SCHOOL DROPOUT PREDICTION OF BANGLADESHI STUDENTS DUE TO COVID-19

BY

Safiqur Rahman Sakkhar

ID: 213-25-042

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

Supervised By

Professor Dr. Md. Fokhray Hossain

Professor

Department of CSE

Faculty of Science and Information Technology

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
January 2023

APPROVAL

This Thesis titled **"School Dropout Prediction of Bangladeshi Students Due to Covid-19"**, submitted by Safiqur Rahman Sakkhar, ID No: 213-25-042 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 M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 17-01-2023.

BOARD OF EXAMINERS

17/01/23

Dr. Touhid Bhuiyan, PhD — **Chairman**
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Ms. Nazmun Nessa Moon — **Internal Examiner**
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Fizar Ahmed — **Internal Examiner**
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Safaet

Md. Safaet Hossain — **External Examiner**
Associate Professor & Head
Department of Computer Science and Engineering
City University

DECLARATION

We hereby declare that, this thesis has been done by us under the supervision of **Professor Dr. Md. Fokhray Hossain, 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:

Hossain

Professor Dr. Md. Fokhray Hossain
Professor, Department of CSE
Daffodil International University

Submitted by:

Sakkhar

Safiqur Rahman Sakkhar
ID: 213-25-042
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty **ALLAH** for His divine blessing makes us possible to complete the thesis successfully.

We really grateful and wish our profound our indebtedness to **Professor Dr. Md. Fokhray Hossain, Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Data Mining*" to carry out this thesis. His endless patience, 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 **Professor Dr. Touhid Bhuiyan**, Head**,** Department of CSE, for his kind help to finish my thesis 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

In year 2020 the whole world faced an unexpected crisis which we known as COVID-19 pandemic. After pandemic declaration by WHO (World Health Organization), every nation around the world started locked down their nations and their communications to other nations to prevent the outbreak of COVID-19. Many sectors hampered by this pandemic, as well as education sector, especially under develop country like Bangladesh are facing a huge loss in education sector because of COVID-19 pandemic. All educational institutes started normalizing their schedules after two years of lockdown. And this made a huge study gap to our students of Bangladesh, especially primary and secondary level students are facing enormous problems to continuing their studies. Primary level students forgot how to spell, how to pronounce how to read and how to write. Some students got addicted with smartphones and online games and some students dropped out from school for various reasons. Some of them dropped out for the financial condition of their family, some of them dropped out because they lost interest in study, some of them got married and some of them started working so they could contribute to their family to get rid of poverty. Peoples who were living their life under poverty or who had stable financial condition before the pandemic but after pandemic financially they are facing loss or became broke have more chances to drop out from school. By being so close of a secondary school it's motivated me to develop a Machine learning model by using machine learning algorithm by seeing vast amount of dropout rate of school students after COVID-19. In this research I applied different machine learning algorithms such as, linear regression, Decision tree, SVM (Support Vector Machine), Random Forest, Naïve Bayes. But from all of them Random Forest got the highest accuracy of 87%. The goal of this research is mining significant facts of being dropout from school, and to predict is any students will be dropout or not. The proposed model was built on Google Colab (python-based ide) and trained on secondary data which was collected through students from different secondary level school. The dataset contains 300 data collected from students with 9 attribute of student data.

TABLE OF CONTENTS

LIST OF FIGURES

.

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Background of the research

The impact of Covid-19 pandemic in whole world is immense, especially developing country like Bangladesh had enormous impact after Covid-19. It took almost 2 years to get back into regular life for Bangladeshi peoples, economically peoples suffered a lot. Middle class families suffered the most. Covid-19 impacted in all sectors in Bangladesh but in education sector it affected a lot. All educational institution opened after almost 2 years, students who was studying in class 6 suddenly they went to school and started studying class 8, there 2 years of education gap took place. Many of the students forgot how to spell, many of the student forgot to read properly. Because of the pandemic almost 80% of the students practiced a bad habit of not studying on regular basis. Conforming to the Household Income Expenditure Survey, 8.4 million student families live in poverty [1]. In the period of pandemic, their income has fallen, they started to think that how could they survive. Many of them overcomes their poverty and many of them couldn't. Who couldn't overcome their poverty simply they stopped their children's education and send them to earn and help their parents to survive? Also, many children lost their parents during these pandemics, many of them don't have any guardian who will bare their educational expenses or send them to the schools. And many low graded students already had aversion on school or on study so they stopped going to schools. There are many possible ways to children's for not going to school. For this reason, I want to predict that how much students dropped out from school because of covid-19 pandemic and what is the reason behind them to dropout school. I want to predict this using secondary data collected from various schools and students who have been dropped out their education. For this kind of research many authors used SVM, Random Forest and Decision tree. I will also predict my desired problem using these machine learning techniques.

1.2 Problem Statement

After Covid-19 lockdown, school student dropout or post lockdown phase of education sector of Bangladesh is facing a huge problem. Losses in education sector we had is irreplaceable. Total 16 million (approximately) students' family fall below poverty line after 3 months of lockdown [1]. Including higher secondary and university students' family. This number must

Be increased more by long term lockdown. If we couldn’t take proper action to solve this problem, our future will face a huge unemployment rate, lack of skilled professionals, beside of all we will face a major economic disaster. The aim of this study is to predict the school student dropout reason after covid-19 pandemic using machine learning for accurate prediction. So that we could identify major reasons behind school student dropout after covid-19 pandemic.

1.3 Aim of the Research

Year 2020 and 2021, education sector of Bangladesh faced a great loss because of Covid-19 pandemic. Many families who were living their life under poverty, they halted their children’s from sending to school. Children, youth, and adults all have the right to education in times of catastrophe, and this right must be prioritized from the outset of all emergency operations [22]. Government of Peoples republic of Bangladesh made primary education free for future development of Bangladesh. That all classes’ people of Bangladesh could get proper education and contribute to the society to develop our country. But most of the peoples of developing country like Bangladesh, who lives under poverty, after Covid-19 breakdown many parents send their children’s to earn so they could contribute to their family to survive, also many children’s lost their father or mother due to covid-19 outbreak, who was the only person to earn for their family, and also many secondary school going girls got married because their parents couldn’t bare more expenses for living under poverty and many more reasons. The objective of my research is to predict the reasons behind student’s school dropout and the rate of school dropout because of Covid-19, by using Machine Learning techniques. There are some similar works has been done before. Such as, predicting school dropout, University dropout etc. in different countries including Bangladesh, using Support Vector Machine, Random Forrest, Decision tree by Machine Learning techniques, before Covid-19 pandemic. My Aim is to collect student data at secondary level and train and test those collected data via Machine Learning techniques to predict with higher accuracy level. So that we could find the main reasons behind student’s school dropout and helps to decrease school dropout rate and contribute a little to make Bangladesh a prosperous country like other developed countries.

1.4 Research Methodology

As school student dropout after Covid-19 lockdown is the most critical problem for upcoming days of education sector of Bangladesh, this topic was selected for research. We will build a prediction model using Machine Learning techniques, to predict dropout and I will find out the most crucial factor for school student dropout after Covid-19 lockdown. Data will be collected from students who already dropped out school and also from some schools and I will use these data to predict the major reasons behind school students drop out after Covid-19 lockdown.

1.5 Proposed Solution

After building the desired prediction model and after successfully train and test collected data through that prediction model, we would be able to find out the most crucial reasons behind this irreplaceable loss. Organizations who are having authority or the government or nongovernmental organizations who are having ability to take action for make a change against this problem could take help from my study, to decrease school student dropout. My study will give the accuracy and assurity to believe the crucial factors behind school student dropout after Covid-19.

1.5.1 School Dropout

A person who has stopped attending school and does not have a high school diploma or its equivalent is referred to as a "school dropout" [23]. Dropping out means leaving school, college or university before graduation. School dropout means leaving school before completion the graduation or before leaving school without getting any certification from education board or equivalent authority. There are many reasons behind school dropout. Financial crisis, family problem, parent’s illiteracy problems, early earning issue, child marriage, bad influence, drugs, depression, mental illness etc. many researchers studied and found enormous amount of reason behind dropout but those facts are depending on level of student. This is a serious issue regarding to development of a nation. School dropout doesn’t affect instantly. It took time to shows its outcome.

1.5.2 Machine Learning

Machine Learning techniques are common techniques used for prediction and analysis using hardware and software [24]. There are many researchers already studied enormous number of topics related to dropout using machine learning techniques. Some of them are related to school dropout from different region, some of them are related to college students and some of them are related to university students even on specific course or subject. Machine learning is a process of using algorithms proven mathematically and scientifically supported statistical models that allow computer-based hardware or software systems accomplish a task by relying on models and predictions rather than particular instructions. It could be classified as a kind of artificial intelligence. On the other hand, rather than being precisely programmed to carry out the task, machine learning algorithms construct a mathematical pattern based on sample data, also termed as "training data", in order to formulate guesses or assessments.

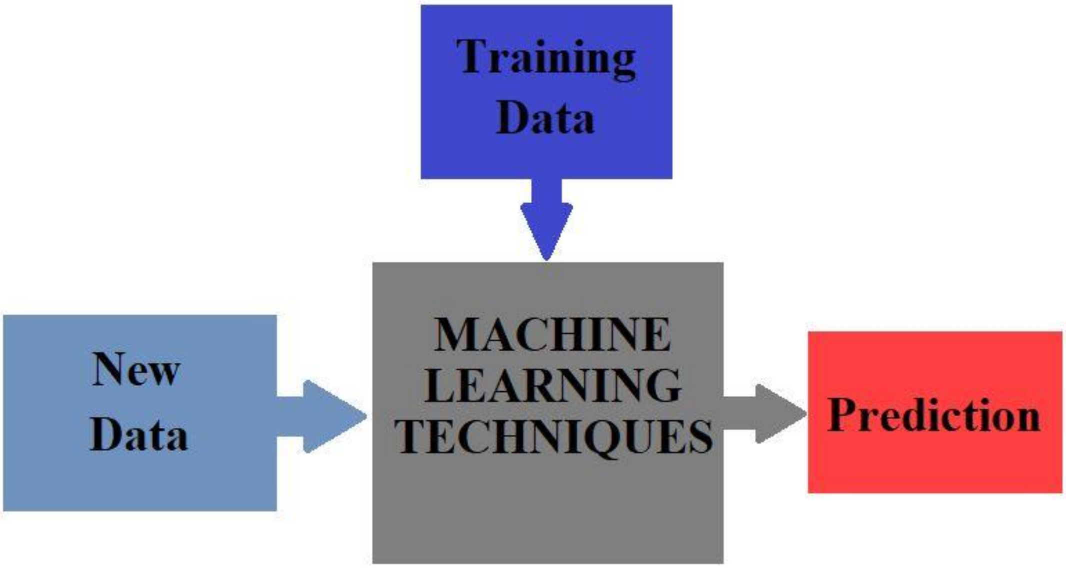


Figure 1.5.2.1: Working flow of machine learning.

By exposing computers to a variety of examples, data sets, and structures that help in the development of their own logic, the subject of artificial intelligence referred to as machine learning teaches computers to learn on their own. To teach computers there are no exact software or system in needed. Two types of learning techniques are used to train machines. One is supervised learning and another is unsupervised learning. Supervised learning is if a mother told to her child that don't play outside while it's raining otherwise you will get ill, then the child learned it from his mother that not to play outside while it's raining. Another scenario is unsupervised learning. In this scene the child goes to play outside while it's raining, get ill and learn that he shouldn't play outside while it's raining, otherwise he will get ill.

1.6 Conclusion

In my research Random Forest got the highest accuracy of 87%. The main goal of my research is to analysis the significant fact behind school dropout after COVID-19 Pandemic and to build a model to predict school student dropout after COVID-19 pandemic by applying different machine learning algorithms.

CHAPTER 2

Literature Review

2.1 Introduction

In data mining sector many researchers use many machine learning algorithms to predict dropout, which is creating a new study zone to develop farther. As a modern-day problem school dropout is a huge unseen problem for the upcoming future. There had been many reasons behind dropping out but in recent we faced a pandemic which causes more financial, mental loss to dropout school either they had their consent or not. There are many researchers already studied various type of dropout cases for enormous number of reasons. This is a modern-day issue which is mostly impacting the young and upcoming generations, which is also increasing the unemployment issue. To analyze this problem and for build a model to predict dropout after COVID-19, there are some studies which encouraged me, inspired me a lot to study further. Why students are dropping out from their study and to predict whether any student are willing to dropping out made many researchers to study on dropout.

2.2 Literature Review

Using Genetic Programming (GP), C. Márquez-Vera et al. [2] developed a model by several experiment to predict dropout within 4-6 weeks at different steps to compare their proposed algorithms with common classic machine learning algorithms to find the early indicators of high school dropout. They compared many traditional unbalanced famed machine learning algorithms with their proposed model.

