

**A MACHINE LEARNING APPROACH TO PREDICT THE CHANCES OF
DROOPING OUT STUDENTS DUE TO COVID-19 IN UNIVERSITY
PERSPECTIVE BANGLADESH**

BY

**MD. AMIRUL ISLAM
ID: 191-15-12123**

**MD. HASANUR RAHMAN
ID: 191-15-12804**

**AND
MOST. SAIRA TABASSUM
191-15-12825**

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

Supervised By

Mr. Raja Tariqul Hasan Tushar
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Mr. Md. Azizul Hakim
Senior Lecturer
Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY


DHAKA, BANGLADESH

JANUARY 2023

APPROVAL

This Project titled "A MACHINE LEARNING APPROACH TO PREDICT THE CHANCES OF DROPPING OUT STUDENTS DUE TO COVID-19 IN UNIVERSITY PERSPECTIVE BANGLADESH", submitted by Md. Amirul Islam, Md. Hasanur Rahman and Saira Tabassum, Student ID No 191-15-12123, 191-15-12804 and 191-15-12825 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfilment 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 26 January 2023.


BOARD OF EXAMINERS



Dr. Touhid Bhuiyan
Professor and Head

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

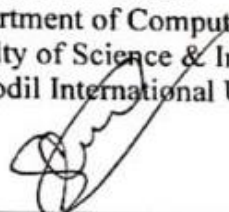
Chairman



Sazzadur Ahmed
Assistant Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

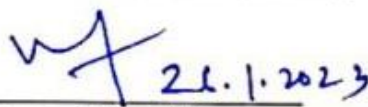
Internal Examiner



Ms. Sharmin Akter
Senior Lecturer

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Dr. Ahmed Wasif Reza
Associate Professor

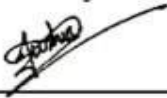
Department of Computer Science and Engineering
East West University

External Examiner

DECLARATION

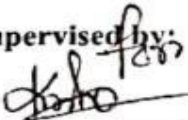
We hereby declare that, this project has been done by us under the supervision of **Mr. Raja Tariqul Hasan Tushar, Assistant Professor, Department of CSE Daffodil International University**. We also declare that neither this project nor any part of this project has been submitted elsewhere for award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

Supervised by:



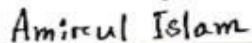
Mr. Raja Tariqul Hasan Tushar
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:



Mr. Md. Azizul Hakim
Lecturer
Department of CSE
Daffodil International University

Submitted by:



Md. Amirul Islam
ID: -191-15-12123
Department of CSE
Daffodil International University

Hasanur Rahman

Md. Hasanur Rahman
ID: -191-15-12804
Department of CSE
Daffodil International University

Saira Tabassum

Most. Saira Tabassum
ID: -191-15-12825
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 **Mr. Raja Tariqul Hasan Tushar, Assistant Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “*Machine Learning*” to carry out this work. His endless scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Dr. Touhid Bhuiyan, Professor and 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

The difficulties of COVID 19 have exposed humanity to some terrible truths. The pandemic's grip has severely harmed a number of industries, including education. Numerous days of school, college, and university closures caused the students to be disengaged from their academics. The amount of students who leave university for practical or financial reasons has become a major source of concern. We successfully investigate the university student dropout rate in our research. We look for the underlying causes of their dropout and work to provide a workable solution. We have gathered information from more than 400 undergraduate Bangladeshi students via an online survey. The most effective techniques for predicting dropout among Bangladeshi students were found after training and testing the dataset with a number of well-known algorithms, including SVM, Logistic Regression, Random Forest, Decision Tree, etc.

TABLE OF CONTENTS

CONTENT	PAGE
Approval	i
Declaration	ii
Acknowledgment	iii
Abstract	iv
List Of Table	vii
List Of Figure.....	viii

CHAPTER

Chapter 1 Introduction..... 1-6

1.1 Introduction.....	1
1.2 Motivation.....	1-2
1.3. Rationale of study	2-3
1.4. Research Questions	3
1.5. Expected Outcomes	3-4
1.6. Report Layout	4-6

Chapter 2 Background 7-10

2.1 Preliminaries	7
2.2 Related Works.....	7-9
2.3 Comparative Analysis and Summary.....	9
2.4 Scope of the Problem	10
2.5 Challenges.....	10

Chapter 3. Research Methodology	11-18
3.1. Research Subject and instruments	11
3.2 Dataset utilized.....	11-13
3.3 Statistical Analysis	13-14
3.4 Applied Mechanism	14-17
3.5. Implementation Requirements	17-18
Chapter 4. Experimental Results and Discussion.....	19-28
4.1: Experimental Setup.....	19
4.2 Experimental Results & Analysis	19-28
4.3 Discussion:	28
Chapter 5: Impact on Society, Environment and Sustainability	29-31
5.1 Impact of Society	29
5.2 Impact on Environment.....	30
5.3 Ethical Aspects.....	30-31
5.4 Sustainability Plan	31
Chapter 6: Summary, Conclusion, Recommendation and Implication for Future.....	32-33
6.1; Summary of the Study	32
6.2: Conclusions.....	32-33
6.3: Implication for further study.....	33
References	34-35

LIST OF TABLES

TABLES	PAGE NO
Table 2.2.1 - Related work comparison	9
Table 4.2.1 – Accuracy, Precision, Recall and F1- score of Algorithms	22
Table 4.2.2 - Confusion Matrix of Bernoulli Naive Bayes	22
Table 4.2.3 - Confusion Matrix of Decision Tree	23
Table 4.2.4 - Confusion Matrix of Random Forest	23
Table 4.2.5 - Confusion Matrix of Logistic Regression	23
Table 4.2.6 - Confusion Matrix of XGBoost Regression	24
Table 4.2.7 - Comparison of results	27

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.2.1 – Percentage of dropped out students during COVID-19	12
Figure 3.2.2 - Reason for dropping the semester	12
Figure 3.2.3 - Drop out percentage of different categories	13
Figure 3.3.1 - Dataset representations	14
Figure 3.4.1 – Missing values of dataset	15
Figure 3.4.2 – Required Python Libraries	15
Figure 3.4.2.1 – Graph for Bernoulli Naïve Bayes	16
Figure 3.4.2.2 – Graph for Random Forest	17
Figure 4.2.1 – Graph for Bernoulli Naïve Bayes	19
Figure 4.2.2 - Graph for Decision Tree Regression	20
Figure 4.2.3 - Graph for Random Forest Regression	20
Figure 4.2.4 - Graph for Logistic Regression	21
Figure 4.2.5 - Graph for XGBoost Regression	21
Figure 4.2.6 - Confusion Matrix for Bernoulli Naive Bayes	24
Figure 4.2.7- Confusion Matrix for Decision Tree Regression	25
Figure 4.2.8- Confusion Matrix for Random Forest Regression	25

Figure 4.2.9- Confusion Matrix for Logistic Regression	26
Figure 4.2.10- Confusion Matrix for XGBoost Regression	26
Figure 4.2.11 – Prediction level of Algorithms	27

CHAPTER 1

Introduction

1.1 Introduction

The worst epidemic in human history was COVID 19. This pandemic hit all sector disastrously. Among those education sectors is one of the most important sectors. Schools and colleges have had their departments shut down an infinite number of times. The institution has not been properly launched. It was so difficult to predict when things will return to normal. We made a significant effort to do a study on the student dropout rate during the COVID-19 period, taking the scenario and the effects of the pandemic into consideration. The study focused mostly on university students in Bangladesh, and the goal was to determine the percentage of students who left school due to financial difficulties or other unnamed problems. We conducted a poll of more than 500 undergraduate Bangladeshi students when the country was in lockdown. Google form was used to aid with the survey. The five algorithms Bernouli Naive Bayes, DecTreeReg, RandomForestRegressor, LogisticReg, and LinearReg were trained and tested on the dataset, and they were shown to be the best approaches for predicting dropout among Bangladeshi students. Situational considerations included the conviction that lockdown was endangering their family's financial stability, academic performance, COVID-19 symptoms, and health issues. The findings showed that hassles were significantly correlated with students' perceptions that lockdown was interfering with their educational pursuits, while anxiety was positively correlated with worries about one's own and one's family's health. Although the public is no longer under lockdown, the epidemic will continue to have a significant influence on academics over the next years. Therefore, it is crucial for authorities to provide them with appropriate emotional, financial, and other forms of support.

