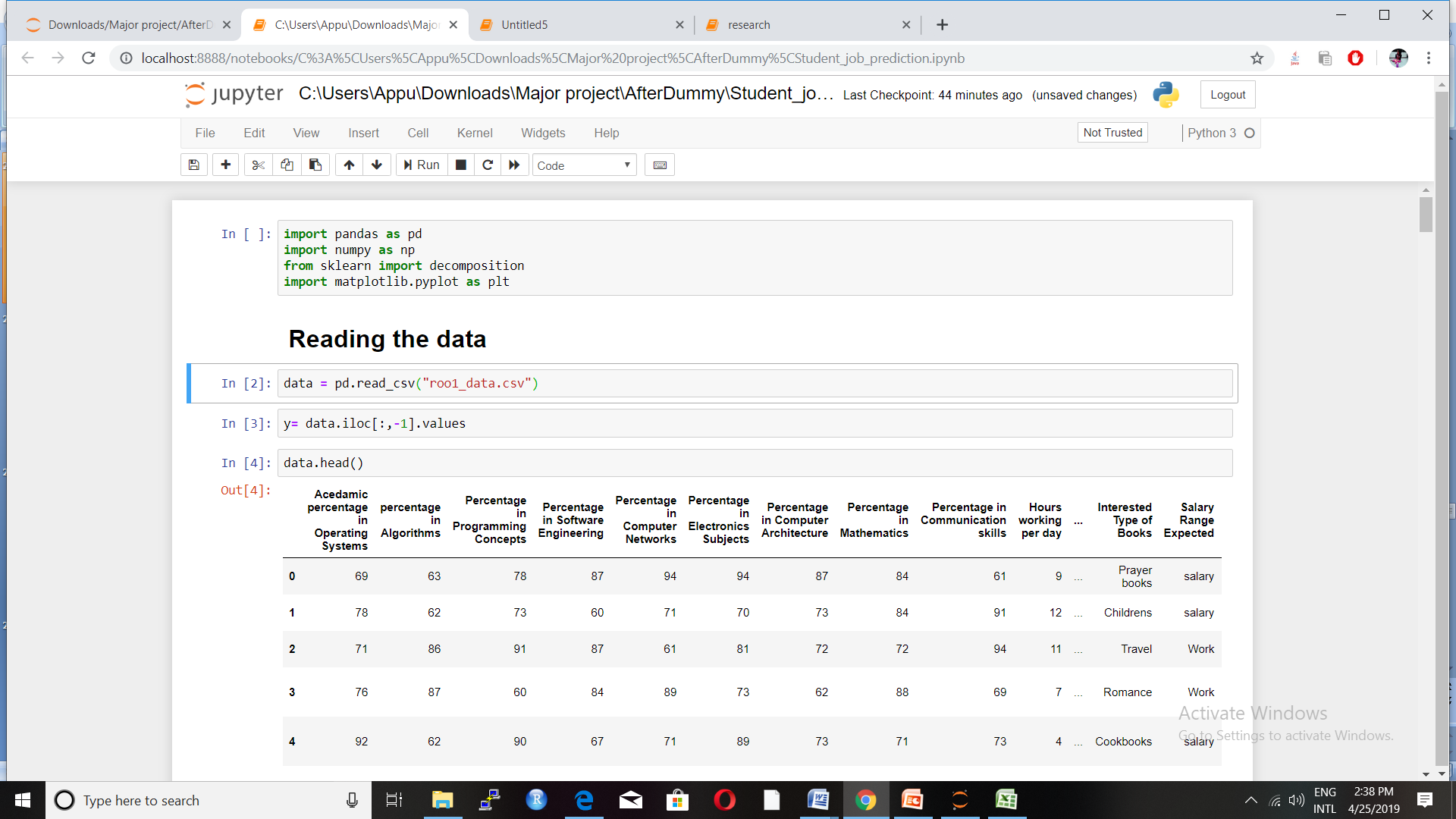
**CHAPTER 5**

**IMPLEMENTATION**

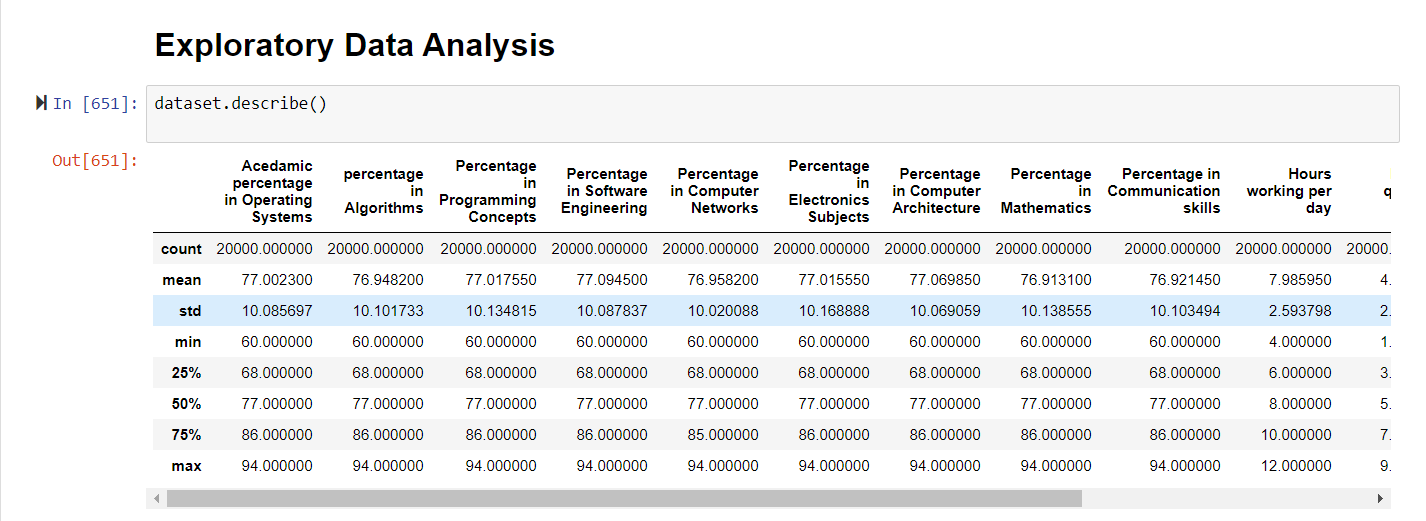
The student job prediction in our project is done using different attributes related to the particular area of interests and the various academic details.

**5.1 GATHERING INFORMATION FROM PRIOR SOURCE**

The information required in the project is the various details of the student including the interests in the various academic fields as well as in the extra-curricular activities,that can be downloaded from various websites and are inputted in the project as excel or CSV files. Since we are using python to code our project the pandas directory is used to input the data. Other directories such as numpy is used for inputting arrays, sklearn is used to test and train the data, apply the SVM algorithm , XG Boost, Random forest(Regressor and classifier) and the matplot is used to plot the graphs and provide a visual result.

