

**STUDENT PERFORMANCE PREDICTION USING MACHINE LEARNING
APPROACH AND DATA MINING TECHNIQUES**

BY

**MD. ANISUR RAHMAN RONY
ID: 162-15-7880**

**NUJHAT TABASSUM AMITHY
ID: 162-15-7750**

AND

**MEHERIN AMIR
ID: 162-15-7922**

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

Supervised By

Md. Azizul Hakim
Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Nusrat Jahan
Sr. Lecturer
Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JULY 2020

APPROVAL

This Project titled “**Student Performance Prediction using Machine Learning Approach and Data Mining Techniques**”, submitted by Md. Anisur Rahman Rony, Nujhat Tabassum Amithy and Meherin Amir 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 09-07-2020.

BOARD OF EXAMINERS



Dr. Syed Akhter Hossain

Professor and Head

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

Chairman



Dr. Sheak Rashed Haider Noori

Associate professor & Associate Head

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

Internal Examiner



Md. Zahid Hasan

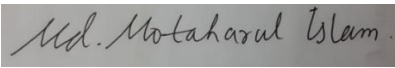
Assistant Professor

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

Internal Examiner



Dr. Md. Motaharul Islam

Professor

Department of Computer Science and Engineering

United International University

External Examiner

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Md. Azizul Hakim, Lecturer, 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. Azizul Hakim

Lecturer

Department of CSE

Daffodil International University

Co-Supervised by:



Nusrat Jahan

Sr. Lecturer

Department of CSE

Daffodil International University

Submitted by:

ANISUR RAHMAN

Md. Anisur Rahman Rony

ID: 162-15-7880

Department of CSE

Daffodil International University

Nujhat Tabassum Amithy

Nujhat Tabassum Amithy

ID: 162-15-7750

Department of CSE

Daffodil International University

Meherin Amir

Meherin Amir

ID: 162-15-7922

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 project/internship successfully.

We really grateful and wish our profound our indebtedness to **Md. Azizul Hakim, Lecturer, Department of CSE, Daffodil International University, Dhaka.** Deep Knowledge & keen interest of our supervisor in the field of “*Machine Learning and 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 our honorable 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 patients of our parents.

ABSTRACT

Nowadays Data Mining and Machine Learning Algorithms have made our work easier by developing prediction capability. We can implement Data Mining and Machine Learning Algorithms in many sectors. Among those, education is one of the most important sectors for the application of machine learning. If we identify the factors that are responsible for student academic performance and apply machine learning algorithms then it will be helpful for students. We can take extra care and necessary steps for students with bad results if we build a prediction model that can predict their performance earlier based on their different types of attributes. In that case those attributes must be correlated with their academic performance. That's why for our research we have collected many instances of different types of students attributes using survey forms which are correlated with academic performance. After that we have selected some important features using different feature extraction algorithms. Then we applied some machine learning algorithms on that preprocessed dataset. Comparison among different algorithms are also shown in our research. Among those algorithms we have chosen that one algorithm for our model which gave the best accuracy. For building our model and visualizing data we used both Python, Jupyter Notebook and Weka. We wanted to deploy our model to the web so that anyone can check their academic performance by giving values of attributes as input. For this we used Python Flask Framework and attached our model with it. Finally, we deployed our Flask App to Heroku Cloud Application Platform. And by using this one student can check their academic performance. Also, the authority and teachers can take necessary steps considering relevant attributes for those students whose performance is very poor. And it will be possible for our prediction model at the beginning of their learning process.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	i
Declaration	ii
Acknowledgement	iii
Abstract	iv
CHAPTER	
CHAPTER 1: Introduction	1-4
1.1 Introduction	1
1.2 Motivation	1
1.3 Rationale of the Study	1
1.4 Research Questions	2
1.5 Expected Output	3
1.6 Project Management and Finance	3
1.7 Report Layout	4
CHAPTER 2: Background	5-9
2.1 Preliminaries/Terminologies	5
2.2 Related Works	5
2.3 Comparative Analysis and Summary	8
2.4 Scope of the Problem	8
2.5 Challenges	9

CHAPTER 3: Research Methodology **10-21**

3.1 Research Subject and Instrumentation	10
3.2 Data Collection Procedure/Dataset Utilized	13
3.3 Statistical Analysis	14
3.4 Proposed Methodology/Applied Mechanism	19
3.5 Implementation Requirements	21

CHAPTER 4: Experimental Results and Discussion **22-43**

4.1 Experimental Setup	22
4.2 Experimental Results & Analysis	29
4.3 Discussion	39

CHAPTER 5: Impact on Society, Environment and Sustainability **44-45**

5.1 Impact on Society	44
5.2 Impact on Environment	44
5.3 Ethical Aspects	45
5.4 Sustainability Plan	45

Chapter 6: Summary, Conclusion, Recommendation and Implication for Future Research **46-47**

6.1 Summary of the Study	46
6.2 Conclusions	46

6.3 Implication for Further Study	47
REFERENCES	48-49
APPENDICES	50-52
PLAGARISM REPORT	53

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.1.1: Google survey form for online data collection	11
Figure 3.1.2: Google survey form for online data collection	12
Figure 3.1.3: Survey form for offline data collection	13
Figure 3.3.1: Statistical visualization of different attributes	16
Figure 3.3.2: Statistical visualization of different attributes	17
Figure 3.3.3: Statistical visualization of different attributes	18
Figure 3.3.4: Statistical visualization of different attributes	19
Figure 3.4: Proposed methodology for research	20
Figure 4.1.1: Top most important features and their feature scores using chi-squared statistical test	24
Figure 4.1.2: Top 10 important features using Extra Trees Classifier	25
Figure 4.1.3: Correlation among features using correlation matrix with heatmap.	26
Figure 4.1.4: Equation of node importance of Random Forest algorithm	27
Figure 4.1.5: Node importance equation components	27
Figure 4.1.6: Calculation of importance of feature	28
Figure 4.1.7: Feature importance equation components	28
Figure 4.1.8: Equation of normalization	28
Figure 4.1.9: Equation of Random Forest feature importance	28
Figure 4.1.10: Components of Random Forest feature importance equation	29
Figure 4.2.1: Accuracy of the model using Random Forest Classifier	30
Figure 4.2.2: Size of train and test data	30
Figure 4.2.3: Model building process using Random Forest Classifier	31
Figure 4.2.4: Model building process using Random Forest Regressor	31
Figure 4.2.5: Accuracy of the model using Random Forest Regressor	31
Figure 4.2.6: Bad and good two classes for applying algorithms using Weka implementation	32
Figure 4.2.7: Visual representation of features for Weka implementation	33

Figure 4.2.8: Visual representation of features for Weka implementation	34
Figure 4.2.9: Applying J48 tree classifier algorithm using Weka software	35
Figure 4.2.10: Visual representation of J48 tree algorithm.	36
Figure 4.2.11: Applying Random Forest tree algorithm using Weka software.	37
Figure 4.3.1: Deploying model to web for predicting SGPA	40
Figure 4.3.2: Deploying model to web for predicting SGPA	41
Figure 4.3.3: Predicted SGPA using Regression	41
Figure 4.3.4: Predicted SGPA using Classification	42

LIST OF TABLES

TABLES	PAGE NO
Table 3.3: Dataset Information	15
Table 4.2.12: Regression Algorithms Accuracy	38
Table 4.2.13: Classification Algorithms Accuracy	38
Table 4.2.14: Classification Algorithms Accuracy using Weka Software	39

CHAPTER 1

Introduction

1.1 Introduction

Data mining can be compared with gold mining. But the difference is instead of gold in data mining, data is mined from huge amounts of unnecessary, garbage, bulk data. These data are collected from different sources such as survey, database, data warehouse, data mart etc. Using data mining techniques, we can extract useful information from large amounts of data. In this case, as a part of extracting meaningful information machine learning algorithms play a great role. We can use data mining techniques in many sectors for many purposes. Some of the sectors are education, web and text mining. And our research is in the education sector. It is predicting student performance in terms of academic results.

In our research, the process and effects of data mining techniques and machine learning algorithms will be shown in detail.

1.2 Motivation

Because of poor performance in the academic area many students face many problems. Such as they lose their valuable time by retaking the course, waste a large portion of their parents' income etc. Besides these they became unable to carry on their higher study. These are the reasons for conducting our research. These are all about our motivation. That's why we are doing this research based project in a computerized way by applying several algorithms of machine learning to solve the problem.

1.3 Rationale of the Study

In our research, we are going to implement different machine learning algorithms using Data Mining techniques. Through our research we will be familiar with machine learning algorithms and Data Mining techniques. By doing these we will be able to solve different similar types of problems using Data Mining and Machine Learning. This is one of the rationales of the study. Besides these another main reason is we can analyze student performance using a dataset in many ways and identify main factors which are responsible for academic performance. Finally

we will be able to make a machine learning model which will be able to predict student performance earlier based on students attributes. These are the rationales of our study.

1.4 Research Questions

The Basis of a research is formed by good research questions. It is the foundation and initial step of a research. There are several steps for writing research questions. Such as a research question should be specific, answerable, should have medium length, complex answer, need focus etc.

In the research work done by Jane Agee [1] we can see the processes for developing research questions with quality. Besides this, we can also see the importance of qualitative research questions for a research.

We can generate research questions in many ways. Through problematization, we can formulate research questions which will lead to influential research. [2]

Research questions can be generated from the goal of study, research objectives and research purposes using mixed methods techniques. [3]

We have several research questions for our research. They are:

How can data mining techniques and machine learning algorithms be implemented in the education sector?

Which type of data is needed for predicting student performance?

Which attributes have correlation with student academic performance?

What are the ways of data collection for predicting student performance?

Which tools are required for implementation?

How feature extraction algorithms are used for finding important features?

According to accuracy which machine learning algorithms are more efficient compared to others?

Why is Weka software easier for machine learning?

How to deploy machine learning models to the web for predicting student performance?

These are the research questions for our research. By answering these questions our research will be conducted.

1.5 Expected Output

After completing our research, our expected outcome will be flask based web app which will be able to predict student academic performance or semester result of university students. And this prediction of semester result or SGPA will be done using student attributes which are correlated with student performance. Moreover, this flask web app will predict SGPA using the machine learning model. This machine learning model will be created by training our model with those data which contain student attributes correlated with academic result. We will get several machine learning models during conducting our research. These machine learning models can be created using both regression and classification algorithms. Our web app can be deployed to Heroku Cloud Application Platform after completing our web app. These are all about our expected outcome.

1.6 Project Management and Finance

For performing multiple activities efficiently and completing our research in time, we followed project management techniques. Project management techniques help us to control our research works. We can schedule our work by using it. We can decide how much time we will spend in each stage of our research works. We can use Gantt Chart, Project Charter etc. for managing our research project. We performed our research in several stages with a specific time in each stage. First of all we chose our research area. It was data mining and machine learning. Then selected it more specifically. That is predicting student performance in terms of academic result. It took one week to do that. Then we prepared our survey form both for online and offline data collection. To do this we needed one week. Then started to collect our data using those forms. It was a long term process. We spent most of our time doing that work. We spent three months doing that work. After that we focused on our coding part. We used python for data preprocessing, applying different algorithms and for making machine learning models. For this work we spent two months. After that we attached our machine learning models with the Flask framework for web implementation. After making a graphical user interface for the web, we deployed our Flask app to the web using Heroku. We did it within two months. We used Weka software for further analysis. To do this we needed one month. And the rest of the time we spent writing our research report. Thus we managed our research project.

For our research we did not need a lot of money. Financial fund was not needed. Only a few amounts of money we needed during our survey form printing for our offline data collection. We managed it by ourselves. Except for that, we did not need any money. These are all about our project management and finance for our research.

1.7 Report Layout

Our report layout contains the following things:

- Title & Cover Page
- Approval Page
- Declaration
- Acknowledgement
- Abstract
- Table of Contents
- List of Tables
- List of Figures
- Introduction
- Background
- Research Methodology
- Experimental Results and Discussion
- Impact on Society, Environment and Sustainability
- Future Work
- Conclusion
- References
- Appendices

These will be described briefly through our whole project.

CHAPTER 2

Background

2.1 Preliminaries/Terminologies

To conduct research there are some steps that should be maintained. First of all research areas and research topics are selected. After that previous research papers related to that topic are read for background study. From background study, ideas are generated and one can know what can be extended from those research works. Then the next part is to implement the research with a research methodology. Such as data collection, data analysis etc. After that experiment results and outputs are analyzed. Then finally it is evaluated. After completing the research work what can be implemented in future is discussed.

