Student Performance Prediction Using Artificial Neural network

Submitted By

Amena Akhter Hira
ID: 172-35-2139
Department of Software Engineering
Daffodil International University

Supervised by

Mr. Md. Anwar Hossen
Assistant Professor
Department of Software Engineering
Daffodil International University

This Thesis report has been submitted in fulfillment of the requirements for the Degree of Bachelor of Science in Software Engineering.

Spring – 2021

APPROVAL

This Thesis titled on "**Student Performance Prediction Using Artificial Neural network**", submitted by Am**ena Akhter Hira**, **ID: 172-35-2139** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Imran

Dr. Imran Mahmud
Associate Professor and Head
Department of Software Engineering
Daffodil International University

Chairman

Md Anwar Hossen
Assistant Professor
Department of Software Engineering
Daffodil International University

Internal Examiner 1

Shakir

Asif Khan Shakir
Senior Lecturer
Department of Software Engineering
Daffodil International University

Internal Examiner 2

Professor Dr M Shamim Kaiser
Institute of Information Technology
Jahangirnagar University

External Examiner

DECLARATION

The Thesis report entitled “**Student Performance Prediction Using Artificial Neural Network**” is done under the supervision **Mr. Md. Anwar Hossen, Assistant Professor**, Department of Software Engineering, Daffodil International University. I declare that this report is my original work for the degree of B.Sc. in Software Engineering and that neither the whole work nor any part has been submitted for another degree in this or any other university.

Submitted by

Hira

Amena Akhter Hira
Id: 172-35-2139
Department of Software Engineering
Daffodil International University

Certified by:

Mr. Md. Anwar Hossen
Assistant Professor
Department of Software Engineering
Daffodil International University

ACKNOWLEDGEMENT

I would first like to thank the almighty Allah for allowing us to accomplish this B.SC study successfully. We are really thankful for the enormous blessings of the Almighty Allah has bestowed upon us, not only during our study period but also throughout our life.

I would also like to express my sincere gratitude to my **Supervisor, Mr. Md. Anwar Hossen,** for the continuous support of my Undergraduate thesis study and research. His guidance helped me in all the time of research and writing of this thesis.

Besides my Supervisor, I would like to thank the rest of my thesis committee, **Senior Lecturer, Asif Khan Shakir, Associate Professor, Nusrat Jahan** and our **Department head, Prof. Dr. Imran Mahmud.**

I would also like to thank some of my best friends, for their inspiration, motivation, encourages helped me a lot to complete this thesis.

Last but not the least, I would like to thank my family: my parents for giving birth to me at the first place and supporting me spiritually throughout my life.

ABSTRACT

Background: Education and student both are correlated to each other. Education quality can be improved by student performance. If we want to lead a good life or productive life, then education is necessary and its quality needs to be improved. Performance evaluation of students is necessary for every educational institute in helping their student and teacher. In this study, we aim to predict the student category based on the performance of the student and propose a workflow of web-based four-tier architecture for the student performance prediction. For this purpose, a survey has been conducted on students in different universities in order to collect data and to analyze and predict the student category based on their performance. We proposed a new predictive model for predict student categories based on their performance and how a trained model can learn from real-time data to predict student performance. For categorized the student based on their performance, used multiple classification models using supervised machine learning algorithms. To get optimum features, we applied different data pre-processing techniques. Some supervised learning algorithm work well with all features yet. Each of the student category categorized by considering the top features. The analysis results indicate that we got the highest performance by using the Decision Tree Classifier (DT) by using op 10 features of Extra Tree Classifier Algorithm and XGBoost show the best performance with Chi-Square Feature Selection technique and other optimum selection features.