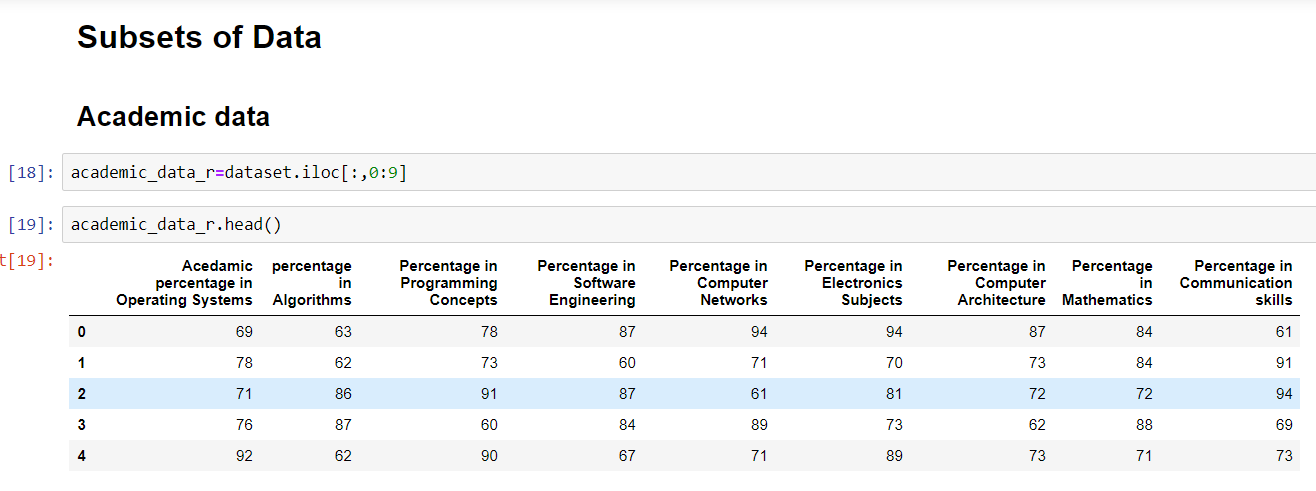
**FIGURE 5.1:** Collection of Data

**Discription of Data**

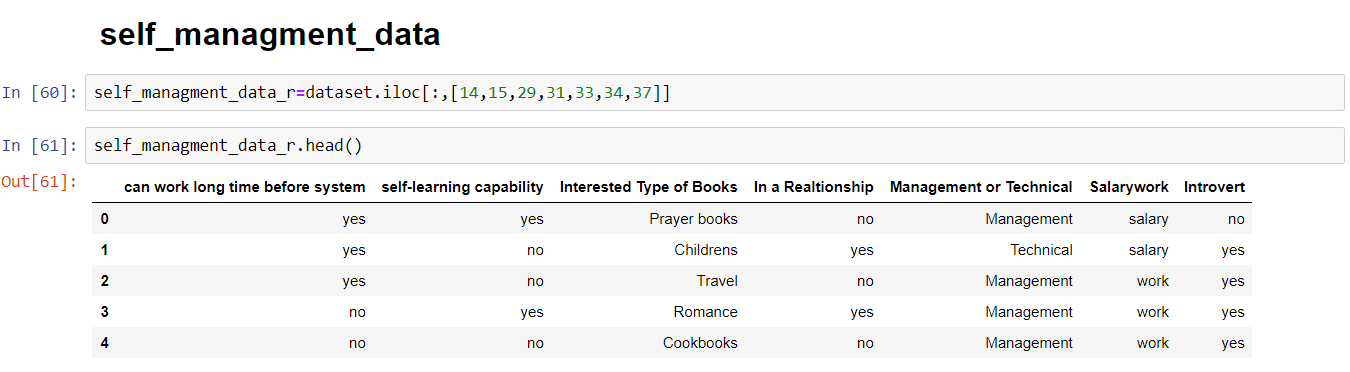
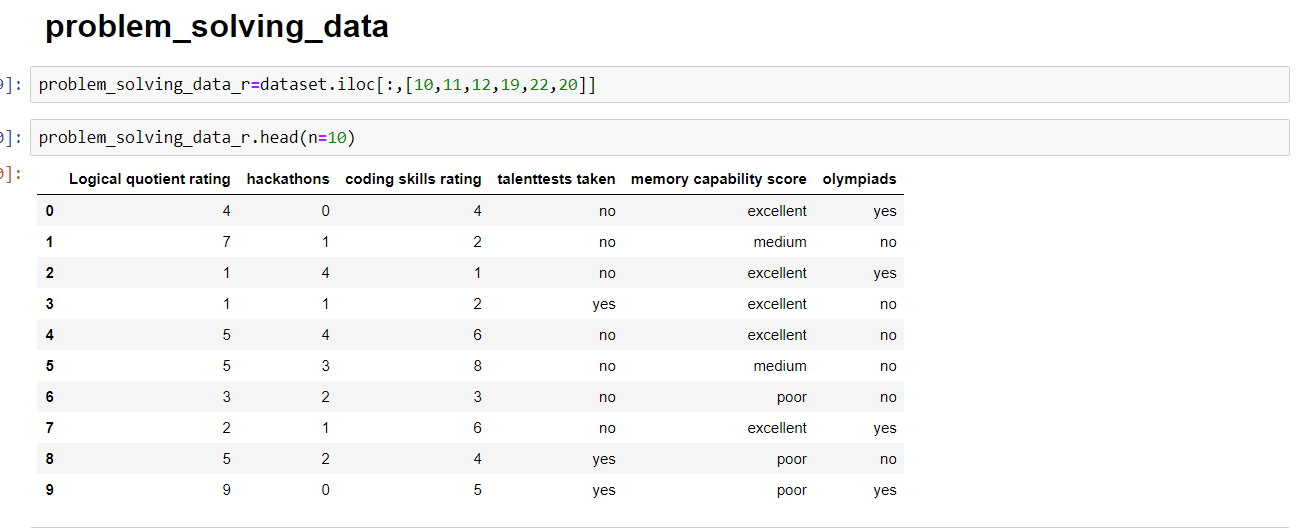
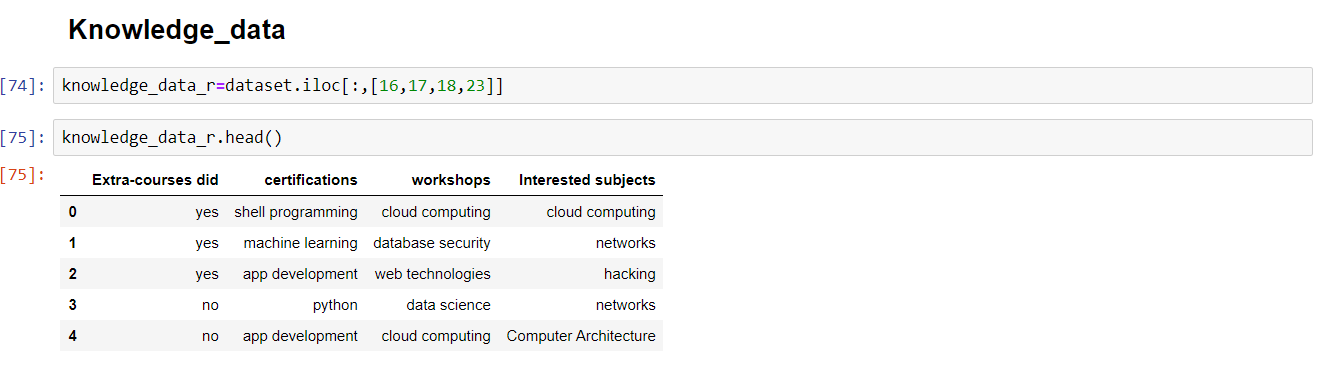
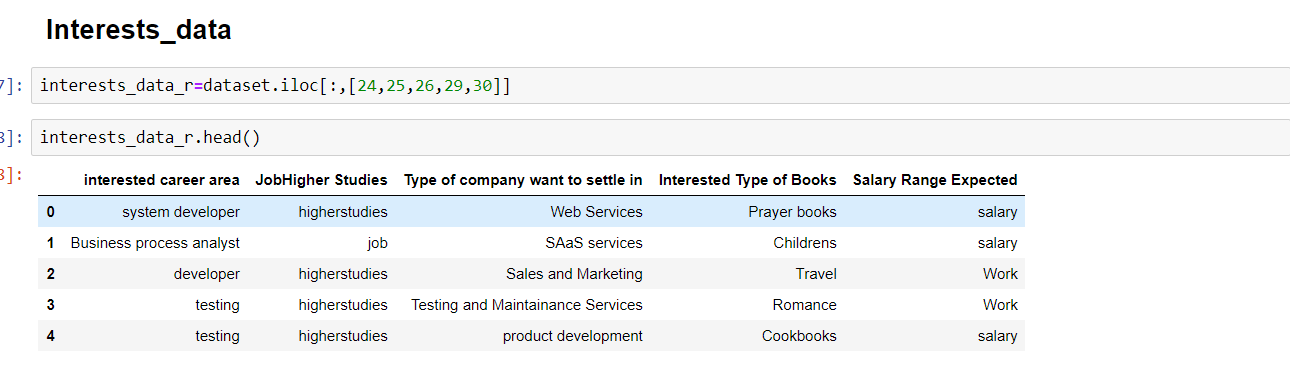
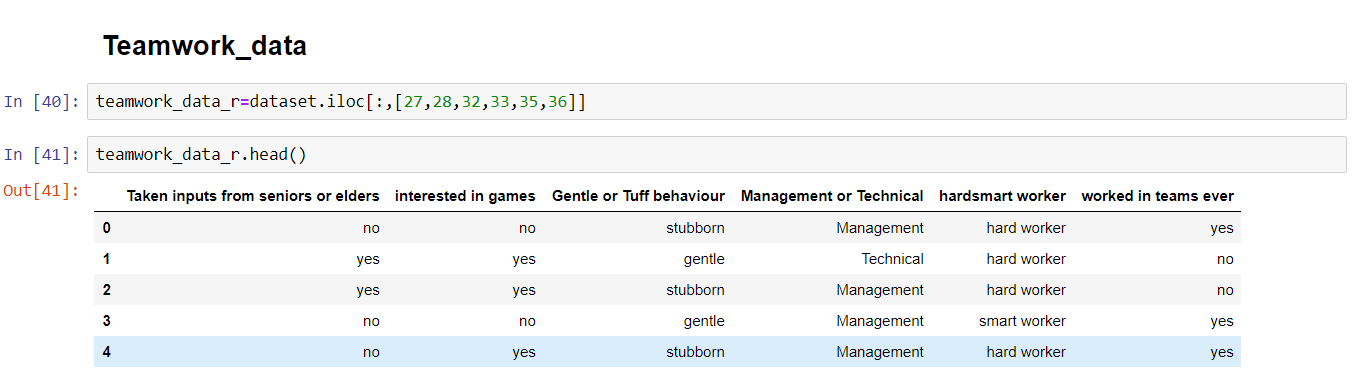
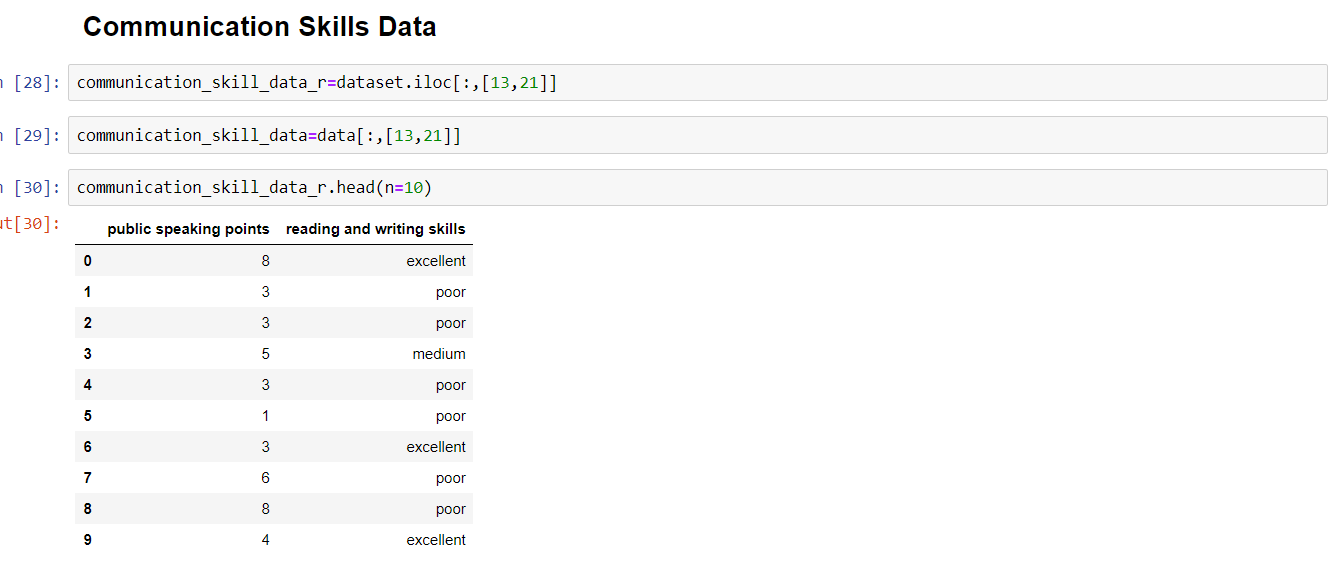
****

