# DETERMINING STUDENTS' PERFORMANCE EFFECTING FACTORS TO PREDICT THEIR PERFORMANCES USING VARIOUS CLASSIFIERS

**BY**

**SHIPANKAR SARKER**
**ID: 162-15-7947**

**LIMI ROY**
**ID: 162-15-7967**

**FARZINA AKTER**
**ID: 162-15-8139**

**ARAFAT AHMED TANZEER**
**ID: 162-15-7895**

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering.

Supervised By

**Md. Sadekur Rahman**
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

**Ms. Farah Sharmin**
Sr. Lecturer
Department of CSE
Daffodil International University

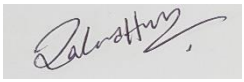

**DAFFODIL INTERNATIONAL UNIVERSITY**
**DHAKA, BANGLADESH**
**OCTOBER 2020**

# APPROVAL

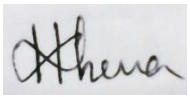
This Project titled **"Determining Students' Performance Effecting Factors To Predict Their Performances Using Various Classifiers"**, submitted by **Limi Roy**, ID No: 162-15-7967, **Shipankar Sarker,** ID No: 162-15-7947, **Farzina Akter,** ID No: 162-15-8139 and **Arafat Ahmed Tanzeer,** ID No: 162-15-7895 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 07/10/2020.

## BOARD OF EXAMINERS

**Dr. Syed Akhter Hossain** — **Chairman**
**Professor and Head**
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



**Md. Zahid Hasan** — **Internal Examiner**
**Assistant Professor**
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



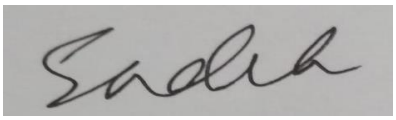
**Most. Hasna Hena** — **Internal Examiner**
**Assistant Professor**
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

**Dr. Mohammad Shorif Uddin** — **External Examiner**
**Professor**
Department of Computer Science and Engineering
Jahangirnagar University

# DECLARATION

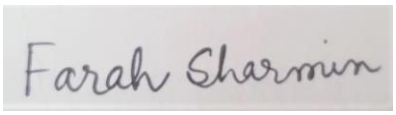
We hereby declare that, this project has been done by us under the supervision of **Md. Sadekur Rahman,** Assistant Professor**, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

**Supervised by:**



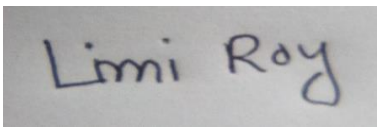
**Md. Sadekur Rahman**
Assistant Professor
Department of CSE
Daffodil International University

**Co-Supervised by:**

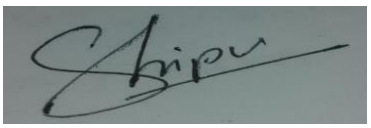


**Ms. Farah Sharmin**
Sr. Lecturer
Department of CSE
Daffodil International University

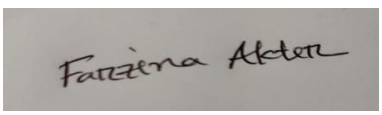
**Submitted by:**



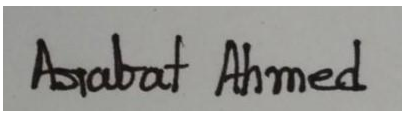
**LIMI ROY**
ID: 162-15-7967
Department of CSE
Daffodil International University



**SHIPANKAR SARKER**
ID: 162-15-7947
Department of CSE
Daffodil International University



**FARZINA AKTER**
ID: 162-15-8139
Department of CSE
Daffodil International University



**ARAFAT AHMED TANZEER**
ID: 162-15-7895
Department of CSE
Daffodil International University

# ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year research project successfully.

We really grateful and wish our profound our indebtedness to **Mr. Md. Sadekur Rahman**, **Assistant Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Data mining*" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Dr. Syed Akhter Hossain, Head, Department of CSE**, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

# ABSTRACT

Student performance in university courses is of great concern to the higher education where several factors may affect the performance. This paper is an attempt to apply the data mining processes, particularly classification, to help in enhancing the quality of the higher education system by evaluating student data to study the main attributes that may affect the student performance in courses. For this purpose, we have used data obtained from Daffodil International University, Dhaka of Department CSE, batch 44. In this research, we will differentiate subjects taken by CSE students according to their performance and organize them. Then we will find the best features that count to student's performance using the information gain attribute of the Decision tree algorithm. Then we will find out the best machine-learning algorithm to predict the performance of students by comparing different classifier algorithms in both the holdout method and k-fold validation. Then we will discuss the impacts of our research on society.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# CHAPTER 1
# INTRODUCTION

## 1.1 Introduction:

Students' performance in the exam and the degree of education depends on different features they get from different organizations across the world. The students' inequality of results is principally due to external-assessment and the internal-assessment given by the educational organizations. The internal-assessment defines the parameters established by the tutors like quiz marks, attendance marks, lab performance, etcetera. That ensures the student's functionality and gives an idea of their terminal grades in the class. On the other hand, external-assessment represents the final outcome of a student.

Performance and quality of learning of students do not depend on educational institutes. It depends on the main perspectives that shape their achievement in the method of education. Many students deal with various conditions like the new syllabus, new academic environment, family difficulties, economic difficulties, etcetera. Those factors affect student's performance throughout university life.

These features help distinguish the student's difficulties and reflect upon how to resolve them to increase student's performance. This study's primary focus is predicting computer science and engineering students' results, discovering the relationship between different features affecting students' performance, and examining the different prediction-algorithms for predicting their final score.

As technology is increasing rapidly, by data analysis and data mining, students and teachers can see through their education practices and provide them recommendations on effectively and efficiently enhancing their education process. Information gain of the Decision tree algorithm finds out which features affect most during their education life and can be more focused on developing their performance.

### 1.2 Motivation:

It is compulsory to know in which topics students are performing properly and in which topics they required special care. So, we needed to perform a survey. Therefore, based on students' experience throughout their programs to understand most subjects they possess in difficulty, the study was performed. After that, different classifier algorithms are implemented to specific dataset to determine the best predicting algorithm that can predict the students' performance most accurately. The analysis was accomplished on predicting the students' performance. However, there is barely any research done where the best-predicting classifier algorithm concluded with the questionnaires' help to predict student performance better. Moreover, according to our knowledge, there is no research on algorithm comparisons with both the Holdout method and the k fold validation method and the best features observed combined.

### 1.3 Rationality of the research:

Our research is focused on detecting out the best-predicting algorithm. For that, we need the complete dataset of students' performance collected by an online survey-poll. Consequently, we identified the topics they performed worst and best, respectively.

Then we implemented several classifier algorithms for comparing their performance, which serves best to determine the students' performance. Our research proportionately contributes for obtaining out the causes that influence the students' performances with the related subject and define the algorithm that performs most useful in this research type. We can select the best features according to their importance in deciding student's performance.