F. Del Bonifro et al. [3] studied dropout by using properly considered a specific type of customary classification algorithms, Linear Discriminant Algorithm(LDA), Support Vector Machine(SVM) and Random Forest(RM). They studied first year under graduate student dropout to find early indicator of dropout so they could build a tool to predict earlier for decrease economic and social cost.

To find a viable solution for put a stop in dropout from E-learning, M. Tan and P. Shao et al. [4] developed a prediction model with large scale (62,375) of data using some famed classification algorithms, Artificial Neural Network (ANN), Decision Tree (DT) and Bayesians Networks(BNs) etc. The aim of the study was to recognize probable dropouts. By the ranking Decision tree (DT) got the higher accuracy of 71.91%.

Some students of Computer science (CS) of Chittagong University also studied dropout prediction. Ahmed, S. A. et al. [5] developed model where they manually created the rules and predict Computer science dropout and students future prospective. And Neural Network shown the highest accuracy of 84.2%. Students who are willing to get admitted in CS on their Under Graduation this study will be very helpful for them to took decision wisely.

A student of Shahjalal University of Science and Technology studied female student attrition from school. Hasan, M. N. et al. [6] compare between Logistic Regression and Linear Discriminant Analysis to predict female student attrition from school. Nowadays female student doing well in education and working sectors, but rate of drop out of female student is not reducing so he develop a model to predict and compare between two machine learning algorithm, LDA and LR. Both algorithms had almost similar accuracy, between them LDA got the higher accuracy of 78.38%.

A study shows that, C4.5 algorithm is better to classify that which characteristic is similar for dropout. A. G. Pertiwi et al. [7] studied dropout in Indonesia using C4.5 algorithm which is a chosen method to predict and analysis dropout. To reduce dropout rate he studied dropout in Indonesia using Decision tree (C4.5) algorithm and got the accuracy of 71.2%.

Research shows that participants who are at risk of dropout from online course could be predict using machine learning model. R. Bukralia, et al. [8] to address the issues behind the dropout from online course, they develop a model to predict student at risk of dropout from online course. They use Linear Regression, Neural Network and Support Vector machine to build their desired model. MLP and SVM got the highest accuracy of 75% to predict dropout in their study.

A study had been done to early predict school dropout. R. S. Baker, et al. [9] used Logistic Regression and Decision Tree algorithm to develop a model for early prediction of school dropout using machine learning techniques. They wanted to build a model for automatic detect whether a student are willing to dropout now or in future using machine learning algorithm, and the found the precision of 75% using Logistic Regression.

2.3 Conclusion

In this chapter some previous researches have been studied and discussed with their model's accuracy. Among them, Ahmed, S. A. et al. [5]. Got the highest accuracy of 84.2%. in the related works authors used various machine learning algorithms to study their desired outcomes. We tried some different prediction models to get better accuracy and better prediction, which will be discussed in further discussions.

CHAPTER 3

Theoretical Model

3.1 Introduction

In this chapter all machine learning algorithms which are used in this research will be described theoretically, with their working procedures. The selected algorithms are classical machine learning techniques which are used for prediction in general. Machine leaning is such technique which is use for make our machines to improve automatically through experience [10]. In this research, five classification algorithms have been used according to previous performances. They are Logistic Regression, Random Forest Classifier, Support Vector Machine (SVM), Decision Tree and Gaussian Naïve Bayes algorithm. Each and every algorithm will be described briefly below and also their working procedures will be discussed theoretically.

3.2 Classification Algorithm

A supervised machine learning technique which we known as the Classification Technique, which uses training data to determine the categorization of new findings. A program in Classification gets to know from a given dataset or assumptions and then categorizes new observations into one of several classes or groups. The goal of classification algorithm is to predict the output for the categorical data which were given to the machine for observation [11].

3.2.1 Logistic Regression

Logistic regression is a popular Machine Learning algorithm that belongs to the Supervised Learning technique. It is used to forecast the categorical dependent variable from a set of independent variables. Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems [19]. A categorical dependent variable's output is predicted using logistic regression. As a result, the outcome must be categorical. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

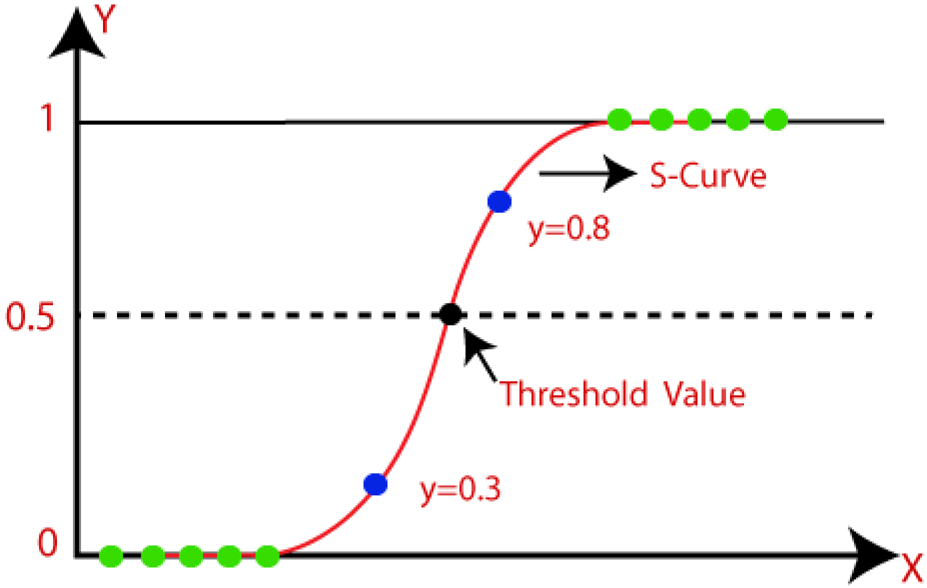


Figure 3.2.1.1: Logistic Regression [19].

This is how it would look if we represent the model visually. Instead of fitting a regression line, we fit a "S" shaped logistic function that predicts two maximum values in logistic regression (0 or 1). The concept of the threshold value is used in logistic regression to define the probability of either 0 or 1. For example, values above the threshold value tend to be 1, while values below the threshold value tend to be 0.

3.2.2 Support Vector Machine (SVM)

Support Vector Machine or SVM is a supervised machine learning algorithm. A supervised machine learning approach called "Support Vector Machine" (SVM) can be applied to classification and regression problems [25]. But most commonly it is used in classification challenges. This algorithm creates hyper-plane and differentiate the classes from the dataset. Here hyper-plane is wall of differentiation for the classes. As example,

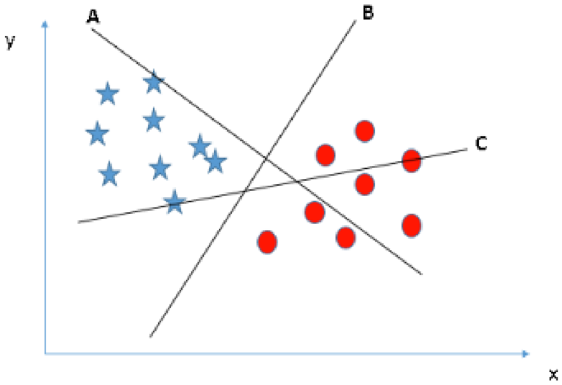


Figure 3.2.2.1: Support Vector Machine [20]

In the above image we can see two different classes one is indicated as star and another one indicated as red dot, and also 3 hyper-plane A, B and C is situated in the features. The goal is to choose the right hyper-plane to dissociate two classes. Using the thumb rule SVM recognize the right hyper-plane which divided the two classes better. A huge plus is to use SVM is that, SVM is robust to the outliers. Which means this algorithm ignore the outliers.

3.2.3 Random Forest

In supervised machine learning techniques, there belongs another algorithm which we known as Random Forest. Both classification and regression challenges could be solved using this algorithm. But it is much suitable for classification challenges. Supervised machine learning algorithms like random forest are frequently employed in classification and regression issues. It creates decision trees from various samples, using their average in the case of regression and majority vote for classification [26]. This algorithm uses ensemble learning technique to dissociate classes. Ensemble learning technique is used to solve more convoluted problem, where various classifiers are used for majority voting to choose the class.

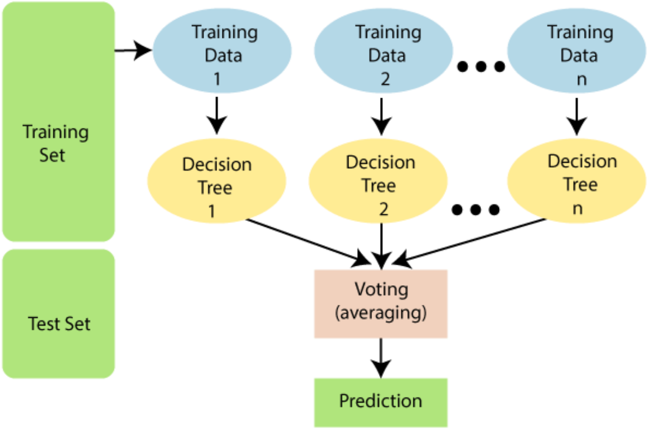


Figure 3.2.3.1: Random Forest [21]

In the above image we can see the working procedure of Random Forest algorithm. Where training data is going through multiple decision tree classifier, and then averaging the vote. In average voting it chooses majority voting for prediction.

Random Forest algorithm has many pros in terms of other algorithms. It takes less time than other algorithms for training and predict with higher accuracy. This algorithm also works with lager datasets efficiently and also with the vast quantity of missing data.

3.2.4 Decision Tree

Decision Tree is a supervised machine learning technique, which is preferred to solve classification problems, but that doesn't mean that this algorithm is only preferred for classification challenges. Also, regression challenges could be solved by this algorithm but often this algorithm is use to solve classification challenges. A decision tree is a supervised machine learning tool that may be used to classify or forecast data based on how queries from the past have been answered. The model is supervised learning in the sense that it is trained on a collection of data containing the desired category and then tested on that data [27]. This algorithm is a tree assemble algorithm. Where features of the dataset are knows as internal node, decision rules are act as branches and the outcome is regarded by the leaf node. To understand it clearly follow the image given below,

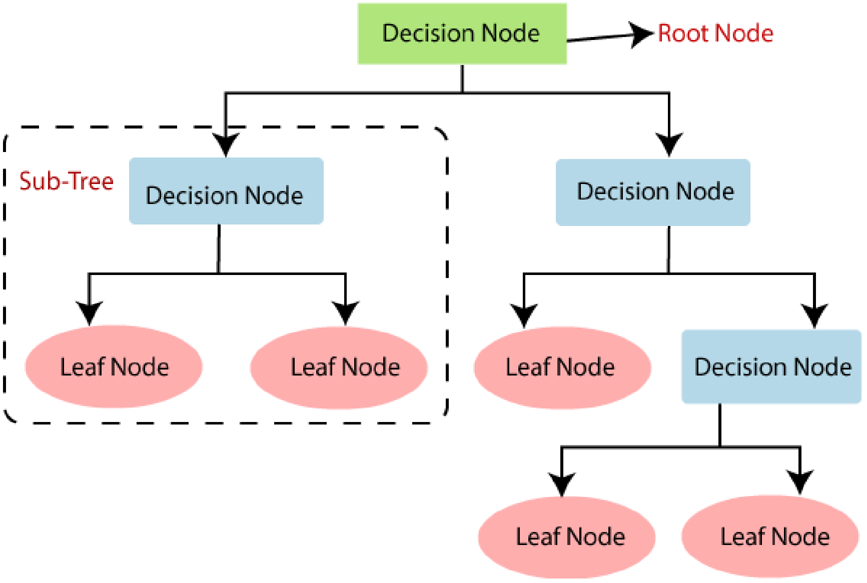


Figure 3.2.4.1: Decision Tree Classifier [29]

This algorithm is very easy to understand, this algorithm took decision as humans used to think while taking any decision. All possible outcomes could be tested by this algorithm. Compared to other algorithms this algorithm needs less data cleaning.

3.2.5 Naïve Bayes

Naïve Bayes algorithm is a supervised machine learning technique to predict classification challenges using Bayes theorem. A group of classification algorithms built on the Bayes' Theorem are known as naive Bayes classifiers [28]. This classification algorithm is highly preferred for high dimensional training data. This algorithm is the most effective and simple classification algorithm which predict quicker than other classification algorithm. This algorithm is a probabilistic algorithm, mostly preferred for probability problems. This algorithm is dependent on Bayes theorem,

$$\boldsymbol{P(A|B) = \frac{P(B|A)P(A)}{P(B)}}$$

This is the formula of Bayes theorem. Where, P can be mentioned as probability and A, B is showing as example there should be attributes of the features from the dataset. The working procedure of Naïve Bayes algorithm is first it will transform the dataset into frequency table, Produce the possibility by searching the chances of given features and then using Bayes theorem calculate the possibility.