5.2 Getting Data Subsets

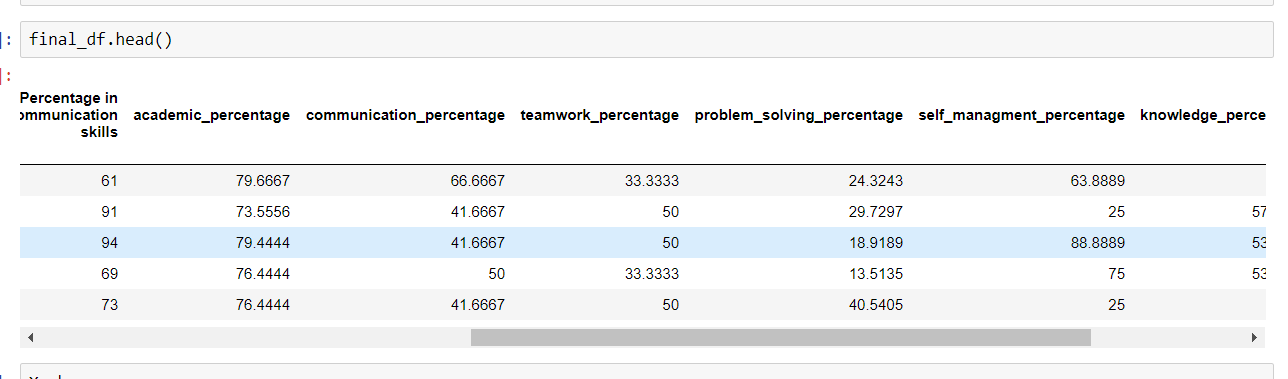
Academic Data



Communication related data

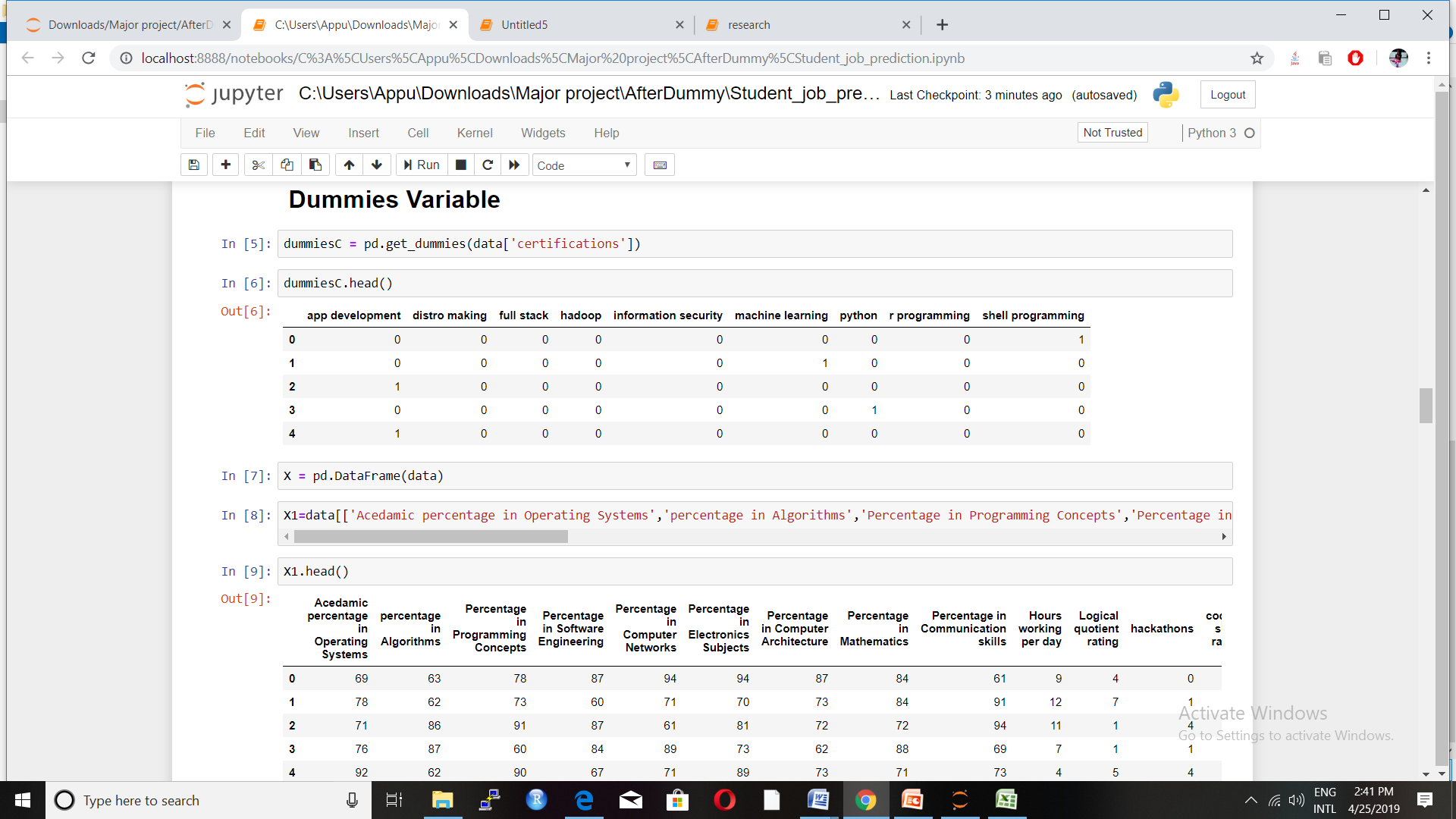


**Final Data from subsetting**

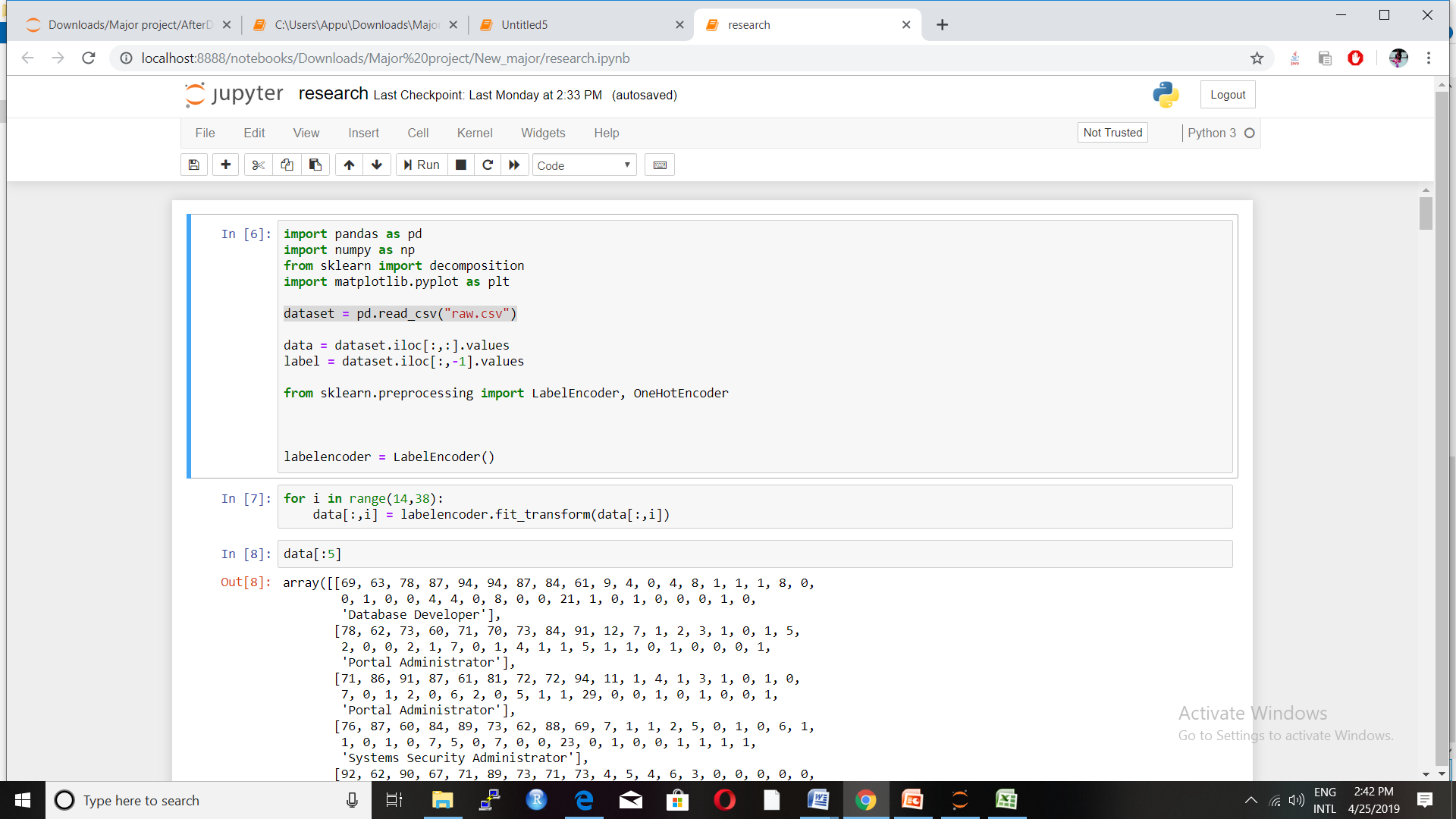