Primary work of a researcher is background study. Study of related works and comparing it with a summary. Background section is an important part of a research work and in this section several parts will be discussed. Review of different research papers which are related to our research works will be discussed shortly in this section. Previous study related to our research, present surroundings will be also focused. Context of our research will be established in the background part. Summary and challenges of research works, scope of problem will be discussed here also. Background study is almost similar to literature review. At the beginning of a research, literature review is one of the most important parts of conducting a research. This literature review should not be faulty. In the research work done by Justus J. Randolph [4] details of literature review has been explained. How to write it, its process, importance, purpose, taxonomy, common mistakes etc. of literature review have been described there. Almost similar research was conducted by David N. Boote and Penny Beile [5] where we can see the role, purpose, necessity, standards and criteria of literature review were briefly explained.

2.2 Related Works

Before us many researchers have conducted their research which are almost similar to our research topic. They showed different techniques of data mining and machine learning algorithms for predicting student performance. Different types of statistics, algorithms and techniques were shown by them to predict student performance.

In the research work done by Amandeep Kaur, Nitin Umesh and Barjinder Singh [6] they conducted research for predicting academic performance of students. They also compared the results of different prediction models. Their goal was to improve education systems by machine learning models. They made their dataset from BTech second year students which contains 1735 instances and 37 attributes. They used Association rule algorithm, Classification algorithm, Machine learning, Fuzzy logic to build their models. They used some classification algorithms such as NaiveBayes, LibSVM, C4.5, J48 tree, Hybrid approach LMT. Logistic Regression was also used. And finally, these algorithms were used to dataset for building models and comparative results of different algorithms were also shown to find an efficient algorithm. Hybrid machine learning classification was best for accuracy. By these predictive models their aim was to decrease rates of the dropout students by taking care after prediction and giving them the right direction at the right time.

In research done by V. Shanmugarajeshwari and R. Lawrance [7] they made a predictive model both for learners and teachers to evaluate student academic performance. It is also a warning model so that students can develop their performance. They collected their dataset from Ayya Nadar Janaki Ammal College, Sivakasi, TamilNadu, India and this data was from the students of Computer Applications. The data contained 12 attributes and 47 records and data was collected through a survey. The attributes were based on their marks, personal information, family background etc. By different feature selection methods, they chose top 6 important features for their implementation. They used the Decision Tree Classification algorithm. Final outcome of their model was predicting pass and reappear students as they classified into two classes. Their future plan was to connect it to the cloud. Their lacking was they could apply other algorithms.

In research done by Ermiyas Birihanu Belachew and Feidu Akmel Gobena [8] they developed a model for predicting performance of students based on 11 attributes which were selected from 34 attributes of the database. To create a model, a dataset was collected from the Wolkite University registry office. 993 student's data of computing and informatics were considered in this case. Classification and clustering machine learning algorithms such as Neural Nets (MLP), Naive Bayesian, Support Vector Machine were used and Naive Bayesian was best for accuracy. Comparative results of those algorithms were also shown and there were three types of experimental results from three experiments. Final output of their research was a predictive model of student performance. Their research started with only one college program but it could

be done through many departments and other colleges. Another efficient algorithm also could be used for better accuracy.

In research done by CH.M.H.Sai Baba, Akhila Govindu, Mani Krishna Sai Raavi and Venkata Praneeth Somisetty [9] they made a model of predicting student performance based on the academic attributes of the students. The data was collected from the educational institute as “csv” format. In that dataset some attributes such as SSC and HSC performance, rank, B tech marks of first, second and third year were kept. There, the decision tree algorithm was used. And using this dataset they developed a model for predicting the chance of getting a job using those academic attributes. They used only a few numbers of attributes to make their prediction model and those attributes were only academic attributes. And this was their limitations because only academic attributes are not sufficient for predicting student performance.

In the research work done by Astha Soni, Vivek Kumar, Rajwant Kaur and D. Hemavathi [10] they conducted research to take necessary steps for those students who get low marks by analysing student performance. They used data mining and classification algorithm. Student performance was measured using four categories of characteristics named academic, behavior, extra-curricular and placement. 48 variables were used as input to the model. Dataset was collected from 2000 graduate and undergraduate students of different universities using a questionnaire survey. There were 45 questions in the survey. They used Decision Tree, Naive Bayes’ and Support Vector Machine algorithm. Clustering algorithms and Data Mining techniques were also used. Finally, they developed their predictive model of pupil performance which gave output as “Good” or “Bad”. In comparison, the Support Vector Machine algorithm was best for accuracy. The research could be done with more components and with other tools for more accuracy.

In the research done by Anal Acharya and Devadatta Sinha [11] they conducted a research for predicting student performance in terms of Grade which was classified into several categories. Their dataset was collected from the students of undergraduate colleges majoring in computer science. They used the C4.5 Decision tree algorithm after selecting appropriate features using different types of feature selection algorithms. Their purpose was to predict students’ performance earlier using machine learning algorithms. They worked with a limited number of domains.

2.3 Comparative Analysis and Summary

There are a lot of works related to our research which were done by many researchers. We have shown it in the related works section of our report. Their works were based on predicting student performance using machine learning algorithms, data mining techniques. For this reason they collected data from different sources such as from educational institutes or through surveys. They collected the data of students for predicting their performance. After collecting data they applied different types of machine learning algorithms. Some of them were regression, classification algorithms. For doing this many of them used Weka software. Important features were identified using different techniques. Comparative results of different algorithms were shown in their research. They made different types of predictive models. Those models classified student performance as “Good” or “Bad”, predicted the chance of getting a job, identified passing and reappeared students and warned students about their performance. Their lacking was taking a limited number of attributes, unable to connect to the cloud etc. From these analyses we can take some ideas which will be helpful for our research. In our research, first of all we will collect our dataset through online or offline surveys among the university students. We will collect those attributes which have effects on student performance or academic results. Then we will preprocess that data. Important features will be identified using feature extraction algorithms. Then several machine learning algorithms will be applied on that preprocessed data. These algorithms might be both regression and classification algorithms. Both python anaconda navigator and weka software will be used for applying machine learning algorithms. Comparison among several algorithms will be shown. Among these experiments efficient algorithms will be identified from their accuracy. Among those machine learning models, the most efficient algorithm will be implemented for web implementation. Python flask web framework might be used in this case. And finally that web app can be deployed to Heroku. And this will be all about our research works. We got these ideas from different similar research works, online resources etc. Some of the research works have been described shortly in the related works section.

2.4 Scope of the Problem

There are several scopes of problem in our research works. Some researchers chose a particular domain for attributes instead of focusing all possible domain of attributes. We have the opportunity to focus in this area. Some researchers only made machine learning models and

comparison among the models were shown. But they did not implement it in cloud, web etc. We can make a system that will be able to predict the student performance and deploy it to the web or cloud. So that anyone can evaluate student performance by giving inputs as their attribute values. More efficient algorithms with higher accuracy can be applied for building our desired machine learning model. Both classification and regression algorithms can be applied in this case. Thus we will be able predict student performance in terms of academic result earlier from their attributes. We will be able to take necessary steps against the probability of bad result from that prediction. Finally, data mining techniques and machine learning algorithms will be helpful for education sectors in this way. Also it will be helpful for students for developing their performance in terms of academic result or SGPA.

2.5 Challenges

For completing our research based project successfully there are some challenges. First challenge is collecting a dataset. As there are no open dataset for our research and we have to collect primary data, there are some challenges regarding this. First of all we have to identify the related factors which are responsible for student academic performance. Then through survey form it will be collected from both online and offline sources. But there is one problem that is whether the students are giving their right information or not. For this we have to keep track of it manually. If we find any anomaly of any instance in our dataset we have to discard that instance. Another challenge is improving our model by increasing the accuracy of it. For increasing accuracy we can follow some rules and techniques. Such as we can take important features using feature extraction algorithms. We can use ensemble learning. Many hybrid and tree algorithms can be used. Deploying our model to the web or cloud is another challenge. All these challenges we have to deal with during our research works.

CHAPTER 3

Research Methodology

3.1 Research Subject and Instrumentation

As our research is based on prediction of student performance first of all, we need the data of students which has effects on student performance. Our research purpose is predicting student performance in terms of academic result or SGPA. Here SGPA means semester result. Data is the most important part of our research. Almost all the work depends on data. If we collect our data and preprocess it then almost above 80% of our research works are done. We collected our dataset using questionnaires. We collected it both through online and offline. In online we used google survey form and in offline or field survey we used hardcopy of word files containing those questionnaires which has effects on student academic results. Details of our data collection procedure will be given in the next part.

The following figures 3.1.1 and 3.1.2 show the google survey form for data online collection.

Fill up the form with correct information about yourself

In the SGPA section you have to give the last/previous one semester SGPA.

Here SGPA is Semester Grade Point Average or Semester GPA.

*** Required**

University Name *

Your answer

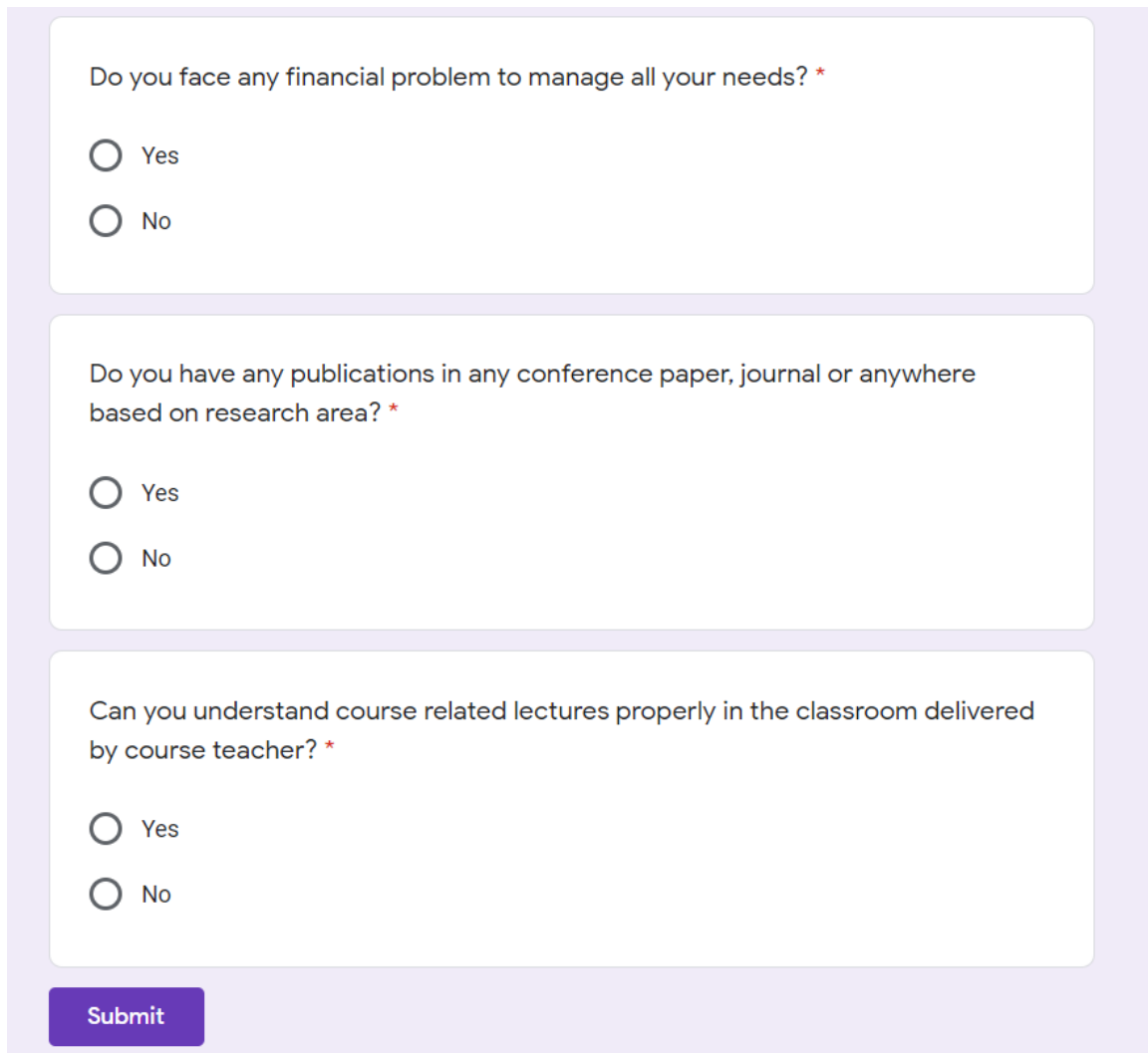
Age (in years) *

Your answer

Department Name *

Your answer

Figure 3.1.1: Google survey form for online data collection.