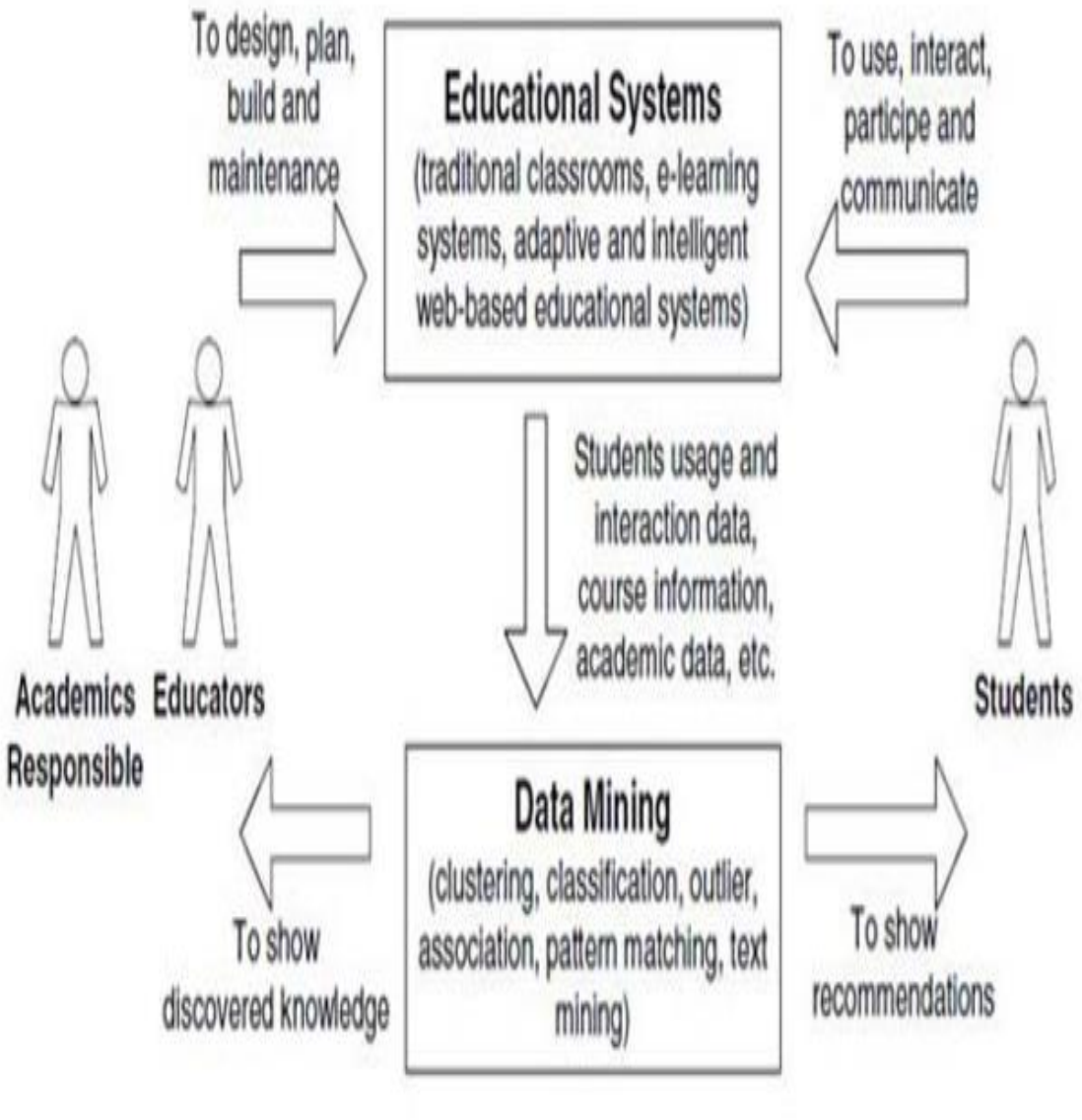



Fig 1.1: Application Cycle of Data-Mining for Educational Systems.

## 1.4 Research Question:

i) Will we be able to encounter the list of the topics in which subjects they did most excellent(best) and most critical(worst)?

ii) Will we be able to encounter the best features according to their importance to decide undergraduate students' results?

iii) Will we be able to find out a specific algorithm that performs better to predict the student results with the k-fold validation method?

iv) Will we be able to find out a specific algorithm that performs better to predict the undergraduate students' results with the Holdout method?

## 1.5 list of expected outcome:

We comprehended the result of this study finding the best features to decide students' performance, finding subjects they did best, and worst, finding the best classifier algorithm to predict students' performance and comparison results between the Holdout method and the k-fold validation method.

## 1.6 Research finance and management:

We had no finance for our research. We just collected datasets from students in our CSE department. The planning of our research was sequential, like at the beginning, ensured the collection of data, then prepare those data for implementing and finally implementation to conclude the complete algorithm evaluation that we worked on. Then, for implementing, we used a platform known as the Jupyter Notebook, and we used python in the Jupyter notebook to implement the algorithms. We also used the Weka software, which is also free software for students.

# CHAPTER 2
# BACKGROUND

## 2.1 Preliminaries:

This section of our research paper illustrates the related research about the performance prediction of the students using data mining techniques, represents the related tasks like related works of the researchers, the algorithms that they have used and the required prediction that is required for the expected outcome of their research. Therefore, this section will reflect about all the researches done previously similar to our topics.

## 2.2 Related research works:

Nahid Hajizadeh, Marzieh Ahmadzadeh worked on using "Association Rule" and "RepTree" and their experiment represented on the retaking subjects by the students, providing the specific courses by instructor and attendance of the students. The best result was provided by RepTree [4].

They found the best result showing algorithm is REPTree algorithm with the following features:

| | |
|---|---|
| **Correctly Classified Instances** | 84.2612 % |
| **Avg. F-Measure** | 0.772 |

Fig 2.1: Consequences of "REPTree" classifier algorithm.

Kamal bunker carried out a research on the decision tree classifiers to determine the performance evaluation of the students in the first year of BA exam. The main focus of their research was to ascertain the students who had chance of being failed and proper counselling can be provided to improve their result [3].

| === Detailed Accuracy By Class === | | | | | | |
|---|---|---|---|---|---|---|
| TP Rate Class | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
| 0.856 | 0.107 | 0.845 | 0.856 | 0.851 | 0.921 | FAIL |
| 0.964 | 0.264 | 0.684 | 0.964 | 0.8 | 0.922 | PASS |
| 0.022 | 0.002 | 0.732 | 0.022 | 0.043 | 0.749 | ATKT |
| 1 | 0.013 | 0.797 | 1 | 0.887 | 0.992 | ABST |
| 0 | 0 | 0 | 0 | 0 | .679 | W.H. |

Fig 2.2: Detailed accuracy by class.

Wahidah Husain and fellow researchers depicted the hiatus in prediction methods, determination of variables in analyzing and comparison of existing prediction models. They considered grade and class tasks as the evaluation standard for performance prediction. Neural network, Decision Tree, Support vector and Naïve Bayes had been implemented for their expected outcome [1].

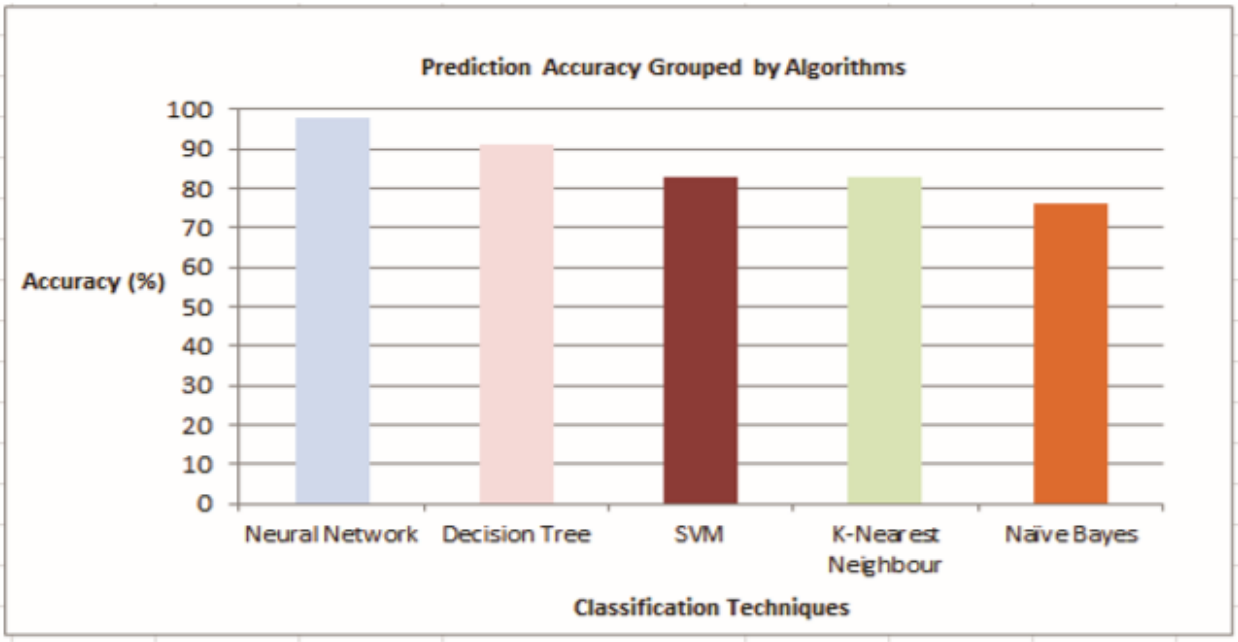



Fig 2.3: Accuracy prediction according to algorithms.

Emad and Mustafa and his fellow members studied the understanding of student's pattern, faculty and managerial decision maker and academic counselling to improve the quality of education who only took C++ courses and compared two methods such as k-cross validation and holdout method [2].

| Algorithm | Hold out | 10-CV |
|---|---|---|
| ID3 | 38.4615 % | 28.3186 % |
| C4.5 | 35.8974 % | 38.0531 % |
| Naive Bayes | 33.3333 % | 38.0531 % |

Fig 2.4: Average Accuracy of three different classifiers [2].

K. Ramar and his team worked on the most influential aspects of the performance prediction of the students for the final examination. They had implemented mainly “Decision Tree” and “Multilayer Perception” for finding out the predictive variable and to find best effective algorithm to determine the results at high school. Their hypothesis was based on type of school students’ study at, private tuition and study at home [5].

| | NAIVE BAYES | MLP | SMO | J48 | REPTREE |
|---|---|---|---|---|---|
| Accuracy | 49.5% | 72.38% | 57.25% | 64.88% | 60.13% |

Fig 2.5: Performance Prediction of five different classifiers [5].