****

**5.2 CALCULATION FOR PLOTTING THE VALUES**

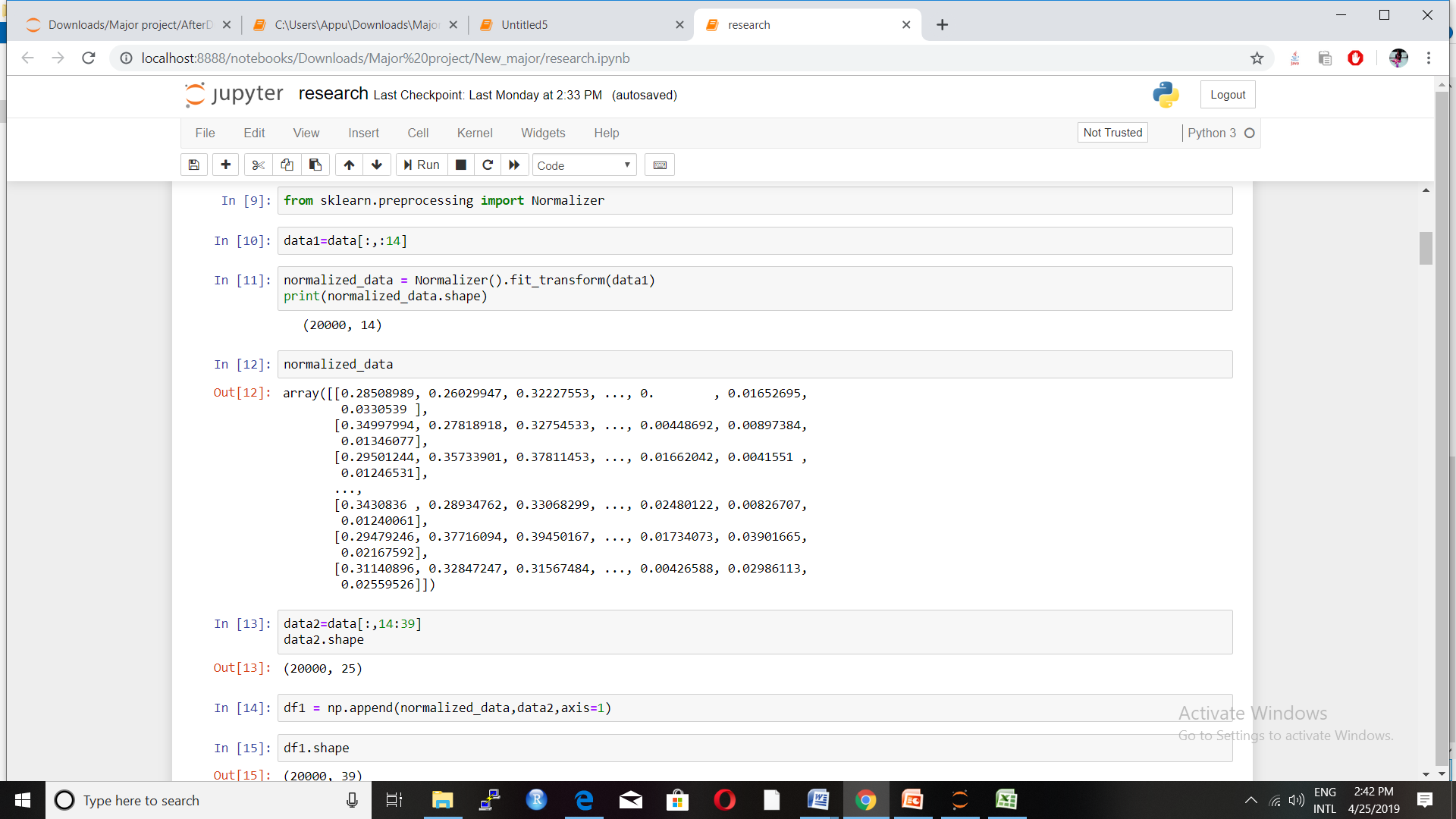
The data present with us –



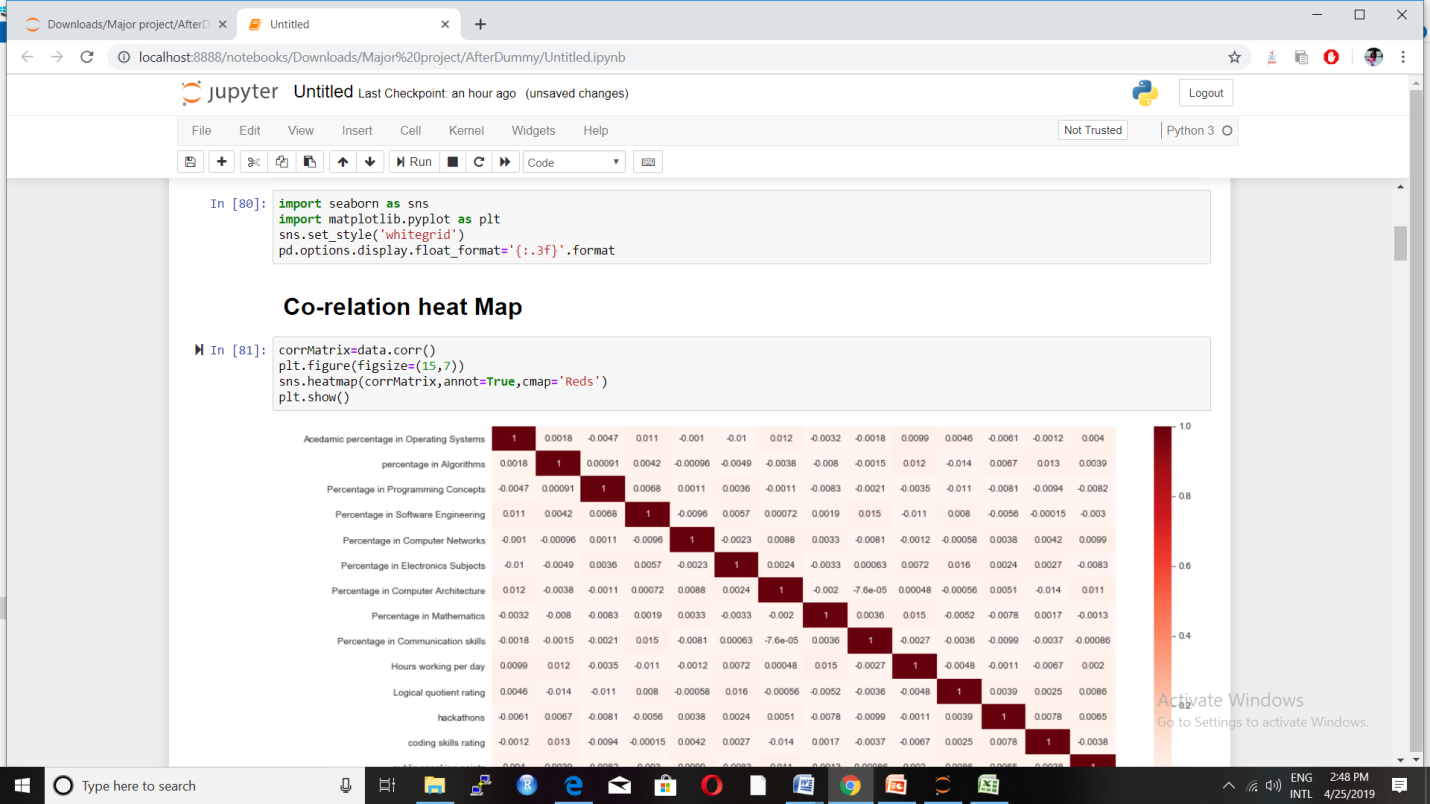
**FIGURE 5.2:** Check for the dummies variable