The image shows a Google survey form with three questions, each with two radio button options: 'Yes' and 'No'. The questions are:

- Do you face any financial problem to manage all your needs? *
- Do you have any publications in any conference paper, journal or anywhere based on research area? *
- Can you understand course related lectures properly in the classroom delivered by course teacher? *

At the bottom of the form is a purple 'Submit' button.

Figure 3.1.2: Google survey form for online data collection.

The following figures 3.1.3 shows the survey form demo for offline data collection.

Age:

Department:

Home District:

Division:

S.S.C GPA:

H.S.C GPA:

SGPA (Last/Previous semester GPA)*:

Current Semester (Such as 1st/2nd/3rd etc.):

Choose the correct option about yourself

1. What is your SGPA range of last semester?

- ☐ 4.00 ☐ 3.75-3.99 ☐ 3.50-3.74 ☐ 3.25-3.49 ☐ 3.00-3.24
☐ 2.75-2.99 ☐ 2.50-2.74 ☐ 2.25-2.49 ☐ 2.00-2.24 ☐ 0.00-1.99

2. Gender

- ☐ Male ☐ Female

3. Relationship Status?

- ☐ Single ☐ In a relationship ☐ Married ☐ Divorced

4. Do you suffer from depression which is harmful for your study?

- ☐ Yes ☐ No ☐ Not sure

5. How much time do you play online/offline games? (Such as PUBG, Ludo Star, Clash of Clans, FIFA etc)

- ☐ More than 15 hours ☐ (10-15) hours ☐ (5-9) hours ☐ (1-4) hours
☐ Don't Play Game

6. Select the distance type of your university from home (distance from where you go to your university daily)?

Figure 3.1.3: Survey form for offline data collection.

3.2 Data Collection Procedure/Dataset Utilized

To collect our dataset first of all we had to identify the relevant attributes which have impacts in student academic performance. Primarily we had selected almost 27 attributes for our data collection. The attributes are age, division, S.S.C and H.S.C GPA, relationship status, depression status, daily gaming time, distance from university, attendance in class, class task complete, number of dropped course, part time or full time job, number of tuitions, daily and weekly study

time, daily sleeping time, syllabus complete before exam, daily leisure time spent, extracurricular activities, skills and project, physical and mental stability, financial condition, research and publication, class lecture understand etc. Though primarily we selected those attributes, later we applied feature extraction algorithms on them to choose more important features among them. Based on the above factors we made an online and offline survey form to collect our dataset. We made some questionnaires so that we can get the desired attributes through the survey form. For online survey we have used google survey form and for offline we used hardcopy of similar questionnaires. We also checked if there is any anomaly while a student fills up that form. If we found it, we discarded that instance. We performed our survey among the university students of different departments. Thus we have collected our desired dataset.

3.3 Statistical Analysis

In our dataset there are 27 attributes and 522 instances. We have discarded those instances manually which contain missing value or anomaly. So there are no missing values in our dataset. Among them some of the attributes contain both nominal and numeric values. While preparing our dataset we kept both data types of the same attributes for analysis deeply. We collected and created our dataset in such a way that our dataset contains both nominal and numeric values for some attributes. We created some ranges and the numeric values were put to that range. Thus both nominal and numeric values of some attributes are kept in our dataset for deep analysis.

Our dataset information containing attributes name and data type are given in the following table 3.3.

TABLE 3.3: DATASET INFORMATION

Serial	Attribute Name	Data Type
1	Age	Numeric
2	Division	Nominal
3	S.S.C. GPA	Nominal and Numeric
4	H.S.C. GPA	Nominal and Numeric
5	Gender	Nominal
6	Relationship Status	Nominal
7	Depression Status	Nominal
8	Playing Games Time	Nominal and Numeric
9	Distance from University	Nominal and Numeric
10	Attendance in Class	Nominal and Numeric
11	Complete Home Task	Nominal
12	Quiz Marks	Nominal and Numeric
13	Dropped Course	Nominal and Numeric
14	Weekly Job Time	Nominal and Numeric
15	Tuition Number	Nominal and Numeric
16	Weekly Study Time	Nominal and Numeric
17	Daily Study Time	Nominal and Numeric
18	Daily Sleeping Time	Nominal and Numeric
19	Complete Exam Syllabus	Nominal
20	Leisure Time Spend	Nominal and Numeric
21	Extra-Curricular Activities	Nominal
22	Focusing On Skill And Project	Nominal
23	Mental Stability	Nominal
24	Financial Problem	Nominal
25	Publication Or Research Paper	Nominal
26	Understanding Class Lecture	Nominal
27	Semester Result or SGPA	Nominal and Numeric

Statistical visualization of different attributes are given in figure 3.3.1, figure 3.3.2, figure 3.3.3 and figure 3.3.4.

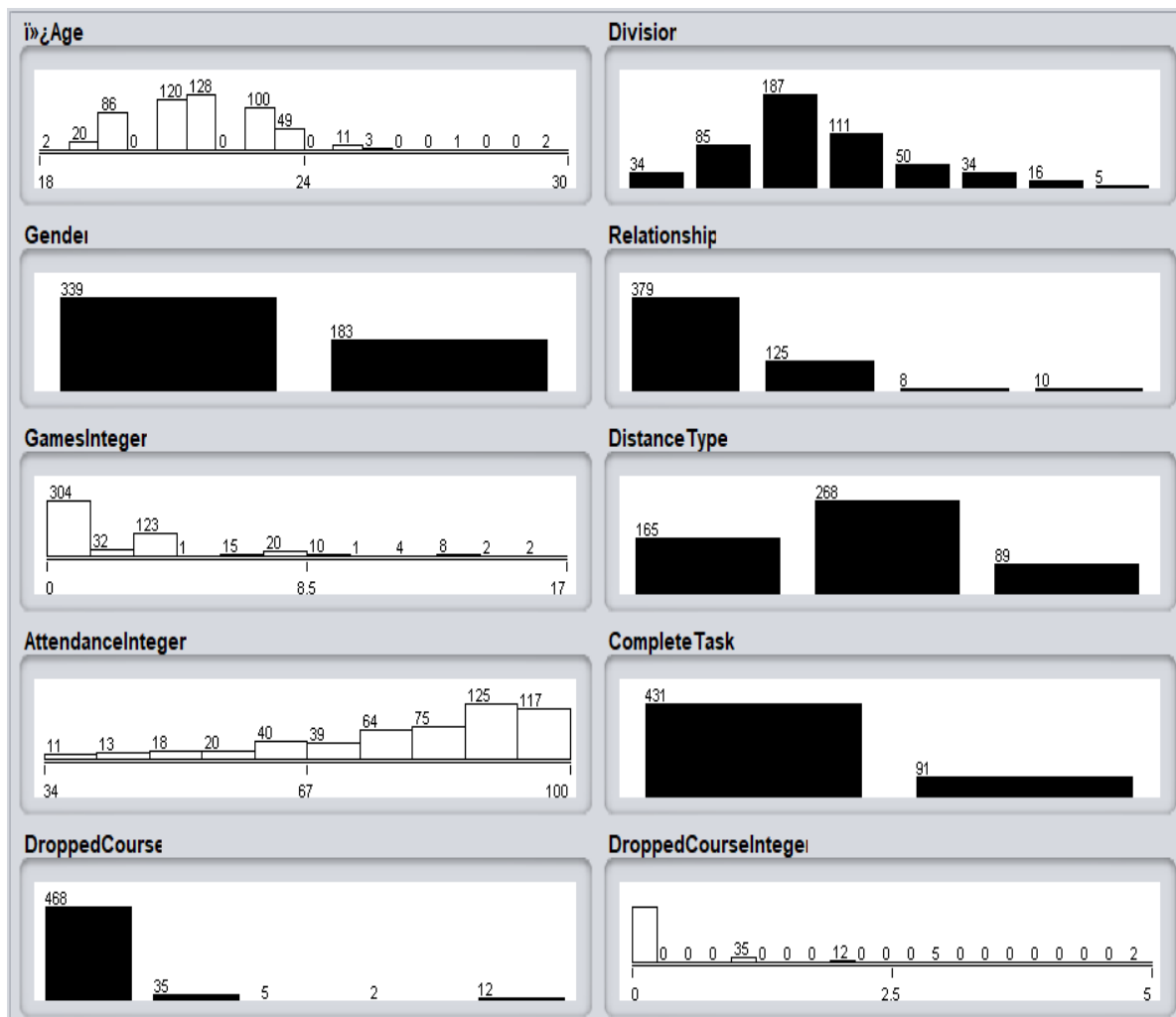


Figure 3.3.1: Statistical visualization of different attributes.

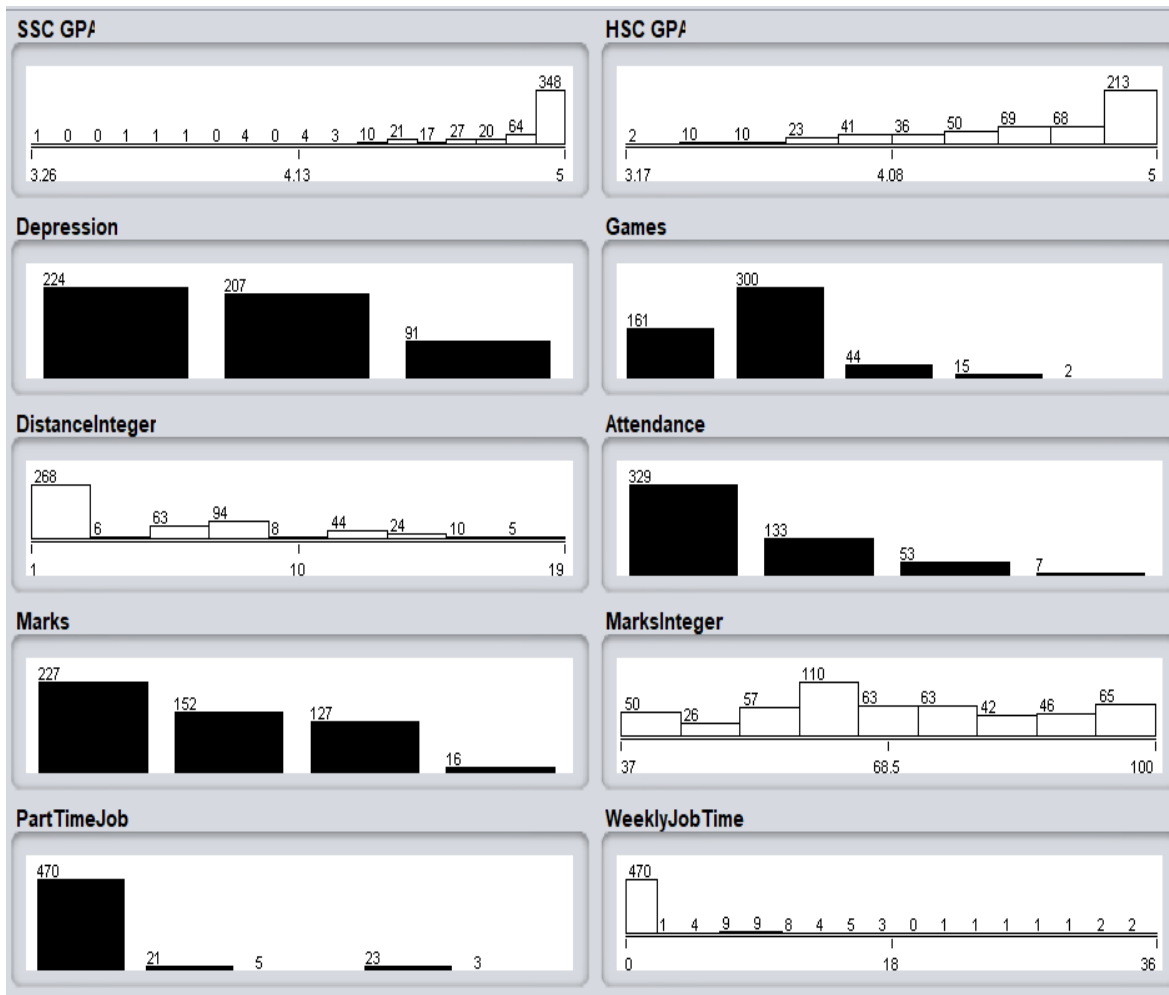


Figure 3.3.2: Statistical visualization of different attributes.