3.3 Conclusion

In this entire chapter all machine learning algorithms are discussed which have been used in this research. All the algorithms are chosen for this research because these algorithms are mostly preferred for classification problems and in different prediction related studies these algorithms performed very good accuracy.

CHAPTER 4

Experimental Model

4.1 Introduction

In this chapter we will able to know all about the data, which have been used in this research. The Method which has been used to complete this study will be introduced in this chapter with data transformation, data preprocessing, data visualization and many more. As per the title of this research no secondary data were available in any source, because this research is based on very recent issue. So, this research must be done by using primary data. Which was collected by the researcher with proper assistance of the supervisor. Some related work was discussed previously. Where researchers worked with different data sets, their issues were different and their situation were different. But this research is for finding out the most effective reason behind getting dropped out after the recent pandemic and also to predict dropout after Covid-19 pandemic. So, only specific questionnaires were asked to collect data for completion of the study.

4.2 Dataset

The dataset which has been used in this research is a primary dataset. As we discussed previously that the research is based in most recent issue of Bangladesh, so there is no data in any source related to this study. For this research primary 300 categorical data collected via 3 secondary school students, who knows anyone who is dropped out from school or maybe he or she will be dropout in future. It is quite difficult to find and collect data from the students who dropped out from school. So, students who know anyone who dropped out or anyone who maybe want to dropped out, they know those students some basic personal information for sure. Which will help to complete this research. Records of the dataset contains 9 attributes where 8 attribute will help to recognize the desired output and all attribute contains categorical data. The attributes are as follow, (Dropout, Student quality, Gender, Financial condition before covid-19, Present financial condition, Parents education, Parents died by covid-19, Got married, started working etc.). This categorical data has been collected from the student by using printed data collection form which was written in Bangla. So that student can easily understand the questionaries of data collection form. The sample image of printed data collection form is attached in the appendix with the acknowledgment of the Head Teachers of those school, from where the data has been collected.

Table 4.2.1: Overview of School Student Dropout after Covid-19 Dataset

Attribute	Type	Description
STD_Q	Discrete	Student quality 1: Very Bad, 2: Bad, 3: Average, 4: Good, 5: Very Good
Gender	Discrete	1: Male, 2: Female
F_C_PRSNT	Discrete	Present financial condition of the student family 1: Very Poor, 2: Poor, 3: Middle Class, 4: Rich, 5: Very Rich
F_C_B_Cvd	Discrete	Financial condition of the student family before Covid-19 1: Very Poor, 2: Poor, 3: Middle Class, 4: Rich, 5: Very Rich
PARENTS_EDU	Discrete	Parents are educated or not
PRNTS_DIED_Cvd	Discrete	Are anyone of the parents died by Covid-19 or not
GOT_MARID	Discrete	is he/she married or not
STRT_WRK	Discrete	Did he/she started working to support their family
DRPOT	Discrete	Are he/she dropped out or maybe wanted to be drop out

4.3 Data Transformation

The collected primary data contains 9 attributes where 1 is output data and other 8 attribute are for recognize the output; all the attribute contains discrete data. The collected data was in form of string type as example if the question is about gender option was male or female. Though all the data was collected through hard copy of data collection form, we had to manually create the data set. All the data was categorical but all data was in form of string we had to convert it to numerical data otherwise the research won’t be completed, because machine only can read numbers to predict and also machine can only give an output via number.

4.3.1 Categorical Transformation of Data

In this research primary data used for the model, as we know that all data was collected through hard copy of data collection form. So, all data was string type. We had to manually create the data set by transforming string type data to categorical data. All the data from 9 attributes are discrete data no continuous data has been used in the data set, so categorical transformation of the data is the best option to prepare the dataset to study. Categorical transformations of all attributes are shown below,

Table 4.3.1.1: Categorical transformation of Student Quality

Column name	**STD_Q**				
Status	Very Bad	Bad	Average	Good	Very Good
Categorical value	1	2	3	4	5

Table 4.3.1.2: Categorical transformation of Gender

Column name	**Gender**	
Status	Male	Female
Categorical value	0	1

Table 4.3.1.3: Categorical transformation of present financial condition of student family

Column name	**F_C_PRSNT**				
Status	Very Poor	Poor	Middle Class	Rich	Very Rich
Categorical value	1	2	3	4	5

Table 4.3.1.4: Categorical transformation of financial condition before Covid-19 of student family

Column name	**F_C_B_Cvd**				
Status	Very Poor	Poor	Middle Class	Rich	Very Rich
Categorical value	1	2	3	4	5

Table 4.3.1.5: Categorical transformation of parent’s educational qualification of student

Column name	**PARENTS_EDU**	
Status	Yes	No
Categorical value	1	0

Table 4.3.1.6: Categorical transformation of parents died by Covid-19

Column name	**PRNTS_DIED_Cvd**	
Status	Yes	No
Categorical value	1	0

Table 4.3.1.7: Categorical transformation of student got married

Column name	**GOT_MARID**	
Status	Yes	No
Categorical value	1	0

Table 4.3.1.8: Categorical transformation of student started working

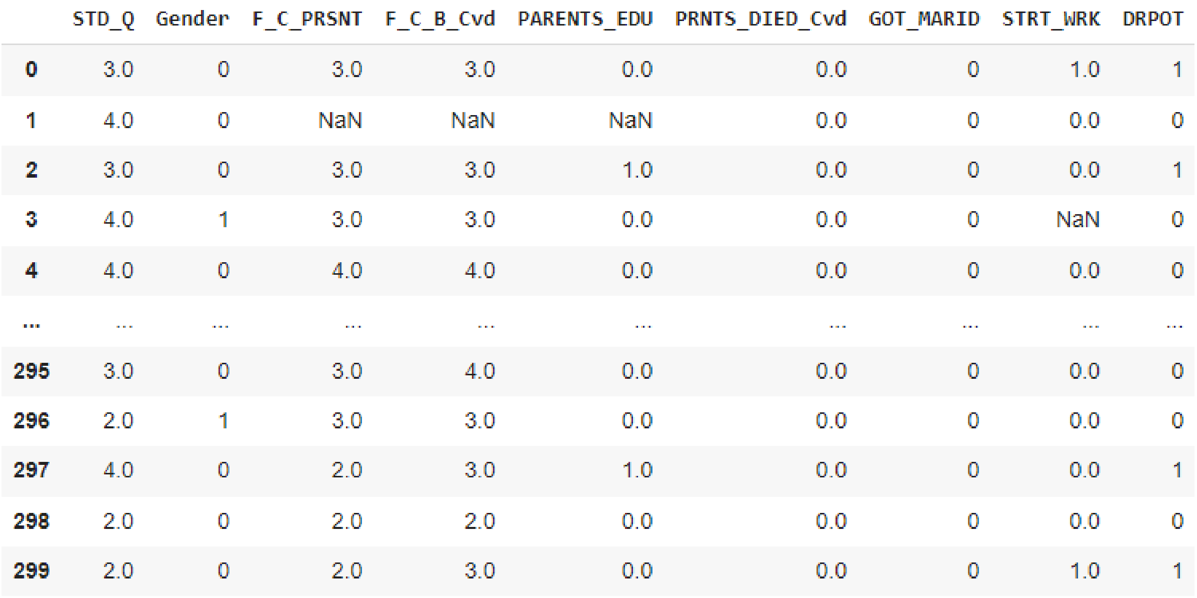
Column name	**STRT_WRK**	
Status	Yes	No
Categorical value	1	0

Table 4.3.1.9: Categorical transformation of student dropout or maybe wanted to dropout

Column name	**DRPOT**	
Status	Dropped Out	Maybe Wanted to Drop Out
Categorical value	1	0

4.4 Data Preprocessing

The dataset which is used in this research was primary data, it has been discussed before, and data were collected via school students of Bangladesh. In the data collection form, some basic personal questionnaires about dropped out student or the students maybe who wanted to be drop out, have been asked to the students to collect desired data. It is possible that some students don’t know all information we want to complete this study so they are asked to let those questions empty, which they don’t know. So, some data are empty, and data set created with missing value. Also, after importing data set into the ide another problem has been found with the dataset. Which is the data type, the data types were in float. So, we had to convert the data type but we had to fill-up the missing value first. Data types float and integer may vary the model accuracy, so without taking any chance converting data type is wise decision. The imported data set is given below,



	STD_Q	Gender	F_C_PRSNT	F_C_B_Cvd	PARENTS_EDU	PRNTS_DIED_Cvd	GOT_MARID	STRT_WRK	DRPOT
0	3.0	0	3.0	3.0	0.0	0.0	0	1.0	1
1	4.0	0	NaN	NaN	NaN	0.0	0	0.0	0
2	3.0	0	3.0	3.0	1.0	0.0	0	0.0	1
3	4.0	1	3.0	3.0	0.0	0.0	0	NaN	0
4	4.0	0	4.0	4.0	0.0	0.0	0	0.0	0
...	...	...	...	...	...	...	...	...	...
295	3.0	0	3.0	4.0	0.0	0.0	0	0.0	0
296	2.0	1	3.0	3.0	0.0	0.0	0	0.0	0
297	4.0	0	2.0	3.0	1.0	0.0	0	0.0	1
298	2.0	0	2.0	2.0	0.0	0.0	0	0.0	0
299	2.0	0	2.0	3.0	0.0	0.0	0	1.0	1

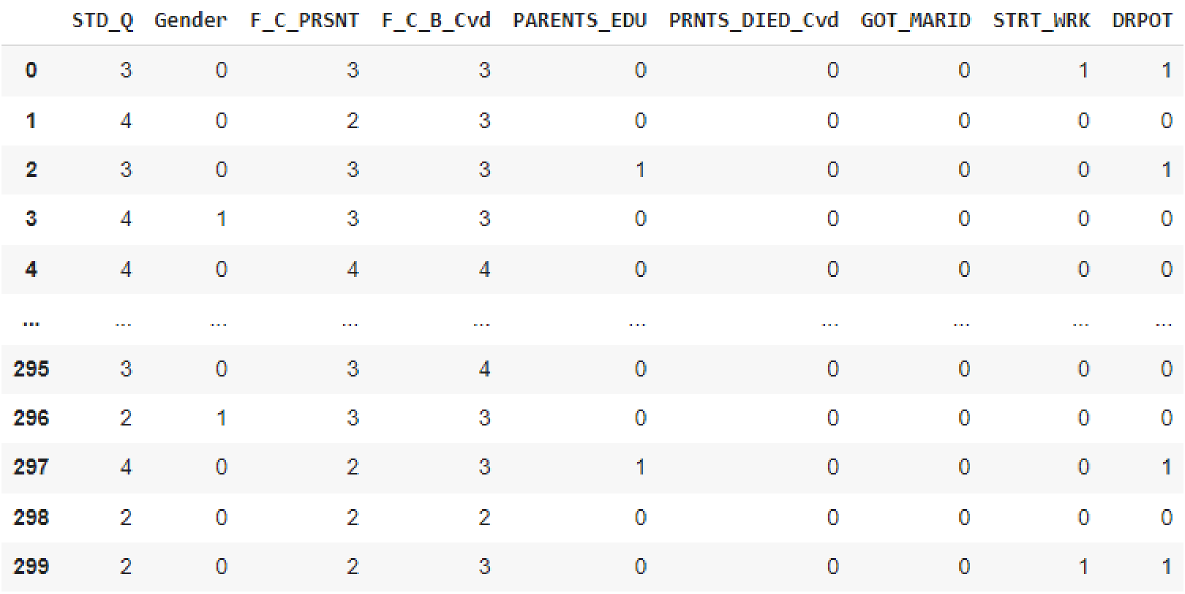
Figure 4.4.1: Overview of Dataset before preprocessing

4.4.1 Handling Missing Value

Missing value handling is a part of data preprocessing. Data preprocessing means prepare the dataset as useful as possible to let the computer to predict more accurately. If the dataset has any missing value the prediction model will predict less accurately. S, before applying machine learning algorithm we have to fill all the missing values. There are 3 ways to handle missing values, mean, median and mode. From 3 of this technique mean is proven most efficient way to handle missing values especially in discrete data set. This mean missing value handling technique is basically calculate the mean of all values of the attribute. The formula of mean is given below,

$$\boldsymbol{Mean} = \frac{X_1 + X_2 + X_3 + \cdots \ldots \ldots + X_n}{n}$$

After calculating mean value of attributes some float mean value comes up, after rounding off those mean value all attributes are converted to integer type so that the prediction model could predict more accurately. After missing value handling and data type conversion the data frame comes like this with 0 missing value,



	STD_Q	Gender	F_C_PRSNT	F_C_B_Cvd	PARENTS_EDU	PRNTS_DIED_Cvd	GOT_MARID	STRT_WRK	DRPOT
0	3	0	3	3	0	0	0	1	1
1	4	0	2	3	0	0	0	0	0
2	3	0	3	3	1	0	0	0	1
3	4	1	3	3	0	0	0	0	0
4	4	0	4	4	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...
295	3	0	3	4	0	0	0	0	0
296	2	1	3	3	0	0	0	0	0
297	4	0	2	3	1	0	0	0	1
298	2	0	2	2	0	0	0	0	0
299	2	0	2	3	0	0	0	1	1

Figure 4.4.1.1: Overview of Dataset after preprocessing

4.5 Data Visualization

As by human nature, all can understand easily understand and also can easily differentiate anything by visualization. The process of creating interactive visuals to understand trends, variations, and derive meaningful insights from data is defined as data visualization [13]. Data visualization is primarily used for data validation and cleaning, exploration and discovery, and communicating outcomes. It helps peoples to understand more quickly and it helps to get proper indication. In this research the topic is based on completely recent issue. There were many related works but situation before was not the same. The goal of this research is predicted school dropout after the pandemic. So, there could be many reasons behind this issue. Trying to evaluate those reason behind dropout. And to evaluate the reason behind data visualization will play a huge role to understand the most effective reason behind student dropout as discussed previously. Data visualization will easily indicate the specific reason behind dropout after Covid-19 pandemic. In the data visualization part, all data will be visualized by bar chart so anyone can easily understand.