1.2 Motivation

Our study's main objective is to ascertain whether they are receiving any benefits from taking online course, as well as whether and to what extent they are happy. As a result of what we discovered, there has to be some study done to anticipate Covid-19 dropout university students and identify the causes of dropout. We have made efforts to find a solution. It's past time to focus on the dropout problem at the graduation level in the Covid-19 scenario and establish measures to

address it as the number of public and private institutions grows daily, as does the number of students. Therefore, in order to create a conceptual framework for lowering dropout rates at the undergraduate and other levels of education in Bangladesh, we will not only examine the elements that contribute to graduation dropout but also the causes for dropout in general. We have chosen a few characteristics to determine if the students may continue their semester and attend their online classes. such as their domicile, type of university, internet access, lack of concentration, type of reading, whether or not they were impacted by COVID-19, and academic outcomes like CGPA, SSC, and HSC GPA, etc. We will be able to comprehend the rate of student dropout during COVID-19 and the primary causes of their dropout by addressing the issues they are not happy with and bringing them to a right method. Our teachers will be able to use this approach to overcome all the difficulties and safeguard the students from being dropped out by correctly organizing the class if we ever face a fresh epidemic like COVID-19 and need an online class again. We believe it is feasible to identify the causes of students stopping their studies.

1.3 Rationale of study

Several definitions of student dropout were created in prior study. The question of whether students will be active through the last week or if this is the last week that they will be active is the one that is covered the most frequently. Early identification of students who are at risk of dropping out reduces the issue and enables the identification of the required conditions. We will be able to comprehend the rate of student dropout during COVID-19 and the primary causes of their dropout by addressing the issues they are not happy with and bringing them to a right method. Our teachers will be able to use this approach to overcome all the difficulties and safeguard the students from being dropped out by correctly organizing the class if we ever face a fresh epidemic like COVID-19 and need an online class again. We believe it is feasible to identify the causes of students stopping their studies. The utilization of trustworthy, precise, and important data is

necessary for a good forecast. The size of the dataset is a frequent error in many research projects in educational technology, learning analytics, and educational data mining. In addition to employing logs, it is difficult to collect enough educational data to properly fulfill the machine learning needs. The challenge of predicting student performance might also be resolved by

classification using Bernoulli Naive Bayes, DecTreeRegression, RandomForest, LogisticRegression, and Linear Regression.

1.4 Research Questions

Research questions is the outcome about related study along with theoretical analysis and data exploration. Here, the query is addressed as to how the task is accomplished and provides a quick overview of the system.

- How will the system predict dropout student's ratio?
- How may system will differentiate raw data and possible dropout students' data?
- How may a regular student can lead to possible dropout student?
- What kind of algorithms are suited to carry out our tasks more precisely?
- How much information do we need to successfully carry out our work?
- How can this model be trained using the right data set?
- Is there any proposed model exists? If so, how was it created and how effectively does it function?
- How may the shortcomings of the current model be fixed?

1.5 Expected Outcomes

The expected outcome of our proposed work is to identify the dropout students of Bangladeshi University at early stage along with better accuracy with selected algorithm and train a model by training to solve the problem

- Early detection of possible dropout students
- The result of examination will be more accurate than the work which has done before
- According to our examined result students will get help to over the situation
- Students' dropout prediction can help universities to reduce dropout rates of students

1.6 Report Layout

This work is based on thesis and has a total of six segments. The work presents a variety of perspectives, each of which is represented as a chapter. Distinct subheadings are divided up into each chapter and represented in an understandable manner. A list of everything in this report is provided below:

Chapter 1

Introduced the effort and discuss its goals, drivers, research questions, and expected outcomes. The topics we have discussed: 1.1 Introduction, 1.2 Motivation, 1.3 Rationale of the Study, 1.4 Research Questions, 1.5 Expected outcome and 1.6 Report Layout. The introduction, including the phases of student's dropout in Bangladeshi University and how they create impact on sociative backdrops. And it has been debated how they develop, disseminate, and affect every individual student. The primary driving force for our thesis study was covered in the section on motivation. The study's major goal was covered in the section on the Rationale of the study part. The Research question chapter explains the major questions that pertain to our investigation. The expected outcome section includes a description of the result we've been working toward. The Report layout structure breaks out our whole project into chapters.

Chapter 2

It provides a summary of earlier work that has been done in this situation. Studying earlier work makes it easier to fully comprehend the work that has to be done for our research. We witness the effects of the authors' choice to draw a line through this field of study later in Chapter 2. The following subjects have been covered: Preliminaries/Terminologies, Related Works, Comparative Analysis and Summary, Scope of the Problem, and Challenges are all included in this report.

Chapter 3

Describes the following steps to build the project. Machine learning algorithms are introduced in this chapter. The following topics are: 3.1 Research Subject and Instrumentation, 3.2 Data Collection Procedure/Dataset Utilized, 3.3 Statistical Analysis, 3.4 Proposed Methodology/Applied Mechanism - 3.4.1 Data Processing, 3.4.2 Proposed Model, 3.5 Implementation Requirements.

Chapter 4

This section includes visualization of graphs, results of the proposed model. Here following topics, we have discussed are: 4.1 Experimental Setup, 4.2 Experimental Results & Analysis, 4.3 Discussion.

Chapter 5

Specifies what is trustworthy to appear in the complete project report and proposal. The chapter comes to a close with a discussion of the limits of our study, which may be used as a springboard for other people's future research. We have talked about the following subjects: Impact on Society, Environmental Impact, Ethical Aspects, and Sustainability Plan are all listed in section 5.

Chapter 6

This section presents the conclusions of this study. The following part we have discussed is: 6.1 Summary of the Study, 6.2 Conclusions, 6.3 Implication for Further Study

CHAPTER 2

Background

2.1 Preliminaries:

One of the most delicate elements to build a nation is education. Education is the key product of success. However, dropout from educational institution as university became more catastrophe in recent years. A significant number of students are dropped through this catastrophe every year. In addition, future demolishing, this catastrophe also causes financial loss and emotional anguish for the affected students family. This crisis situation known as dropout is caused due to the lack of proper acknowledgement and care of education sector. It may be fatal and result in the destruction of the loss of life due to anxiety, depression etc. Therefore, dropout catastrophe can enhance survival when it gets detected early, and infected student's rates will drop as a result. In order to identify this at the primary level and to diagnose that early on, different algorithm is utilized properly. Since this condition is discovered by time by time, there is a chance for error during a normal process of examination, which might cause tragedy for the affected students family. Automated prediction which is made through algorithm is currently gaining popularity. because the detection accuracy rate is so precise. There are certain restrictions on the prior research in this area. Therefore, we are working within those constraints to fix problems that will improve the accuracy of our system. We analyzed 15 publications that were relevant to our study out of the more than 35 that we investigated in order to provide a better outcome that would directly benefit affected students and their families.

2.2 Related Works:

The Covid-19 pandemic has exposed many incapability and injustice to our education system. Our Main goal of this research is predicting dropout rate of students during COVID-19 where we can use this model to reduce dropout rate. Globally speaking, the issue of drop out is becoming a significant issue day by day. Higher education institutions have tried to use technology to ensure continuity of education in spite of the lockdown and offers online classes and learning experiences

as a Alternatives to in-class time. But due to poor internet facility, students disrupt their studies and this internet problem is one of the reasons for drop out.

G.sallan and S.Behal et al. used enhanced machine learning algorithms. They predict the reasons for dropout students based on the AI expert's system. They used Decision Stump, NDTREE, EMLA algorithms. Their model shows 78.37% accuracy [1]

Presently, the most popular study issue is predicting student dropout. Using several algorithms including Logistic Regression, Decision Tree, Random Forest, Naive Bayes, and Support Vector Machine, Janka Kabathova et al. demonstrated the student dropout rate. Random Forest methods demonstrated improved accuracy. [2]

Marcell Nagy and Roland Molontay et al represent the research paper on dropout rate based on Secondary School Performance. They gathered 15,825 data from undergraduate students which is a large dataset. They applied the following machine learning models on the datasets such as Decision Tree, Boosted Tress, Deep learning etc. Deep learning provided the best accuracy which is 73.5%. In this work, they used advanced machine-learning techniques [3]. Mingjie Tan et al. proposed a model that identifies potential dropouts in e-learning programs. In this work, they used an artificial neural network algorithm that consists of an input layer unit, output layer unit, and hidden layer unit. They also used Decision Tree(DT), Bayesian Networks. Among of them Decision Tress(DT) algorithm gives 94.63% higher accuracy [4].

Costa et al. focused on course data to predict potential dropouts [5].