Dr. Shailendra Narayan Singh with his research mates along with Leena Khan researched on the quality of the area of the concern in the field of higher education. Throughout their research they focused on the aggrandizing the existing student model and focused on to identify the amount of concentrate they put on their course syllabus. The algorithm they used are classification, linear regression, association rule etc. [6].

| | Level Of Analysis | User Benefitted |
|---|---|---|
| Educational Data Mining | Course Level: Design and development of Intelligent Curriculum, Conceptual Enhancement and development | Students, Faculty |
| | Departmental: Predictive modeling of Success/Failure across the departments | Students, Faculty |
| | Institutional: Academic Performance and overall Institution Growth and Reputation | Administrators, Funders |
| | Regional: Comparisons between various Systems | Administrators, Funders |
| | National and International | Educational Authorities ,Governments |

Fig 2.6: Educational Data Mining Model.

## 2.3 Summary and comparative analysis:

Table 2.1: Summary of previous researches

| AUTHOR NAME | ALGORITHM USED | OBJECTIVE OF WORK |
|---|---|---|
| Dr. Shailendra Narayan Singh, Dr. Mansaf Alam, Leena khan | Association rule mining classification regression clustering | Studied domain model behaviour modeling trend analysis |
| Rajesh Bunkar Umesh Kumar Singh Bhupendra Pandya, Kamal Bunkar | ID3 C4.5 CART | Focused on the likely to failing students |
| K. Ramar | MLP SMO RepTree Naïve Bayes | Find out the most influential aspects for detection of student performance |
| Shahiria, Wahidah Husaina Amirah Mohamed, | K-nearest neighbour Neural Network Decision Tree Support vector machine Naive Bayes | On the basis of CGPA and tasks of students |
| Qasem A. Al-Radaideh | C4.5 ID3 Naïve Bayes | k-cross validation and holdout method has been used for finding best fit algorithm |
| Marzieh Ahmadzadeh Nahid Hajizadeh | Association rule RepTree | Worked on three aspects such as attendance, non-retaking of courses and way of teaching of the instructor |

### 2.4 Scope for the problem:

Several researches had been done on the basis of performance of students applying classifiers of datamining. Throughout our journey of four yeas in the BSC courses of CSE, we had seen the fluctuations in the student's performance regarding their grades. Some of the research represented the facts putting different reference such as grades, attendance, quiz, assignment etc. by applying classifier algorithm. But less amount of works has been completed which reflects overall journey of varsity courses and what difficulties student faced and regarding that subjects. Therefore, we had conducted a survey based on different questionnaires so that we can find out the dominating aspects of the study that leads to determine reason behind their fluctuated result in the university. So, we have applied different classifiers to evaluate the algorithm efficiency in order to find out the best fitted algorithm using two methods such as k-cross validation and holdout method.

### 2.5 Research challenges:

#### 2.5.1 Collection of data:

It was some time-consuming piece to gather every section's data on a survey-form. After uploading that survey-form, it was challenging to get detailed replies from the scholars. Next, we divide the survey sheet among understudies and physically input the .exel sheet information.

#### 2.5.2 Combining Data:

As we also collected our data by hand, we needed to combine those data by hand. There was so much room for the human mistake—moreover, data collection performed by a single sheet. We had to separate them to find the worst and the best data set and then again combine them.

#### 2.5.3 Implementation Challenges:

Before implementation, we need to preprocess those data. We used several platforms. So we had to take our data portion-wise to use them carefully. We also required to use the same dataset for different purposes. For that, we carefully used our data for not overlapping.

# CHAPTER 3
# METHODOLOGY

## 3.1Preliminaries:

Here we will discuss the data collecting process, data-preprocessing, implementation-requirements, and proposed methodology about our research.

This fragment will mostly ponder the best deciding features that fundamentally influence the understudies' exhibition all through their classes with various classifiers' assistance. Therefore, we will see the topics that understudies did more regrettable, and students did the most satisfying to utilize the informational index we have.

## 3.2 Instrumentations and topics for the research:

With an enormous amount of datasets, it is badly arranged to discover any typical result; in such cases, the Data-Science displays a method to develop the explication. With the expansion of innovation, machine learning includes a superior path inside the information science that teams up in surrendering and assessing the info information and settling on viable choices dependent on the provided information. In this research, we have considered features that directly link to determining a student's performance with the guidance of the decision-tree algorithm's information-gain attribute, presented the Venn-diagram representing the subject's students did most excellent(best) or more unsatisfactory(worst) and showed a pie chart of gender-analysis.

At that point, we analyze various sorts of classification-algorithms, then we will discover which operates best inside this sort of condition.

To choose highlights that impact understudy's outcomes legitimately, we will utilize data to investigate the WEKA software's decision-tree algorithm. For different practices, we will utilize data-mining and machine-learning in the Jupyter Notebook of Anaconda framework.

## 3.3 Collection of data and procedure:

We had made an online-survey structure for the assortment of information from the outset, which had separated into two areas. The section indicated that students did the most excellent(best) or most unsatisfactory(worst) where and why. Notwithstanding, the polls were

the equivalent to both the areas [7]. The most challenging part occurred when we surveyed the form on the web, but we had not received good reactions as we anticipated.

Accordingly, we assembled our CSE department's everyday routine of 44-batch; at that point, we shared the form's link in the classes so every understudy can fulfill the survey to get an adequate measure of information.

Table 3.1: Questionnaires asked in the online survey.

| ATTRIBUTE NAMES | QUARIES | DESIRABLE VALUES |
|---|---|---|
| Q-1 | Did the students' class start on schedule? | { no(false), yes(true) } |
| Q-2 | Did it seem difficult for the student as it was investigative(analytical)? | { no(false), yes(true) } |
| Q-3 | Was the students' course systematic(analytical)? | { no(false), yes(true) } |
| Q-4 | Was the students' course teacher supportive? | { no(false), yes(true) } |
| Q-5 | Was there coding involved? | { no(false), yes(true) } |
| Q-6 | Did coding appear challenging for you? | { no(false), yes(true) } |
| Q-7 | Were there enough needed accumulations available? | { no(false), yes(true) } |
| Q-8 | Were the co-curricular exercises such as quiz-performance, assignment etcetera on time? | { no(false), yes(true) } |

| | | |
|---|---|---|
| Q-9 | Does the student require counselling hour for the subject? | { no(false), yes(true) } |
| Q-10 | Does the student require a retake of the subject for the development of their result? | { no(false), yes(true) } |
| Q-11 | How was the students' involvement throughout the curriculum? | { partial, bad, good } |
| Q-12 | Did the student need any required education before attending the program(course)? | { no(false), yes(true) } |
| Q-13 | Did the student have previous education about the program(course)? | { no(false), yes(true) } |
| Q-14 | Was the student serving any part-time employment throughout the course? | { no(false), yes(true) } |
| Q-15 | Does the student receive sufficient study-time? | { no(false), yes(true) } |
| Q-16 | Did the mentioned subject relatable to CSE (computer science and engineering)? | { no(false), yes(true) } |
| Q-17 | Was the student self-motivated throughout the curriculum? | { no(false), yes(true) } |
| Q-18 | What was the students' final result? | { Best ,Worst } |

There were likewise a few fake placebo type inquiries these have no association with our examination like,

Table 3.2: Dummy Questionnaires Asked In The Online Survey [7].

| ATTRIBUTE NAMES | QUARIES | DESIRABLE VALUES |
|---|---|---|
| Q-1 | Is there some other explicit explanation for the most noticeably terrible presentation in this specific course? | { free hand writing within 300 words } |
| Q-2 | Is there whatever other explanation that inspired yourself to accomplish the most desirable score in this specific course? | { free hand writing within 300 words } |

Besides, we likewise got some information about sexual orientation, the course names of the most noticeably terrible outcome, and the course names of the most desirable outcome in the survey form for statistical-analysis.

### 3.3.1 Loading of data:

After the recovery of the information, we should enter the information into the data set for additional execution. After that, import the crude dataset in the form of a .csv document file later we place the .csv record to the Jupyter Notebook.