Keywords: machine learning, supervised learning, decision tree classifier, XGBoost, extra tree classifier, Chi-Square, artificial neural network**;**

Table of Contents

CHAPTER 1

INTRODUCTION

1.1 Background

A student's performance defines a student's academic achievement and his/her responsibility for the study. Another side, improve education quality depends on students' performance and their education ability. Drop out is one of the main reason for the decrease in education quality. Many of the students drop out at many education levels but most of the drop out happen at secondary or tertiary level. In 2018, students of the ages 16 to 24 are dropout almost 2.1 million and the overall dropout status rate was 5.3 percent [1]. In our country, 48% of students drop out at secondary cycles and 25% of students drop out at higher secondary cycles [2]. Some programs search right students with the right message who is more relevant to the program. In paper [3], the Author used a machine learning approach to predict the dropout student and prioritize students who may be at risk of not graduating on time, and suggest a particular student for going off-track. They also predict student retention and give a student ranked list based on the possibility of dropout rate. For this reason, they adapted the ML techniques such as Random Forest Search, Logistic regression. Student performance prediction is one of the best emphasis for reducing the dropout rate and failing issues. The student dropout rate is regarded as a negative factor that impacts the quality of education. So, analyzing the student performance can help in this matter. ANN is coming up to human biological inspiration, so it has been more effective for clustering and classification tasks. As well it gives better accuracy compare to another Machine Learning techniques [4].

Therefore, Data collected through a survey questionnaire from different universities inspired by the Kaggle dataset. Features like institutional name, nationality, current result cgpa, time of group study, absent in a semester, amount of drop semester, parents satisfied with result, due amount, etc. etc. added to the dataset. After collecting the data, used binning Machine Learning technique to label the data. And use data pre-processing techniques to convert categorical features to numerical features. After that, we are run many feature reducing technique such as chi 2 and Extra Tree Classifier which show that institutional name is less important for a categorized student model. Those techniques help to reduce the feature and help to increase accuracy in the model. So finally, we get a dataset with 17 attributes from which the top of 10 important features have been selected after encoding so that we could be able to compare the effectiveness of the attributes on the accuracy of the model. Although, data has been processed and initial features were selected to create training and test sets. After that we run different supervised machine learning algorithm e.g. XGBoost, Decision Tree Classifier and ANN [5]. XGBoost show the best accuracy with the missing values. XGBoost is one of the best model of ensemble method because of it can handle both data characteristics [6]. Ensemble learning methods are an extremely effective model for training on student performance and increase the predictive accuracy of student performance [7]. Therefore, XGBoost algorithm can be applied to provide a prediction model so that the student will be categorized in level [high, medium, low] based on their performance. And this predicted model trains on predicted store data after a period of time for increasing the model accuracy level.

1.2 Motivation of the Research

Improving the quality of education has long been of interest topic to academicians. The dropout rate in our country has increased day by day. So its impact on the quality of education and on a better life and on building a better country. Measuring academic performance and classifying it on the basis of their performance is a challenging subject because student performance is a product of socio-economic, physiological and environmental factors. The main motivation of the research comes from my real-life experience and the previous research work of my university senior brother. In my batch, I have seen many students who didn’t continue their graduate study but are active in 1st year then after day by day they fall up and don’t overcome their position. So, I wanted to explore them through their performance and suggest to them what’s their position with the categorization level. It will be effective for our teacher to decide which students need counseling for retention of their study cycle.

As an intermediate Data Scientist, I wanted to help improve the quality of education by categorizing students based on their performance with a predictive model that can accurately classify student level (High, Medium, Low) and increase student retention capacity of any institute with the remedy solution. Also, a predicted model how to store the predicted data for training purpose and after a period of time predicted model trained on that stored data and increase the accuracy of the predicted model.

1.3 Problem Statement

Student dropout is a problem that impacts all kinds of educational institutes and reduces the quality of education. Students are the valuable assets of any educational institute. But when they didn't continue their study with their career course that time it impacts on social economic. According to a 2019 report of Bangladesh Bureau of Educational Information and Statistics (BANBEIS), there are 17.9% of students' dropout at the primary level, and 36% of students' dropout at the secondary level. Alongside, some students study with their choosing subject but when they continue their studies on their subject they feel that it's not too relevant for their career. So after that, they changed their subject or lost their interest in that subject and they just disappear slowly which gradually increases the dropout rate. A pretend model's accuracy didn't increase after making the model, always the model accuracy stands at a point. Thus why, sometimes a model prediction didn't give the right prediction, because the model predicts some confusion value which clearly visualizes with a confusion matrix.