4.5.1 Student Quality of Dropped out Student

In this visualization part student quality of dropout students and maybe wanted to dropout students' data is visualized. This data is visualized in 3 parts, all students, only male students and only female students. The visualization shows that students with student quality 'Bad' and 'Average' are dropped out the most. And after gender separation it shows that, male students with student quality with ‘Very Bad’, ‘Bad’ and ‘Average’ are the most dropped out. On other hand female students with the student quality of ‘Bad’ and ‘Average’ are the dropout most.

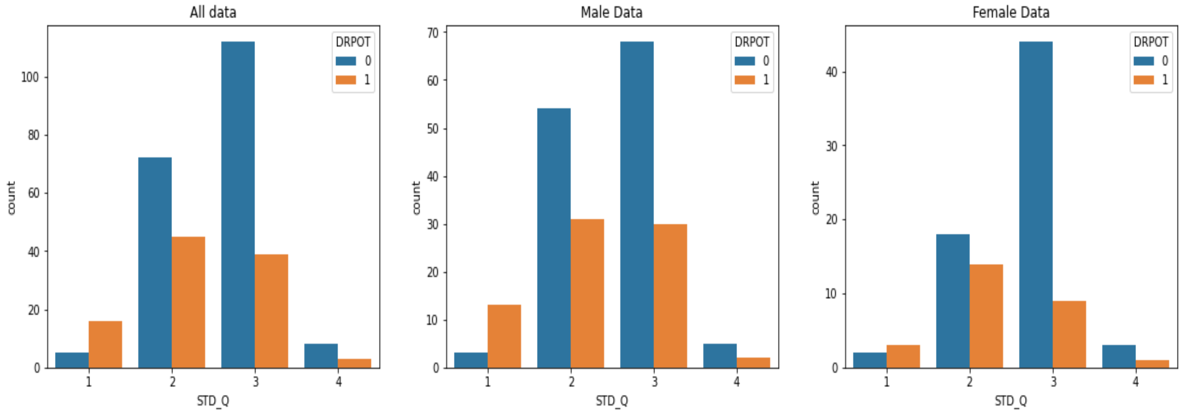


Figure 4.5.1.1: Visualization of dropout student data According to student quality

4.5.2 Financial Crisis of Dropout Students Family

In this research some basic informative data of dropped out student and maybe wanted to dropout student was collected. To study any link in between student dropout and student’s family’s financial condition, student’s financial condition before pandemic and after pandemic was collected. To differentiate the relation between financial crisis and student dropout. But in the dataset which is used in this study has no comparison of financial condition. So, a little modification has been done to properly visualize the desired relation. By subtracting ‘Financial Condition Present’ from ‘Financial condition Before’ formatted a new column name ‘Financial Decrement’. In column ‘Financial Decrement’ value 0 indicates there is no change in their financial condition in present and before pandemic, on the other hand value > 0 indicates that financial decrement occurred on those student family’s financial condition. Then visualized the data using bar graph to understand the relation between financial crisis and student dropout.

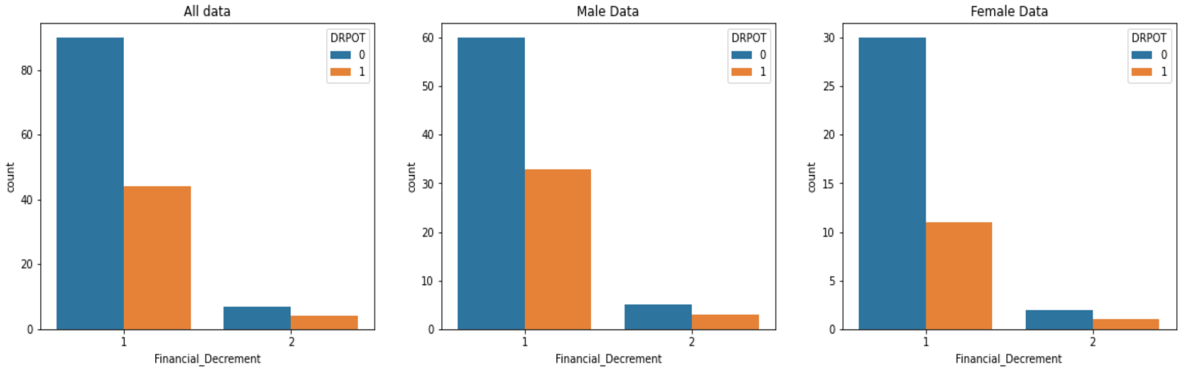


Figure 4.5.2.1: Data visualization of dropout student according to financial crisis

It shows 1 and 2 as financial decrement. Options of financial condition was as follows, (Very Poor, Poor, Middle Class, Rich and Very Rich). 5 steps were selected to determine financial condition. As per as the steps, 1 means financial condition decreased 1 step and 2 means financial condition decreased 2 steps. As example if financial condition was ‘Middle Class’ after 1 step decrement means their present financial condition is ‘Poor’.

4.5.3 Parents Educational Qualification

Parent’s educational qualification plays a huge role in every student’s life. Educated parents understand that education is much important than anything. On the other hand, maximum uneducated parents think earning money is much important than education. In this part of data visualization chapter, relation between parents education and student dropout will be visualized, to understand the relation quickly. In the bar it shows that both educated and uneducated parents children are dropped out, though the number of uneducated parents is much greater than the educated parents. This bar again proved that education is important to give proper importance to the education. The bar divided into 3 part (All Data, Male Data, Female Data). Which is dropout from all genders, only male dropped out student and only female dropped out students. According to the bar graph male student’s parents are more uneducated than female students. One thing is most noticeable, student number of maybe wanted to be dropout got the higher uneducated parent rate.

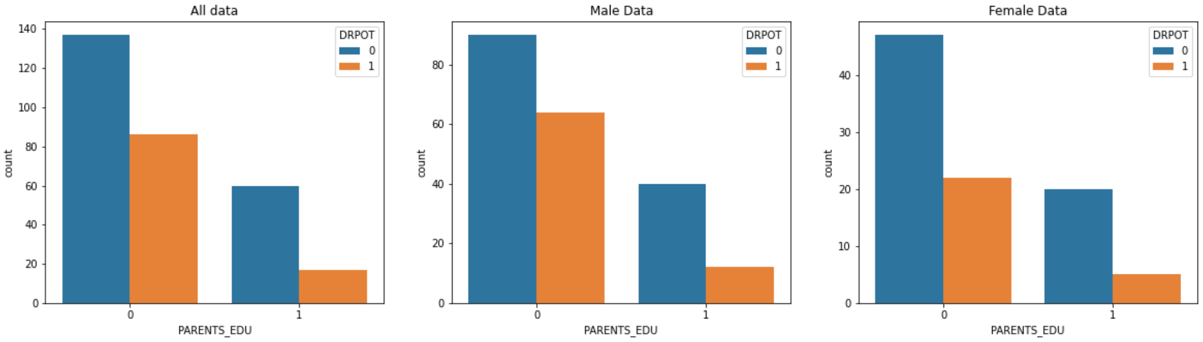


Figure 4.5.3.1: Visualization of dropout student data According to parent’s educational qualification

4.5.4 Parents Death

Any of the parent’s death in early age put a heavy mental also financial impact in a student life. School dropout is regularly occurring but because of covid-19 the rate of school dropout is increased. There is various reason to be dropout from school. Parents death could be one of them, students could be mentally broke so he or she don’t want to continue study or maybe family's financial condition fallen so it's hard for them to continue study. In this visualization part we will have the proper visualization of parent's death and student dropout.

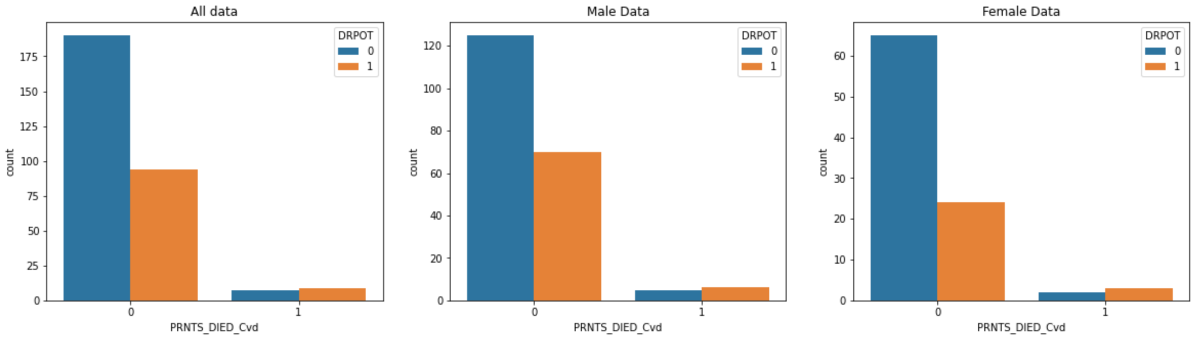


Figure 4.5.4.1: Visualization of dropout student data According to death of their parents

In the graph it's clear that parent's death doesn't impact much to student dropout. Very lower number of students parents died by covid-19 who are dropped out from school. Student whose parents didn't died by covid-19 are dropout the school more.

4.5.5 Students Marital Status and School Dropout

In a country like Bangladesh many students got married in under age. According to law no can get married under age of 18 but in remote and rural area also in urban areas peoples under age of 18 got married secretly by the consent of their parents. Some of those marriage happens by their consent and some of them happens forcefully. Which is destroying the young talents. Many society in Bangladesh still believes that all married women should be a housewife and only men should be earn for their families livelihood. So, child marriage and the thinking of the society is killing more dreams and destroying more young talents, who could be contribute much more to the society. After covid-19 also in the period of pandemic many parents quickly give marriage to their daughters for security, which is completely unnecessary thought. In cause of this in laws of the female students don't want much education for the bride. In this part of data visualization, it will clear that if student's marital status does impact to their education or not.

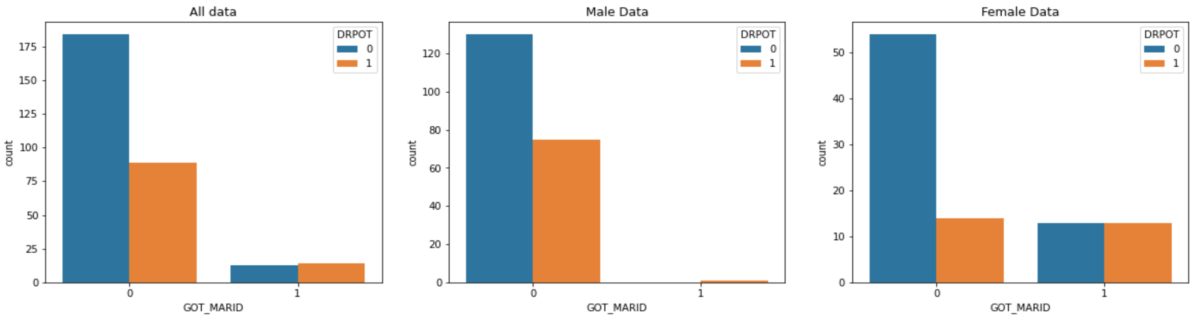


Figure 4.5.5.1: Visualization of dropout student data According to student’s marital status

In the given bar chart, it shows clearly that marital status doesn’t impact much in the student dropout in case of all gender, but in the case of female student’s marriage impacted almost similarly on the school dropout without marriage. In the case of male student dropout marriage impacted nearly 0. Which shows clearly under age marriage still happening in our country and it is causing school dropout to the female school students.