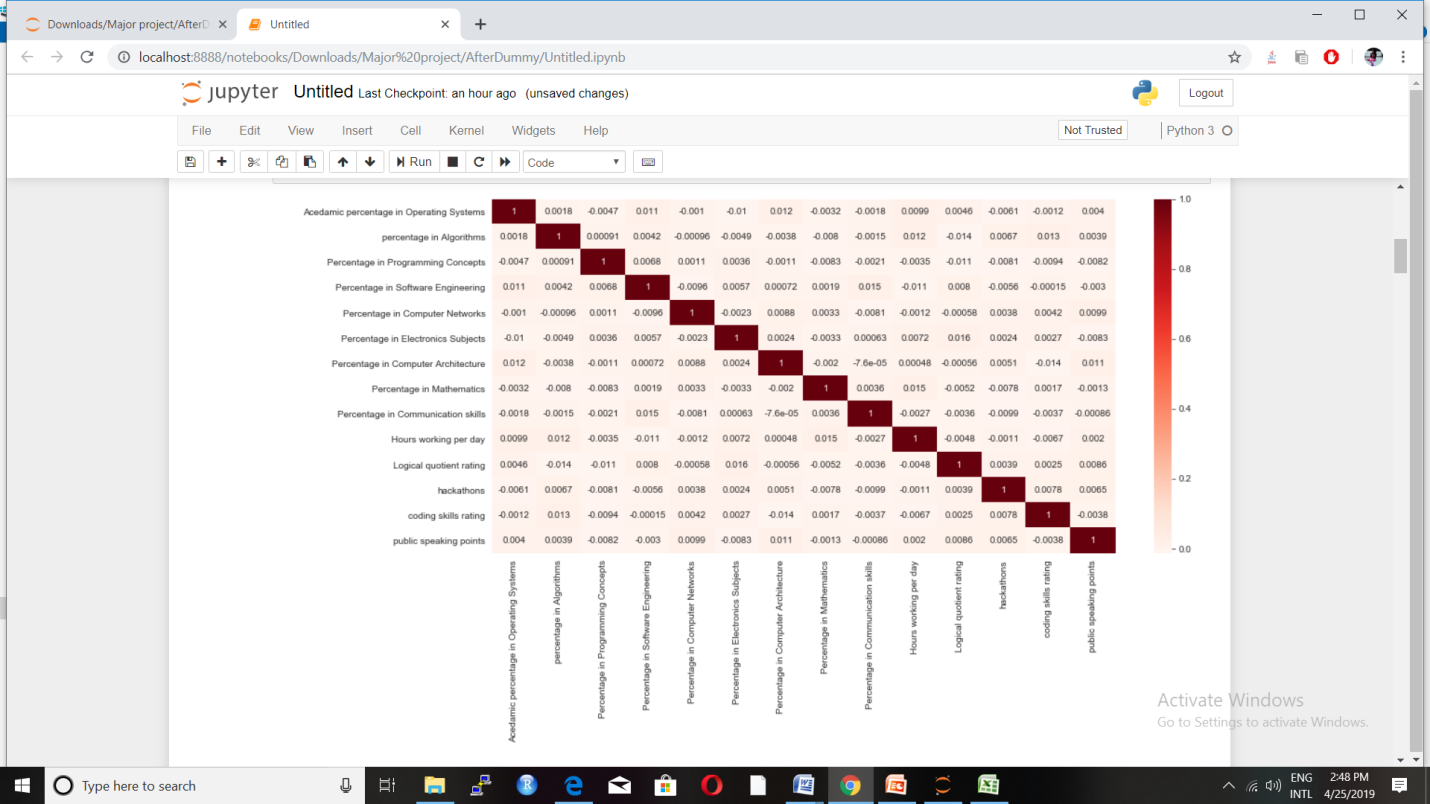


**FIGURE 5.3:** Encoding done here for the categorical coloumns

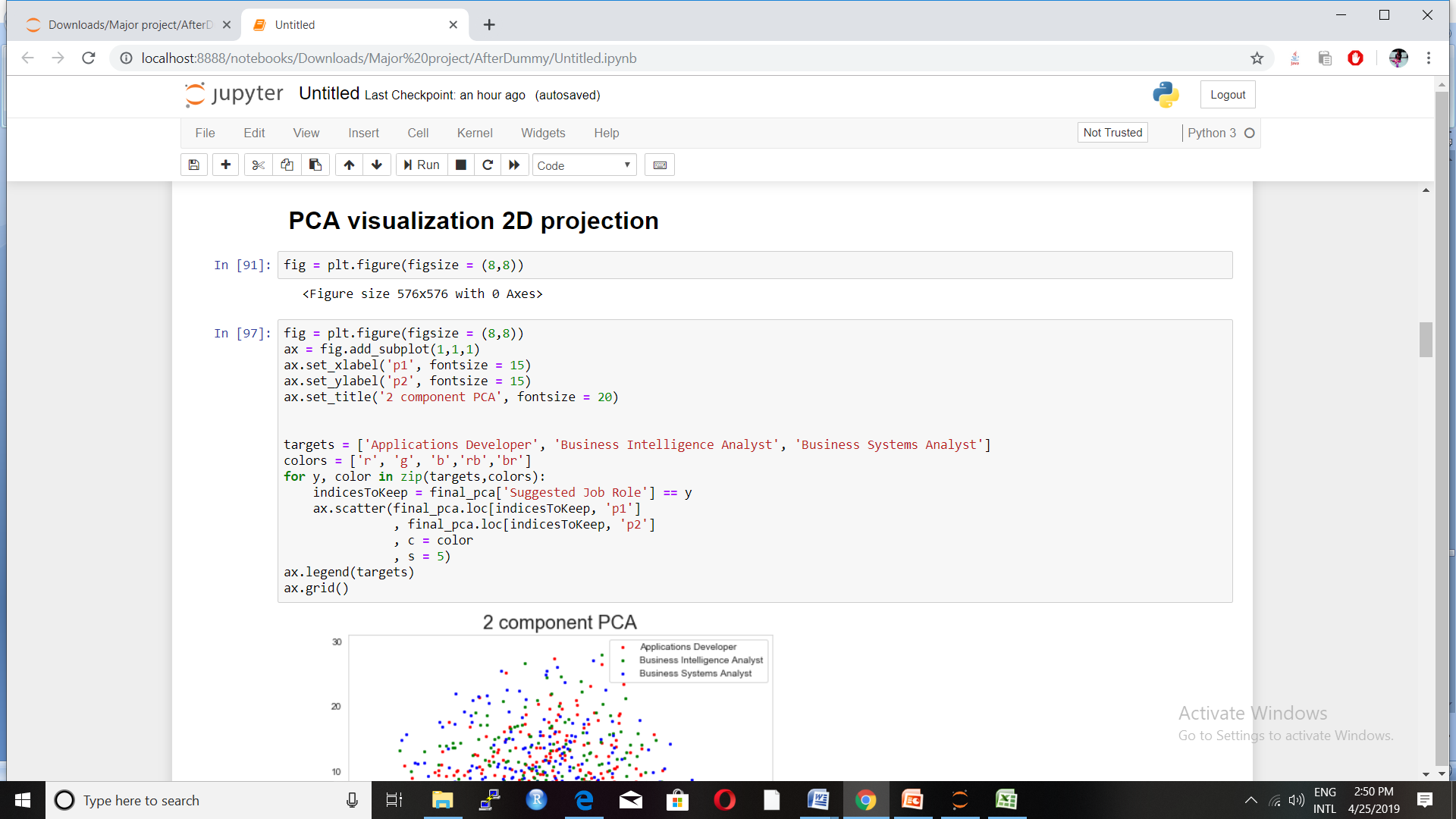


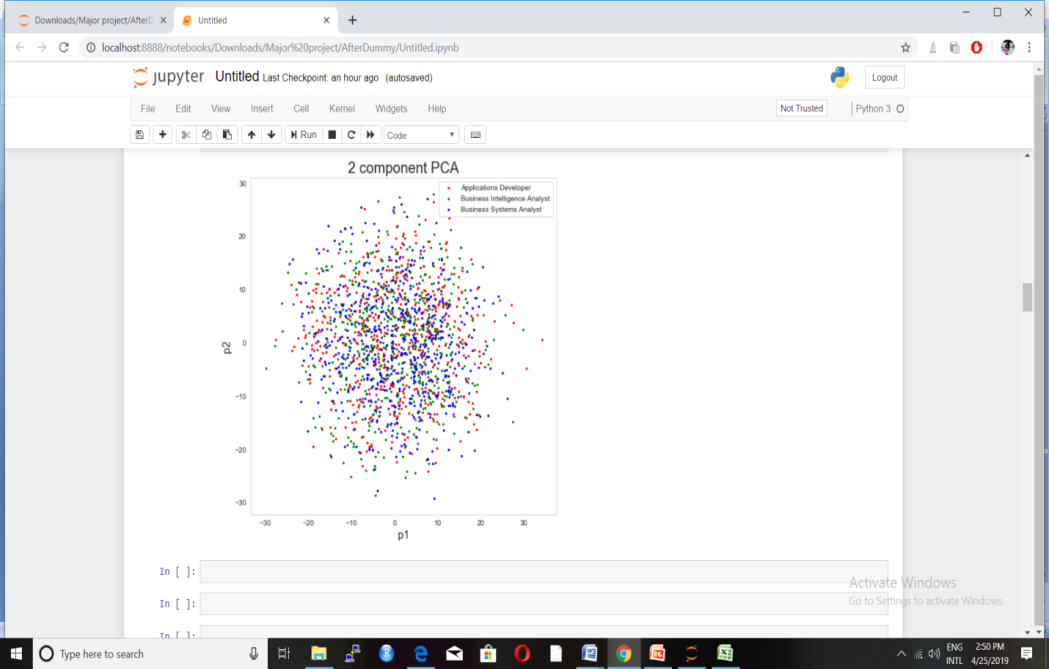
**FIGURE 5.4:**Normalisation for the first 14 coloumns