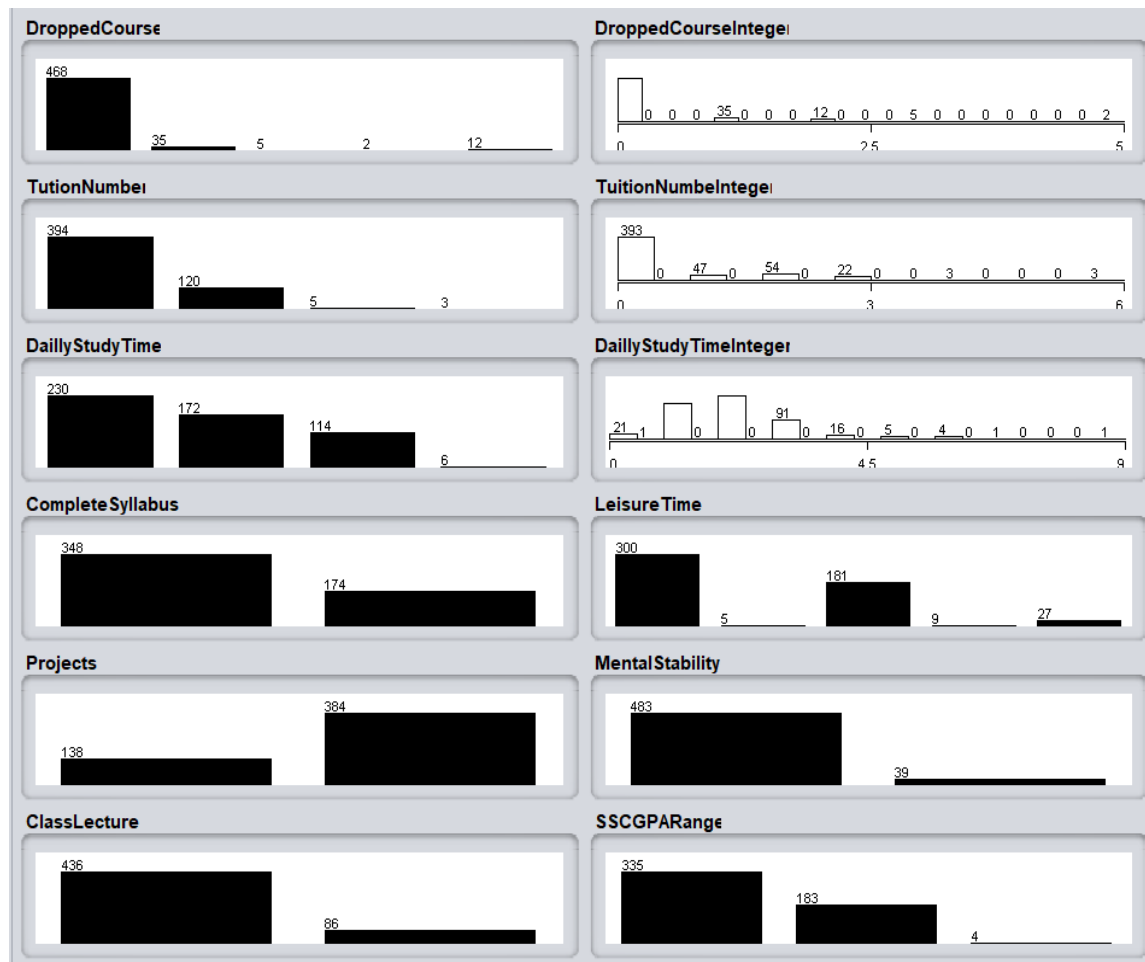


Figure 3.3.3: Statistical visualization of different attributes.



Figure 3.3.4: Statistical visualization of different attributes.

3.4 Proposed Methodology

We have done our research work step by step. For predicting student performance in terms of academic results primarily we have selected some relevant attributes of a student. Then we made a survey form based on those attributes. We have designed our survey form with some questions in such a way that if students fill up the form, we can get our desired attributes. With that survey form we have collected our desired data. After that we have preprocessed our dataset applying different methods. In this case one important step was feature selection. Applying different types of methods, we have selected our important features for our model. Then we applied machine learning algorithms on that preprocessed dataset. Finally, we built our desired model. We also used weka software for further analysis. We have implemented our model with Flask for web

implementation and we deployed our final Flask app to the web using Heroku. These are our working processes.

The following figure 3.4 shows the proposed methodology of our research work.

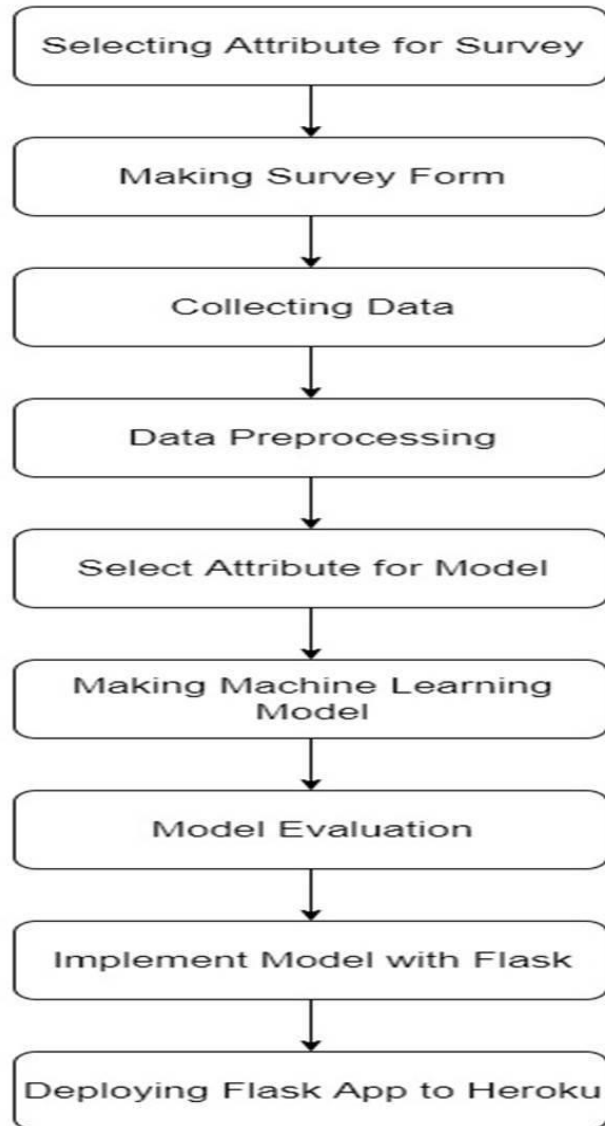


Figure 3.4: Proposed methodology for research.

3.5 Implementation Requirements

For implementing our research work there are some requirements. To make our machine learning model we have used anaconda navigator. We implemented our research works in python language. We used Jupyter Notebook, Spyder of anaconda navigator for making our machine learning model. They are very much useful for applying machine learning algorithms, training machine learning models, visualizing data, preprocessing data etc. In short all types of data analysis related works become easier using those tools. By importing libraries we can complete many complicated and difficult tasks of machine learning easily. After making the machine learning model we focused on deploying our model to the web. For this purpose we have used the Python Flask Framework. Our model is made using python and Flask is also a framework of python. So it makes our work easier. We used HTML, CSS and Bootstrap for making our web app of SGPA prediction. Finally we deployed our app to Heroku Cloud Application Platform. If we list our requirements for implementation, then the list we will be:

- Python
- Anaconda Navigator
- Jupyter Notebook
- Spyder
- Flask Web Application Framework
- HTML
- CSS
- Bootstrap
- Heroku

CHAPTER 4

Experimental Results and Discussion

4.1 Experimental Setup

After collecting the dataset our next work was to preprocess it. In this case we used a python jupyter notebook. First of all, using the pandas data frame we loaded our dataset. While implementing with python we loaded our dataset as a csv format which was collected from a survey form. With all columns and rows, we displayed our whole dataset so that we could get the overall view about our dataset. In case of any missing value in the survey form or if we found any anomaly, we have deleted that instance. For this reason, we did not need to be concerned about missing values during the data preprocessing step as we already deleted the whole instance from the csv file which contained missing values. Unnecessary columns were dropped which had no effect on our model. There were two types of data in our dataset. They are nominal and numerical data. For nominal data we need to encode it as it is impossible to work with nominal value. So, we converted our nominal data into numeric data using Label Encoder. We encoded our data using Label Encoder which was imported from sklearn.preprocessing. Then the column which contains nominal data was fitted and transformed of that LabelEncoder object. By doing this we were able to get the corresponding numeric value of nominal data. We also used discretization for replacing multiple nominal values into one nominal value. Discretization is used to replace similar type value into one value. And these are all about our data preprocessing steps.

Before applying data mining techniques and machine learning algorithms irrelevant attributes should be filtered or removed using different types of feature selection techniques such as wrapper, filter, embedded etc. [12]

Using feature subset selection methods such as wrapper selection methods we can identify and remove irrelevant and redundant features which will reduce the dimensionality of data and lead our algorithm in an efficient way. [13]

Feature selection is one of the important steps. In the feature selection step, unnecessary attributes are removed from the dataset. Besides it, the most important attributes are also identified. Primarily we had many unnecessary attributes in our dataset which had no effects on our model. Those attributes did not have any correlation with predicting student performance.

That's why we needed feature selection. There are several techniques and algorithms in feature selection. Such as:

1. Forward Selection: Forward selection follows the iteration technique where in each iteration the most important feature is added. This feature is best for the model and this work is done from an empty feature.
2. Backward Elimination: Backward elimination technique removes the least important attribute in each iteration. It starts with all attributes and by removing least important attributes finally it keeps the most important attributes.
3. Recursive Feature Elimination: In Recursive Feature Elimination, a subset of features is created which gives the best performance of a model. In each iteration it identifies best or worst performing features for a particular model and creates subset.

Some feature selection techniques [14] which were implemented in our dataset are given below:

1. Univariate Selection:

The features which have the strongest relationship with student performance can be found. In this case, SelectKBest, chi2 are used to do that and these are imported from sklearn.feature_selection. The following figure 4.1.1 shows top most important features and their feature scores using chi-squared statistical tests.

Specs	Score
WeeklyStudyTimeInteger	947.610584
AttendanceInteger	867.204486
GamesInteger	457.251590
WeeklyJobTime	241.967243
DroppedCourseInteger	179.326960
LeisureTimeInteger	167.314793
DistanceInteger	129.986481
DaillyStudyTimeInteger	91.448732
CompleteSyllabus	51.807711
SleepingTimeInteger	39.683553
Depression	30.934625
TuitionNumbeInteger	30.495961
CompleteTask	21.536361
FinancialProblem	11.895685
ExtraCurriculum Activities	10.537419
Relationship	6.465927
Gender	5.940735
Division	4.468921
ClassLecture	4.185459
Age	2.606806
Projects	2.108320
PartTimeJob	1.609043
Publications	1.039918
HSC GPA	0.760512
MentalStability	0.746193
SSC GPA	0.141712

Figure 4.1.1: Top most important features and their feature scores using chi-squared statistical tests.

2. Feature Importance:

Using feature importance, we can find the score of features and its importance according to prediction attributes. Higher the score of features, higher the importance of features. And this work is done by Extra Trees Classifier which is imported from sklearn.ensemble.

The following figure 4.1.2 shows top 10 important features using Extra Trees Classifier.

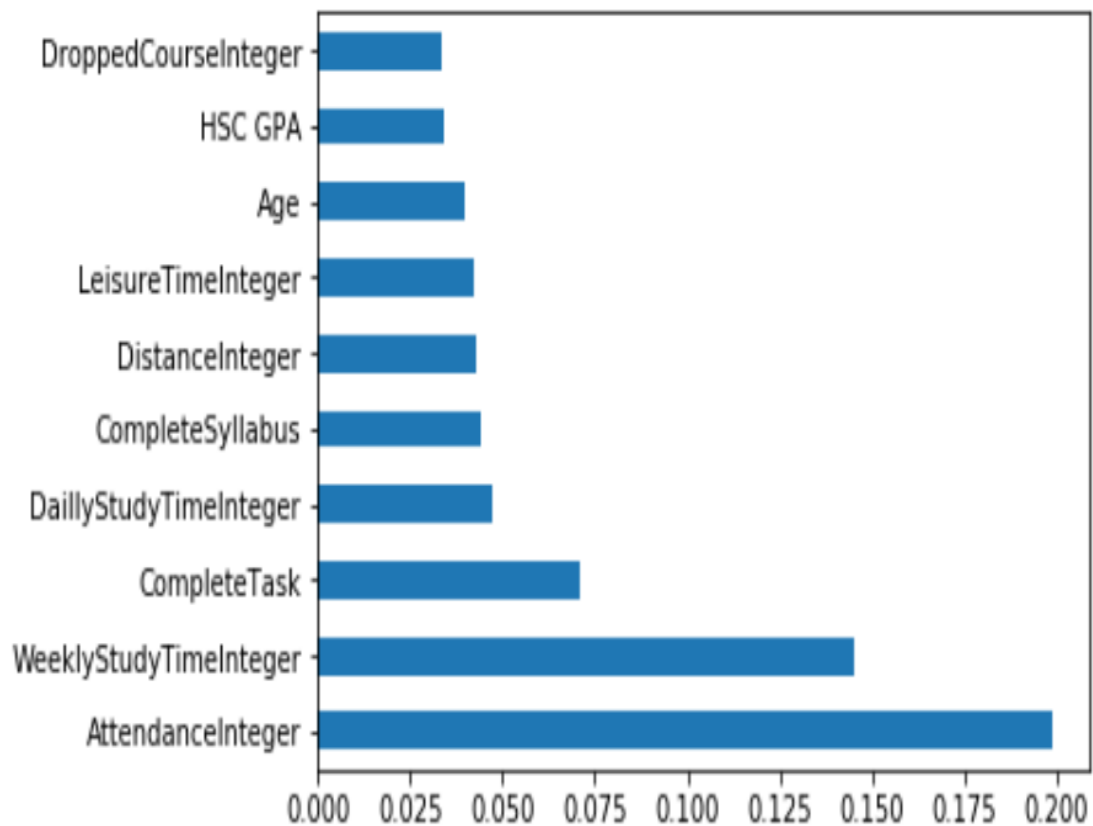


Figure 4.1.2: Top 10 important features using Extra Trees Classifier.