4.5.6 Early Earning and School Dropout

In an under develop country like Bangladesh the number of family living under poverty line is huge. For poverty many parents wanted that their children helps them in agricultural work or earn some money by doing any small job or business to financially support them. Though the government of Bangladesh created many scope so that students who cannot afford money to get educated could study with almost minimum amount of money, the amount which could be easily afford by any poor family. But still students dropping out from school and doing agricultural works, small jobs or doing small business to financially support their family. Some of them eagerly wanted to study but situation is not by their side. In this part of data visualization it shows clearly that many students dropped out from school and started earning money and the rate is almost similar students who started work and who is not started yet in case of all gender. On the other case of female student dropout rate for started earning is nearly 0 but in the case of male student the number is higher. Which shows that most of the male students dropped out school and started earning by themselves.

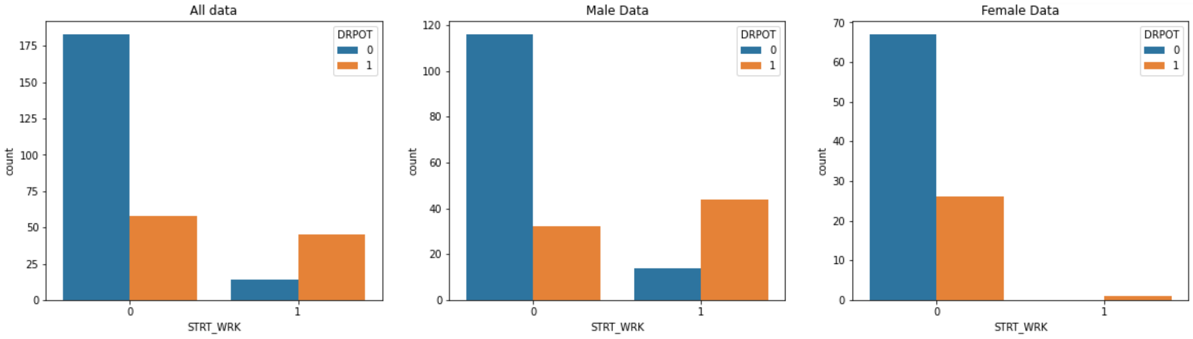


Figure 4.5.6.1: Visualization of dropout student data According to early earning

4.6 Conclusion

In the entire chapter, all data mining approaches are discussed, like data transformation, preprocessing, visualization and many more. In data transformation part data of the dataset used in this research is prepared to study by converting all string data into numeric data. In the data preprocessing part all missing value handled so that proposed model could predict more accurately. And in the data visualization part relations between various factors and student dropout is discussed. These approaches helped to understand dataset properly and to prepare the dataset for further study.

CHAPTER 5

Result & Discussion

5.1 Introduction

In this chapter, all process of applying Machine Learning algorithms and results will be discussed. For this cause some standard should be discussed before so that the outcomes will be very clear. After applying the Machine Learning algorithms some performance measurement will be discussed, which will help to appraise the finest result of applied Machine learning algorithms. After discussing this chapter of this research, it will be very clear that which machine learning algorithm performed better.

5.2 Applying Machine Learning Algorithms

After studying researches which are related to this research 5 Machine Learning Algorithms have been used to acquire greater accuracy rate for the classification to classifying the target class (Dropout). The applied algorithms are (Random Forest, Gaussian Naïve Bayes, Decision Tree Classifier, Support Vector Machine and Logistic Regression). These Machine Learning Algorithms Performed outstanding in previous student dropout related works.

5.2.1 Preprocessing and Declaration

To apply machine learning algorithm in the collected dataset of this research, the machine must need to be trained or experienced to predict the dropout. For training computers or making experienced to predict desired outcome percentage split have been used in this research. Where n% and (100-n) % of data is split to use n% of data to train the machine and (100-n) % of data to test the model. The accuracy of the model will be determined by the (100-n) % of data which was reserved to be tested. In this research 80-20 data split is used, which means 80% rows from the dataset is used for training and 20% rows from the dataset is used to test the model, and the split data was selected randomly by the machine.

5.3 Analysis of Performance Measurement

Performance measure means measure the model performance, that who good it’s working or predicting. Performance measures are important in machine learning because they are used to evaluate learning algorithms. These measures are sometimes used as heuristics to build learning models [14]. For measuring performance of the applied machine learning algorithms (Random Forest, Support Vector Machine, Decision Tree Classifier, Gaussian Naïve Bayes, Logistic Regression) which were used to predict student dropout, was measured by (accuracy, precision, recall, F-score, confusion matrix, ROC curve) performance measurement techniques. In the rest of the chapter, results of performance measurement parameters will be discussed.

5.3.1 Accuracy Rate of Classification

In this research 5 machine learning algorithms are used, to determine best algorithm among 5 of the algorithms accuracy rate is the most important performance measurement parameters. In this performance measurement parameter exact categorized sample is divided by the total number of sample and then multiply by 100 to get the accuracy rate of the model. Here, exact categorized sample is total number of True Positive (TP) and True Negative (TN).

$$\boldsymbol{ACCURACY\ RATE = \frac{TP + TN}{Total} \times 100}$$

The obtained accuracy rates of the algorithms are shown below,

Table 5.3.1.1: ACCURACY RATE OF APPLIED MACHINE LEARNING ALGORITHMS

Applied Algorithms	Accuracy Rate
SVM	83%
Decision Tree Classifier	83%
Random Forest	87%
Gaussian Naïve Bayes	80%
Logistic Regression	82%

According to the table 5.3.1.1, all 5 machine learning algorithms which are used in this research scored above 80% except Gaussian Naïve Bayes and Random Forest scored the highest accuracy.

5.3.2 Precision

Precision is one measure of a machine learning model's performance because it measures the accuracy of a positive prediction made by the model. Precision is calculated by dividing the number of true positives by the total number of positive predictions (i.e., the number of true positives plus the number of false positives) [15].

$$\boldsymbol{PRECISION = \frac{TP}{TP + FP}}$$

$$\mathbf{TP + FP = Predicted\ Yes}$$

The achieved Precisions of applied machine learning algorithms are given below,

Table 5.3.2.1: PRECISION SCORE OF MACHINE LEARNING ALGORITHMS

Applied Algorithms	Precision
SVM	0.83
Decision Tree Classifier	0.84
Random Forest	0.87
Gaussian Naïve Bayes	0.81
Logistic Regression	0.81

In the shown table 5.3.2.1, all machine learning algorithms scored precision of above 0.8. Gaussian Naïve Bayes and Logistic Regression got the same precision of 0.81, among all of them Random Forest scored precision value of 0.87 which is highest scored precision value.

5.3.3 Recall

The recall is calculated as the proportion of Positive samples that were correctly classified as Positive to the total number of Positive samples. The recall of the model measures its ability to detect positive samples. The more positive samples detected, the higher the recall [16]. The formula is,

$$\boldsymbol{Recall = \frac{TP}{TP + FN}}$$

$$\boldsymbol{TP + FN = Actual\,Yes}$$

Acquired recall values of applied machine learning algorithms are shown below,

Table 5.3.3.1: RECALL VALUE OF APPLIED MACHINE LEARNING ALGORITHM

Applied Algorithms	Recall
SVM	0.83
Decision Tree Classifier	0.83
Random Forest	0.87
Gaussian Naïve Bayes	0.80
Logistic Regression	0.82

In the table 5.3.3.1, Random Forest scored highest recall value among all 5 machine learning algorithms.

5.3.4 F-score

The harmonic mean of a system's precision and recall values is used to calculate an F-score. It is calculated using the formula: 2 x [(Precision x Recall) / (Precision + Recall)] [17]. The use of F-score values to determine the quality of a predictive system has been criticized because a moderately high F-score can be the result of an imbalance between precision and recall and thus does not tell the entire story. Systems with a high level of accuracy, on the other hand, struggle to improve precision or recall without negatively impacting the other.

$$\boldsymbol{F - SCORE = \frac{2 \times Precision \times Recall}{Precision + recall}}$$

F-score values of applied machine algorithms are shown below,

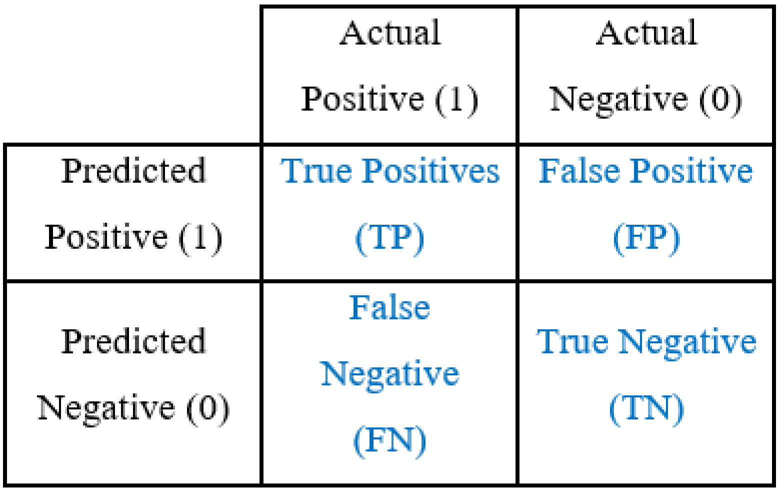
Table 5.3.4.1: F-SCORE VALUE OF MACHINE LEARNING ALGORITHMS

Applied Algorithms	F-Score
SVM	0.83
Decision Tree Classifier	0.84
Random Forest	0.87
Gaussian Naïve Bayes	0.80
Logistic Regression	0.81

As other performance measure parameter F-Score value also score the highest for Random Forest.

5.3.5 Confusion Matrix

Confusion matrix is a model to know how a prediction model is performing. A confusion matrix is a summary of classification problem prediction results. The number of correct and incorrect predictions is summarized with count values and divided by class. This is the confusion matrix's key. It provides information not only about the errors made by your classifier, but also about the types of errors made [17]. A prediction model predicts 4 types of prediction which are True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN). By seeing the number of these values, it is easy to understand that how good or bad the model is performing. A 2×2 matrix is used to visualize this matrix.



	Actual Positive (1)	Actual Negative (0)
Predicted Positive (1)	True Positives (TP)	False Positive (FP)
Predicted Negative (0)	False Negative (FN)	True Negative (TN)

Figure 5.3.5.1: Basic Diagram of Confusion Matrix [30].

Heat map plotting of confusion matrix of all applied machine learning algorithms will be given below. It will help to visualize confusion matrix of all machine learning algorithms which was used in this research.a

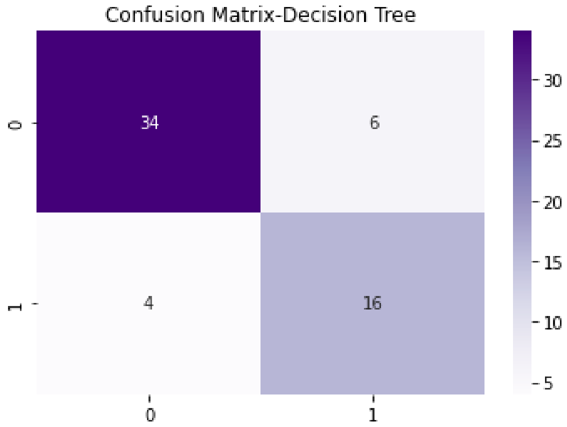


Figure 5.3.5.2: Confusion Matrix for Decision Tree

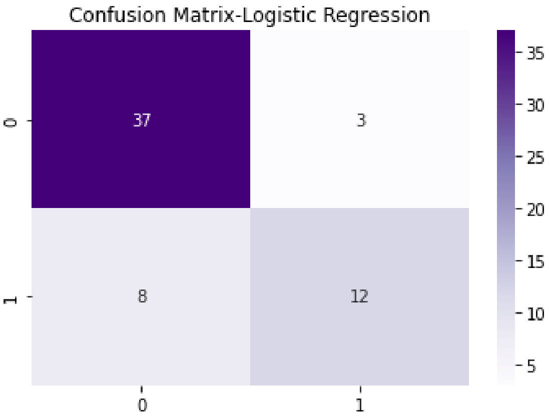
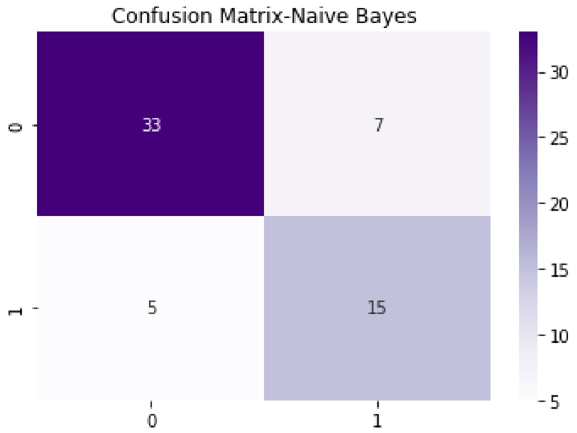


Figure 5.3.5.3: Confusion Matrix for Logistic Regression



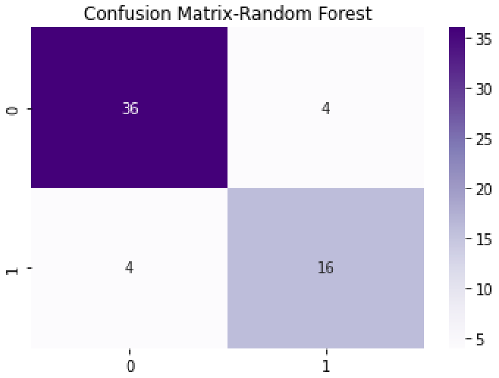


Figure 5.3.5.4: Confusion Matrix for Naïve Bayes

Figure 5.3.5.5: Confusion Matrix for Random Forest

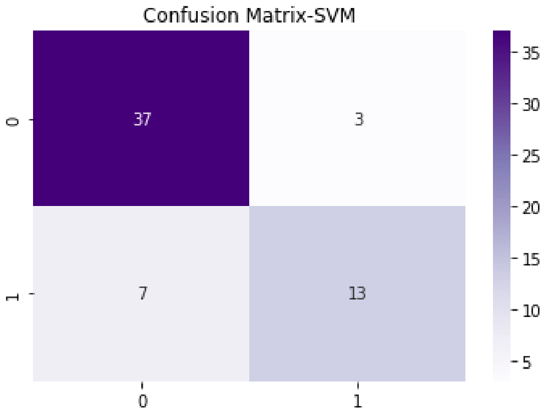


Figure 5.3.5.6: Confusion Matrix for Support Vector Machine.