Abir and Rasel et al. They developed a model using the Random Forest algorithm that showed 97.4% accuracy. In this model, they collected real data of 2100 students and generated 70 features in the dataset.[6]

Mia Hossain and Labib et al, worked on seven algorithms and their main objective was to identify whether students are benefiting from online classes or not. They used different machine learning algorithms. Svm gives the best result which is 94% accuracy [7]

Table 2.2.1 - Related work comparison

Author	Area of study	Machine Learning Models	Accuracy
G.sallan and S.Behal et al	Students Dropout	EMLA (Enhanced Machine Learning Algorithm)	78.37%
Janka Kabathova et al	Student's Dropout in University Courses	Random Forest	93%
Marcell Nagy et al	Dropout based on Secondary School Performance	Deep Learning	73.5%
Mingjie Tan et al	Student Dropout in E-Learning	Neural Network, Decision Tree, Bayesian Networks	94.63%
Abir and Rasel et al	University Student current semester dropout risk	Random Forest	97.4%

2.3 Comparative analysis and Summary:

In our proposed work, we have suggested deep learning method in order to more effectively identify along with predict dropout students. We have suggested a Deep Convolutional Network (CNN) for entirely automated dropout students in raw data that can address major causes and issues in order to meet our objectives. We evaluated our proposed approach using a dataset that is openly accessible. The publicly viewable data set we utilized was acquired from Kaggle. Data scientists may still engage with one another on a range of topics using the Kaggle platform. Dataset preprocessing and augmentation were undertaken after sampling. It boosted the dataset's quality and length. Resnet-50 and Denset-121 have both been used to achieve this. Based on their level of accuracy, the two algorithms have been examined. Another model merging U-net architecture using Inception v3 has been established so as to contrast overall achievement. Then, the accuracy standards of the two models were compared. To evaluate our proposed model, Accuracy, Precision, Recall, and F1-score have also been generated.

2.4 Scope of the Problem:

When student start to discontinue from attending educational institution on the regular basis, then it's known as dropout. There are basic kinds of this catastrophe depending upon where it developed. Primary phases develop from the school and primary school level where they first manifest. Inside of the educational institution, dropouts can cause several different societal issues depending on the number of dropouts. This is deadly to students and their family from different perspectives. Not to mention the financial and mental hardships the student and their family must endure while pursuing this catastrophe. The probability of survival for the students and their family with dropout catastrophe can be improved with early identification and prediction. Additionally, it might aid the mental, social position. Automated dropout prediction and detection can make it simpler and more accurate to find the sensitive issue for what this is happening for. To achieve this, machine learning techniques can be applied. We concentrated on applying deep learning techniques to better accurately and precisely detect this social and national catastrophe. We also concentrated on finding ways to address the problems that arise while employing deep convolutional neural networks. This dramatically improves the student's chances of surviving.

2.5 Challenges:

The major obstacle we confronted while working was accumulating and assessing the information. Due to unforeseeable circumstances, it was difficult for us to gather data in person while we couldn't go to the primary schools and universities. We were forced to gather the information for our research online as a result. The next step was to choose the datasets that would work best for our research. Our data was gathered via Kaggle. This is a platform where data scientists may compete with one another on the chosen field of specialization. Another trouble started once the data had been compiled. To appropriately have used the artificial intelligence, we needed to restructure our raw data into digital data. We have to push ourselves to work constantly in to achieve superior quality and higher accuracy.

CHAPTER 3

Research Methodology

3.1 Research Subject and instruments

Detecting early, the chances of someone dropping their studies during Covid-19 is sometimes lifesaving, for some families. Visiting the accounts office of the organization is one of the practical ways to resolve the issue, for the students, who are in need. This procedure sure does minimize the problem to a certain extent but the organization still lacks the proper research to take any methodological steps, toward this issue. Using machine learning can be an excellent way to conduct this sort of research. The principal goal of this study is to predict the dropout rate of university students during COVID-19 based on drop-out reasons, family income, and CGPA results. To accomplish our objective, we've employed various machine learning algorithms such as Random Forest Regression, Bernoulli Naive Bayes, Decision Tree Regression, Logistic Regression, XGBoost Regression, etc. To execute and run the code We have used google colab. The code is written using the Python programming language with NumPy, pandas, matplotlib, scikit-learn, seaborn, statsmodels libraries. A highly configured PC with a high GPU is used to execute the machine learning algorithms and machine learning models in a smooth way.

3.2 Data Collection Procedure/Dataset Utilized

For our research, we collected data from university students through online Google survey forms and offline survey forms. About 418 students' information is collected from the online and offline surveys. After pre-processing, we train our dataset to estimate the accuracy of the model.

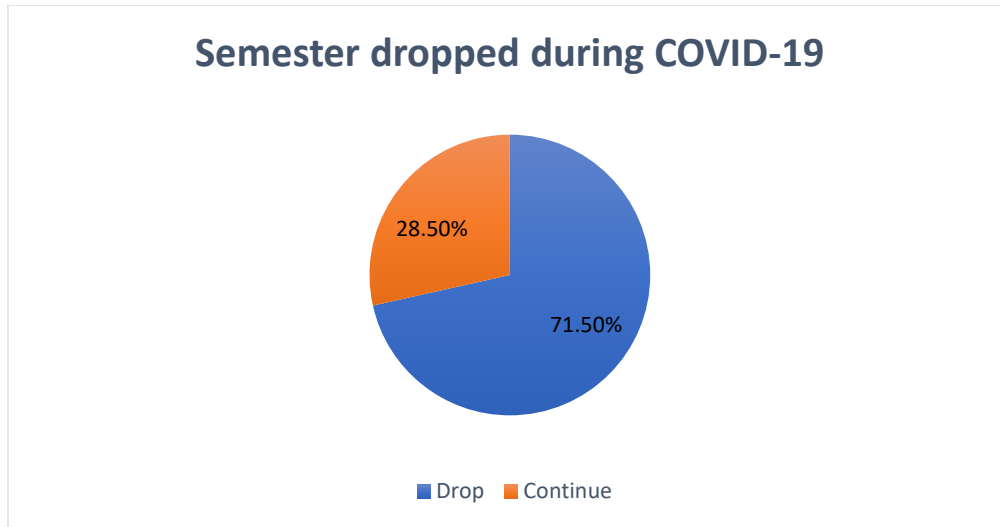


Figure 3.2.1 – Percentage of dropped out students during COVID-19

Figure 3.2.1 shows that 71.5% of students did not drop out and 28.5% of students dropped out during the Covid-19 pandemic.

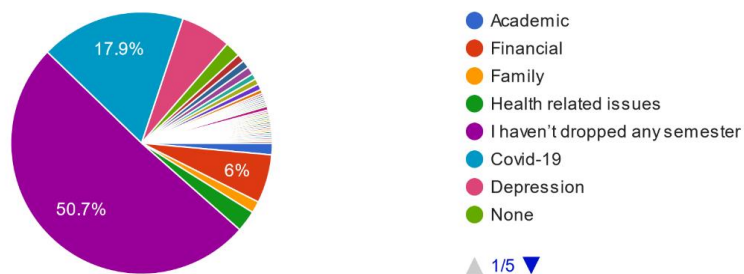


Figure 3.2.2 - Reason for dropping the semester

From Figure 3.2.2, 50.7% of the students preferred that they did not skip any semester. The rest chose other reasons. Among them, 17.9% of students dropped out due to Covid-19. 6% for financial problems. 6.2% due to depression. 1.4% for family problems. 2.6% because of health-related issues. 1.4% is caused by academic results.

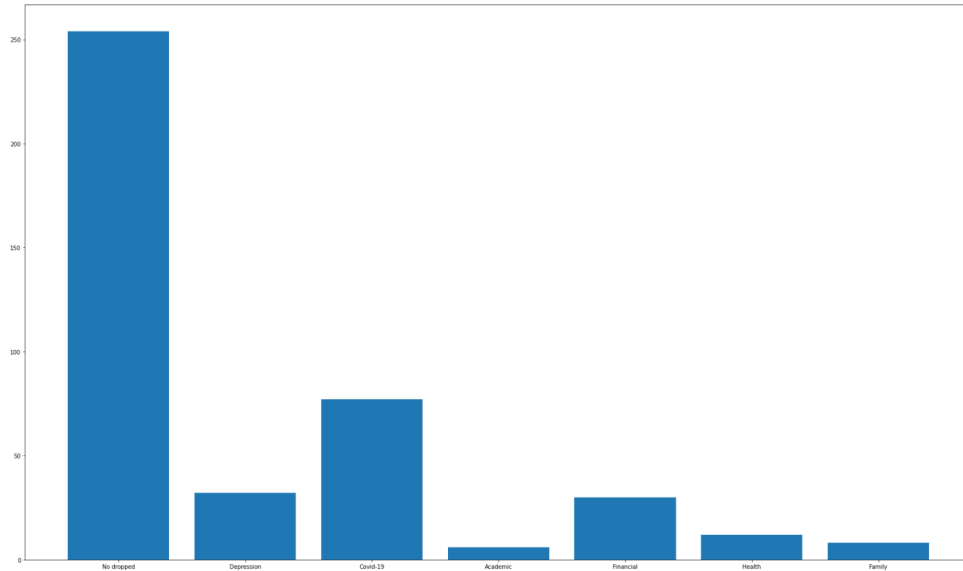


Figure 3.2.3 - Drop out percentage of different categories

Using unique function from panda's libraries, we calculate the dropout percentage of different categories. From Figure 3.2.3, the No dropped out rate is the highest. After that, the section on covid-19 is higher in percentage. The main reason of drop out is the Covid-19 pandemic.