3. Correlation Matrix with Heatmap:

Using a correlation matrix with a heatmap we can show the correlation among features. We can identify the features which are correlated with the targeted attribute. This work is done using sns which is imported from seaborn.

The following figure 4.1.3 shows the correlation among features using correlation matrix with heatmap.

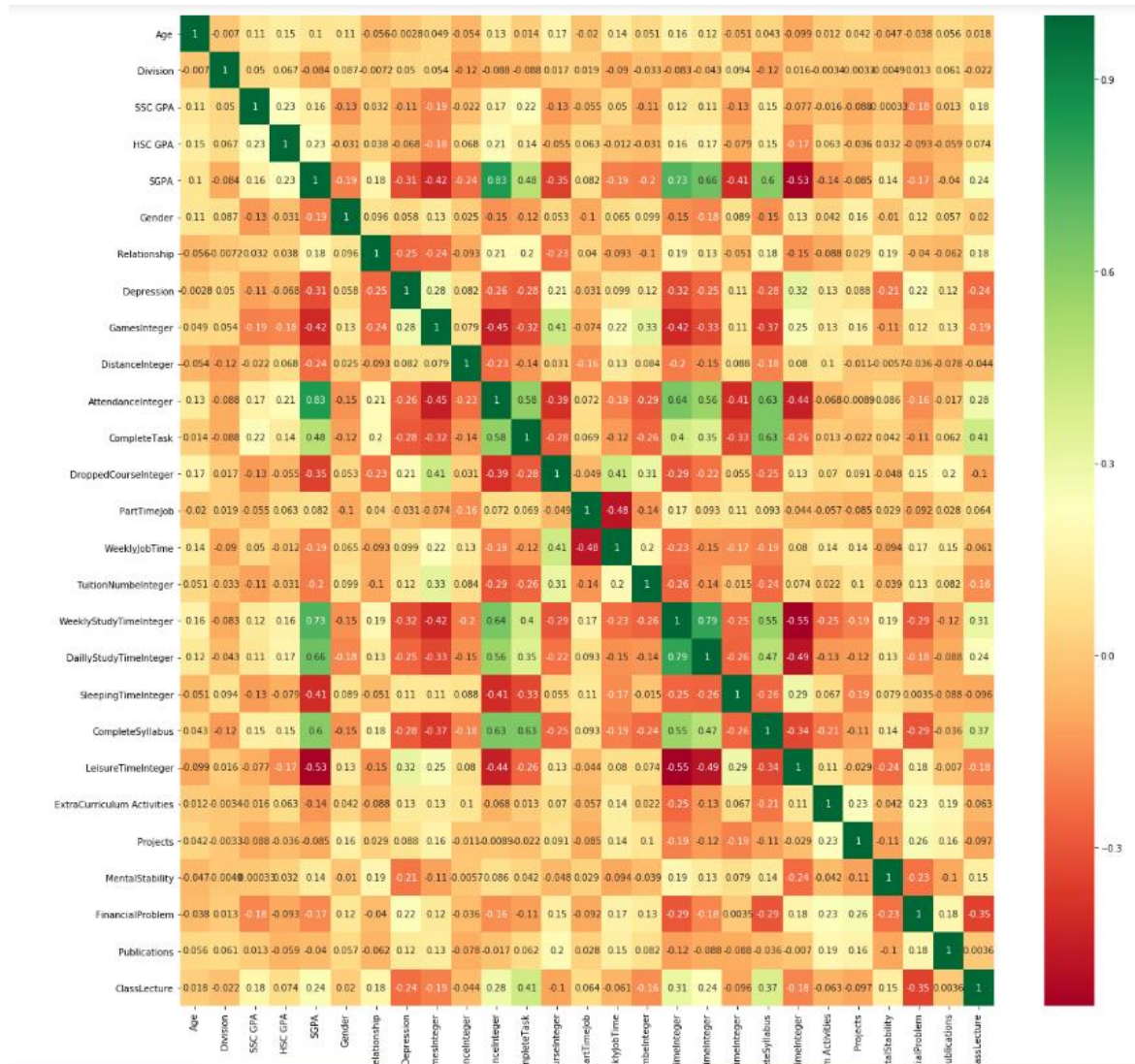


Figure 4.1.3: Correlation among features using correlation matrix with heatmap.

We prepared our dataset in such a way that both classification and regression algorithms can be applied. We made our model in terms of SGPA, which is the semester result of a university student. While collecting data and preparing our dataset, semester results were kept as float values which was perfect for the regression model. Based on that float valued result, corresponding classes were also kept so that we can apply classification algorithms. Thus both regression and classification algorithms were used. We have applied several classification and regression algorithms which will be shown in experiment and output sections. Among those random forest algorithms gave higher accuracy compared to other algorithms. This was helpful for web implementation using python flask framework and it was deployed to heroku. Now I am going to describe the model and the basic process of making a model. After finishing data processing steps our next work was to make a model. To deploy our model to the web, we used a

random forest classifier algorithm because it works better than other classification algorithms. After processing data, we splitted our dataset into a train set and a test set. In this case we kept 80% data as train set and 20% data as test set. Work of splitting the dataset was done using `train_test_split` and it was imported from `sklearn.cross_validation`. Then we imported `RandomForestClassifier` from `sklearn.ensemble`. Then we created an object of it. To make the object we have several parameters inside the Random Forest Classifier. We have changed the value of parameters in such a way that we can get higher accuracy. Ensemble random forest algorithms give better accuracy and performance compared to other algorithms, especially when the data is imbalanced. [15]

Random Forest algorithms are implemented in different ways in different places. In Scikit-learn it finds the node importance by Gini Importance guessing two child nodes only. [16]

The following figure 4.1.4 shows the equation of node importance of Random Forest algorithm:

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)}$$

Figure 4.1.4: Equation of node importance of Random Forest algorithm.

The following figure 4.1.5 shows node importance equation components of the Random Forest algorithm.

- $ni_{sub(j)}$ = the importance of node j
- $w_{sub(j)}$ = weighted number of samples reaching node j
- $C_{sub(j)}$ = the impurity value of node j
- $left(j)$ = child node from left split on node j
- $right(j)$ = child node from right split on node j

Figure 4.1.5: Node importance equation components.

The following figure 4.1.6 shows calculation of importance of features of the Random Forest algorithm.

$$fi_i = \frac{\sum_{j: \text{node } j \text{ splits on feature } i} ni_j}{\sum_{k \in \text{all nodes}} ni_k}$$

Figure 4.1.6: Calculation of importance of feature.

The following figure 4.1.7 shows feature importance equation components of the Random Forest algorithm.

- $fi_{sub}(i)$ = the importance of feature i
- $ni_{sub}(j)$ = the importance of node j

Figure 4.1.7: Feature importance equation components.

Normalization occurs from the range 0 to 1 by dividing the sum of feature importance values. The following figure 4.1.8 shows the equation of normalization of Random Forest algorithm.

$$normfi_i = \frac{fi_i}{\sum_{j \in \text{all features}} fi_j}$$

Figure 4.1.8: Equation of normalization.

Finally, the Random Forest feature importance is calculated as the average of all trees. The following figure 4.1.9 shows the equation of Random Forest feature importance of the Random Forest algorithm.

$$RFfi_i = \frac{\sum_{j \in \text{all trees}} normfi_{ij}}{T}$$

Figure 4.1.9: Equation of Random Forest feature importance.

The following figure 4.1.10 shows components of the Random Forest feature importance equation of Random Forest algorithm.

- $RFfi_{sub(i)}$ = the importance of feature i calculated from all trees in the Random Forest model
- $normfi_{sub(ij)}$ = the normalized feature importance for i in tree j
- T = total number of trees

Figure 4.1.10: Components of Random Forest feature importance equation.

This is a random forest classification for SGPA with different classes. And applying this random forest classification we can get our desired model. Similarly for float valued SGPA we used Random Forest Regressor for making our regression model.

When our model making process is finished then it is eligible for predicting semester results or SGPA of a student based on the attributes of a student. And the attributes are those attributes by which the model was trained and made.

4.2 Experimental Results & Analysis

We applied different types of Machine Learning algorithms to build our model. Among them, random forest algorithms gave the best accuracy. We applied both classification and regression algorithms. For classification we applied Random Forest Classifier and for regression we applied Random Forest Regressor. We prepared our dataset in such a way that we can apply both classification and regression algorithms. And we did by keeping both nominal and numeric values of SGPA. While applying classification algorithms we used SGPA containing nominal values. While applying regression algorithms we used SGPA containing numeric values.

While using the Random Forest Classifier the accuracy of our model was almost 86%.

The following figure 4.2.1 shows the accuracy of the model using Random Forest Classifier.

```
In [44]: rdf.score(X_test, y_test)
```

```
Out[44]: 0.861271676300578
```

```
In [45]: #creating and training a model  
#serializing our model to a file called model.pkl  
import pickle  
pickle.dump(rdf, open("model.pkl", "wb"))
```

```
In [46]: from sklearn.metrics import accuracy_score
```

```
In [47]: y_pred_gini = rdf.predict(X_test)  
print ("Random Forest Classifier using Gini Index\nAccuracy is ", accuracy_score(y_test,y_pred_gini)*100 )
```

```
Random Forest Classifier using Gini Index  
Accuracy is 86.1271676300578
```

Figure 4.2.1: Accuracy of the model using Random Forest Classifier.

In this case we classified student results into two classes. They are good and bad. We kept 33% as test data. Rest of them were train data.

The following figure 4.2.2 shows the size of train and test data

```
In [39]: from sklearn.model_selection import train_test_split
```

```
In [40]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=42)
```

Figure 4.2.2: Size of train and test data.

The following figure 4.2.3 shows the model building process using Random Forest Classifier.

```

In [39]: from sklearn.model_selection import train_test_split

In [40]: X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=42)

In [41]: #Import Random Forest Model
from sklearn.ensemble import RandomForestClassifier

In [42]: #Create a Gaussian Classifier
rdf=RandomForestClassifier(n_estimators=100)

In [43]: #Train the model using the training sets
rdf.fit(X_train, y_train)

Out[43]: RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                                max_depth=None, max_features='auto', max_leaf_nodes=None,
                                min_impurity_decrease=0.0, min_impurity_split=None,
                                min_samples_leaf=1, min_samples_split=2,
                                min_weight_fraction_leaf=0.0, n_estimators=100, n_jobs=None,
                                oob_score=False, random_state=None, verbose=0,
                                warm_start=False)

```

Figure 4.2.3: Model building process using Random Forest Classifier.

For numeric SGPA we applied regression using Random Forest Regressor.

The following figure 4.2.4 shows the model building process using Random Forest Regressor.

```

from sklearn.ensemble import RandomForestRegressor

# create regressor object
regressor = RandomForestRegressor(n_estimators = 100, random_state = 0)

In [17]: # fit the regressor with x and y data
regressor.fit(X, y)

Out[17]: RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=None,
                                max_features='auto', max_leaf_nodes=None,
                                min_impurity_decrease=0.0, min_impurity_split=None,
                                min_samples_leaf=1, min_samples_split=2,
                                min_weight_fraction_leaf=0.0, n_estimators=100, n_jobs=None,
                                oob_score=False, random_state=0, verbose=0, warm_start=False)

```

Figure 4.2.4: Model building process using Random Forest Regressor.

In this case accuracy was 96%.

The following figure 4.2.5 shows the accuracy of the model using Random Forest Regressor.

```

In [18]: regressor.score(X, y)

Out[18]: 0.9608365901273316

```

Figure 4.2.5: Accuracy of the model using Random Forest Regressor.