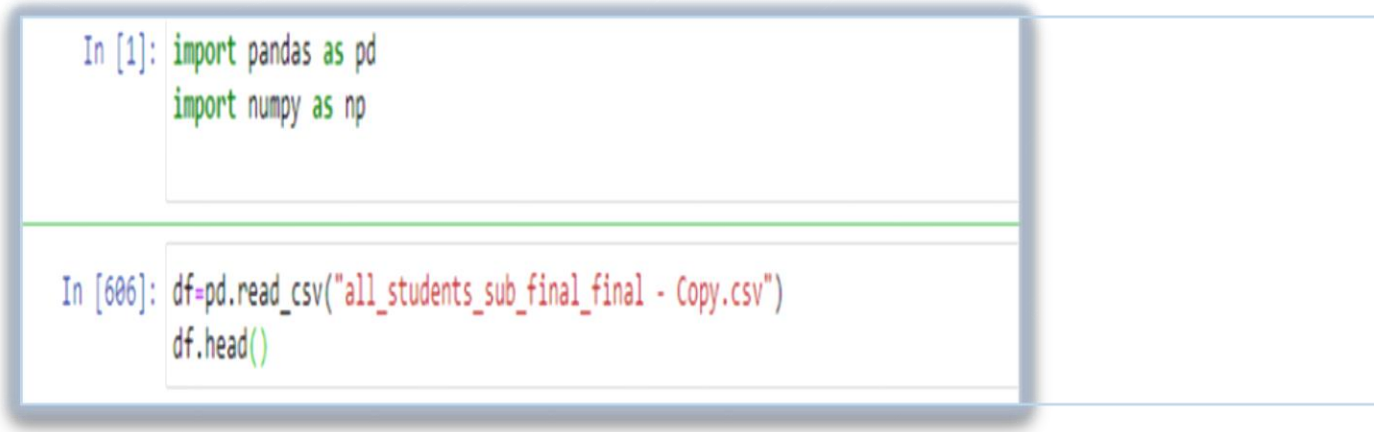



Fig 3.1: Data Loading for implementation in the Jupyter Notebook.

### 3.3.2 Preprocessing of data:

For the algorithm performing in the best way, it is essential to retain the information in a standard-form. By experiencing Data-preprocessing, we will locate the supervised dataset loaded up with categorical values.

#### 3.3.2.1 Conversion of upper case:

We will change over all the letters in capitalized design in this dataset and eliminate undesirable spaces with the goal that our machine can see each particular information appropriately.

```
In [25]: df2 = df.apply(lambda x: x.astype(str).str.upper())

In [7]: df2.to_csv("student.csv", sep=',', encoding='utf-8',index=False)
```

Fig 3.2: The whole dataset converting in the upper case.

#### 3.3.2.2 Cleaning of data:

It is difficult fixing the ruined, copy, and fragmented information physically. Hence, we erase void information sources.

```
In [4]: df.isnull().values.any()
Out[4]: True

In [5]: df=df.dropna()

In [6]: df.isnull().values.any()
Out[6]: False

In [24]: df.head()
```

Fig 3.3: In our dataset Data cleaning process.

We saw a portion of the course's name have a few additional rooms at the outset and in the tailpiece. We eliminated those spaces for presenting each subject merely distinct.

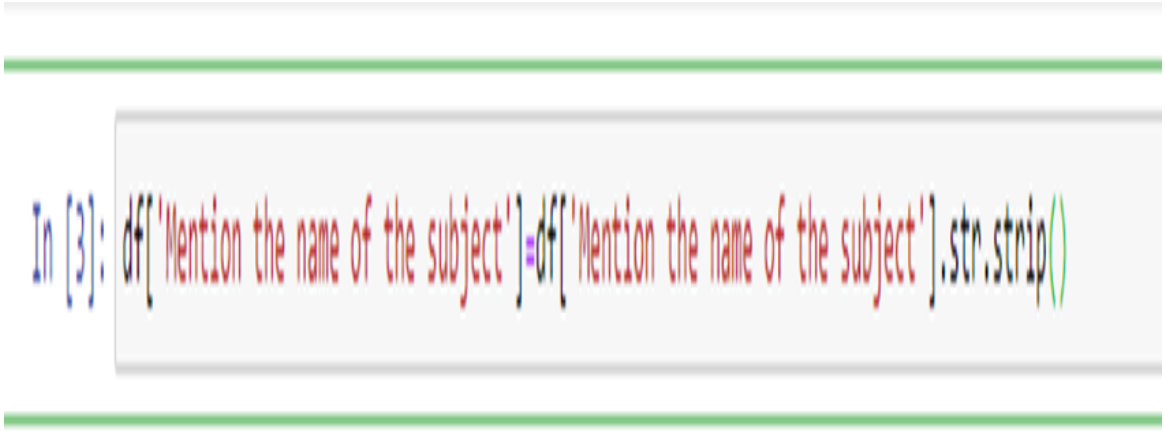



Fig 3.4: Extra spaces removed from those subjects' names.

### 3.4 Evaluated statistical analysis:

In this segment, we will observe which factors are generally valuable to choose understudy's outcomes and several graphical portrayal of statistical-analysis.

#### 3.4.1.1 The venn diagram:

Venn-Diagram is a productive method to speak to the correlation or convergence outwardly among various arrangements of information. So also, we have executed a Venn-chart in the Jupyter Notebook utilizing. matplotlib library for showing the courses in which understudies did most exceedingly awful, and they did best.

We discovered four courses with just the most exceedingly awful outcomes, eight courses with merely the best outcomes, and other twenty-seven courses wherein understudies did both most desirable and most exceedingly awful outcomes inside an aggregate of 35 particular subjects.

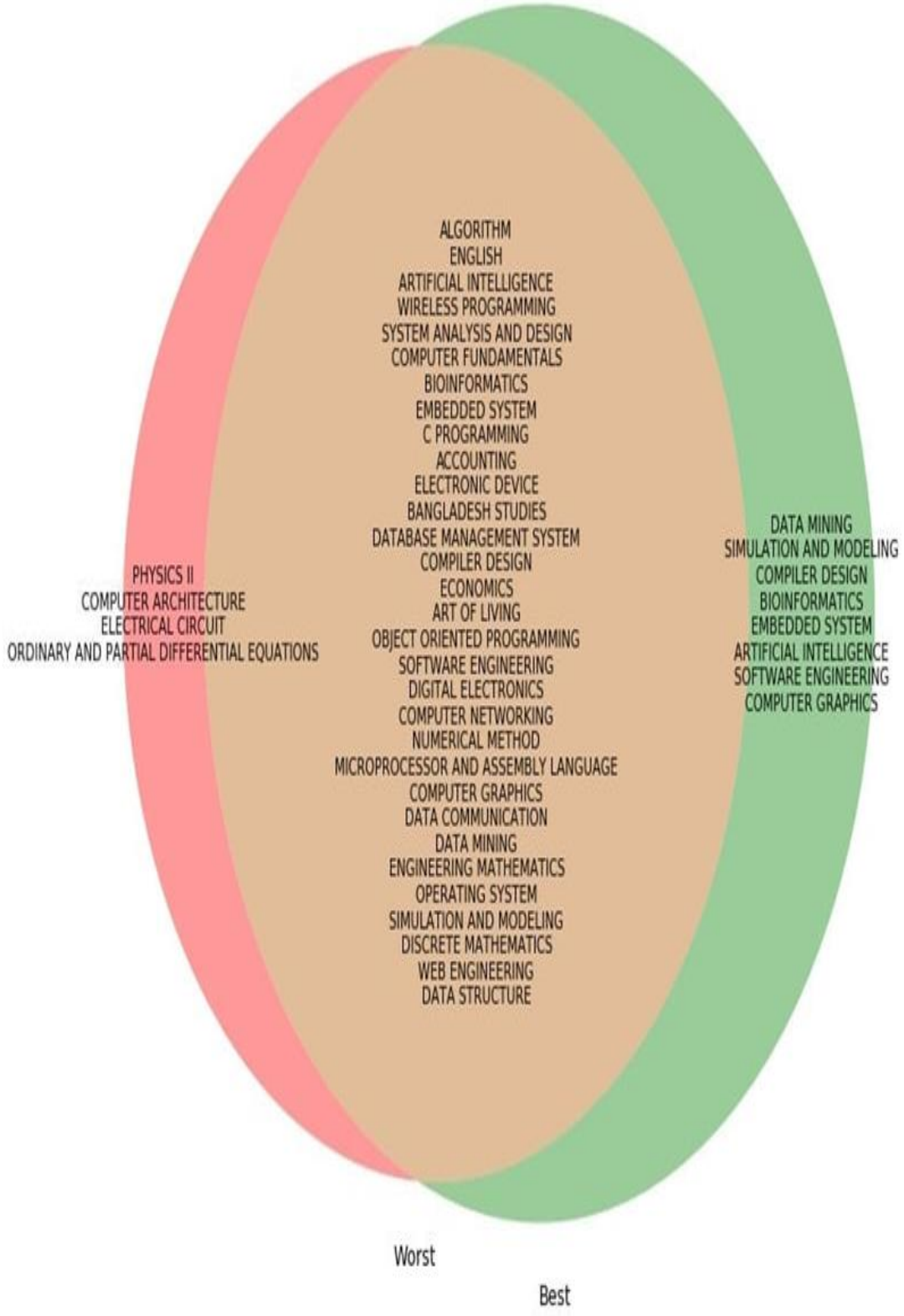



Fig 3.5: different subjects showed in Venn Diagram based on understudies' performance

### 3.4.1.2 The gender analysis:

We have found about 30.4% female and 69.6% male in our survey.