3.3 Statistical Analysis

From figure 3.3.1 given below, we can see the visualization of the dataset distribution where the data is divided into two categories. One is the dropout count. Another is the no drop-out calculation. A total of 418 data we have used in our study. From this figure, we can show that 299 students did not drop out throughout the time of COVID-19 but 119 students dropped out during the COVID-19 pandemic.

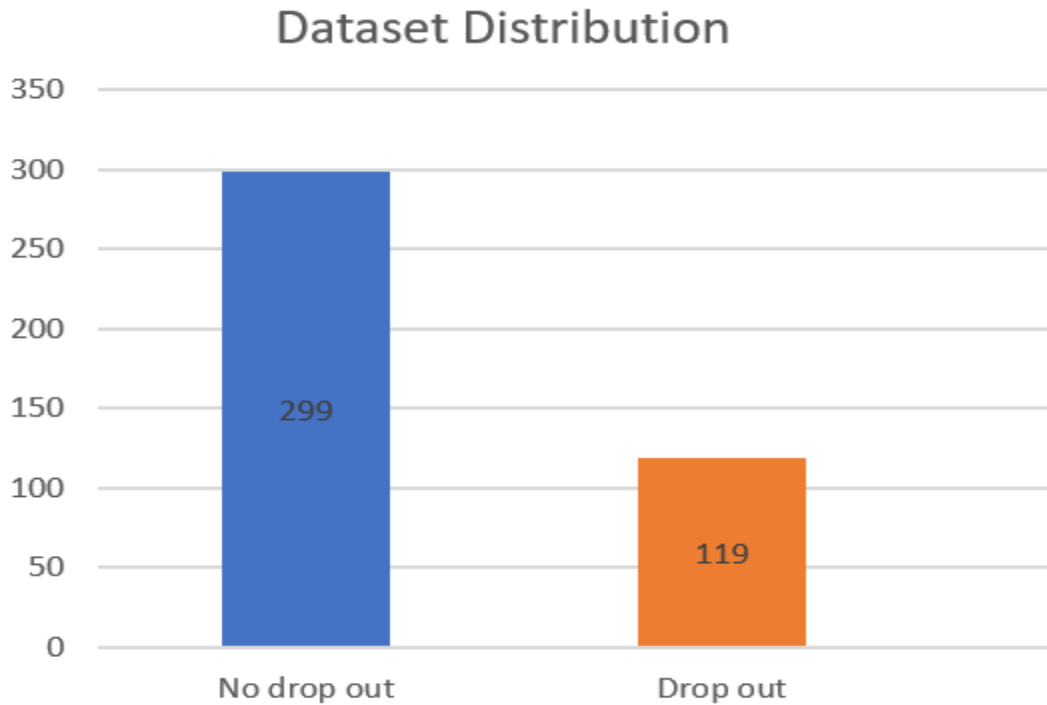


Figure 3.3.1 - Dataset representations

3.4 Proposed Methodology/Applied Mechanism

3.4.1 Data Processing

The process of our data collection happened, from online and offline Google Forms for our research. The survey was conducted online and offline through Google Forms. 18 questions were included in our Form for subjects to answer. Being that's the case, we have collected a decent amount of data physically during this process. A total of 418 students provided information in our Google Form. After the collection of the data, we clean some while dropping unnecessary columns. Project architecture involves multiple steps to achieve the required output. We rename some variables so that we can get the expected result. For that, we have selected some characteristics from the dataset which are gender, current CGPA, university type, reason, family income, and internet facility. For x_{train} we selected these features and for y_{train} we selected drop semester features. This helped us to check whether there are any missing values in our dataset or not.


```
[ ] data.isnull().sum()
```

```
Gender          0
Current_cgpa    0
University_types 0
Drop_semester   0
Reason          0
Living_area     0
Family_income   0
Internet_facilities 0
dtype: int64
```

Figure 3.4.1 – Missing values of dataset

We have removed the character ',' from the family income values so that we can get clear prediction results. And for some students entering integer number in cgpa section converts cgpa value to float number. Next, we calculate household income based on 'no drop' and 'drop semester'.

3.4.2 Proposed Model

Step 1 - Importing the necessary libraries

```
#import library
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.datasets import make_classification
from sklearn.model_selection import train_test_split
import seaborn as sns
import sklearn as sk
import sklearn.metrics as skl
from sklearn import metrics
from sklearn.metrics import classification_report
from sklearn.metrics import r2_score
import statsmodels.api as sm
import statsmodels.formula.api as smf
```

Figure 3.4.2 – Required Python Libraries

- Numpy
Numpy is used to manipulate multi-dimensional arrays and matrices. It can solve many high mathematical problems.
- Pandas
Pandas is a type of Python package that can handle large datasets and is an open-source Python package. It is mostly used for data manipulation.
- Matplotlib
Matplotlib is used to create interactive and animated visualizations of datasets.
- Scikit – learn
scikit-learn has some methods that help make decisions with different types of algorithms
- Statsmodels
The Statsmodels is a Python module that contains some classes and functions for statistics.

In the model, we have built to achieve high accuracy. First, we used Bernoulli Naïve Bayes, Decision Tree Regression, Random Forest Regression, Logistic Regression, and XGBoost regression algorithms. We iterate the algorithm on the training data and make decisions.

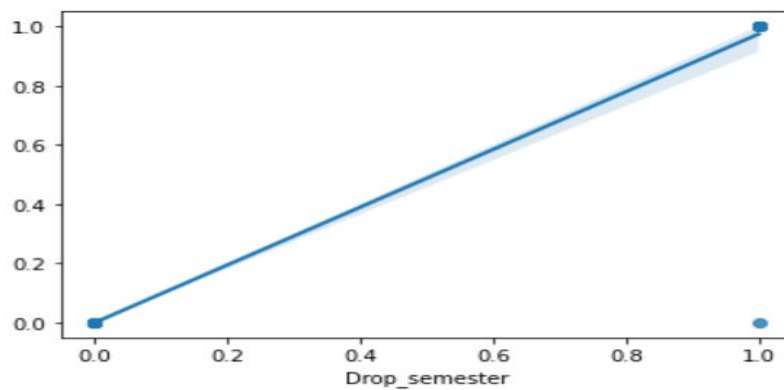


Figure 3.4.2.1 – Graph for Bernoulli Naïve Bayes

1. Bernoulli Naïve Bayes

In this process, we have used 20 percent of data for testing and 80% of data for training. After training and testing the model, we acquire 98 percent accuracy. In Bernoulli Naïve Bayes algorithm, we cannot use words in features instead of words we have to use binary numbers so that the model can give us high accuracy results. We convert our target values to binary numbers (0 and 1).

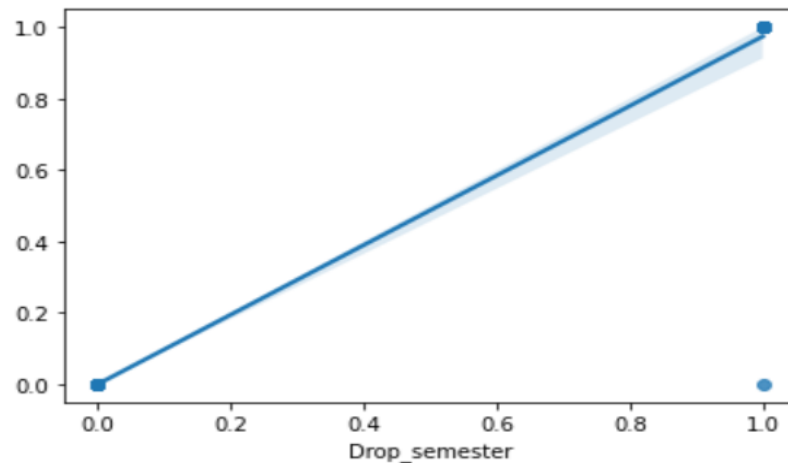


Figure 3.4.2.2 – Graph for Random Forest

2. Random Forest

Random forest is an algorithm which is used for predicting many industries related issues. Training and testing data are divided into two categories, we used 20% data for testing dataset and 80% for model training. This model showed 98% accuracy which is similar to the Bernoulli Naïve Bayes algorithm. From Figure 3.4.2.2, the blue line depicts the predicted values.

3.5 Implementation Requirements

In the Platform, we've employed google colab to run the code, train the dataset and test the model that we build to find the predictions we make. Other IDEs like jupyter notebook can be used for development. To build machine learning models and run applications, a sufficient amount of random-access memory (RAM), and fast performance central processing unit (CPU) are required.