1.4 Research Questions

Question 1: The main question that this research revolves around is, how a trained model can learn from real time data to predict the level of students and increase the accuracy of the model?

Question 2: What are the key factors that lead to classify the students based on their performance?

Question 3: Can ML model accurately predict the Student Classify Level and will it help to decrease the dropout rate?

1.5 Research Objectives

The main objective of this research is to store the prediction data which's predict the model and train again the model with the stored data after a period of time and increase the accuracy of the model.

The next objective is to perform different data pre-processing techniques and make the dataset clean and tidy. Also, find out the best features to predict student classification and which feature selection technique suggest the top of the features among the dataset.

The last objective is to make an ML model and evaluate with optimum dataset and tuning model parameters until the maximum accuracy model is found.

1.6 Research Scope

The Scope area of this research is under the survey of student performance dataset. This research explores the best key factors to classify the student level and tries to visualize all the key factors through different modern plots/graphs. This research also pulls out the best predictive model for classifying the student level by several parameter tuning and evaluation steps. Finally, this research performs how to a prediction model trained with predicted store data and increase the accuracy model.

1.7 Thesis Organization

This thesis contains of five different section, the first section presented the whole topic and work of this thesis. In the second section, we reviewed some of the literature that are related to this thesis. Section 3, describes the methodology, different pre-processing steps and visualize the key factors of student classification. In section 4, we put the experiments result and discussion. And finally, conclusion described in section 5.

CHAPTER 2

LITERATURE REVIEW

A topic that has long been interesting for academicians and practitioners in student performance prediction with classification. The reason for the increasing dropout and reduces education quality. Analysis of student Performance can help both teacher and student. Teachers can counsel student based on student performance and improve their performance in examinations. Also, it can be helped to examination department to choose appropriate faculty and teaching meditation to enhance student education results (DR. R Senthil Kumar, Jithin Kumar.K.P, et al. 2018) [8]. They use the MapReduce MongoDB framework model to analyze student performance. Key factors are a supreme priority to analyze student performance. Analyze student performance models provide the finest accuracy depends on foremost key factors (Syed Tahir Hojazi et al. 2016) [9]. They offer a hypothesis and test their hypothesis with simple linear regression. Intercession can help those that are falling behind their educational goals, but given limited resources, such programs must specialize in the proper students, at the proper time, and with the proper message. (Everaldo aguiar, Himabindu Lakkaraju et al. 2015) [10] They represent an incremental approach which will be wont to select and prioritize students who could also be in danger of not graduating on time and to advise what could also be the predictors of specific students going off-track. Analyze student performance differ to online based education and offline-based education. (P.Cuijpers, A. Van den Beemt et al. 2018) [11] They are using correlations, multiple regressions, and process mining to analyze student performance in a blended MOOC. Many research works are observed student performances, to enhance their grades and to prevent them from throwing in the towel from school by employing a data-driven approach. Most of the authors used a data-driven approach to speculate student performance (Jie Xu, Kyeong Ho Moon et al. 2017) [12]. The ensemble method helps to provide the best accuracy of the model which predicts student performance on their academic history. Student behavior and their academic achievement relationship is a strong relationship that proved by Bagging boosting and Random Forest algorithm (Elaf Abu Amrieh, Thair Hamtini et al. 2016) [13]. They are using Data mining to select best key features with data mining. Their proposed model gives the best accuracy with behavioral features to achieve up to 25.8% from 22.1%. The machine learning model evaluates 3 or 4 evaluation metrics, which isn't proved the particular performance altogether the case. As an example, the performance of the model can't judge with the accuracy in the imbalance dataset problem. To compare the performance of supervised learning methods are used to evaluate the performance with a large number of performance evaluation metrics. [14](Şeyhmus Aydoğdu et al. 2019) The author detects that number of attendance, time spent in the course material, number of attendance to archived courses are contributed most to predict student performance with ANN. It is difficult to predict how much samples variables in ANN contribute to predict output target. Therefore, Feature Selection Technique will be help to predict the best features which are relevant to predict the performance of student. Evaluation matrix can help to compare the model performance to another model and select the best prediction model. From this study, we used an evaluation matrix to evaluate the model's performance more precisely.