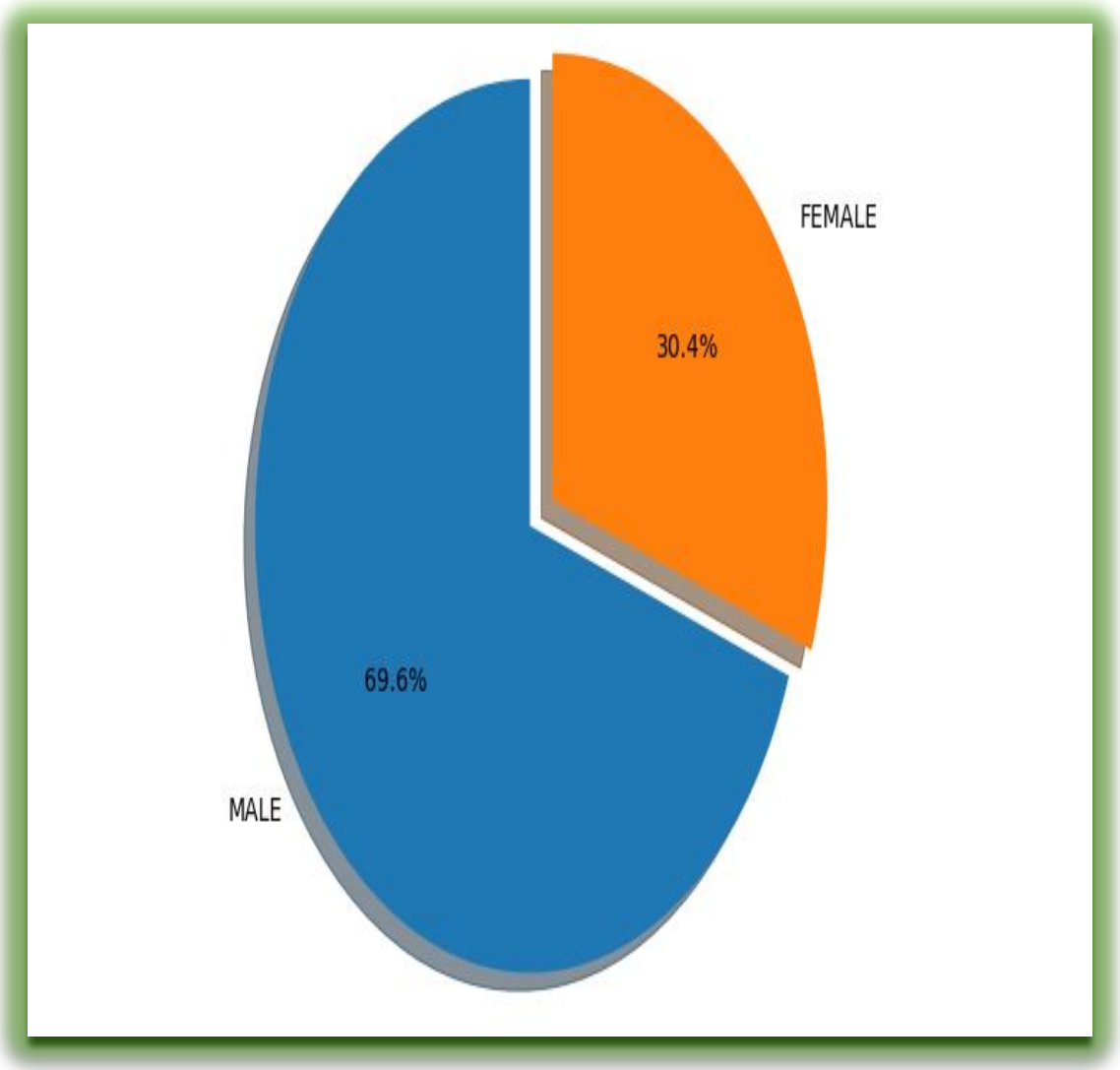



Fig 3.6: From the survey-dataset, Gender analysis.

### 3.4.2 Selection of top features:

Throughout this study, we have discovered the best features to decide the student's outcomes using information-gain. Information gain refers to a strategy for feature-evaluation that relies upon the measure of data gave which is utilized to include determination. In our study, there are many questionnaires dependent on the highlights for which understudies did best or most noticeably awful. Along these lines, we have assessed the features utilizing information-gain by the Weka apparatus. Utilizing the Weka software, we have discovered the information-gain

where the determination characteristic was "InfoGainAttributeEval," and the pursuit technique was "Ranker" with the edge - 1.7976931348623157E308.

```
Ranked attributes:
 0.28216  11 How was your involvement during the course?
 0.23186   2 Was your course teacher helpful?
 0.19012  17 Were you self-motivated during the course?
 0.10506  13 Did you have prior knowledge about the course?
 0.07823   3 Was your course analytical?
 0.07211   7 Was there required resources available?
 0.05892  10 Did you need to retake the course for improvement of your result?
 0.04851   8 Were the co-curricular activities such as quiz,assignment etc on time?
 0.04154  15 Did you get enough study time?
 0.02998   9 Did you need counselling hour for the course?
 0.02789   6 Did coding seem hard to you?
 0.02761  14 Were you doing any part-time job during your course?
 0.02223   1 Did you class start on time?
 0.00801  16 Was the course relatable to Computer Science and Engineering?
 0.00748   5 Was coding involved?
 0.00507   4 Did it seem hard to you as it was analytical?
 0.00325  12 Did you require any prerequisite knowledge before taking the course?

Selected attributes: 11,2,17,13,3,7,10,8,15,9,6,14,1,16,5,4,12 : 17
```

Fig 3.7: Top Features that decide understudies' performance according to the ranking.

We can observe that "student involvement" and "teacher's helpfulness" are the most significant features for understudy's exhibition.

### 3.5 Research proposed methodology:

Here is our research model spoke to in the following flowchart:

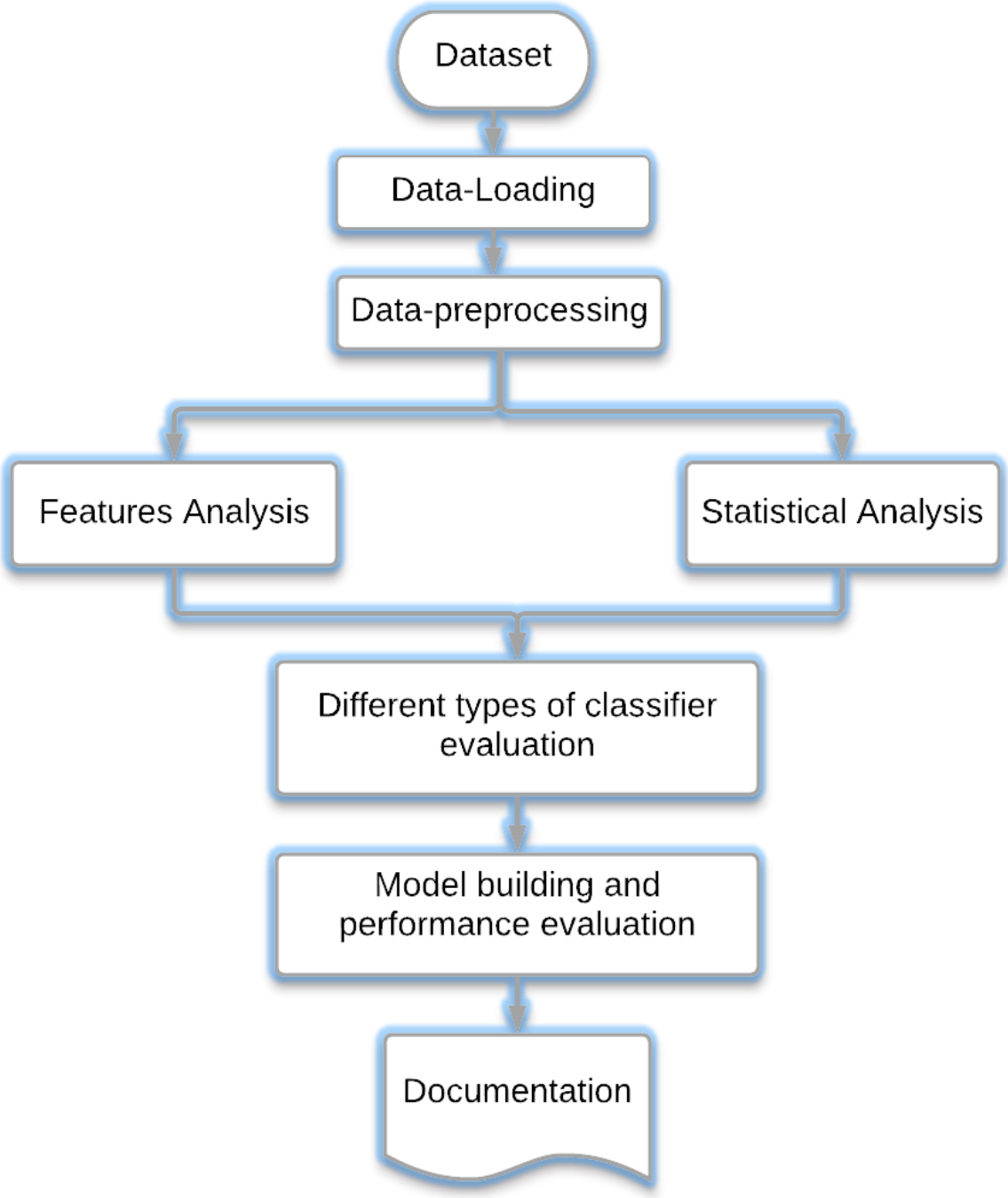



Fig 3.8: Our work model showed in the flowchart.