Implementation Requirements are:

Hardware Requirements-

- x64-based processor
- 64-bit operating system
- CPU having @2.10GHz 2.59 GHz
- Intel(R) Core (TM) i3-10110U processor
- 8.00 GB RAM

Software Requirements-

- Windows 10 Pro operating system

Coding tools-

- Google Colab editor
- Python environment setup

CHAPTER 4

Experimental Results and Discussion

4.1 Experimental Setup

We collected data for this investigation using both online and offline Google Forms. The dataset was prepared, and a CSV file was subsequently created. The file size of the dataset is 63kb. Since, we have trained many machine learning models, by mounting Google Drive to Google Colab, we can then use the dataset CSV file to perform our research work, we used a wide range of PCs with 64-bit operating systems and x64-based processors. We used Windows 10 Pro operating system to compile and test the code. To run the code, we have used google colab which is a notebook. The Python programming language, to create the machine learning models, we utilized. We used to google colab and a jupyter notebook to execute the code. It takes RAM: 1.36 GB/ 12.68 GB and Disk: 22.94 GB/ 107.72GB.

4.2 Experimental Results & Analysis

Figure 4.2.1 shows the Actual and Predict data based on Income. Bernoulli Naive Bayes has shown 98.8% accuracy on test data. Before applying Bernoulli Naive Bayes to the dataset, we have to confirm that all values on the dataset are binary numbers. We convert target variable values into 0 and 1. This algorithm gives a 98.8% accuracy score on test data.

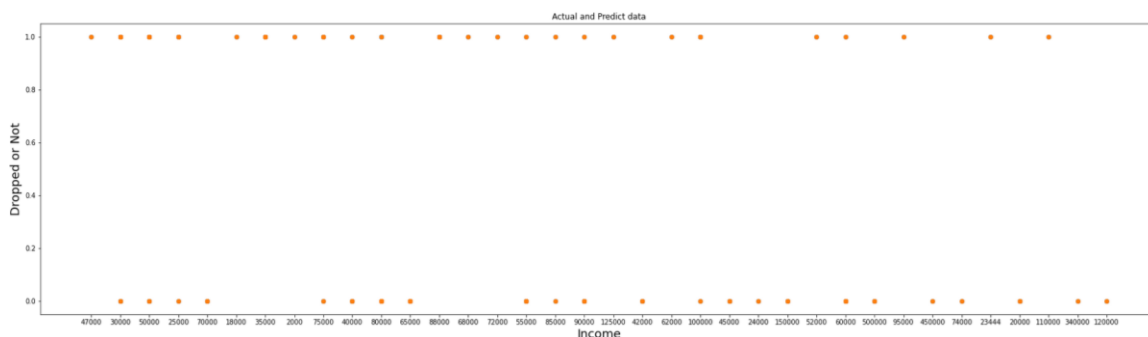


Figure 4.2.1 – Graph for Bernoulli Naïve Bayes

In figure 4.2.2, Decision Tree Regression gives 98.8% accuracy on training data. Decision Tree is used to resolve regression tasks with training data.

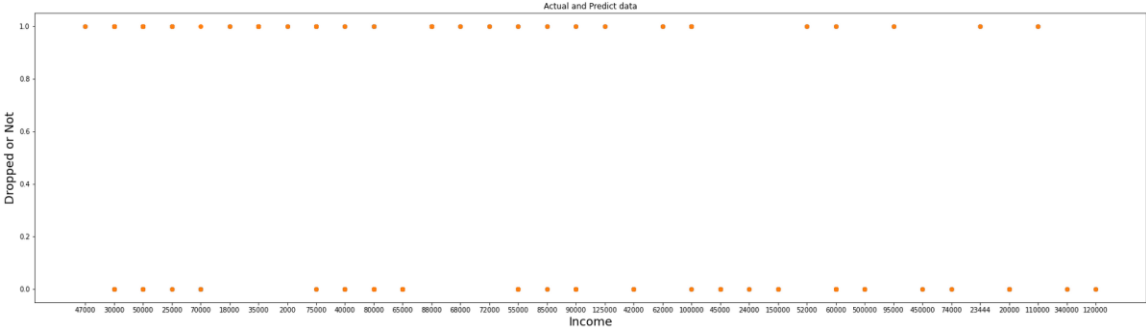


Figure 4.2.2 - Graph for Decision Tree Regression

From figure 4.2.3 given below, Random Forest Regression with 98.8% accuracy is similar to Decision Tree regression.

The dataset contains a small amount of data and we select 6 features. Random forest regression by fewer multiple trees that give accurate results.

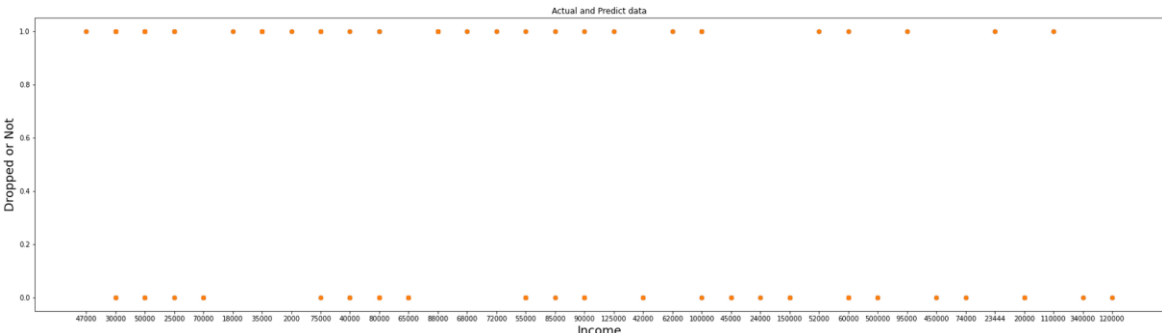


Figure 4.2.3 - Graph for Random Forest Regression

From Figure 4.2.4, Logistic Regression has shown 94.0% accuracy. The association between the dependent and independent variables is established through logistic regression. It gives 94.0% accuracy with a 5.95% accuracy error.

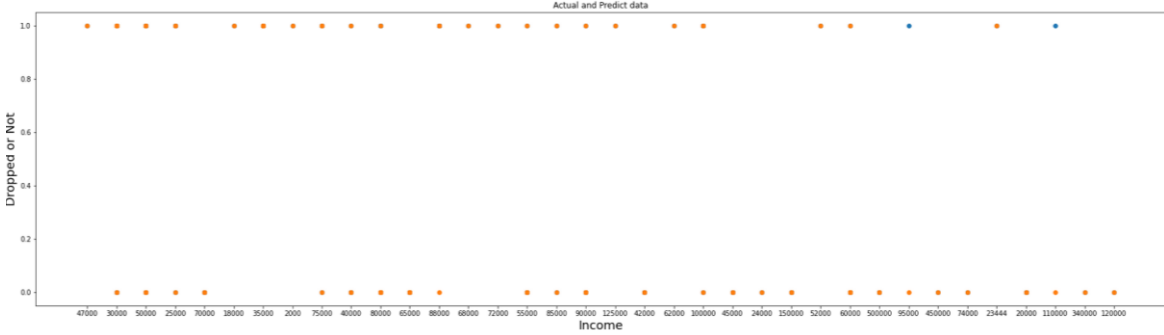


Figure 4.2.4 - Graph for Logistic Regression

From the given bellow Figure 4.2.5 has shown that XGBoost Regression gives 75% accuracy with a 93% score of this model.

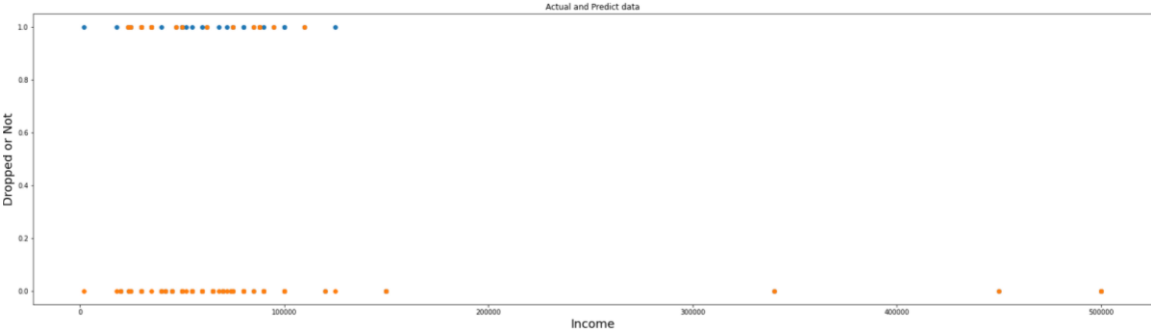


Figure 4.2.5 - Graph for XGBoost Regression

Table 4.2.1 – Accuracy, Precision, Recall and F1- score of Algorithms

Algorithms	Accuracy	Precision	Recall	F1-score
Bernoulli Naive Bayes	0.98	0.98	1.00	0.99
Decision Tree	0.97	0.98	0.98	0.98
Random Forest	0.98	0.98	1.00	0.99
Logistic Regression	0.94	0.90	1.00	0.95
XGBoost	0.75	0.68	1.00	0.81

Table 4.2.2 shows the True Positive, True negative, False Positive, and False Negative of Bernoulli Naive Bayes which is used for making predictions.