We also applied the K-Nearest Neighbors, Support Vector Machine etc. algorithms. But their accuracy was less than the Random Forest algorithm. So we decided to keep the Random Forest algorithm for our further web implementation process.

For further analysis we have used Weka Software. We have applied different types of algorithms on our dataset. In this case we converted our dataset from csv to arff format using Weka Software. We analyzed our data in different ways. One of them was we have classified students into two categories in terms of semester results or SGPA. There were two types of results. One is good and another is bad.

The following figure 4.2.6 shows bad and good two classes for applying algorithms using Weka implementation.

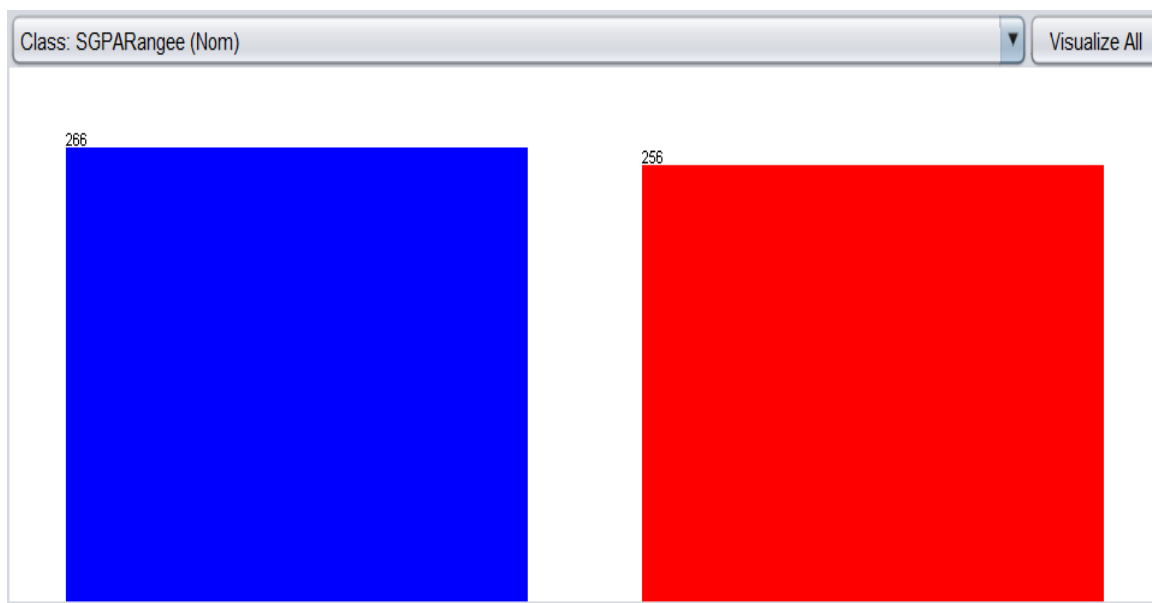


Figure 4.2.6: Bad and good two classes for applying algorithms using Weka implementation.

The following figure 4.2.7 and figure 4.2.8 shows the factors we kept after preprocessing our data and for applying algorithms to predict SGPA.

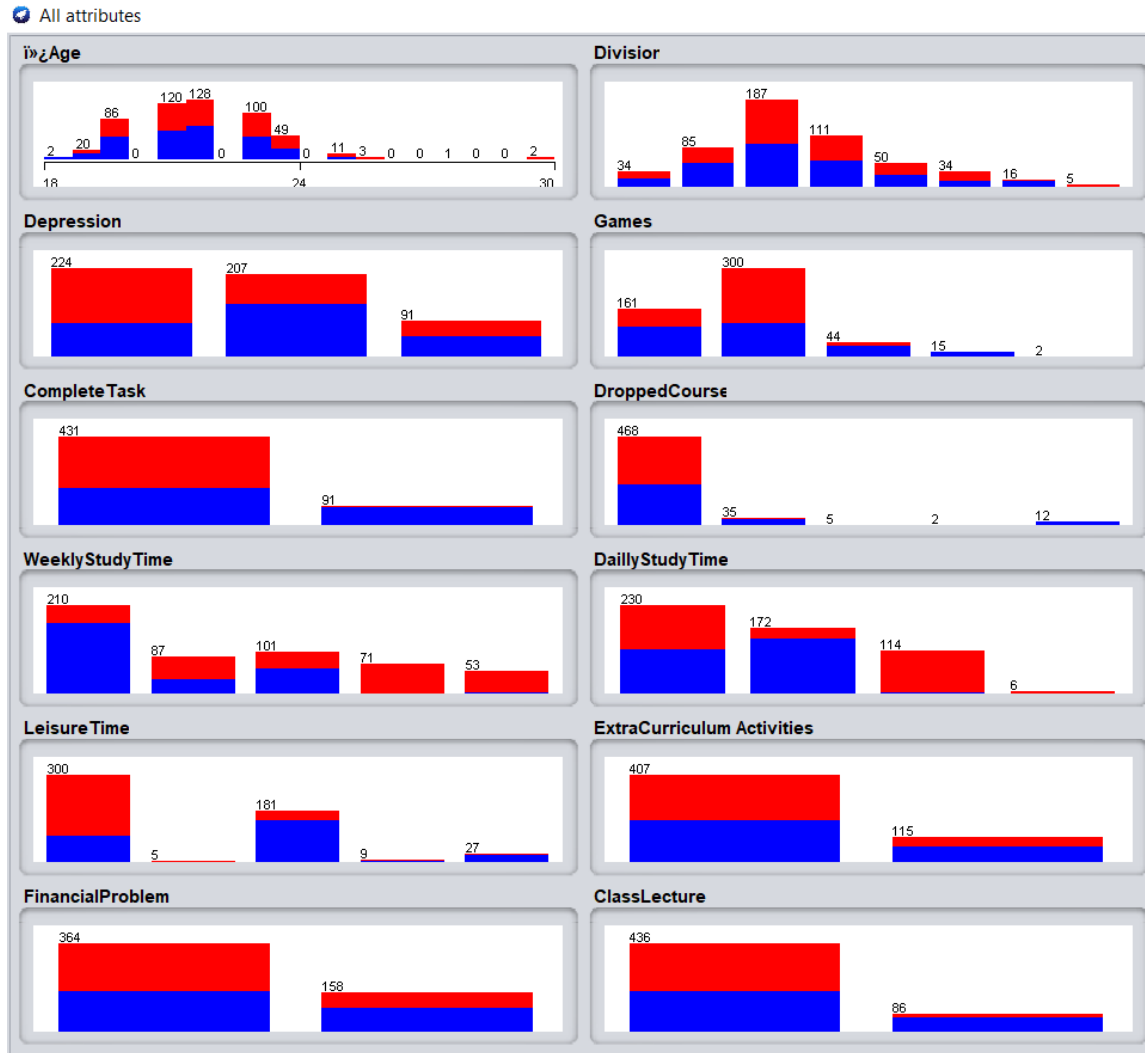


Figure 4.2.7: Visual representation of features for Weka implementation.

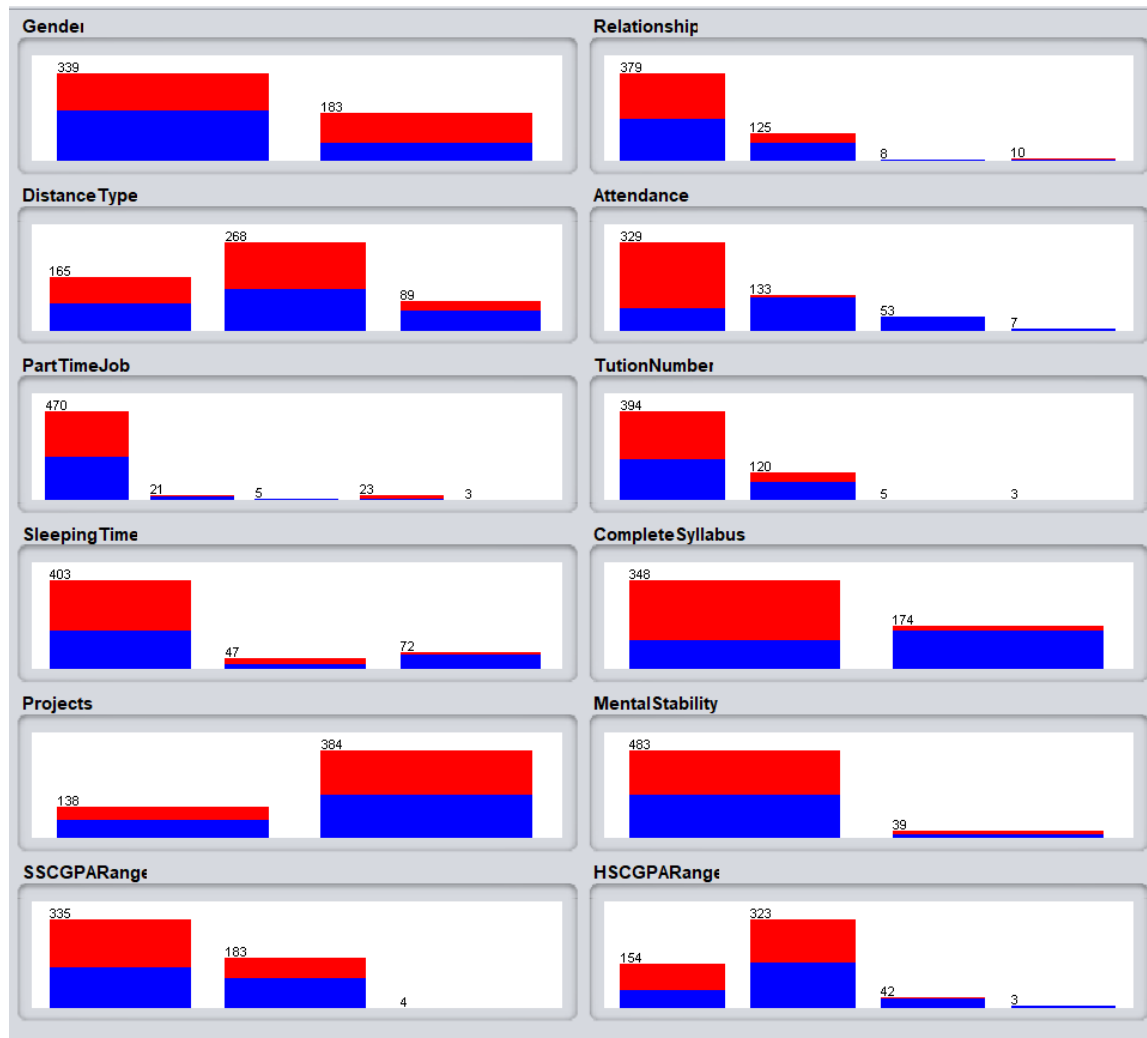


Figure 4.2.8: Visual representation of features for Weka implementation.

After the preprocessing using Weka, we have applied different types of classification algorithms. Tree J48 classifier is one of them.

The following figure 4.2.9 shows the implementation of J48 tree classifier algorithm using Weka software:

```

Correctly Classified Instances      432           82.7586 %
Incorrectly Classified Instances    90           17.2414 %
Kappa statistic                    0.6553
Mean absolute error                 0.2124
Root mean squared error             0.3597
Relative absolute error             42.4861 %
Root relative squared error         71.9418 %
Total Number of Instances          522

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
      0.808    0.152    0.846    0.808    0.827      0.656    0.887    0.850    Bad
      0.848    0.192    0.810    0.848    0.828      0.656    0.887    0.869    Good
Weighted Avg.   0.828    0.172    0.828    0.828    0.828      0.656    0.887    0.859

=== Confusion Matrix ===

  a  b  <-- classified as
215  51 |  a = Bad
 39 217 |  b = Good

```

Figure 4.2.9: Applying J48 tree classifier algorithm using Weka software.

Here we see that accuracy is 82.75%.

The following figure 4.2.10 shows the visual representation of the J48 tree algorithm.