### 3.6 Requirements for implementation:

After the proposed-methodology, we choose to take and introduce all hardware. We require some equipment prerequisites and several software necessities. We have utilized Weka-

Software to discover the features. Weka-software is a renowned situation for applying machine-learning calculations and information analysis. After that, We have utilized the Anaconda constrictor to apply machine-learning calculations that are an available open-source dispersion for the python programming language, and a package-manager Anaconda constrictor comprises of Anaconda constrictor guide, Jupyter Notebook, and bugs. We have utilized python language for the usage of the various classifiers, which is an elevated level programming-language. Python contains countless libraries for information investigation, AI, information science, picture preparing, information mining, etcetera.

**Requirements of Hardware:**

1. A computer
2. Hard drive (minimum 512 GB )
3. A monitor to display
4. Ram (minimum 4 GB )

**Requirements of Software:**

1. OS: Windows 10 or newer
2. The Weka software 3.8.4
3. Python 3.8 or above
4. AnacondaJupyter Notebook 6.0.3
5. Pandas 1.0.5
6. Scikit-learn 0.23.1Jupyter Notebook 6.0.3
7. Pandas 1.0.5
8. Scikit-learn 0.23.1

# CHAPTER 4
# RESULTS AND DISCUSSION OF THE EXPERIMENT

## 4.1 Preliminaries:

In this study, we have talked about various classifiers of machine-learning. Machine-learning can be characterized as the investigation of calculations that analyze the conditions where we require the result. We utilized the accompanying classifiers to figure out which one operates proficiently to anticipate the exhibition of understudies.

## 4.2 Experimental setup:

We installed python on our computer at first. Then we installed Anaconda platform. After that, we installed our required libraries to implement algorithms. Most of the libraries came with Anaconda, but we had to install some extra third-party libraries to run our algorithms such as .graphviz and .pydotplus for saving decision tree as .pdf or .png file. We also had installed WEKA software for best feature selection.

## 4.3 Results and analysis of the experiment:

Our primary objective was to discover the most significant features. We did this in the earlier part by Weka instruments. We should utilize top features just to machine-learning calculations for preparing purposes. Because there were just eighteen features left in the wake of eliminating insignificant segments, we utilized the entirety of those features for enhancing the exactness of our calculations and because that was not such enormous information to deal with. Subsequently, we utilized the entirety of the eighteen features as opposed to the top ten sections.

Our dataset includes categorical values (non-numerical value) only. However, the Machine-learning operates with only numerical-values. For converting categorical values in numerical ones, o primary strategies had been utilized. They are “OneHotEncoder” and “LabelEncoder”, which have a place with the aspect of the. sklearn library. Under “LabelEncoder”, each estimation of the section is changed over into a specific number.

Practically the entirety of those features has one of a kind binary values. That adjusts upon us that then, extraordinary colinearity among features after Label-encoding will have happened.

That can produce the 'Dummy variable trap' while regression analysis. 'Dummy variable trap' is a situation wherein the independent-variables are multilinear, which will lose us in profound since factors will be anticipated from each other.

To evade the circumstance of "Dummy Variable Trap," we utilized "OneHotEncoder(drop='first')" to improve our conversion from categorical-value to numerical-values.

We first implement the Decision-Tree Classifier algorithms, with eighty of our training dataset by "DecisionTreeClassifier" from the scikit-learn library. This shows around 0.819444% accuracy. We produced a decision-tree from it by using '.graphviz' and '.pydotplus' libraries.

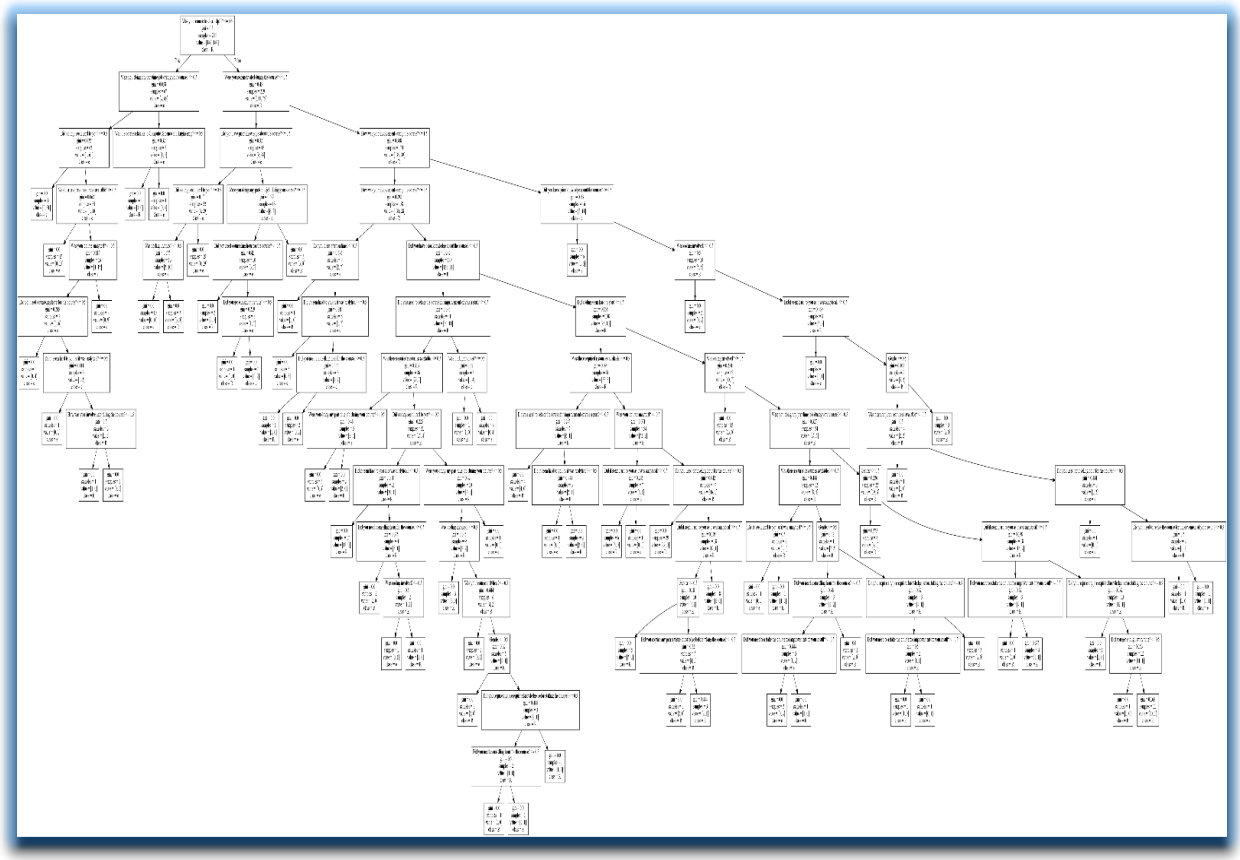


Fig 4.1: By applying “DecisionTreeClassifier”, we created the above Decision tree.

We applied machine-learning algorithms to the dataset in a couple of different manners. We first utilize the K-fold cross-validation (k-cv) technique to utilized five classifiers and ten folds to analyze. From that point forward, we utilized the Holdout strategy with those similar five classifiers and compared them.

**Formulas:**

From a confusion matrix,

Accuracy=(TP+TN)/(TP+FN+FP+TN); the ratio of correctly predicted observation to the total predicted observation.

Recall = TP/(TP+FN); the ratio of correctly predicted positive observations to the all predicted observation in class "yes".

Precision = TP/(TP+FP); the ratio of correctly predicted positive observation to the total predicted positive observation.