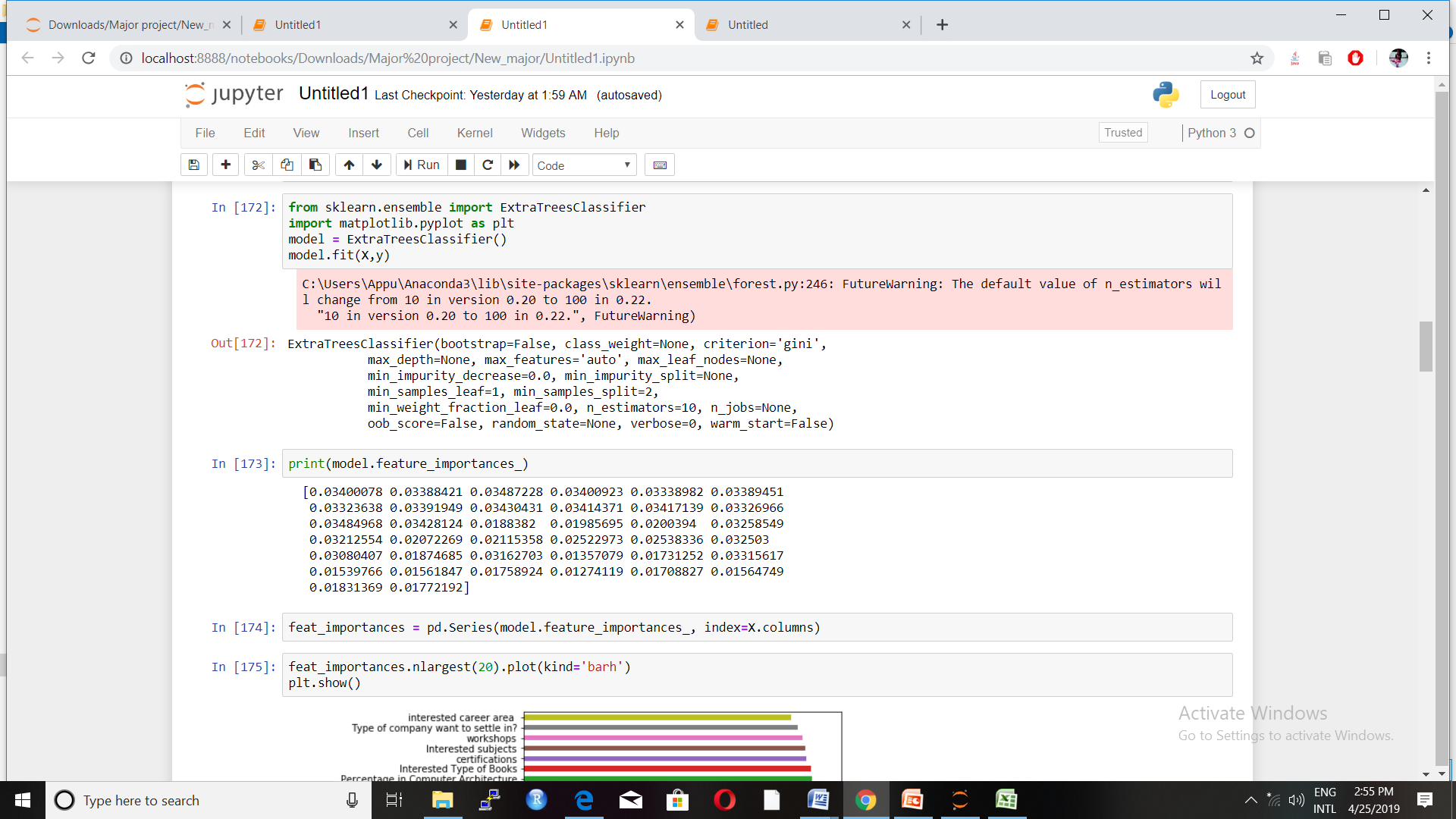


**FIGURE 5.5:** Correlation heatmap

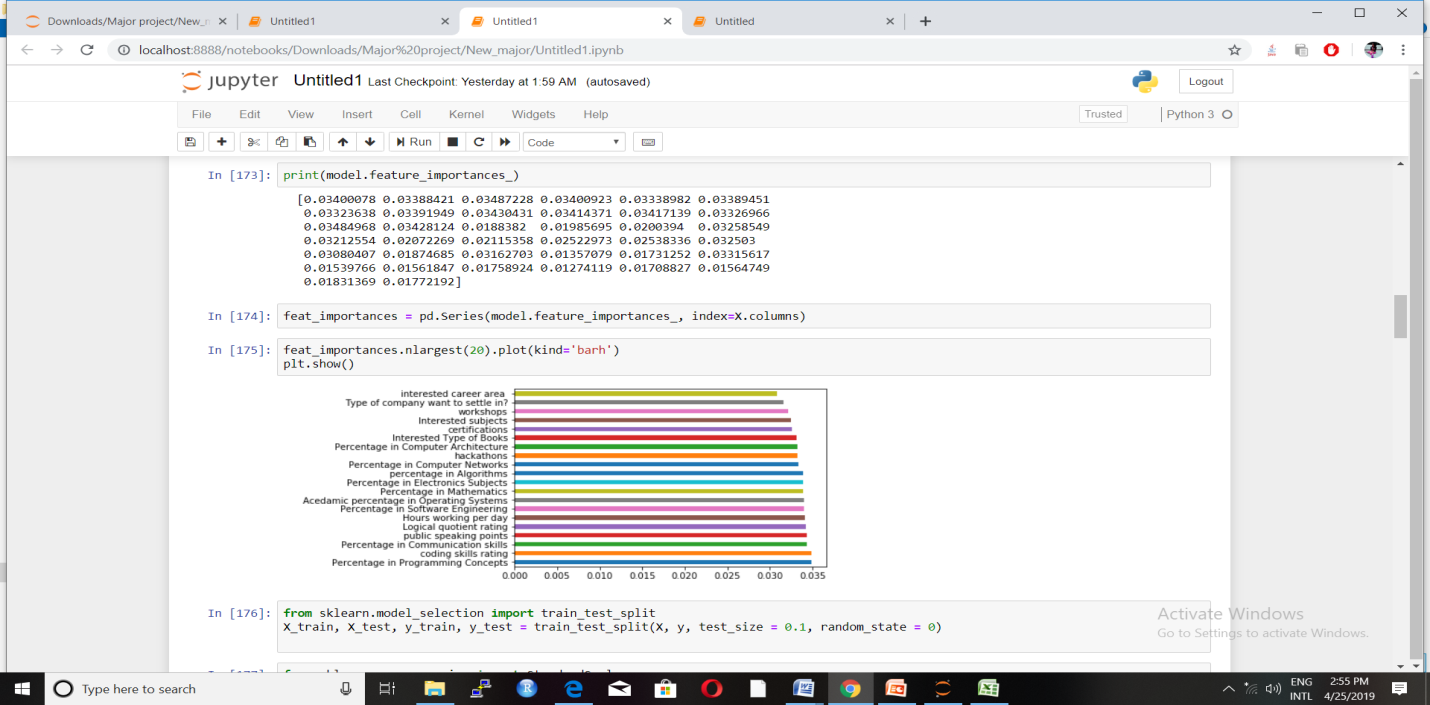




**FIGURE 5.6:** PCA visualization with the two compressed data sets having P1 and P2



**FIGURE 5.7:** Feature Importance

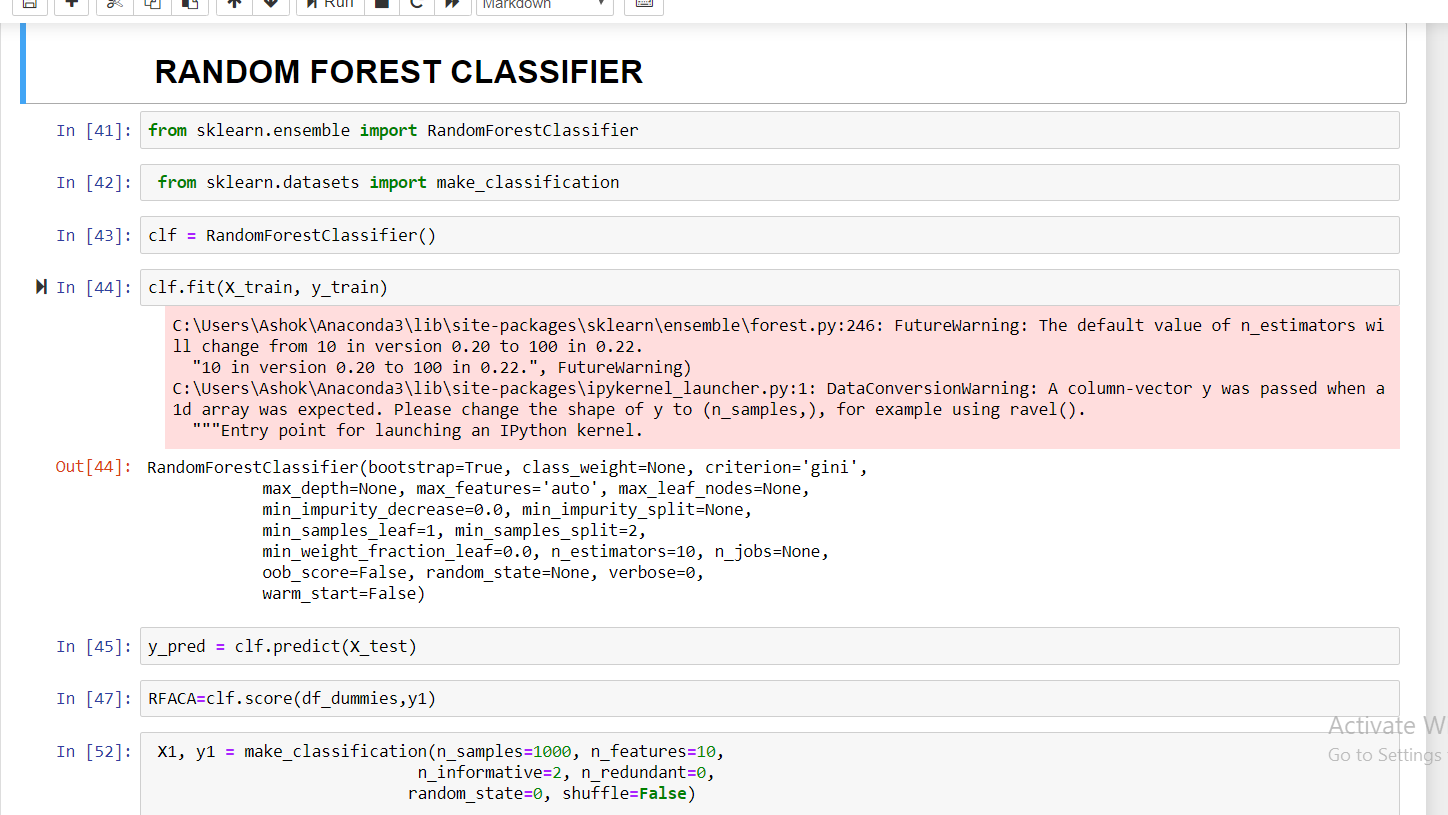


**FIGURE 5.8:** Feature Importance visualization

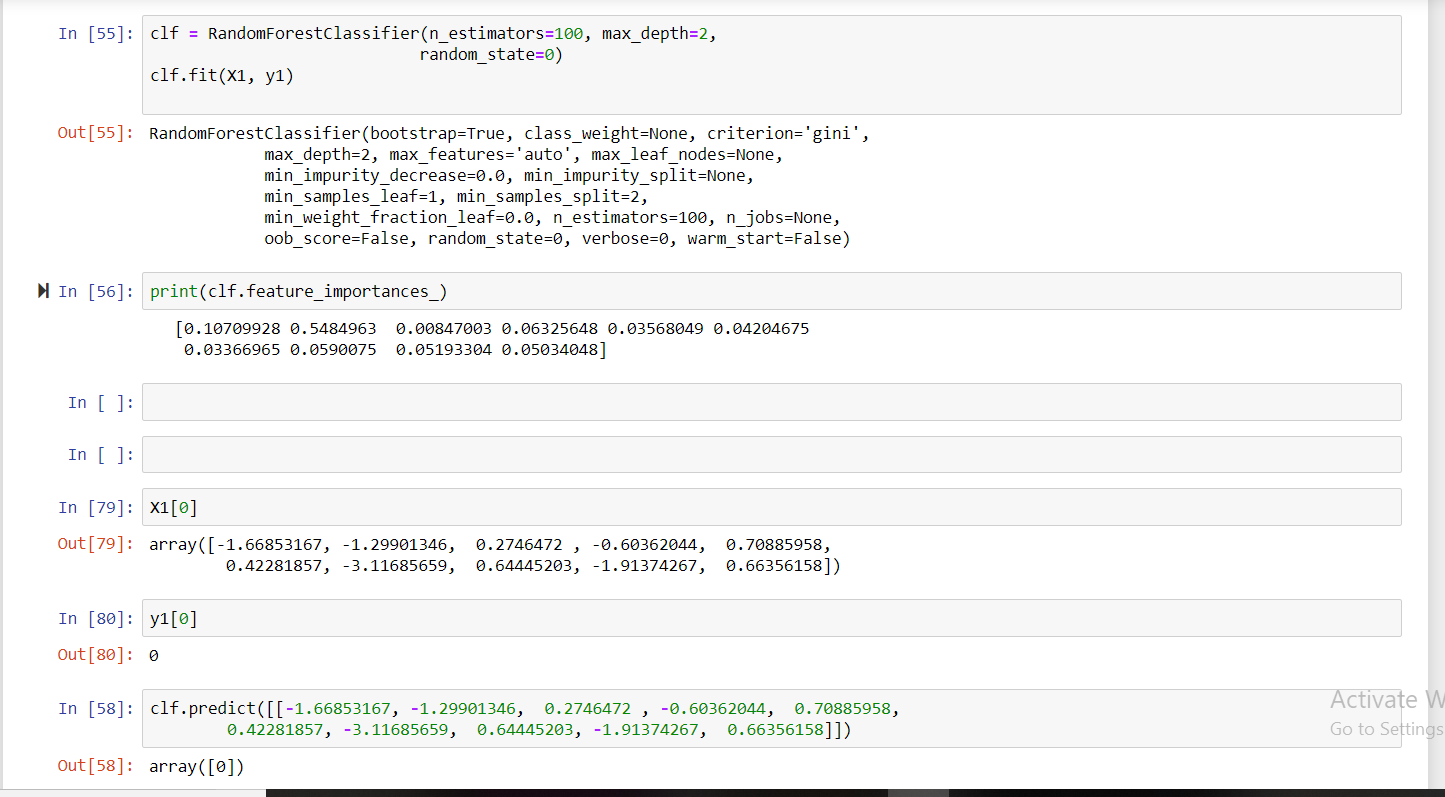
**5.3 APPLYING ALGORITHMS FOR BEST ACCURACY**

**5.3.1 RANDOM FOREST**

Random forest is useful in classification , this algorithm helps in lots of decision tree which recursively breaking and creating decision supports reaches down to calculate results.Finally results are compared and calculated. Before implementing Random Forest proper parameters are designed and defined, then one can train model more efficiently and increasing accuracy



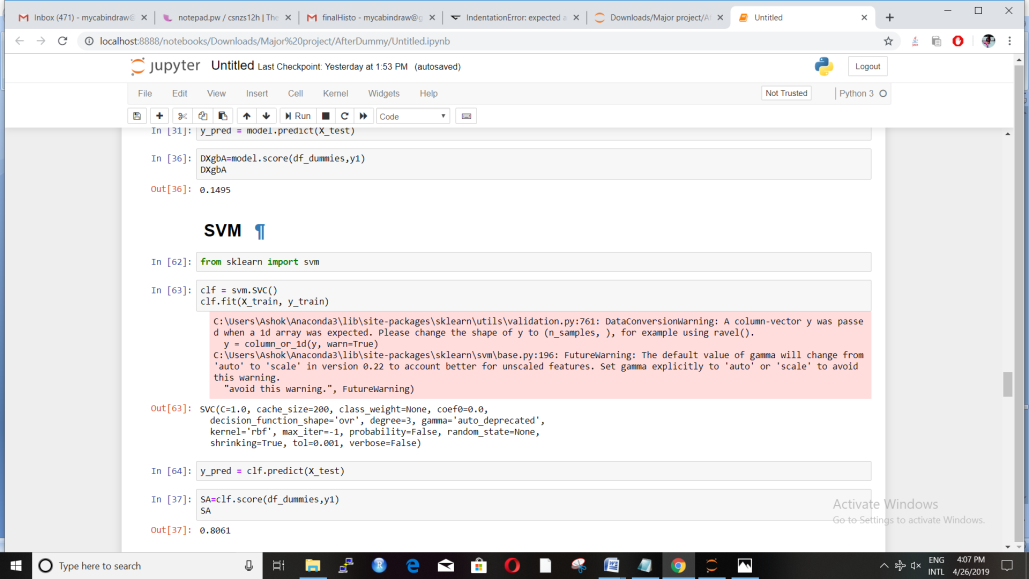
**FIGURE 5.9:** Prediction by RFA



**FIGURE 5.10:** Random Forest Algorithm

**5.3.2 SUPPORT VECTOR MACHINE**

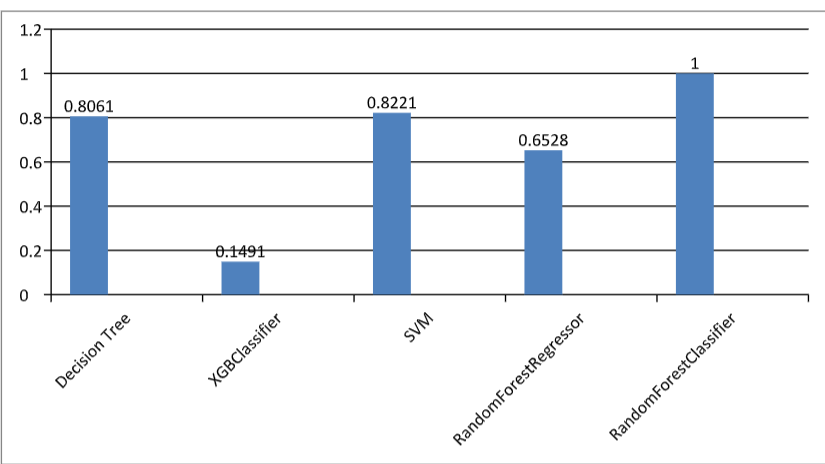
The SVM or the support vector machine algorithm is used for both classification and regression. This is used to make clusters of similar data and with the help of a hyper-plane separating that similar types of data.



**FIGURE 5.11:** Prediction by SVM

**5.4 COMPARISON OF ALGORITHMS**

The efficiency of all the algorithms are compared and RANDOM FOREST CLASSIFIER accuracy came to be highest i.e. 100 percent.