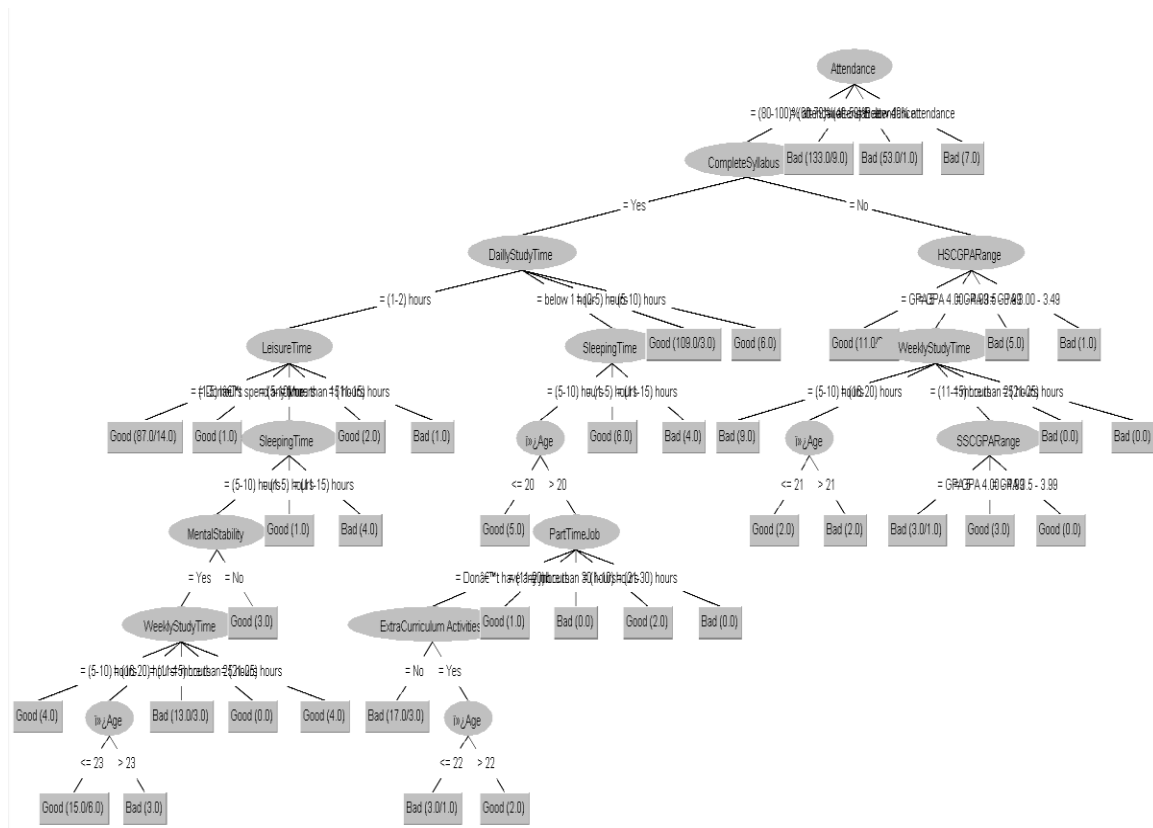


Figure 4.2.10: Visual representation of J48 tree algorithm.

We also applied the Random Forest tree algorithm.

The following figure 4.2.11 shows the implementation of the Random Forest tree algorithm using Weka software.

```

Correctly Classified Instances      437           83.7165 %
Incorrectly Classified Instances    85           16.2835 %
Kappa statistic                    0.674
Mean absolute error                 0.2371
Root mean squared error             0.3338
Relative absolute error             47.4453 %
Root relative squared error         66.7632 %
Total Number of Instances          522

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
      0.857    0.184    0.829    0.857    0.843      0.674    0.925    0.928    Bad
      0.816    0.143    0.846    0.816    0.831      0.674    0.925    0.930    Good
Weighted Avg.   0.837    0.164    0.837    0.837    0.837      0.674    0.925    0.929

=== Confusion Matrix ===

  a  b  <-- classified as
228 38 |  a = Bad
 47 209 |  b = Good

```

Figure 4.2.11: Applying Random Forest tree algorithm using Weka software.

Here the accuracy was 83.71%. It is the nearest to the J48 algorithm.

In both cases Cross-validation Folds were kept 10.

Other classification algorithms were also applied. But the above two algorithms gave the best accuracy compared to other algorithms.

We also analysed our data using Weka in different ways. Such as classifying them into several categories. We also used several regression algorithms using Weka as we have kept a column in our dataset which contains numeric values of SGPA.

In table 4.2.12, table 4.2.13 and table 4.2.14 accuracy of different regression and classification algorithms are shown:

TABLE 4.2.12: REGRESSION ALGORITHMS ACCURACY

Algorithm	Accuracy
Linear Regression Algorithm with All Nominal and Numeric Attributes	64.86 %
Random Forest Regressor Algorithm with All Nominal and Numeric Attributes	97.21 %
Linear Regression Algorithm with only Numeric Attributes	62.55 %
Random Forest Regressor Algorithm with only Numeric Attributes	97.11 %
Linear Regression Algorithm with only Nominal Attributes	41.37 %
Random Forest Regressor Algorithm with only Nominal Attributes	96.06 %

TABLE 4.2.13: CLASSIFICATION ALGORITHMS ACCURACY

Algorithm	Accuracy
Random Forest Classifier Algorithm with All Nominal and Numeric Attributes	86.71 %
KNeighbors (KNN) Classifier Algorithm with All Nominal and Numeric Attributes	82.65 %
Support Vector Classifier (SVC) Algorithm with All Nominal and Numeric Attributes	82.65 %
Random Forest Classifier Algorithm with only Numeric Attributes	99.42 %
KNeighbors (KNN) Classifier Algorithm with only Numeric Attributes	83.81 %
Support Vector Classifier (SVC) Algorithm with only Numeric Attributes	97.11 %
Random Forest Classifier Algorithm with only Nominal Attributes	87.28 %
KNeighbors (KNN) Classifier Algorithm with only Nominal Attributes	79.19 %
Support Vector Classifier (SVC) Algorithm with only Nominal Attributes	84.39 %

TABLE 4.2.14: CLASSIFICATION ALGORITHMS ACCURACY USING WEKA SOFTWARE

Algorithm	Accuracy
Random Forest Tree Algorithm with only Nominal Attributes	83.71 %
J48 Tree Algorithm with only Nominal Attributes	82.75 %

4.3 Discussion

From our experimental result we can see that, Random Forest algorithm gave more accuracy compared to other algorithms. It is efficient in both classification and regression. So we decided to use the Random Forest algorithm for the web implementation. So that we can predict a student performance in terms of semester result or SGPA from their attributes as input.

For implementing our model with python flask and deploying it to the web using Heroku first of all we saved our model. We saved our machine learning model as model.pkl by dumping. This work was done using pickle which was imported. Then our model began eligible for predicting student performance from the attributes of students by which it was trained. We used html, css, JavaScript, bootstrap etc. for the markup and design of the web. We used two html files located in the templates folder. One is index.html and another is result.html file. In this case, index.html file was used to take input using form. These inputs are those attributes which were used to train our model and by these inputs student performance will be predicted. There are two types of inputs. One is a real number and another input is for nominal value. First type input supports integer and float type value. For nominal inputs there are several options. These options have a corresponding value. And these values are those values which were generated during building our model using dictionary mapping. During using label encoder, we have created numerical values of each categorical label. Ultimately while selecting any nominal valued option from html form actually we are selecting the categorical values by which they were trained. We used bootstrap as well as css to make our web interface more attractive. Finally, we kept a submit button in the index.html file for taking those inputs. Performance of students in terms of SGPA is shown in the result.html file which is predicted by our model using those attributes. Some screenshots of output of index and result html files are given below.

The following figure 4.3.1 and figure 4.3.2 show deploying models to the web for predicting SGPA.

Fill Up The Form For Predicting SGPA

SGPA Prediction Form

Enter your age

Range: (18-30) years

Division:

Dhaka

Gender:

Male

Relationship Status:

Single

Depression Status:

Yes

Playing Games Time:

(1-4) hours

Figure 4.3.1: Deploying model to web for predicting SGPA.

Mental and Physical Stability:

Yes ▼

Facing Financial Problem?

Yes ▼

Have publications or research paper?

Yes ▼

Class Lecture Understand:

Yes ▼

SSC GPA:

3.5 - 3.99 ▼

HSC GPA:

3.00 - 3.49 ▼

Submit

© 2020 Copyright: Group DELTA

Figure 4.3.2: Deploying model to web for predicting SGPA.

The following figure 4.3.3 shows predicted SGPA using Regression.

Result After Prediction

Your Predicted SGPA is: 3.88 (Float upto 2 decimal places)

SGPA: 3.8829500000000006 (In details)

© 2020 Copyright: Group DELTA

Figure 4.3.3: Predicted SGPA using Regression.

The following figure 4.3.4 shows predicted SGPA using Classification.



Figure 4.3.4: Predicted SGPA using Classification.

Another important file during building our Flask App is script.py. All main parts are done inside this file. All-important machine learning, flask, web libraries are imported in this file. Two routes are created inside this file. One is for index.html and another is for result.html. Two customized functions are also created inside script.py file. One is for result prediction in terms of SGPA where work of prediction is done. It takes input as a list, makes the list into an array and loads the model for prediction. It returns the ultimate result which is predicted using predict function. Another function is the result function in which inputs from the form are collected. Also, it makes a list of those values and passes it another function in which work of prediction is done. When the function returns the value of result then it keeps it in a variable and passes it to result.html form using render_template for showing the result. Finally the performance of students is predicted.

After making the flask app and attaching our model with it our next step was to deploy it to the web. To do this we had to follow several steps. We wanted to deploy our flask app to Heroku which is a platform as a service. To do this first of all we had to install git and Heroku CLI. We had to install gunicorn to the virtual environment where the flask was installed. Gunicorn does the work of handling requests and complicated tasks are also handled by it. Then we created a requirements.txt file. We used many libraries and files while building our flask app. By creating requirements.txt file we inform Heroku that our project needs these for running our flask app. Then we created a text file named Procfile to inform Heroku what is essential for starting our flask app. We also created a file named .gitignore which helps to remove unnecessary files while deploying our app to heroku. After creating the above files using some commands, finally our flask app was deployed on Heroku.

We deployed 3 types of Flask web apps for predicting student semester results or SGPA. One of them was deployed using regression algorithm. [17]

Another was deployed using classification algorithm. [18]

And the last one was deployed using a regression algorithm with numeric attributes. [19]

In our experiment, first of all we preprocessed our data for applying machine learning algorithms and making our desired machine learning model. We applied different feature extraction algorithms for finding important features for our machine learning model. We applied different types of regression and classification algorithms on that preprocessed data. We applied K-Nearest Neighbors, Support Vector Machine, Random Forest, J48 Tree algorithms. We used Python Jupyter Notebook, Weka Software for applying those algorithms. Among those algorithms, the Random Forest algorithm gave the best accuracy compared to others. As the Random Forest algorithm gave the best accuracy, we used the Random Forest algorithm for web implementation. We attached our machine learning model which is made using the Random Forest algorithm. We used the Python Flask framework for making our web app. And finally deployed our app to Heroku. Thus we made a web app using data mining techniques and machine learning algorithms which can predict the semester result of a student using their attributes as input.

CHAPTER 5

Impact on Society, Environment and Sustainability

5.1 Impact on Society

Our research has a great impact on our society. There is a positive impact on our society. From our research students will get benefits. They can predict their performance earlier. From their prediction they will be aware if the prediction is bad. They can develop their attributes which are correlated with their academic result. They can improve their academic result from their prediction. Necessary steps can be taken and authority can take several steps for those students whose results are predicted as bad. By doing these they will be eligible for higher studies both in our country and abroad. They need not to retake their courses. They can save their money and valuable times. Moreover their parents will be proud of their brilliant results. By developing their result they will be regarded as assets both for our society and country. They will be able to contribute for our society and country positively. These are the positive impacts from our research.

5.2 Impact on Environment

There is no negative impact of our research on the environment. Actually our research does not have any correlation with the environment. Basically our research is for the students. How to check their performance based on their attributes. After predicting their academic results they will be warned if their predicted result is very much bad. Moreover different necessary steps can be taken by authorities from their predicted result. Students can develop their attributes which have effects on their academic results if the predicted results are poor. In our research we analyzed the data of students in different ways. We tried to implement data mining techniques and machine learning algorithms in the education sector. So we can see that there was not any impact on the environment in our research.

5.3 Ethical Aspects

Major ethical terms have been maintained during conducting our research works. Informed consent has been followed during our research.

While working with data, ethical aspects should be maintained and we should be aware of the data related terms such as personal data, data ownership, consent, purpose of use, trustworthiness etc. [20]

We collected our dataset by informing the students about our purpose. And for which their data will be used. Nothing was kept hidden in this case. Our research was not harmful to any person, community or the world. Rather it was beneficial for the students. It was helpful for the students to develop their performance. Our research was helpful for their self judgement. They can judge themselves earlier and improve their academic result according to that. There was respect for anonymity and confidentiality in our research. Protection for privacy was also kept in our research. We did not collect any personal information of a student by which our research was conducted. We designed our dataset in such a way that it doesn't contain any personal information of anyone. Only common attributes of students were kept in our research. Privacy was strictly maintained in our research. Skills for conducting our research were obtained before doing it as a researcher. All ethical issues were taken into consideration in our research so that it cannot harm anyone.