F-Measure = (2 * Precision * Recall) / (Precision + Recall); the weighted average of precision and recall.

**K-fold cross-validation comparing:**

We used cross-validation with a fold count of ten. We Found "Gaussian Naïve Bayes" as the most efficient algorithm in the F1 score, recall, and accuracy. The "Support Vector Classifier" algorithm was best according to precision only.

| | Logistic Regression | Support Vector Classifier | Decision Tree | Random Forest | Gaussian Naive Bayes | Best Score |
|---|---|---|---|---|---|---|
| Accuracy | 0.851905 | 0.854603 | 0.807460 | 0.854921 | 0.865794 | Gaussian Naive Bayes |
| Precision | 0.887199 | 0.892979 | 0.827068 | 0.873064 | 0.881212 | Support Vector Classifier |
| Recall | 0.810458 | 0.810131 | 0.788235 | 0.837908 | 0.854575 | Gaussian Naive Bayes |
| F1 Score | 0.843560 | 0.846523 | 0.804223 | 0.850323 | 0.864979 | Gaussian Naive Bayes |

Fig 4.2: Comparison of algorithms gained by cross-validation where k=10.

**The holdout approach of comparing:**

We analyzed those five algorithms independently, with 80% of our training dataset. We applied the .score() function that profits just the average of exactness (precision). Here we likewise can see that "Gaussian Naïve Bayes" performed the most beneficially with 0.84722 per cent accuracy.

| | |
|---|---|
| Logistic Regression | 0.763889 |
| Support Vector Classifier | 0.777778 |
| Decision Tree | 0.763889 |
| Random Forest | 0.805556 |
| Gaussian Naive Bayes | 0.847222 |

Fig 4.3: Comparison of algorithms gained by the Holdout approach.

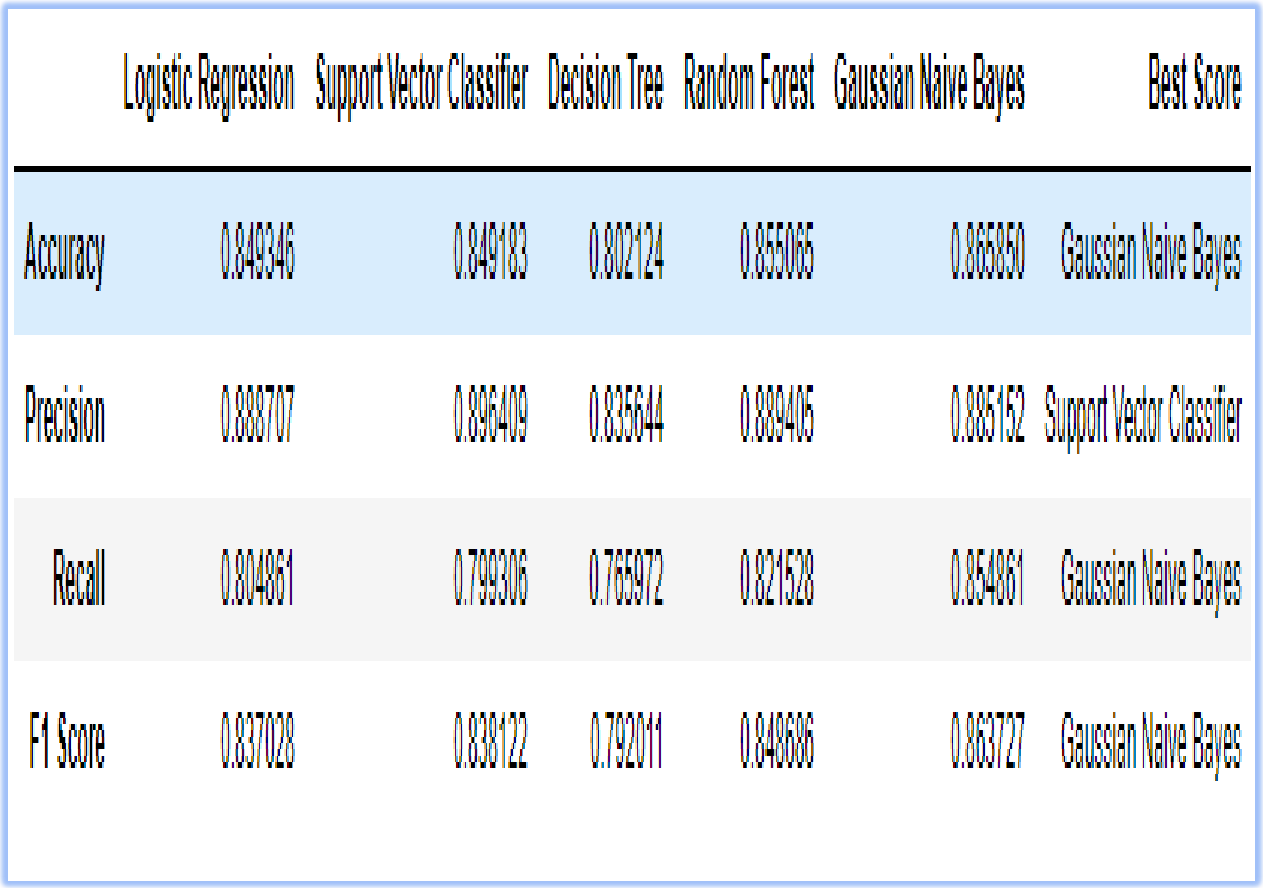
**4.4 Experimental discussion:**

As we applied 80% information to train the algorithm in the holdout approach, and there is only 360 information altogether, it might concede our outcomes by expanding or decreasing the fold counts in the k-folds segment.

We can observe from fig-4.2 and fig-4.3 there was quite a bit of reverence with the average accuracy. We made two additional comparisons with k=5 folds and the k=20 folds (we used ten folds before; we fancied to use double and half of that for comparison).

| | Logistic Regression | Support Vector Classifier | Decision Tree | Random Forest | Gaussian Naive Bayes | Best Score |
|---|---|---|---|---|---|---|
| Accuracy | 0.838106 | 0.860368 | 0.812989 | 0.857746 | 0.868740 | Gaussian Naive Bayes |
| Precision | 0.869575 | 0.905475 | 0.815260 | 0.869784 | 0.885342 | Support Vector Classifier |
| Recall | 0.799206 | 0.810159 | 0.815714 | 0.843810 | 0.854762 | Gaussian Naive Bayes |
| F1 Score | 0.829424 | 0.852271 | 0.813319 | 0.853520 | 0.867764 | Gaussian Naive Bayes |

Fig 4.4: Comparison of algorithms gained by cross-validation where k=5.



| | Logistic Regression | Support Vector Classifier | Decision Tree | Random Forest | Gaussian Naive Bayes | Best Score |
|---|---|---|---|---|---|---|
| Accuracy | 0.849346 | 0.849183 | 0.802124 | 0.855065 | 0.866850 | Gaussian Naive Bayes |
| Precision | 0.888707 | 0.896409 | 0.835644 | 0.889405 | 0.885152 | Support Vector Classifier |
| Recall | 0.804861 | 0.799306 | 0.765972 | 0.821528 | 0.854861 | Gaussian Naive Bayes |
| F1 Score | 0.837028 | 0.838122 | 0.792011 | 0.849686 | 0.863727 | Gaussian Naive Bayes |

Fig 4.5: Comparison of algorithms gained by cross-validation where k=20.

Between those cross-validations, there are minuscule variations that we can ignore.

K-fold cross-validation attempts to restore the best outcome between the folds. Then again, the Holdout approach attempts to respond to the average accuracy. That is the reason the k-fold approach appears to provide the best outcomes. In this manner, we can infer the way that in the Holdout approach, there is an absence of training dataset comparatively (80% here) that outcomes in a lower accuracy percentage. The Holdout approach should work better if we manage to increase our training dataset.

Feature significance is the inbuilt objective that accompanies Tree-Based Classifiers only. We utilized here ExtraTree Classifier for obtaining the main ten features for our dataset. Nonetheless, we performed the most useful feature discovering with the Weka instruments by information-gain attribute. Here we utilize the "ExtraTreesClassifier" algorithm to locate the most beneficial features.