5.3.6 Receiver Operating Characteristic (ROC) Curve

The receiver operating characteristic, or ROC, curve is a popular plot for displaying the tradeoff between the true positive rate and the false positive rate for a binary classifier at various classification thresholds at the same time [18]. According to the confusion matrix, this curve has two parameters. The first is known as the True-Positive Rate, while the second is known as the False-Positive Rate.

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

A model with no skill is created when a classifier illustrates at the point (0.5, 0.5) or by a diagonal line that runs from the bottom left of the ROC Curve to the top right and has an AUC of 0.5. A model with perfect competence is generated when a classifier draws a line from the bottom left of the ROC Curve to the top left of the ROC Curve and moves the top right of the ROC Curve.

The ROC curve shown below, where all 5 machine learning algorithm applied in this research is plotted in one plot,

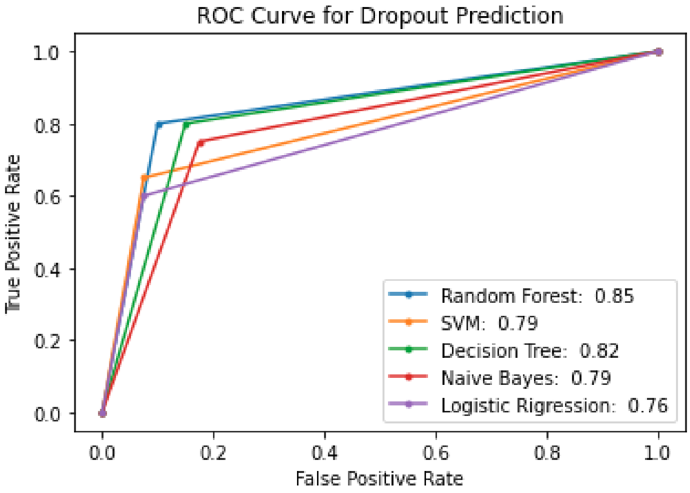


Figure 5.3.6.1: ROC for 5 Machine Learning Algorithms.

5.4 Conclusion

In this chapter, completely discussed that Support Vector Machine, Decision Tree Classifier and Random Forest performed well at every performance measurement (Precision, Recall, F-Score, ROC Curve, Accuracy rate). Support Vector machine scored 0.83 precision value and Decision Tree Classifier scored precision value of 0.84 which is almost similar to the precision value of Support Vector Machine. On other hand Random Forest scored well in every performance measure with the accuracy rate of 87%. Overall, the outcome and performance of this model met the requirements of the study.

CHAPTER 6

Strategic Plan to Re-enrollment of Students

6.1 Introduction

In this chapter some strategy will be discussed for re-enrollment of the dropout students or consultation strategy for students who wants to dropout school. By applying these strategies number of school student dropout could be reduce. Dropout impact directly in the socio economic condition of a nation, instead of act instantly this damages takes time to act on social and economic condition of a country. Maybe discussed strategies in this chapter will help to reduce dropout so that damages near future could be decrease as much as possible.

6.2 Strategic Plan

The most effective strategic plan for reduction of the issue school student dropout could be consultation and logistic support. Also lack of proper motivation is liable for dropout. Parents who are not enough educated they found no values in education, those parents must need proper consultation so that they could be motivated to send their children’s to the school. Student who dropped out or wanted to dropout for their family’s financial condition, they are needed logistic supports. There are many NGO’s and government organizations who works for under privileged peoples, to ensure good food, good health and education. Without these organizations every MPO enlisted school has poor funds to provide full free scholarship to those students who are not able to provide minimum amount of fees to the school. Under age marriage is also liable for school dropout, many female students got married before turn 18 years old with or without their consent and still many parents don’t want to send their daughter in law to school or to get educated and become independent. By the constitution of Peoples Republic of Bangladesh government marriage before 18 is punishable crime. Both the parents could be go to jail for this kind of acts. These parents’ needs consultation and awareness should be raised to prevent this kind of crime. In year 2023 not only male, females are also contributing equally to our nations as males are doing. Moral awareness should be spread on those area where school dropout rate with child marriage is very common. In city area parents are more educated and more aware about their children’s educations, in fact they try to spend more money to provide better education to their children’s according to their financial ability. On the other hand in rural area where almost maximum number of people depends on agriculture and working as a day labour, don’t know the value of education. So, in rural areas peoples need to be more motivated so that they could be more aware about their children’s education.

6.3 Conclusion

There are various reasons to dropout school or wishing to dropout. We have to analysis the main reason for being dropout and have to take action as per the reason. If the reason is financial it could be solved by financing them but if the reason is not financial then the reason is motivational. We have to motivate them so that they re-enroll to the school again and those who want to dropout they changed their mind that they will never dropout school.

CHAPTER 7

Critical Appraisal

7.1 Introduction

The Potentiality, Languish and Opportunity of this research will be discussed in this chapter. Nothing could be developed or invent which have no limitations. Everything has limitations this research is not different from other studies, this research also has some limitations and also has some pros. Which will be discussed in the entire chapter.

7.2 SWOT Analysis

Swot is a strategic planning and management approach that can be used to assist a person or organization in identifying their research planning strengths, weaknesses, opportunities, and threats. It's also referred to as scenario analysis or situational assessment. This research also has some strengths, weaknesses, opportunities and threats which will be discovered in the discussion of SWOT analysis.

7.2.1 Strength

In this research, the strongest part is the motive of this research. This research is limited into Bangladesh but student dropout is a very common problem all over the world. Student dropout especially school, college and university dropout impact the most in a nation’s economy and future development. Many nations in the world provide free education to their residents, which means the government of those nations are investing on the students to get educated and contribute towards the nation. If any student dropout in the middle of the school all previous investment gone wasted. An uneducated resident cannot contribute more than an educated resident to the nation. The school dropout rate of Bangladesh was decreasing but sudden pandemic and the lockdown for covid-19 increased the school dropout rate again. To resolve this school dropout after pandemic related issue this research have been studied. This research could also develop a connection between education sector and data mining sector cause we are getting digitalized and use of computer technology in every sector are increasing day by day.

7.2.2 Weakness

If we want to discuss the weakness of this research, then we must say that the dataset is not so large. In the other researches researchers used a vast amount of secondary data which was collected through a source of education sector or a big institution, but in a short time collect a huge numbers of primary data is a very big deal, which is quite tough and expensive. Though this small set of dataset shows a very impressive accuracy, in larger dataset accuracy could be grown more. Cause larger the training data means more experience for the machine and more accurate prediction. This is the only weakness of this research, which could be solve by gathering vast amount of data. So, this cannot be a big concern beyond this study.

7.2.3 Opportunity

This research has a huge opportunity, education sectors or the administration of education sector of Bangladesh or any organization who have the authority to take a step to prevent or decrease illiteracy issue could be helped by this research. Education sector of Bangladesh getting digitalized day by day, use of computer technology is increasing rapidly. So if the authority want to prevent school student dropout after any hard situation they could take help of this research. In my point of view this study could be great opportunity for the education sector and also the data mining researchers.

7.2.4 Threat

In this research, most burdensome threat is predict the school student dropout accurately. This model proposed in this research can predict school student dropout accurately 87% which means 87 predictions out of 100 is accurate, which means there is a chance of 13% of wrong prediction. But this threat can be resolve by the further research near future.

7.3 Conclusion

After discussing SWOT analysis of this research, we must say that the predictive model derived in this research could help the education sector, social and commercial sector also by preventing school dropout.

CHAPTER 8

Conclusion

8.1 Conclusion

In this research, a predictive model using machine learning technique is introduced by combining data mining technique. The applied classification techniques which were used to develop desired prediction model of this research are, Logistic Regression, Decision Tree Classifier, Random Forest, Gaussian Naïve Bayes and Support Vector Machine (SVM). To determine which algorithm is more effective to predict school student dropout after covid-19 pandemic more accurately. The Random Forest conquered the highest accuracy of 87% in the final result.

Without accuracy rate other performance measurement parameters are used in this research to determine which algorithm could predict school student more effectively. They are precision, recall F-score and ROC curve, this performance measurement parameter helped to determine the algorithm which performs the best in this research. In precision, recall and F-score, Decision Tree Classifier, Support vector Machine and Random Forest scored the highest value, but among them Random Forest scored highest value in every performance measurement. The ROC curve helped to visualize so that we could determine most effective machine learning algorithm, which predict the student dropout more accurately then other machine learning algorithms. The proposed model was developed by Google Colab, which is a python based online IDE. The machine was trained by primary data which was collected from school students who know students who dropped out from school and any students who shows willingness to dropout school.

In education sector of Bangladesh school dropout is a very common term in rural area or in the area where most of the people lives under the poverty line. Specially after pandemic that dropout number increased more. To determine the issues behind their dropout this research was proposed. This research was base on 2 outputs one is dropout and another one is students who shows willingness to dropout school. If the prediction from the proposed model shows 0 that means there is a bit chance to be dropped out from school and is the prediction comes out 1 then the students got the higher chance to get dropped out from the school. Authorities who want to prevent school dropout could take help of this research to decrease school dropout in early stage.

8.2 Further Suggested Work

In future, this research could be studied more to conquer higher accurate prediction and more effectiveness by combing more machine learning algorithms and more data mining techniques. Different data sets for dropout prediction could be used in the proposed model or new techniques could be apply by combining various new or complex methods. Large number of data set could perform better in the proposed model of this research. This research also shows that proposed model used in this study is not only used for predict school student dropout, other dropouts from various educational institute or source could be predict by the proposed model.

References

[2] C. Márquez-Vera, A. Cano, C. Romero, A. Y. M. Noaman, H. Mousa Fardoun, and S. Ventura, “Early dropout prediction using data mining: a case study with high school students,” *Expert Systems*, vol. 33, no. 1, pp. 107–124, Nov. 2015, doi: 10.1111/exsy.12135.

[3] F. Del Bonifro, M. Gabbrielli, G. Lisanti, and S. P. Zingaro, “Student Dropout Prediction,” *Lecture Notes in Computer Science*, pp. 129–140, 2020, doi: 10.1007/978-3-030-52237-7_11.

[4] M. Tan and P. Shao, “Prediction of Student Dropout in E-Learning Program Through the Use of Machine Learning Method,” *International Journal of Emerging Technologies in Learning (iJET)*, vol. 10, no. 1, p. 11, Feb. 2015, doi: 10.3991/ijet.v10i1.4189.

[5] Ahmed, S. A., Billah, M. A., & Khan, S. I., “ A machine learning approach to performance and dropout prediction in computer science: Bangladesh perspective”. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-6). IEEE.

[6] Hasan, M. N., “A Comparison of Logistic Regression and Linear Discriminant Analysis in Predicting of Female Students Attrition from School in Bangladesh”. In 2019 *7th Int. Conf. on Data Science & SDGs*, pp 129-134

[7] A. G. Pertiwi, T. Widyaningtyas, and U. Pujianto, “Classification of province based on dropout rate using C4.5 algorithm,” *2017 International Conference on Sustainable Information Engineering and Technology (SIET)*, Nov. 2017, doi: 10.1109/siet.2017.8304173.