5.4 Sustainability Plan

We can keep a maintenance plan for our research project. We have completed our primary version of research. If we want to extend it and maintain it in future, we need a sustainability plan. Several steps can be taken in this case. We can update our web app which was deployed to the web by adding additional features day by day. We can make our web app more attractive in future with attractive design using html, css, javascript, bootstrap etc. We can add some additional latest features which will make our prediction based web app more efficient and easier. We can apply more efficient and hybrid algorithms to make our model more efficient. Other attributes can be taken into consideration for web implementation. More research can be conducted on our topic. These works and similar works can be taken into consideration for our maintenance in future. These are all about our sustainability plan.

CHAPTER 6

Summary, Conclusion, Recommendation and Implication for Future Research

6.1 Summary of the Study

In our research, we wanted to implement data mining techniques and machine learning algorithms in education sector. Our aim was to predict student performance from their attributes. So that at the beginning of their study necessary steps can be taken to improve their academic performance. For this reason, first of all we have collected primary data of students. We have collected those data which have correlation with student academic performance in terms of semester result or SGPA. We collected our data through online google survey form and offline survey form. Data was collected from university students. After collecting our dataset we preprocessed it. After preprocessing our data, we applied different machine learning algorithms for predicting semester results. Among those algorithms, the Random Forest algorithm gave the best accuracy compared to other algorithms. That's why we made our machine learning model using the Random Forest algorithm. We attached our machine learning model to the Python Flask framework for web implementation. In our web app we made a form to collect inputs. These inputs are those attributes of students by which our machine learning model was trained and created. That means these are the attributes of students which have correlation with student academic performance. We loaded our model with these inputs to predict student performance. Thus our web app can predict student performance in terms of semester result or SGPA from the attributes of a student. After making our web app, we deployed it to the web using Heroku. Thus our web app can predict student results from their attributes at the beginning of their learning process. If their results are predicted as bad, necessary steps can be taken at the beginning of their learning process to develop their result. These were all about our research.

6.2 Conclusions

Data mining and machine learning algorithms can play a great role in the education sector if we can use it in an efficient way. If the performance of a student can be predicted from the beginning of their teaching from their attributes, we can take extra care of those students from

the prediction. Day by day data is increasing rapidly in the world. And by using and analyzing this data, especially the data of students we can develop our education sector which will be helpful for students. We can identify the weakness of a student and the related factors of bad results and can take necessary steps regarding these factors. We can arrange counsel hours for the students with the probability of bad performance and we can give suggestions and the right direction from earlier. In our research, we tried to explain all processes and implementation of data mining and machine learning algorithms which are helpful for developing the education sector.

6.3 Implication for Further Study

In our research, we collected the data of university students. In future another type of data can be collected with different attributes. Another domain of students can be chosen for implementation of data mining and machine learning algorithms. Other attributes which have effects on student performance can be identified and collected. Large amounts of data can be collected from many departments. If possible, the student database of any educational institute or university can be used for further analysis. Another machine learning algorithm can be used to see the differences among algorithms and which algorithm gives best accuracy can be identified. Data can be analyzed in different ways with additional tools in future.

REFERENCES

- [1] J. Agee, "Developing qualitative research questions: a reflective process", *International Journal of Qualitative Studies in Education*, vol. 22, no. 4, pp. 431-447, 2009. Available: 10.1080/09518390902736512.
- [2] M. Alvesson and J. Sandberg, "GENERATING RESEARCH QUESTIONS THROUGH PROBLEMATIZATION.", *Academy of Management Review*, vol. 36, no. 2, pp. 247-271, 2011. Available: 10.5465/amr.2011.59330882.
- [3] Onwuegbuzie, Anthony J., and Nancy L. Leech, "Linking Research Questions to Mixed Methods Data Analysis Procedures", *The qualitative report*, vol. 11, no. 3, pp. 474-498, 2006.
- [4] J. Randolph, "A Guide to Writing the Dissertation Literature Review", *Practical Assessment, Research, and Evaluation*, vol. 14, no. 1, p. 13, 2009. Available: <https://scholarworks.umass.edu/pare/vol14/iss1/13>.
- [5] D. Boote and P. Beile, "Scholars Before Researchers: On the Centrality of the Dissertation Literature Review in Research Preparation", *Educational Researcher*, vol. 34, no. 6, pp. 3-15, 2005. Available: 10.3102/0013189x034006003.
- [6] A. Kaur, N. Umesh and B. Singh, "Machine Learning Approach to Predict Student Academic Performance", *International Journal for Research in Applied Science and Engineering Technology*, vol. 6, no. 4, pp. 734-742, 2018. Available: 10.22214/ijraset.2018.4125.
- [7] V. Shanmugarajeshwari and R. Lawrance, "Analysis of students' performance evaluation using classification techniques," 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16), Kovilpatti, 2016, pp. 1-7. Available: 10.1109/ICCTIDE.2016.7725375.
- [8] E. Belachew and F. Gobena, "Student Performance Prediction Model using Machine Learning Approach: The Case of Wolkite University", *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 7, no. 2, pp. 46-50, 2017. Available: 10.23956/ijarcse/v7i2/01219.
- [9] C. Baba, A. Govindu, M. Sai Raavi and V. Somisetty, "STUDENT PERFORMANCE ANALYSIS USING CLASSIFICATION TECHNIQUES", *International Journal of Pure and Applied Mathematics*, vol. 115, no. 6, pp. 1-7, 2017.
- [10] A. Soni, V. Kumar, R. Kaur and D. Hemavathi, "PREDICTING STUDENT PERFORMANCE USING DATA MINING TECHNIQUES", *International Journal of Pure and Applied Mathematics*, vol. 119, no. 12, pp. 221-227, 2018.
- [11] A. Acharya and D. Sinha, "Early Prediction of Students Performance using Machine Learning Techniques", *International Journal of Computer Applications*, vol. 107, no. 1, pp. 37-43, 2014. Available: 10.5120/18717-9939.
- [12] S. Beniwal and J. Arora, "Classification and Feature Selection Techniques in Data Mining", *International Journal of Engineering Research & Technology (IJERT)*, vol. 1, no. 6, pp. 1-6, 2012.
- [13] M. Karagiannopoulos, D. Anyfantis, S. Kotsiantis and P. Pintelas, "Feature selection for regression problems", *Proceedings of the 8th Hellenic European Research on Computer Mathematics & its Applications*, Athens, Greece, vol. 2022, 2007.
- [14] Feature Selection Techniques in Machine Learning with Python, available at <<
<https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e>>>, last accessed on 27-03-2020 at 10:22 PM.
- [15] W. Lin, Z. Wu, L. Lin, A. Wen and J. Li, "An Ensemble Random Forest Algorithm for Insurance Big Data Analysis," in *IEEE Access*, vol. 5, pp. 16568-16575, 2017.

[16] The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-Learn and Spark, available at <<<https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3>>>, last accessed on 16-03-2020 at 02:27 PM.

[17] SGPA Prediction using Regression, available at <<<https://sheltered-wave-71161.herokuapp.com/>>>, last accessed on 23-03-2020 at 12:15 AM.

[18] SGPA Prediction using Classification, available at <<<https://pure-shelf-21646.herokuapp.com/>>>, last accessed on 23-03-2020 at 12:30 AM.

[19] SGPA Prediction Regression (With Numeric Attributes), available at <<<https://gentle-stream-91665.herokuapp.com/>>>, last accessed on 23-03-2020 at 12:48 AM.

[20] D. Hand, "Aspects of Data Ethics in a Changing World: Where Are We Now?", Big Data, vol. 6, no. 3, pp. 176-190, 2018. Available: 10.1089/big.2018.0083.

APPENDICES

Appendix A: Offline Survey or Field Survey

Age:

Department:

Home District:

Division:

S.S.C GPA:

H.S.C GPA:

SGPA (Last/Previous semester GPA)*:

Current Semester (Such as 1st/2nd/3rd etc.):

Choose the correct option about yourself

1. What is your SGPA range of last semester?

- ☐ 4.00 ☐ 3.75-3.99 ☐ 3.50-3.74 ☐ 3.25-3.49 ☐ 3.00-3.24
☐ 2.75-2.99 ☐ 2.50-2.74 ☐ 2.25-2.49 ☐ 2.00-2.24 ☐ 0.00-1.99

2. Gender

- ☐ Male ☐ Female

3. Relationship Status?

- ☐ Single ☐ In a relationship ☐ Married ☐ Divorced

4. Do you suffer from depression which is harmful for your study?

- ☐ Yes ☐ No ☐ Not sure

5. How much time do you play online/offline games? (Such as PUBG, Ludo Star, Clash of Clans, FIFA etc)

- ☐ More than 15 hours ☐ (10-15) hours ☐ (5-9) hours ☐ (1-4) hours
☐ Don't Play Game

6. Select the distance type of your university from home (distance from where you go to your university daily)?

- ☐ Large Distance (Above 10 KM) ☐ Medium Distance (4-10 KM) ☐ Small Distance (0-3 KM)

7. What is the percentage of your attendance in your class?

- ☐ (80-100) % ☐ (60-79) % ☐ (40-59) % ☐ Below 40%

8. Do you complete your home tasks, assignment, presentation, quizzes etc. properly?

- ☐ Yes ☐ No

9. How much marks (percentage) do you get in the exam (such as quizzes and mid-term)?

- ☐ (80-100) % ☐ (60-79) % ☐ (40-59) % ☐ Below 40%

10. How many times have you ever had to drop a course for any reason?

- ☐ 4 or more times ☐ 3 times ☐ 1 time ☐ Never drop any course

11. How many hours per week do you spend in a part-time or full-time job?

- ☐ More than 30 hours ☐ (21-30) hours ☐ (11-20) hours ☐ (1-10) hours
☐ Don't have any job

12. How many tuitions do you have for earning money?

- ☐ More than 5 tuitions ☐ (3-5) tuitions ☐ (1-2) tuitions ☐ No tuition

13. How many hours do you study per week?

- ☐ (5-10) hours ☐ (11-15) hours ☐ (16-20) hours ☐ (21-25) hours

14. How much time do you study daily?

- ☐ More than 10 hours ☐ (6-10) hours ☐ (3-5) hours ☐ (1-2) hours ☐ Below 1 hour

15. What is your daily sleeping time?

- ☐ (1-5) hours ☐ (6-10) hours ☐ (11-15) hours ☐ more than 15 hours

16. Can you complete your whole syllabus properly before every exam?

- ☐ Yes ☐ No

17. How much time do you spend in your leisure time daily? (Such as in Facebook, WhatsApp, YouTube, Television Program or any other similar platform)

- ☐ More than 15 hours ☐ (11-15) hours ☐ (6-10) hours ☐ (1-5) hours ☐ Don't spend any time

18. Are you involved in any extra-curriculum activities? (Such as singing song, dancing, attending different type of competitions, cultural activities etc.)

- ☐ Yes ☐ No

19. Do you spend more time in doing projects, increasing skills rather than academic study?

- ☐ Yes ☐ No

20. Are you mentally and physically stable/fit for your study?

- ☐ Yes ☐ No

21. Do you face any financial problem to manage all your needs?

- ☐ Yes ☐ No

22. Do you have any publications in any conference paper, journal or anywhere based on research area?

- ☐ Yes ☐ No

23. Can you understand course related lectures properly in the classroom delivered by course teacher?

- ☐ Yes ☐ No

Appendix B: Online Survey Link

<https://forms.gle/sCGBWxDE13nQrYmr6>

PLAGARISM REPORT

Turnitin Originality Report

Processed on: 01-May-2020 21:52 +06

ID: 1313219995

Word Count: 8471

Submitted: 1

Student Performance Prediction By Md.
Anisur Rahman Rony

Similarity Index	Similarity by Source	
	Internet Sources:	N/A
7%	Publications:	N/A
	Student Papers:	7%

exclude quoted	exclude bibliography	exclude small matches	mode: quickview (classic) report
Change mode	print	refresh	download

2% match (student papers from 11-Feb-2018)

Submitted to Daffodil International University on 2018-02-11

<1% match (student papers from 02-Sep-2017)

Submitted to University of Sheffield on 2017-09-02

<1% match (student papers from 16-Sep-2019)

Submitted to Deakin University on 2019-09-16

<1% match (student papers from 28-Jun-2019)

Submitted to KDU College Sdn Bhd on 2019-06-28

<1% match (student papers from 11-Apr-2015)

Submitted to Nanyang Technological University, Singapore on 2015-04-11

<1% match (student papers from 09-Dec-2019)

Submitted to Higher Education Commission Pakistan on 2019-12-09