CHAPTER 3

METHODLGY

The proposed model contains seven steps that start from collecting raw data and ends with highest accuracy model and show prediction result. At first, we collect the raw data with some questionnaires answer from different universities. After getting data, we run the data pre-processing technique for obtaining clean and tidy data. And then we find out the best features to lead the student classification.

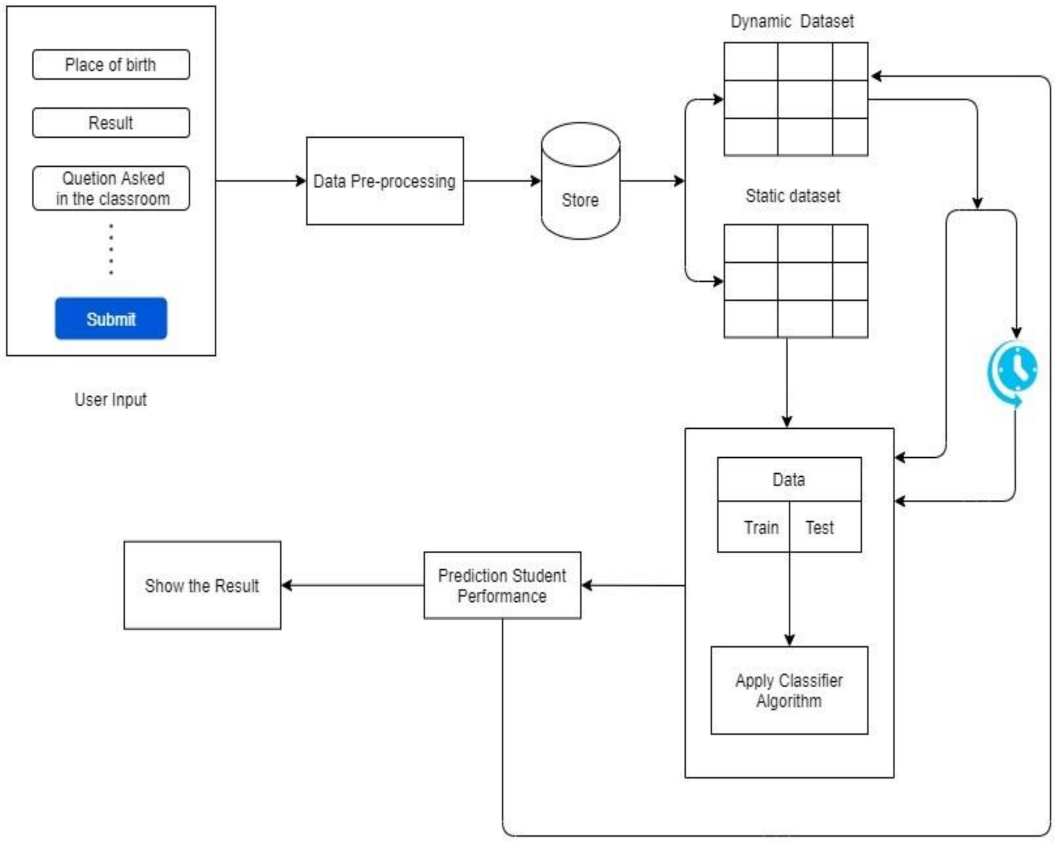


Fig: 3.1: Proposed model Workflow.

Subsequently, we select four different supervised classification algorithms – XGBoost, Decision Tree, Artificial Neural Network, Polynomial regression for solving the student classification problem. Different feature selection technique conveys optimum features, reduced features with different standpoints. Afterward, we select 10 best features from those feature selection techniques and fit with the different classification models which show different accuracy on the training set. After getting accuracy, we plot a confusion matrix for visualization. Then, we evaluate the model with different evaluation matrix. Then all models tested with test data which are independent test records and compare their accuracy and select the prediction model which is giving the best accuracy. The proposed model workflow represents in **Figure 3.1.**

3.1 Data Pre-Processing

In this study, we used the collected dataset from different universities with some questionnaires answer. This dataset Contain 19 attributes and 740 observations. The data we have collected, needs some preprocessing before moving further.