[8] R. Bukralia, “Predicting Dropout in Online Courses: Comparison of Classification Techniques,” *MWAIS 2010 Proceedings*, May 2010, Accessed: Nov. 22, 2022. [Online]. Available: https://aisel.aisnet.org/mwais2010/19

[9] R. S. Baker, A. W. Berning, S. M. Gowda, S. Zhang, and A. Hawn, “Predicting K-12 Dropout,” *Journal of Education for Students Placed at Risk (JESPAR)*, vol. 25, no. 1, pp. 28–54, Oct. 2019, doi: 10.1080/10824669.2019.1670065.

[10] M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects” Science 349, VOL 349, ISSUE 6245 (2015), doi: 10.1126/science.aaa8415.

[11] “Classification Algorithm in Machine Learning - Javatpoint,” *www.javatpoint.com*. https://www.javatpoint.com/classification-algorithm-in-machine-learning

[12] H.-I. Lim, "A Linear Regression Approach to Modeling Software Characteristics for Classifying Similar Software," in 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), 2019, pp. 942-943

[13] A. Das, “Data Visualization in Data Science,” *Medium*, Aug. 23, 2020. https://towardsdatascience.com/data-visualization-in-data-science-5681cbdde5bf

[14] J. Brownlee, “Train-Test Split for Evaluating Machine Learning Algorithms,” *Machine Learning Mastery*, Jul. 23, 2020. https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/

[15] “Precision,” *C3 AI*. https://c3.ai/glossary/machine learning/precision/#:~:text=Precision%20is%20one%20indicator%20of

[16] “Precision and Recall in Machine Learning - Javatpoint,” *www.javatpoint.com*. https://www.javatpoint.com/precision-and-recall-in-machine-learning#:~:text=What%20is%20Recall%3F

[17] J. Brownlee, “What is a Confusion Matrix in Machine Learning,” *Machine Learning Mastery*, Nov. 17, 2016. https://machinelearningmastery.com/confusion-matrix-machine-learning/#:~:text=A%20confusion%20matrix%20is%20a%20summary%20of%20prediction%20results%20on

[18] “Receiver Operating Characteristic (ROC) Curve,” *C3 AI*. https://c3.ai/glossary/data-science/receiver-operating-characteristic-roc-curve/

[19] “Logistic Regression in Machine Learning - Javatpoint,” www.javatpoint.com. https://www.javatpoint.com/logistic-regression-in-machine-learning

[20] “SVM Algorithm | Working & Pros of Support Vector Machine Algorithm,” *EDUCBA*, Sep. 12, 2019. https://www.educba.com/svm-algorithm/

[21] “machine-learning-random-forest-algorithmhttps:www.javatpoint.commachine-learning-random-forest-algorithm - Javatpoint,” *www.javatpoint.com*. https://www.javatpoint.com/machine-learning-random-forest algorithmhttps://www.javatpoint.com/machine-learning-random-forest-algorithm.

[22] W. Consulting, “#CoronaVirus: Don’t Let Our Children Down!,” *Global Campaign for Education*. https://campaignforeducation.org/en/press-centre/coronavirus-dont-let-our-children-down?gclid=EAIaIQobChMI567l3prH_AIVpTdyCh3vpACeEAAYAiAAEgKfF_D_BwE (accessed Jan. 14, 2023).

[23] “School dropout Definition,” *Law Insider*. https://www.lawinsider.com/dictionary/school-dropout

[24] E. Burns, “What Is Machine Learning and Why Is It Important?,” *SearchEnterpriseAI*, Mar. 2021. https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML

[25] R. Sunil, “Understanding Support Vector Machine algorithm from examples (along with code),” *Analytics Vidhya*, Mar. 11, 2019. https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

[26] S. E R, “Random Forest | Introduction to Random Forest Algorithm,” *Analytics Vidhya*, Jun. 17, 2021. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/#:~:text=Random%20forest%20is%20a%20Supervised

[27] “Decision Tree,” *CORP-MIDS1 (MDS)*. https://www.mastersindatascience.org/learning/machine-learning-algorithms/decision-tree/#:~:text=A%20decision%20tree%20is%20a

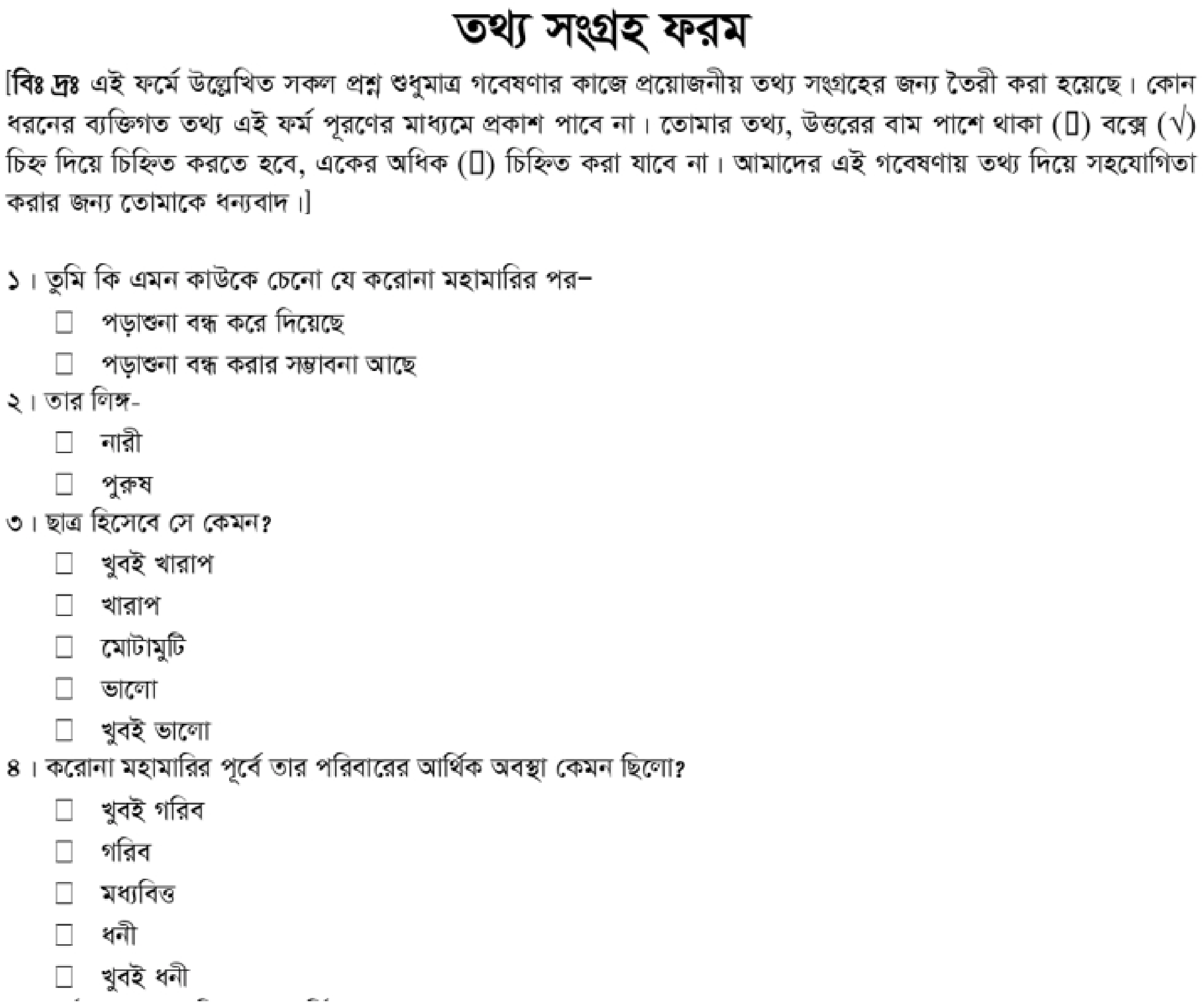
[28] N. Kumar, “Naive Bayes Classifiers - GeeksforGeeks,” *GeeksforGeeks*, Jan. 14, 2019. https://www.geeksforgeeks.org/naive-bayes-classifiers/

[29] javaTpoint, “Machine Learning Decision Tree Classification Algorithm - Javatpoint,” *www.javatpoint.com*. https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

[30] “Measuring Performance: The Confusion Matrix,” *Glass Box*, Feb. 17, 2019. https://glassboxmedicine.com/2019/02/17/measuring-performance-the-confusion-matrix/

Appendices

Data Collection Form



তথ্য সংগ্রহ ফরম

[**বিঃ দ্রঃ** এই ফর্মে উল্লেখিত সকল প্রশ্ন শুধুমাত্র গবেষণার কাজে প্রয়োজনীয় তথ্য সংগ্রহের জন্য তৈরী করা হয়েছে। কোন ধরনের ব্যক্তিগত তথ্য এই ফর্ম পূরণের মাধ্যমে প্রকাশ পাবে না। তোমার তথ্য, উত্তরের বাম পাশে থাকা (□) বক্সে (√) চিহ্ন দিয়ে চিহ্নিত করতে হবে, একের অধিক (□) চিহ্নিত করা যাবে না। আমাদের এই গবেষণায় তথ্য দিয়ে সহযোগিতা করার জন্য তোমাকে ধন্যবাদ।]

১। তুমি কি এমন কাউকে চেনো যে করোনা মহামারির পর–

□ পড়াশুনা বন্ধ করে দিয়েছে
□ পড়াশুনা বন্ধ করার সম্ভাবনা আছে

২। তার লিঙ্গ-

□ নারী
□ পুরুষ

৩। ছাত্র হিসেবে সে কেমন?

□ খুবই খারাপ
□ খারাপ
□ মোটামুটি
□ ভালো
□ খুবই ভালো

৪। করোনা মহামারির পূর্বে তার পরিবারের আর্থিক অবস্থা কেমন ছিলো?

□ খুবই গরিব
□ গরিব
□ মধ্যবিত্ত
□ ধনী
□ খুবই ধনী

৫। বর্তমানে তার পরিবারের আর্থিক অবস্থা কেমন?

- □ খুবই গরিব
- □ গরিব
- □ মধ্যবিত্ত
- □ ধনী
- □ খুবই ধনী

৬। তার বাবা-মা কী শিক্ষিত?

- □ হ্যাঁ
- □ না

৭। করোনায় আক্রান্ত হয়ে তার বাবা অথবা মা কেউ কী মৃত্যু বরণ করেছে?

- □ হ্যাঁ
- □ না

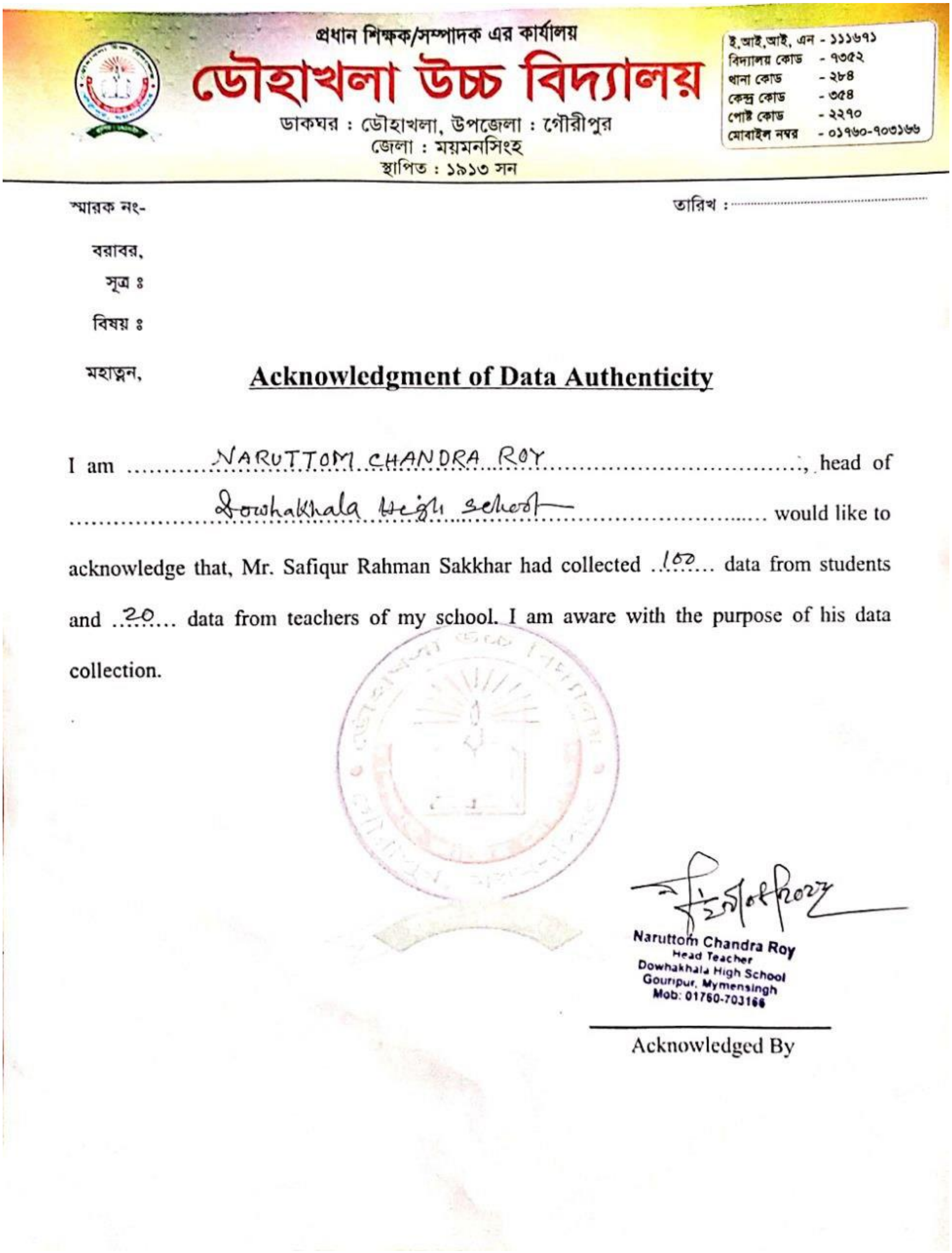
৮। সে কী বিবাহিত?