Table 4.2.2 - Confusion Matrix of Bernoulli Naive Bayes

Algorithm	True Positive	True Negative	False Positive	False Negative
Bernoulli Naive Bayes	45	38	0	1

Table 4.2.3 shows the True Positive, True negative, False Positive and False Negative of Decision Tree which is used for making predictions.

Table 4.2.3 - Confusion Matrix of Decision Tree

Algorithm	True Positive	True Negative	False Positive	False Negative
Decision Tree	44	38	1	1

Table 4.2.4 shows the True Positive, True negative, False Positive and False Negative of Random Forest which is used for making predictions.

Table 4.2.4 - Confusion Matrix of Random Forest

Algorithm	True Positive	True Negative	False Positive	False Negative
Random Forest	45	38	0	1

Table 4.2.5 shows the True Positive, True negative, False Positive and False Negative of Logistic Regression which is used for making predictions.

Table 4.2.5 - Confusion Matrix of Logistic Regression

Algorithm	True Positive	True Negative	False Positive	False Negative
Logistic Regression	45	34	0	5

Table 4.2.6 shows the True Positive, True negative, False Positive and False Negative of XGBoost Regression which is used for making predictions.

Table 4.2.6 - Confusion Matrix of XGBoost Regression

Algorithm	True Positive	True Negative	False Positive	False Negative
XGBoost	45	18	0	21

Confusion matrix are useful for classifying problems. It provides us with data visualization of correct and incorrect number prediction. Figure 4.2.6 to Figure 4.2.10 represents the Confusion matrix of the algorithms that are used for our model.

Let's examine the confusion matrix for our model:

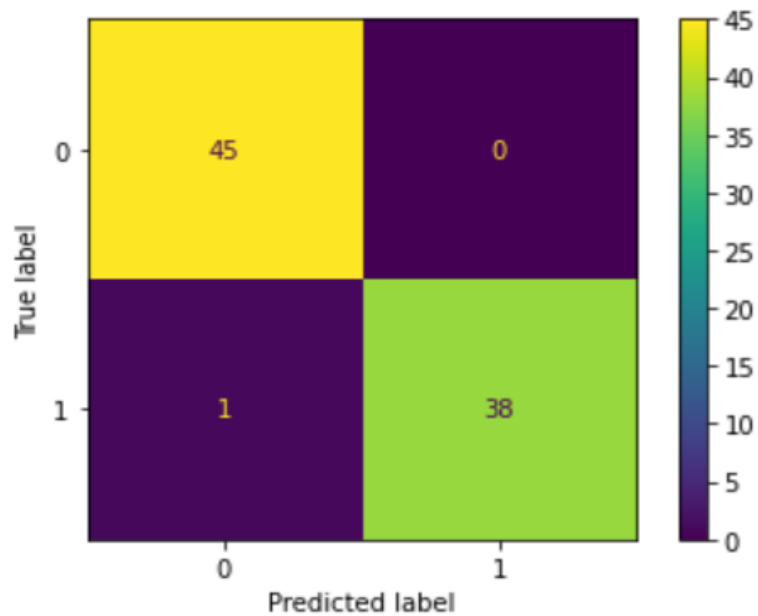


Figure 4.2.6 - Confusion Matrix for Bernoulli Naive Bayes

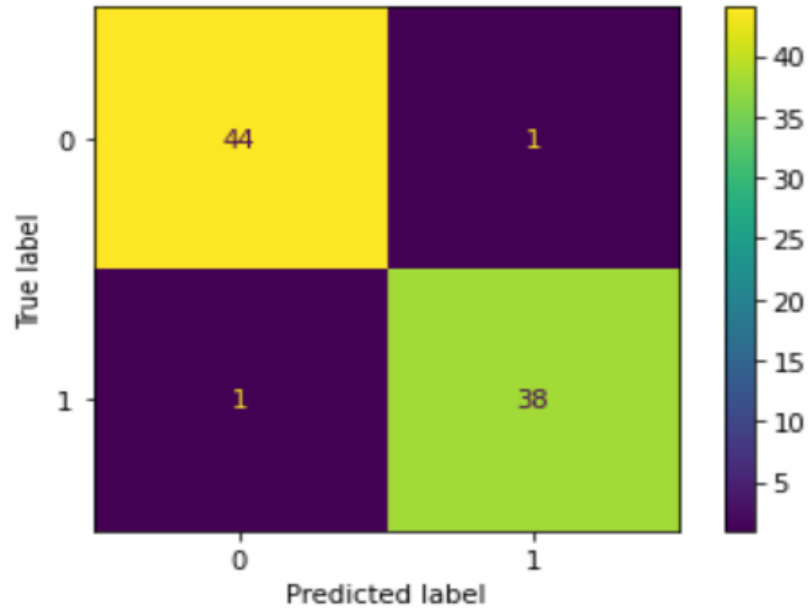


Figure 4.2.7- Confusion Matrix for Decision Tree Regression

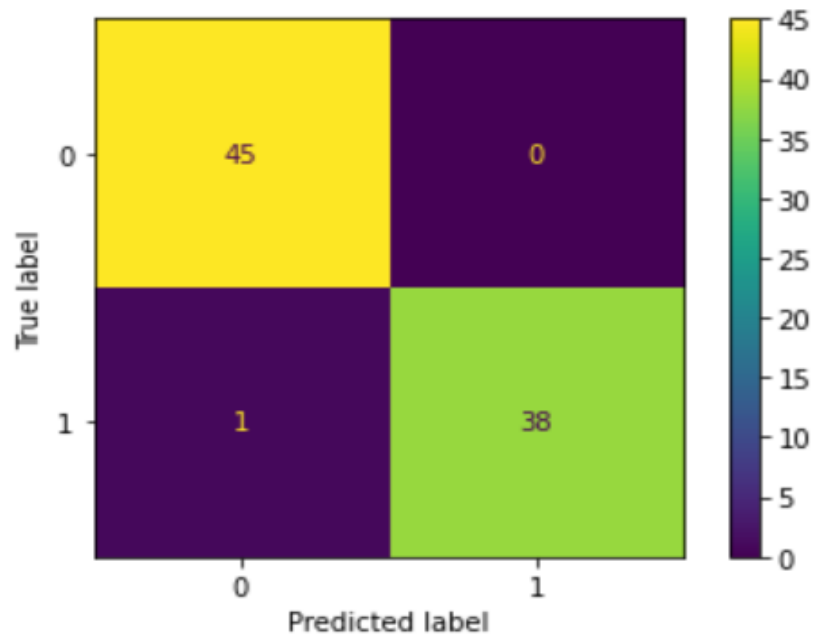


Figure 4.2.8- Confusion Matrix for Random Forest Regression

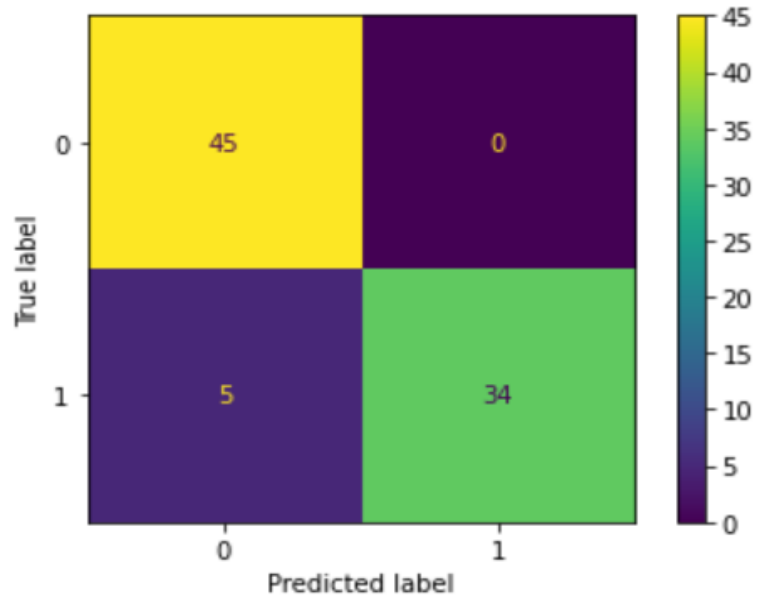


Figure 4.2.9- Confusion Matrix for Logistic Regression

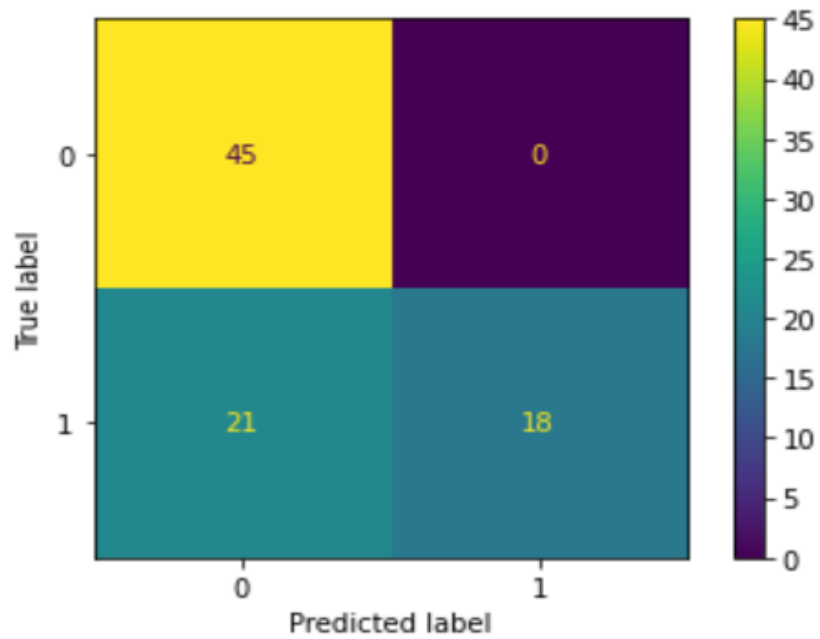


Figure 4.2.10- Confusion Matrix for XGBoost Regression

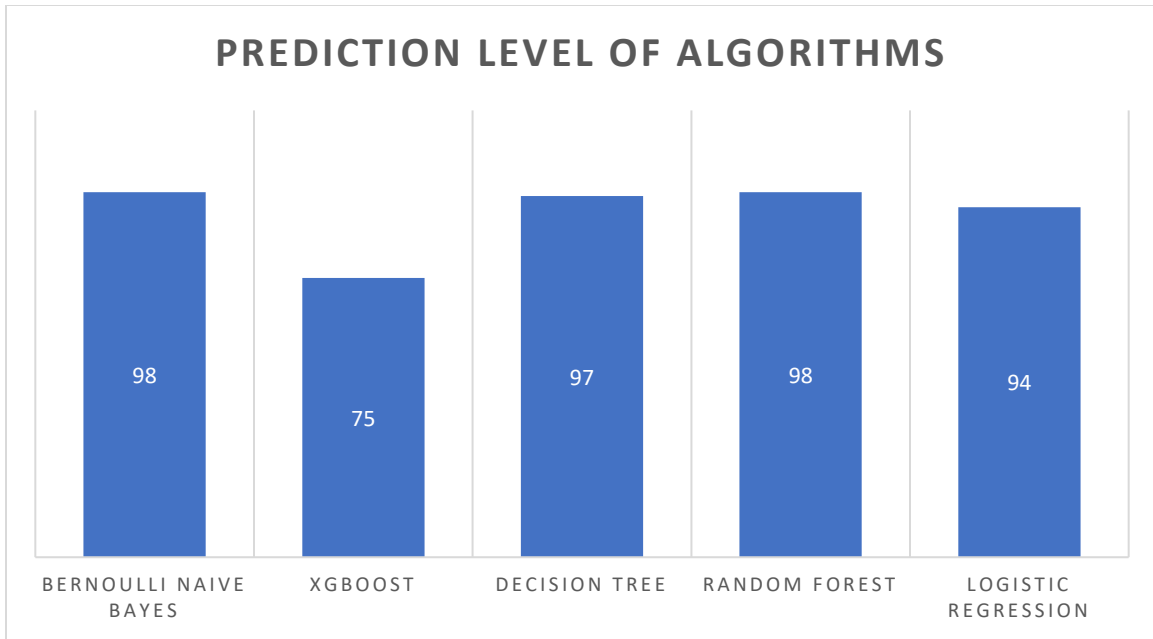


Figure 4.2.11 – Prediction level of Algorithms

Table 4.2.7 - Comparison of results

Algorithms	Accuracy
Bernoulli Naive Bayes	98%
Decision Tree Regression	97%
Random Forest Regression	98%
Logistic Regression	94%
XGBoost Regression	75%

As can be seen from the data above, the Bernoulli Naive Bayes method provides the best accuracy result, with a score of nearly 98%. Accordingly, we employ this technique to create our model.

4.3 Discussion

After pre-processing the data set, we build our model and train the data on this model. After that, we got our expected result. In this segment, we made an analysis of the expected results that we test and analyze. Bernoulli Naïve Bayes, Decision Tree Regression, Random Forest Regression, Logistic Regression, and XGBoost Regression algorithms are applied to test data. Later than we compare the result, accuracy, precision, and error values. Here we can see from the following table 4.2.6 that the accuracy percentage of Bernoulli Naïve Bayes is high compared to other algorithms which is 98.8%. In contrast, random forest regression gives somewhat similar accuracy but with some point differences. The random forest algorithm has shown 98% accuracy. XGBoost gives lower accuracy than other algorithms which is 75%. Various other machine learning algorithms can be used to increase the accuracy rate like KNN and Neural Network. Logistic regression gives us an accuracy of 94.04%. The medium accuracy results are seen in Decision Tree Regression which is 97% respectively. If we compare these results, we can decide which algorithm we should use to build our model. In the end, we can say that the Bernoulli Naïve Bayes algorithm gives the best result.

CHAPTER 5

Impact on Society, Environment and Sustainability

5.1 Impact of Society:

It is possible to view dropout of students as a curse for education society. This sickness has raised as a major problem for many students throughout the past few years over the whole country. One of the most delicate elements to build society is education. Basically, dropout of students causes due to the weakness of educational infrastructure and lack of proper echeloned. Since there is currently no treatment for this disease; so, many students getting drops away from proper institutional education every year. Preventing it is the answer, then. However, the work might be exceedingly challenging at times.

The rate of dropouts will be much reduced if we can quickly identify the cause of this crisis scenario. We will do it with our recommended effort. The long-term economic effects of dropping out of education have been worse and are still getting worse. Youth dropping out of university without graduating have had a detrimental impact on their social and economic life. The economic insignificance of university dropouts has gotten worse in recent years as a result of technological advancements that have increased the need for a highly educated labor force and made a university education a prerequisite for employment.

Those who have completed their education can obtain employment in business and other industrial fields, where they will be adequately compensated later on for their daily necessities to the degree that they can save them.

Additionally, obtaining a university degree helps people develop fundamental abilities that allow them to join the military or police or be admitted to a more advanced level of schooling. A university degree, for example, is emphasized as being important in the Human Capital Theory because it affects "economic success in life."

5.2 Impact on Environment:

Around Bangladesh, more than hundred-thousands of students is getting dropout in each year. Thus, the whole nation is going under darkness only because of fooleries. When they far away from proper educational knowledge, they of course can't know how to utilize every single thing in a proper way. As they have lack of experimental education knowledge so, this can drive them to waste of wealth instead of proper utilization. According to the modern world, Bangladesh is also doing the finest progress. It requires more educational knowledge and certificates to stay equally to the advance technology and others implementation sites. But dropout problem is making crisis situation here in this perspective. Except acquiring proper institutional knowledge our domestic technology site cannot be developed and we can get fail to save ourselves from different aspects as, invention of new technology, military sophisticated technology etc. Developing different types of technology like radar system for both weather and military, vehicle design and prototype, land cultivating tech etc. can lead our daily life so easily. So, unless we are acquiring knowledge this cannot be done.

Impact on environment due to dropout of students can create a strong change. On the contrary, when an individual student get dropout from institutional education, they become depressed due to fall from a stable educational environment.

5.3 Ethical Aspects:

Our initiative aims to make students dropout predict in more effective way and solve this problem or reduce the percentage of dropout quicker in the educational sector. More and more students will benefit as a consequence, ultimately saving their carrier. Because the likelihood of getting benefits will rise the earlier the problem is discovered. Additionally, it will benefit the student's and their family psychological and economic conditions. The sooner the student get recover that crisis situation, the less risk will arise to their carrier, and the quicker the student may resume back to study. Every student will be able to receive precise results to our project. Testing errors can occasionally have severe results. Since improper care might result in several consequences. It

sometimes also causes a number of unexpected situations. From it, financial and very serious psychological issues may develop. We utilized public domain datasets to finish our study. It was taken from Kaggle. And the information contained in the datasets was likewise gathered with permission.

Additionally, we intend to gather real-time data from educational institutions like university. These figures are going to be more precise. The precision of our study will rise when more precise datasets are used. Because of this, we can provide students better methods for spotting dropout, which will benefit them more. As a result, everything of our labor is morally acceptable.