Such as, both the features and labels contained categorical data and we have handled that by encoding both. We applied total six different data pre-processing technique to make our data optimum. The pre-processing steps are shown in the **Figure 3.2,** and describe step by step in sequence.

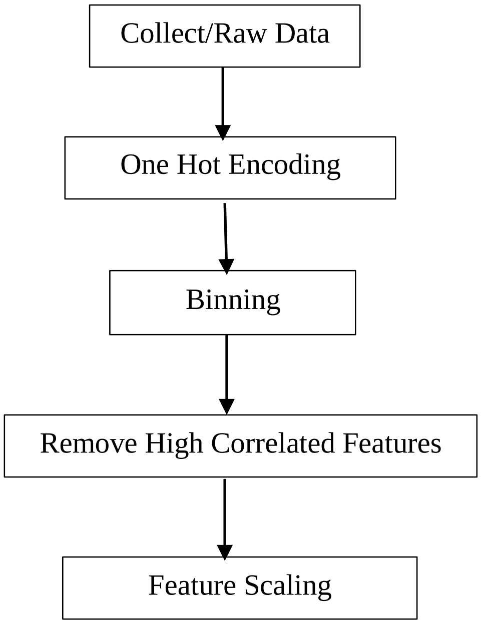


Figure 3.2: Data Pre-Processing Steps

3.1.1 Dataset Collecting

In this study, we collect the dataset with some questionnaires survey from different universities. This dataset contains 740 student's record with 19 attributes.

Attributes	Type	Description
gender	Categorical	male, female
nationality	Categorical	Bangladesh, Somalia
Place_of_birth	Categorical	
department	Categorical	SWE,CSE,MCT, ESDM
current_year	Categorical	1^{st}, 2^{nd}, 3^{rd}
time_of_group_study	Numerical	
absent_in_a_semester	Numerical	
last_semester_result	Float	
current_result_cgpa	Float	
question_ask_in_class	Categorical	Yes, No
amount_of_drop_semester	Numerical	
use_additional_course_material	Categorical	Yes, No
drop_reason	Numerical	Yes, No
meet_with_advisor	Categorical	Yes, No
absent_in_a_semester	Numerical	
due_amount	Categorical	
parents_satisfied_with_result	Categorical	Yes, No
fathers_education	Categorical	
mothers_education	Categorical	

3.1.2 One Hot Encoding

Categorical variables are qualitative data in which the values are assigned to a set of distinct groups or categories such as drop_reason, due_amount, institutional_name and 14 other features are in dataset. These variables are stored as categorical string value which represent various value of data. Some instance includes drop_reason('personal', 'financial', 'irregularities', 'depression' 'poor_academic_result', 'drop_reason_0'), gender(male, female). In ML, there have different approach available to deal with categorical variables. In this study we used one hot encoding to allocate with categorical variables. One hot encoding converts categorical variables into dummies variables (Scikit-learn.org, 2018). For example: drop_reason feature has 6 different values: ‘personal’, ‘financial’, ‘irregularities’, ‘depression’ ‘poor_academic_result’, ‘drop_reason_0’. After applying one-hot encoding, values will encode into numbers like this:

drop_reason	drop_reason_personal	drop_reason_financial	drop_reason_poor_academic_result	drop_reason_irregularities	drop_reason_depression	drop_reason_0
personal	1	0	0	0	0	0
financial	0	1	0	0	0	0
poor_academic_result	0	0	1	0	0	0
_irregularities	0	0	0	1	0	0
_depression	0	0	0	0	1	0
NO	0	0	0	0	0	1

3.1.3 Binnig

Binnig method is used for data smoothing and sorted value of data by consulting its “neighborhood”, that is, the values around it. We are collected raw data with some questionnaires answer from different universities. But there has no label in the dataset, so we used binning method for labeling the data and categorized the data based on students’ given information. In the binning method, we retention a range based on the current result cgpa with a categorical name. After getting the categorical name with labeling and fixed them into a target variable. For instance: the current result cgpa contains 0 to 4.0 floating values, at first kept a range with these values such as 2.5-3.0 = “Low”, 3.0-3.5 =” Medium”, 3.5-4.0 =” High” and other remaining values goes to Low categorical value.