- □ হ্যাঁ
- □ না

৯। সে কী পরিবারকে আর্থিক ভাবে সহযোগিতা করার জন্য অর্থ উপার্জন করে?

- □ হ্যাঁ
- □ না

Acknowledgment of Data authentication



প্রধান শিক্ষক/সম্পাদক এর কার্যালয়

ডৌহাখলা উচ্চ বিদ্যালয়

ডাকঘর : ডৌহাখলা, উপজেলা : গৌরীপুর
জেলা : ময়মনসিংহ
স্থাপিত : ১৯১৩ সন

ই.আই.আই, এন	- ১১১৬৭১
বিদ্যালয় কোড	- ৭৩৫২
থানা কোড	- ২৮৪
কেন্দ্র কোড	- ৩৫৪
পোষ্ট কোড	- ২২৭০
মোবাইল নম্বর	- ০১৭৬০-৭০৩১৬৬

স্মারক নং-

তারিখ : ....

বরাবর,

সূত্র ঃ

বিষয় ঃ

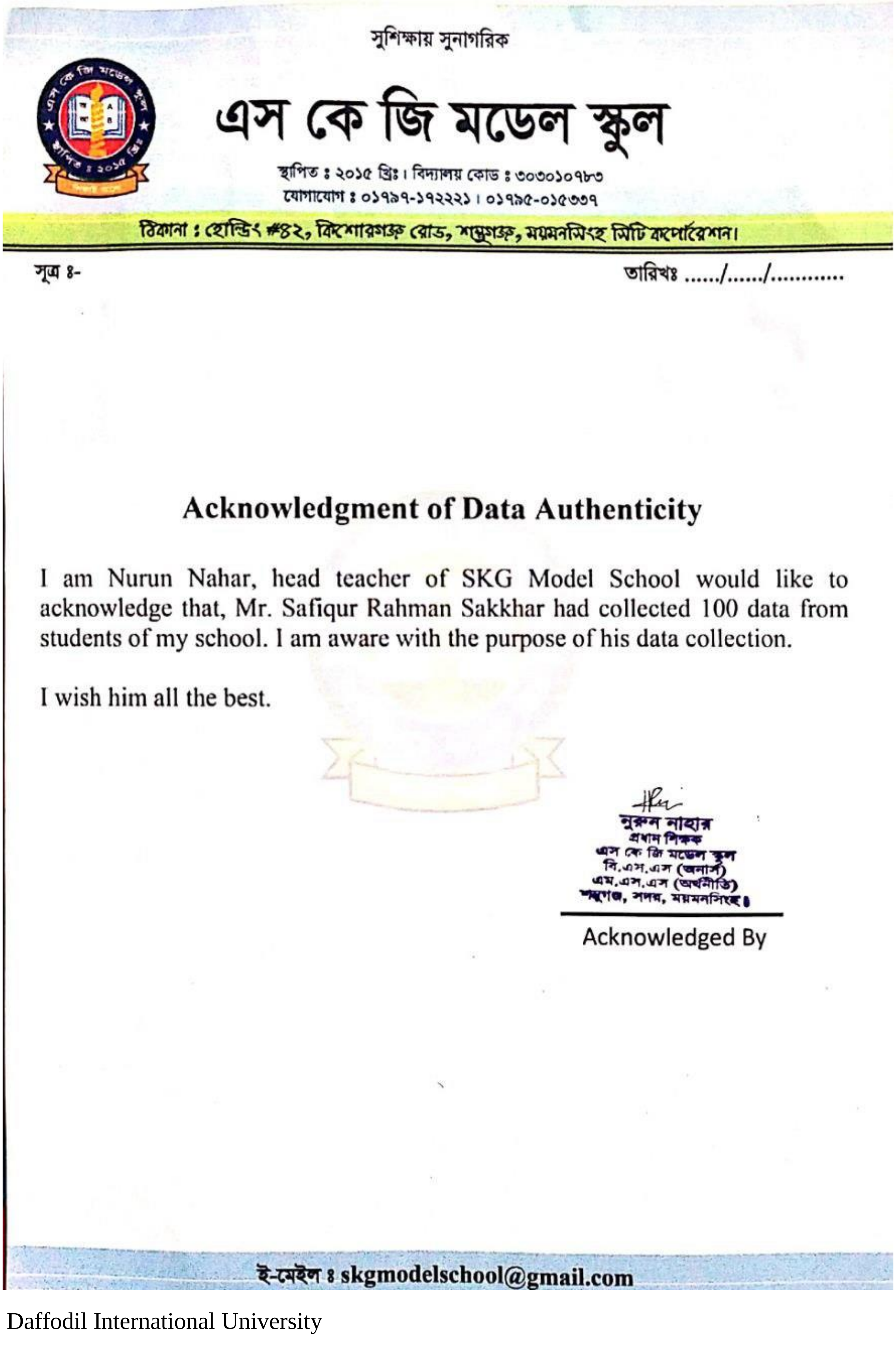
মহাত্মন,

Acknowledgment of Data Authenticity

I am ....NARUTTOM CHANDRA ROY.., head of ....Dowhakhala High school.. would like to acknowledge that, Mr. Safiqur Rahman Sakkhar had collected ..100... data from students and ..20.... data from teachers of my school. I am aware with the purpose of his data collection.

Naruttom Chandra Roy
Head Teacher
Dowhakhala High School
Gouripur, Mymensingh
Mob: 01760-703166

Acknowledged By



সুশিক্ষায় সুনাগরিক

এস কে জি মডেল স্কুল

স্থাপিত ঃ ২০১৫ খ্রিঃ। বিদ্যালয় কোড ঃ ৩০৩০১০৭৮৩
যোগাযোগ ঃ ০১৭৯৭-১৭২২২১ । ০১৭৯৫-০১৫৩৩৭

ঠিকানা : হোল্ডিং #৪২, কিশোরগঞ্জ রোড, শম্ভুগঞ্জ, ময়মনসিংহ সিটি কর্পোরেশন।

সূত্র ঃ- তারিখঃ ....../....../............

Acknowledgment of Data Authenticity

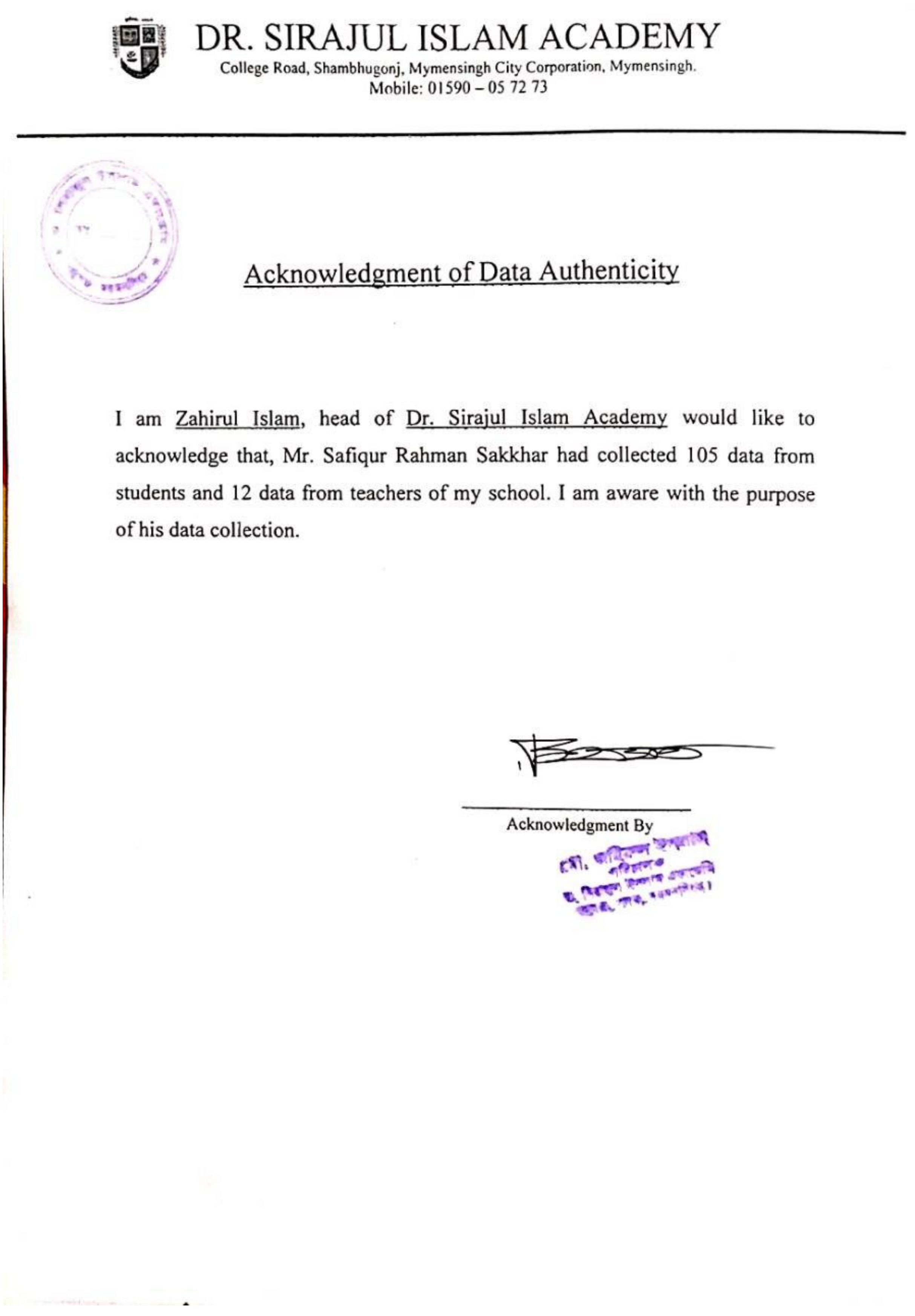
I am Nurun Nahar, head teacher of SKG Model School would like to acknowledge that, Mr. Safiqur Rahman Sakkhar had collected 100 data from students of my school. I am aware with the purpose of his data collection.

I wish him all the best.

নুরুন নাহার
প্রধান শিক্ষক
এস কে জি মডেল স্কুল
বি.এস.এস (অনার্স)
এম.এস.এস (অর্থনীতি)
শম্ভুগঞ্জ, সদর, ময়মনসিংহ।

Acknowledged By

ই-মেইল ঃ skgmodelschool@gmail.com



DR. SIRAJUL ISLAM ACADEMY

College Road, Shambhugonj, Mymensingh City Corporation, Mymensingh.
Mobile: 01590 – 05 72 73

Acknowledgment of Data Authenticity

I am Zahirul Islam, head of Dr. Sirajul Islam Academy would like to acknowledge that, Mr. Safiqur Rahman Sakkhar had collected 105 data from students and 12 data from teachers of my school. I am aware with the purpose of his data collection.

Acknowledgment By

মো. জহিরুল ইসলাম
পরিচালক
ড. সিরাজুল ইসলাম একাডেমি
[illegible]

SCHOOL DROPOUT PREDICTION OF BANGLADESHI STUDENTS DUE TO COVID-19

ORIGINALITY REPORT

16%	14%	7%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	3%
2	Submitted to Daffodil International University Student Paper	3%
3	Submitted to Napier University Student Paper	1%
4	"Implications of Meta Classifiers for Onset Diabetes Prediction", International Journal of Innovative Technology and Exploring Engineering, 2020 Publication	1%
5	Submitted to University of Hertfordshire Student Paper	1%
6	Submitted to University of Northumbria at Newcastle Student Paper	<1%
7	academic-accelerator.com Internet Source	<1%