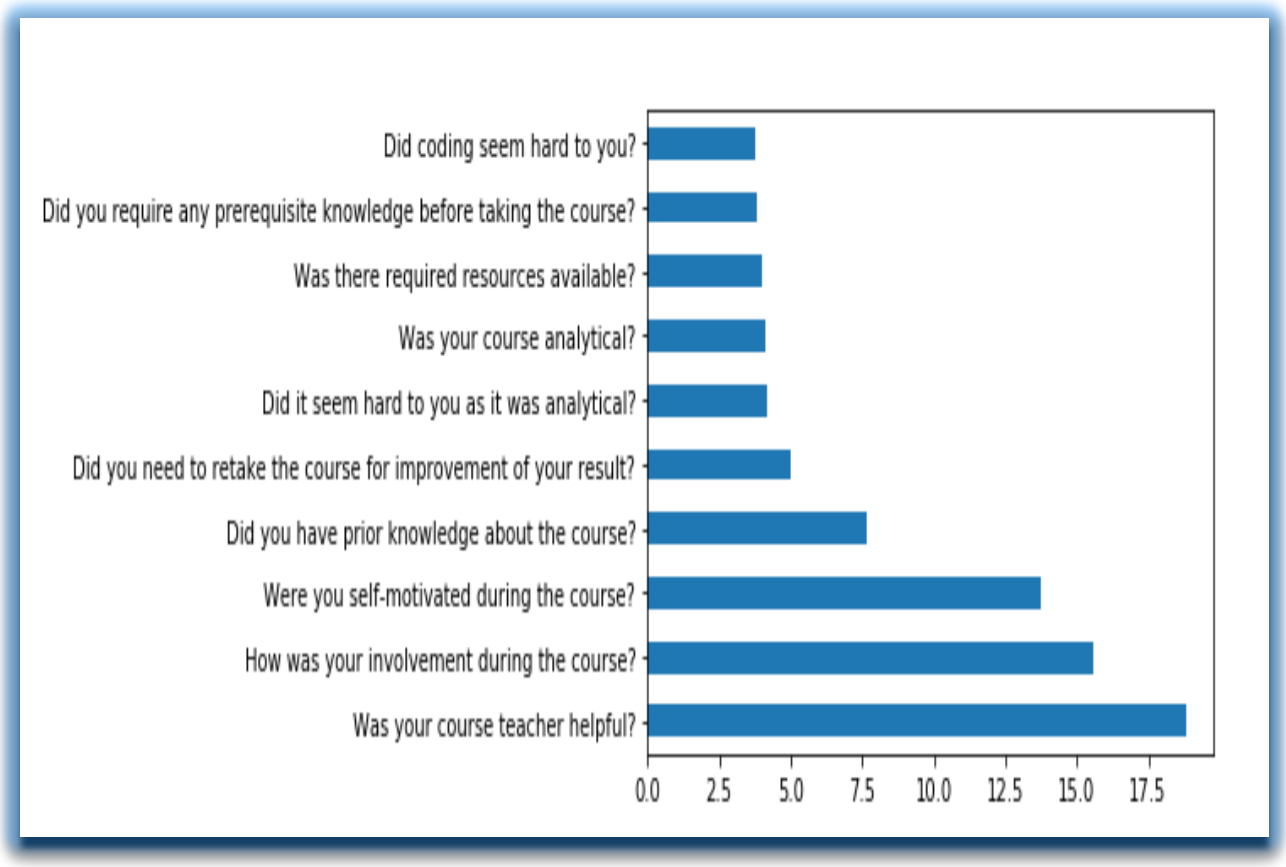



Fig 4.6: Top Features ("ExtraTreesClassifier" algorithm).

As should be obvious, there are a few contrasts between the outcomes concerning their positioning.

# CHAPTER- 5

# INFLUENCE ON ENVIRONMENT, SOCIETY AND SUSTAINABILITY

## 5.1 Influence on society:

Elaborate study of any topic helps to understand the underlying problems. In such case, we have conducted a survey to figure out the reasons of an inconsistent performance of the pupils. Our overall research mainly regards to the great standard version of education that is imparted by the university and extracting the information from students help us to impose different algorithms that can perpetrate best to predict the performance of students so that their performance bar can be increased highly. As a result, an improvised and proper education helps to improve the society which is the main matter of concern for us.

## 5.2 Influence on environment:

To ensure and aggrandize the value of education, it is highly mandatory to carry out all researches in a good and suitable environment so that it becomes convenient to design and set a new goal for offering a tentative solution to every problem. Our research illustrates on the performance evaluation of students and tries to avoid uncertainties that could hamper our research.

## 5.3 Ethical perspectives:

An accurate outcome is likely to be expected through the legitimacy of the information. In such cases, moral features of every elaborate study must be taken into consideration otherwise unexpected, uncertain and erroneous outcome may cause adverse effect to our expected outcome. Our research has been done on a raw data which is taken from the students through online survey. Our outcome can be helpful to evaluate the performance of the students and helpful to evaluate the best classifier algorithm using two methods such as “K-Cross Validation” and “Holdout Method”.

## 5.4 Sustainability design:

The main purpose of our enormous study was to offer a solution to the students so that their education can be saved from being hampered. This research of ours not only helps to determine the student’s performance, it also represents a solution to any questionnaire-based research from being known as time consuming research. Opinion of people is the initial solution to every problem that we face in our daily life. As a part of our process, we had gathered the related

information from the students on the basis of our regular assessment which are given by the varsity, pre-process those data for finding out the dominating aspects of their fluctuated performance in their varsity life and also we compared different classifier with the help of two methods such as K-cross validation method and holdout method. As an outcome we had been successful in getting the performing algorithm which can be applied to similar research like this. As a result, time of researchers would be saved and they can directly use this classifier.

# CHAPTER 6

# CONCLUSION, RECOMMENDATION, SUMMARY AND FUTURE RESEARCH IMPLICATION

## 6.1 Summary of the study:

From the study of our research, it has been evident the topmost features of students which affect their performance. We have implemented total of five classifiers which includes “Decision Tree”, “Gaussian Naïve Bayes,” “Support Vector Classifier”, “Logistic Regression” and “Random Forest”. After a comparative analysis of these classifiers with the help of two methods such as “K-Cross Validation” and “Holdout” method respectively, “Naïve Bayes” classifier works efficiently with the accuracy of 84%. We have also identified that the number of subjects that student did good is four and the number of subjects that students did worst is eight out of total thirty-five subjects.

## 6.2 Conclusions:

As per the result of our study of research, it can be concluded that the best and effective algorithm which can be used both with “K-Cross Validation” and “holdout” method is the Naïve Bayes for the determination of the student’s performance.

## 6.3 Implication of further study:

As per the outcome of our research, it is observed that in case of questionnaire-based research, “Gaussian Naïve Bayes” is best effective algorithm that can be taken into consideration while doing similar research if any research is done to find out student’s performance. Directly this classifier can be used without doing any comparison with other classifiers.

# REFERENCES:

[1] Shahiri, A. M., & Husain, W. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, *72*, 414-422.

[2] Al-Radaideh, Q. A., Al-Shawakfa, E. M., & Al-Najjar, M. I. (2006, December). Mining student data using decision trees. In *International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan*.

[3] Bunkar, K., Singh, U. K., Pandya, B., & Bunkar, R. (2012, September). Data mining: Prediction for performance improvement of graduate students using classification. In *2012 Ninth International Conference on Wireless and Optical Communications Networks (WOCN)* (pp. 1-5). IEEE.

[4] Hajizadeh, N., & Ahmadzadeh, M. (2014). Analysis of factors that affect the student's academic performance-Data Mining Approach. *arXiv preprint arXiv:1409.2222*.

[5] Ramesh, V. A. M. A. N. A. N., Parkavi, P., & Ramar, K. (2013). Predicting student performance: a statistical and data mining approach. *International journal of computer applications*, *63*(8).

[6] Khanna, L., Singh, S. N., & Alam, M. (2016, August). Educational data mining and its role in determining factors affecting students academic performance: A systematic review. In *2016 1st India International Conference on Information Processing (IICIP)* (pp. 1-7). IEE.

[7]https://docs.google.com/forms/d/e/1FAIpQLScHa0e4QPE8NLEmZbOE7q70ieWyrgOdUwF3PFBUKtUuPVgPdA/viewform?fbclid=IwAR09XFaOnEJMKxhEWiYx1I2JIv6azcY3twQaVg4LEpqQyYoEN33ivlwjB8A

[8] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," Expert Systems with Applications, vol. 33, no. 1, pp. 135–146, Jul. 2007

Determing Students' performance effecting factors to predict their performances using various classifiers

ORIGINALITY REPORT

| 14% | 3% | 3% | 13% |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | Submitted to TechKnowledge<br>Student Paper | 10% |
| 2 | Submitted to Daffodil International University<br>Student Paper | 2% |
| 3 | Submitted to National College of Ireland<br>Student Paper | <1% |
| 4 | Submitted to Queensland University of Technology<br>Student Paper | <1% |
| 5 | Ho-Suk Lee, Hong-Rae Lee, Jun-U Park, Yo-Sub Han. "An Abusive Text Detection System based on Enhanced Abusive and Non-Abusive Word Lists", Decision Support Systems, 2018<br>Publication | <1% |
| 6 | Chih-Wen Cheng, May D. Wang. "Chapter 13 Healthcare Data Mining, Association Rule Mining, and Applications", Springer Science and Business Media LLC, 2017 | <1% |