3.1.4 Removing Highly Correlated Feature

Correlation defines how one or more variables are co-related to each other. In statistics, it defines how one variables changes depended on the other variable. So it’s a mutual relation for two or more variable which are contains categorical or numerical value. To measure, how strong the relationship between two variables. We use Correlation and Pearson’s Correlation (Pearson,1920) method. The PCC equation is –

Equation 3.1:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,

r = Pearson Correlation Coefficient

x_i = x variable samples

y_i = y variable sample

$\bar{x}$ = mean of values in x variable

$\bar{y}$ =mean of values in y variable

In PCC, values of ‘r’ ranges from -1 to +1 where 0 means there is no relation between the two variables. Positive value means if one variables value will increase then other variable value will increase. And negative value means opposite of positive values such one variable value will increase then other variable value will decrease. So we find highly correlated features in our dataset after calculating PCC Table 3.1, where the Correlation value range between 0.8 to 1.0.

Table 3.1: Feature with Highly Correlated value

Feature Name	Correlated with	r
last_semester_result	Current_result_cgpa	0.83

3.1.5: Feature Scaling

Feature Scaling is a technique to standardize the independent features present within the data in a fixed range. It applied during the data pre-processing to handle highly varying frequency or values or units. To overcome this problem, we need to bring the varying features to the same level of magnitude. If feature scaling isn't applied, then a Machine Learning algorithm show low accuracy compare to feature scaling applied dataset. There are two common methods are available to perform best Feature Scaling –

- Standardization
- Min Max Scaler

In this study, we used standardization feature scaling technique to select our features and standard our dataset. Standardization is a very effective technique to rescale features value so that it has distribution with 0 mean value and variance equals to 1.

Before Feature Scaling: Figure 3.3, boxplot represent us the magnitudes of each feature before standardization.

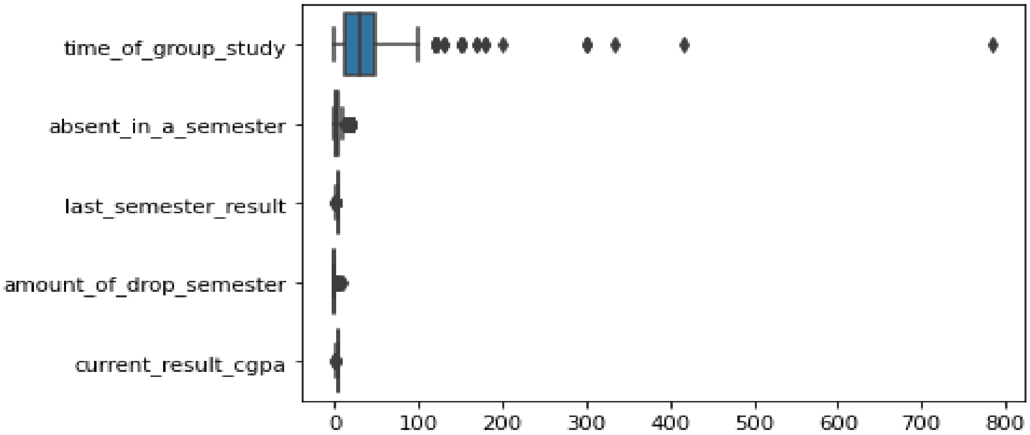


Figure 3.3: Features magnitudes before Scaling

After Feature Scaling: Figure 3.4, boxplot represent us the magnitudes of each feature after standardization.