5.4 Sustainability plan:

Deep learning algorithms are being used in our study to predict dropout students. Making use of Deep Learning technology Our study can assist educational researcher and teacher to detect and predict dropout students frequently. Our study focuses on both the detection and prediction of dropouts. Our research will enhance the process of identifying them which starts at early stage from raw data set employing several types of algorithms. Our study will enable the education sector to distinguish between students with different stages of dropout occurrence with ease. If the education sector, education researcher, teacher is aware of the dropout stage, they can treat it effectively. In order to sustain this research, we will gather additional information from other databases. We'll visit the various university as well to speak with the teacher, students and principles. Knowing that we will have a better chance of obtaining accurate information if we can speak with them directly.

CHAPTER 6

Summary, Conclusion, Recommendation and Implication for Future

6.1 Summary of the Study

During covid, study drop was a very contemporary issue, in this region of the globe. Only based on the sheer wave of the pandemic, many reports on study drops were on the verge of getting unnoted. As it was an epidemic, consequences were generalized to all individuals within the economic framework. However, on the contrary, we've seen a very low percentile of drop rates, in the actual number of student drops. As there wasn't any existing solid research on the topic on that timeline, we had to gather the information through fieldwork, only to get percentile results by machine learning, accurately wise-speaking. which was a 28.5% drop with the ninety-eight percentiles of accuracy. Various machine learning algorithms are used such as Bernoulli Naive Bayes, Decision Tree Regression, Random Forest Regression, Logistic Regression, and XGBoost Regression. Among them, Bernoulli Naive Bayes gives the best accuracy results which are ninety-eight percentiles.

Not to mention, the percentile of accuracy, is a great achievement in itself. However, fetching the data won't resolve the issue. Which will solve the issue, and is structured through pre-determined steps by the authority. If authorities consider the research, they could have provided and come up with viable financial solutions for those students, who are in need. This will resolve the issues of a study break, which usually leads one to a very uncertain financial domain and career, especially in this capitalistic society, therefore, more burden on per capita. To avoid these sorts of general effects, this sort of research will work as an excellent choice of resources.

6.2 Conclusions

During the last two years, we have seen a very radical shift in the rates of student dropping, as one of the direct by-products of Covid-19 in Bangladesh. Fortunately, we were able to capture this phenomenon in the form of data using machine learning. In this research, we train the dataset on different algorithms and compare the accuracy results to get the best accuracy results. We used five different algorithms to predict dropout based on family income, specific reasons, and CGPA

results. Bernoulli Naive Bayes algorithm has higher accuracy percentage, which is 98%. Random Forest Regression algorithm also gives the best accuracy. Afterward, the decision tree regression algorithm showed an accuracy of 97%. XGBoost Regression algorithm provides the lowest result which is 75%. We choose a household income, drop-out reasons, and drop-semester or not characteristics to predict. If we compare the model of other research like SVM, KNN, and our model, our model shows 2 or 3 percent more accuracy, Therefore, based on these factors, for prediction, our proposed model was used.

6.3 Implication for Further Study

To enhance the quality of prediction, in near future, we will try to gather similar sorts of research. An enriched dataset will definitely be able to help us to get more precise accuracy. Data from both schools and colleges will be used in our research for better accuracy, in the future. We only make predictions with the data of university students. In the future, predictions can be made with the data of school and college students. In a way to achieve the best accuracy result, we can use neural network algorithms, large datasets, and other techniques. we can increase our accuracy results by collecting more data, adding more variables, and choosing better features.

References:

- [1] Francesca Del Bonifro, Maurizio Gabbrielli, Giuseppe Lisanti and Stefano Pio Zingaro, "Student Dropout Prediction" AIED 2020: Artificial Intelligence in Education, vol.12163, pp. 129-140, 30 June 2020.
- [2] Jovial Niyogisubizo, Lyuchao Liao, Eric Nziyumva, Evariste Murwanashyaka, Pierre Claver Nshimyumukiza, "Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization", vol.3, 2022.
- [3] Fisnik Dalipi, Ali Shariq Imran, Zenun Kastrati, "MOOC Dropout Prediction Using Machine Learning Techniques: Review and Research Challenges", DOI:10.1109/EDUCON.2018.8363340. April 2018.
- [4] Carlos Márquez, Alberto Cano, Cristóbal Romero, "Early Dropout Prediction using Data Mining: A Case Study with High School Students", DOI:10.1111/exsy.12135, February 2016.
- [5] Nicolae-Bogdan Şara, Rasmus Halland, C. Igel, Stephen Alstrup, "High-School Dropout Prediction Using Machine Learning: A Danish Large-scale Study.", ID: 215749827, 2015.
- [6] V.Hegde, P.P.Prageeth, "Higher education student dropout prediction and analysis through educational data mining", DOI:10.1109/ICISC.2018.8398887, ID: 49540813, 2018.
- [7] Neema Mduma, Khamisi Kalegele, Dina Machuve, "A Survey of Machine Learning Approaches and Techniques for Student Dropout Prediction", Data Science Journal, 17 April 2019.
- [8] Ali Shariq Imran, "MOOC Dropout Prediction Using Machine Learning Techniques: Review and Research Challenges", IEEE Global Engineering Education Conference (EDUCON), 2018.
- [9] Lee, L., Martínez, S., Castán Rocha, J., Terán Villanueva, J., Menchaca, J., Treviño Berrones, M. and Rocha, E. (2020) Evaluation of Prediction Algorithms in the Student Dropout Problem. Journal of Computer and Communications, 8, 20-27. doi: 10.4236/jcc.2020.83002.
- [10] Luis Earving Lee¹, Salvador Ibarra Martínez¹, José Antonio Castán Rocha¹, Jesús David Terán Villanueva¹, Julio Laria Menchaca¹, Mayra Guadalupe Treviño Berrones¹, Emilio Castán Rocha², "Evaluation of Prediction Algorithms in the Student Dropout Problem", DOI: 10.4236/jcc.2020.83002, Vol.8 No.3, 2020.
- [11] Meseret Yihun Amare, Stanislava Simonova, "Global challenges of student's dropout: A prediction model development using machine learning algorithms on higher education datasets", pp. 1-10, 2021.
- [12] João Gabriel Corrêa Krüger, Alceu Britto, Jean Paul Barddal, "An Explainable Machine Learning Approach for Student Dropout Prediction", pp. 1-19, 19 October 2022.

- [13] Jay S.Gil, Allemar Jhone P.Delima2, Ramcis N.Vilchez, "Predicting Students' Dropout Indicators in Public School using Data Mining Approaches", vol. 9, January 2020.
- [14] Hee Sun Park, Seong Joon Yoo, "Early Dropout Prediction in Online Learning of University using Machine Learning", vol. 5, 2021.
- [15] Ahmed Alamri, Mohammad Alshehri, Alexandra Cristea, Filipe D.Pereira, Elaine Oliveira, Lei Shi, Craig Stewart, "Predicting MOOCs Dropout Using only two easily obtainable Features from the First Week's Activities", vol. 1-10, 2020.
- [16] Haarika Dasi, Srinivas Kanakala, " Student Dropout Prediction Using Machine Learning Techniques", vol. 10, 2022.
- [17] Xingqiu Tang, Hao Zhang, Ni Zhang, Huan Yan, Fangfang Tang, Wei Zhang, "Dropout Rate Prediction of Massive Open Online Courses Based on Convolutional Neural Networks and Long Short-Term Memory Network", vol. 2022, Article ID 8255965, pp. 1-11, <https://doi.org/10.1155/2022/8255965> 16 may 2022
- [18] Marina Segura, Jorge Mello, Adolfo Hernández, "Machine Learning Prediction of University Student Dropout: Does Preference Play a Key Role?", <<<https://doi.org/10.3390/math10183359>>>
- [19] Sunny Behal, Gautam Sallan, "prediction of student dropout using enhanced machine learning algorithm", DOI:10.37418/amsj.9.6.61, May 2020.
- [20] Khalid Oqaidi, Sarah Aouhassi, Khalifa Mansouri, "Towards a Students' Dropout Prediction Model in Higher Education Institutions Using Machine Learning Algorithms", DOI:10.3991/ijet.v17i18.25567, September 2022.
- [21] Marcell Nagy, Roland Molontay, "Predicting Dropout in Higher Education based on Secondary School Performance", vol. 1-6.

ORIGINALITY REPORT

25%

SIMILARITY INDEX

23%

INTERNET SOURCES

9%

PUBLICATIONS

16%

STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	7%
2	Submitted to Daffodil International University Student Paper	3%
3	Submitted to Manchester Metropolitan University Student Paper	2%
4	www.springjournals.net Internet Source	1%
5	meral.edu.mm Internet Source	1%
6	scholarworks.aub.edu.lb Internet Source	1%
7	www.hindawi.com Internet Source	1%
8	m.scirp.org Internet Source	1%
9	Submitted to University of Sheffield Student Paper	<1%