**FIGURE 5.13:**Efficiency of algorithms

**5.5 Flask API**

Flask is platform t creat and deploy python api and intigraton with front –end such as html,csspags or Java script .Once finalized model is build and ready to predict, an API is needed to predict through python function a model is saved in PKL formet through pickle library, furtur its is loaded into flask api. Flask supports python so cross platform app can be easily crated. To create a flask api , a function is crated , that includes loding of model from directory in server, then api request front end to get required input data . in our project, json POST request is created,that takes input in JSon format through http request to java Script. Input data is now formatted into python dataframe , input is passed to function . An output is stored at SQL Database with user ID and as well as output for front end, again output data is formatted into Json format to send the api output.

When Frontend request for api, api invokes through javascript function

Input= UserID

Details={ UserId, accadmica\_details data[],extrac\_activity data2[] }

If data is not not provided well,, it invoked api try to get data from database accordingly user,

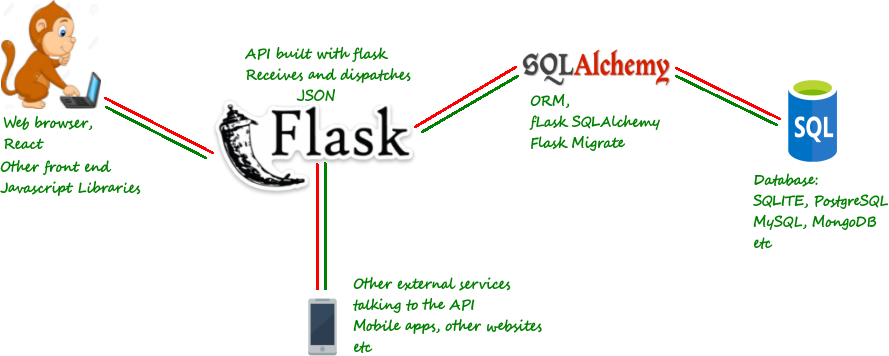
If it is also not available, then front end will ask for updating details,

Flask makes this error handing robust,and gives control over data and related queries,

Flask api supports JSON data format so its is easy to define dataframe in python acoordingly the input.

DataInput Format in JSON= {

UserId=’’val’’,acadmicDetails1=’’’val2’’,name=’string1’’,lastname=’’string2’’ }



**FIGURE 5.14: FLASK DIAGRAM**

**CHAPTER 6**

**SOFTWARE TESTING**

**6.1 OBJECTIVES OF SOFTWARE TESTING**

To finalize the developed application for meeting the business requirements, it is necessary

to test the software created. This gives a great test inclusion to the software and increases the

performance level of the application as per specifications.

Uses of software testing:

1. It is used to meet the end user business requirements and to do the system specification.
2. Quality of the application is improved for the better customer satisfaction.
3. For the delivery of the best quality product or the software application ,testing is important so that it results into more reliable, accurate and consistent results .
4. It is used to remove all the errors and bugs which might have come when the application was being developed.
5. To check whether all the functionalities of the developed product or application is met.

**6.2 TYPES OF TESTING**

1. Unit Testing
2. Integration Testing
3. Functional Testing
4. System Testing
5. White Box Testing
6. Black Box Testing
7. Acceptance Testing
8. POSTManApitestings

**6.2.1 UNIT TESTING**

**Unit Testing is method to check individual units or components of the software. Its main purpose is to validate the functionalities of the developed product. It will validate all key design specifications and test whether all things are working properly. The minimum part of a software which can be tested is referred to be as a unit.**



**RESULTS**

**MODULE TO BE TESTED**

**SOFTWARE ENGINEER**   **TEST CASES**

**FIG: 6.1 UNIT TESTING**

**6.2.2 INTEGRATION TESTING**

The phase in which singular programming modules are consolidated and tried is called Integration Testing. This phase comes after Unit Testing. The input will be unit tested prior to applying test cases for integration Testing. Grouping into larger piles is done and then defined test scenario is applied to aggregates which outputs the overall integrated system.

**6.2.2.1 PURPOSE**

Integration Testing verifies performance functional, non-functional and reliability requirements based on the main design items. These items are applied through their respective interfaces with the help of black box testing. Success and error cases are derived through applicableinput data and parameters.Construction of test cases are done to check whether all components interact correctly. This is done after each module has been individual tested i.e. Unit Testing. The idea of “building blocks” is applied over here.

Ttypes of integration testing:-

1. Big Bang
2. Top-Down
3. Bottom-Up

There are some other integration Patterns which are

1. Collaboration
2. Backbone
3. Layer
4. Client/Server
5. Distributed
6. High Frequency

**6.2.3 FUNCTIONAL TESTING**

It is a testing where the developed system is tested against functional requirements. It is a quality assurance process checking the quality of the functionalities under scrutiny. The testing is done by providing them input and then checking whether the desired output is achieved or not. It basically gives us the answer to what the system does.

Steps in Functional Testing:

1. It identifies the kind functions the software can perform.
2. Generate input data with the help of blueprint.
3. Determine the output based on the specifications.
4. Test Case Execution.
5. Compare certain and expected outputs.

If the user itself creates test conditions prioritizing the actual requirements, then Functional testing works more effectively. It is not concerned as in how the processing is happening but with the output which the processing results. It does not make any system assumptions but simulates the actual system design.

**6.2.4 SYSTEM TESTING**

System Testing comes where a complete software is tested. The test confirms system compatibility with the specified requirements. It falls within the sphere of Black Box testing where the tester is unknown of the logic of the system.

The input are all components of the system that have passed integration Testing. The main advantage of system testing is detection of any kind of inconsistencies between assemblages and the hardware It attempts to identify absconds both inside the framework and framework overall.

System Testing is performed on the whole system keeping in mind the FRS and SRS. It not only tests the design but also goes beyond boundaries defined in the requirements specification.

Various types of testing which are considered during System testing:

1. Graphical UI testing
2. Usability testing
3. Software performance testing
4. Compatibility testing
5. Exception handling
6. Load testing
7. Scalability testing

**6.2.5 WHITE BOX TESTING**

White Box Testing testicles inner structures of the application or checks whether the entire framework is working effectively or not. In this experiments are structured utilizing the assistance of programming aptitudes just as interior viewpoint of the framework. The analyzer decides yields appropriately while picking the contributions to practice ways through code.

White Box Testing is done at unit level but can also be applied at integration and system level of software testing. It can function within a unit, between units while integration, and also between subsystems.

**6.2.6 ACCEPTANCE TESTING**

Well Acceptance Testing tests for the system for acceptability is well defined. System’s is checked with both users acceptance and developer acceptance, so it is required to test the system for acceptance to present to higher authorities and stake holders.

This kind of testing is finished remembering client and business necessities. It is performed to check whether the framework fulfills all acknowledgment benchmarks and to empower the client or any approved organization to acknowledge the framework. Maximum,Acceptance testing utilizes Black Box testing structure strategy. Testing is impromptu and does excluded any severe technique. It is for the most part performed at the last after the framework testing and before the last sending.

**TYPES OF ACCEPTANCE TESTING:**

**1-** Alpha Testing.

2- External Acceptance Testing

3- Beta Testing

4- Customer Acceptance Testing

**6.2.7 POST MAN API testing**

Once a api is build in flask, it is saved as python file with .py extension, it is now uploaded to postman software to apitesting,this software gives a integration of front end input without building front end , to test the api it is loaded to file Destination , then it the choose option to give input data , which is in our case Json Input with userId,

Once you give input it send input to api, fetch the results and shows it, if the results are as desired, api working fine.

**6.2.8 BLACK BOX TESTING**

There are the various uses of black box testing:-

1- Testing of functional validity.

2- It is used to check the system behaviour and the testing performance.

3- If the system is sensitive to the values which we are giving as a input.

4- How much the system can bear the volume of the data.

**requirements**

**events**

**input**

**output**

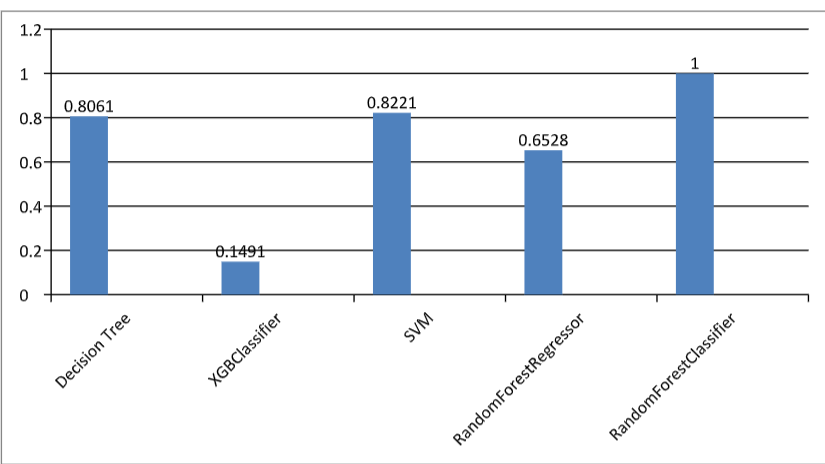
**FIG:-6.2 BLACK BOX TESTING**

**CHAPTER 7**

**CONCLUSION**

As per the algorithms used that are random forest regressor, support vector machine, logistic regression, random forest classifier, XG Boost,Decision tree, we found that **Random forest classifier**has the best output for doing the prediction with the best precision and accuracy.

We will use random forest classifier to train the data sets and to predict the best job suitable for the candidate.



**FIGURE 7.1:** Conclusion of different algorithm

**CHAPTER 8**

**REFERENCES**

1- https://www.sciencepubco.com/index.php/ijet/article/view/11738/4565

2- https://www.analyticsvidhya.com/blog/2018/08/dimensionality-reduction-techniques-python/

3- P.KaviPriya, “A Review on Predicting Students’ Academic Performance Earlier, Using Data Mining Techniques”, International Journal of Advanced Research in Computer Science and Software Engineering

4- Ali Daud, NaifRadiAljohani, “Predicting Student Performance using Advanced Learning Analytics”, 2017 International World Wide Web Conference Committee (IW3C2).

5- Marium-E-Jannat,SaymaSultana,MuniraAkther, “A Probabilistic Machine Learning Approach for Eligible Candidate Selection”, International Journal of Computer Applications (0975 – 8887)Volume 144 – No.10, June 2016

6- SudheepElayidom, Dr. Sumam Mary Idikkula, “Applying Data mining using Statistical Techniques for Career Selection”, International Journal of Recent Trends in Engineering, Vol. 1, No. 1, May 2009.

7- Dr. Mahendra Tiwari ,Manmohan Mishra, “Accuracy Estimation of Classification Algorithms with DEMP Model”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 11, November 2013. 8- Ms. RoshaniAde,Dr. P. R. Deshmukh, “An incremental ensemble of classifiers as a technique for prediction of student’s career choice”, 2014 First International Conference on Networks & Soft Computing

**CHAPTER 9**

**FUTURE ENHANCEMENTS**

A quite powerful web app can be built wherein inputs are not given directly instead student

parameters are taken by evaluating students through various evaluations and examining.

Technical, analytical, logical, memory, psychometry and general awareness, interests and

skills based tests can be designed, parameters are collected through them so that the results

will certainly be accurate and the system will be morereliable to use. Decision trees have few

limitations like overfitting, no pruning, lack of capabilityto deal with null and missing

values and few algorithms have problem with the amplenumber of values. All these can be

taken into consideration and even more reliable andmore accurate algorithms can be used.

Also, the use of **MLP(Multi-Layer Perceptron)** algorithm will be quite performance

enhancing technique for our work in future.

And majorly, the course prediction part, that too only those courses which goes around the

area of interest of the candidate will be the highest achievable goal, for which the system has

to be trained more and more.

Then the project will be more powerful to dependupon and even more efficient to depend upon.