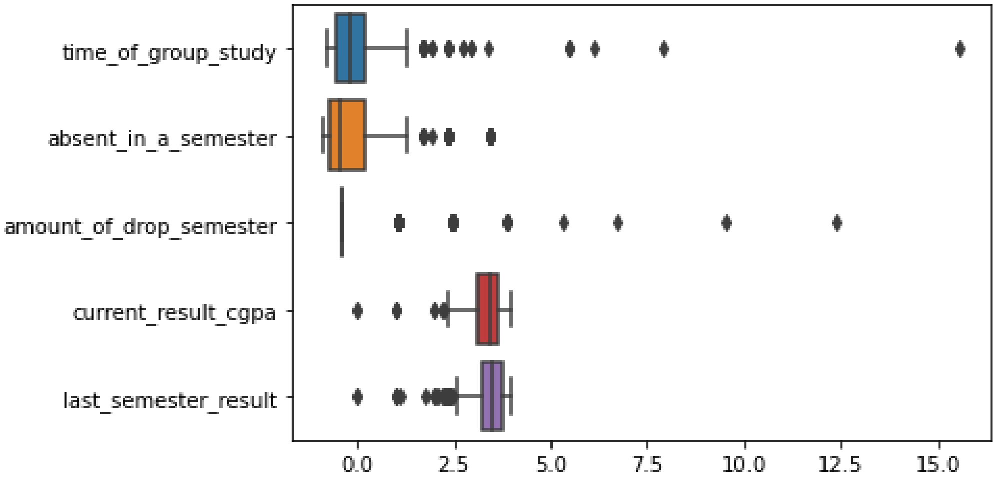


Figure 3.4: Features magnitudes after Scaling

3.2 Key Factors of Student Classification

We discover some of the key factors that are highly responsible for classification student.

Time of group study: According to Chi2 Square and Extra Tree Classifier model feature importance (Figure 3.5), time_of_group_study has the highest effect on Student Classification. (Figure 3.5) , present that student with time_of_group_study of 0-200 hours has the highest performance rate.

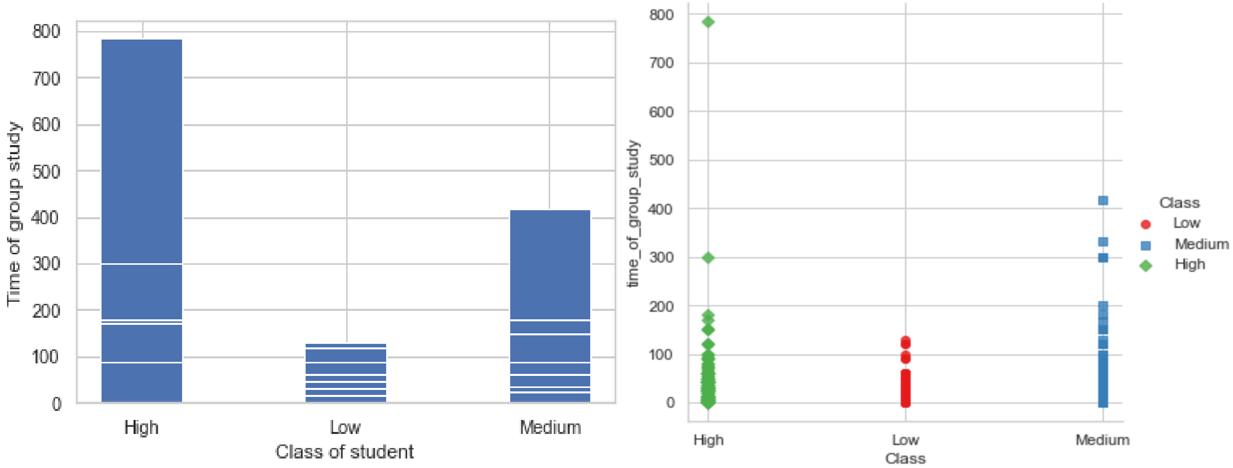


Figure 3.5: Time of Group Study plot

Absent in a semester: According to Chi2 Square and Extra Tree Classifier model feature importance (Figure 3.6), absent_in_a_semester has the highest effect on Student Classification. Figure 3.6, present that medium students of dropout rate much more than low students of drop out rate.

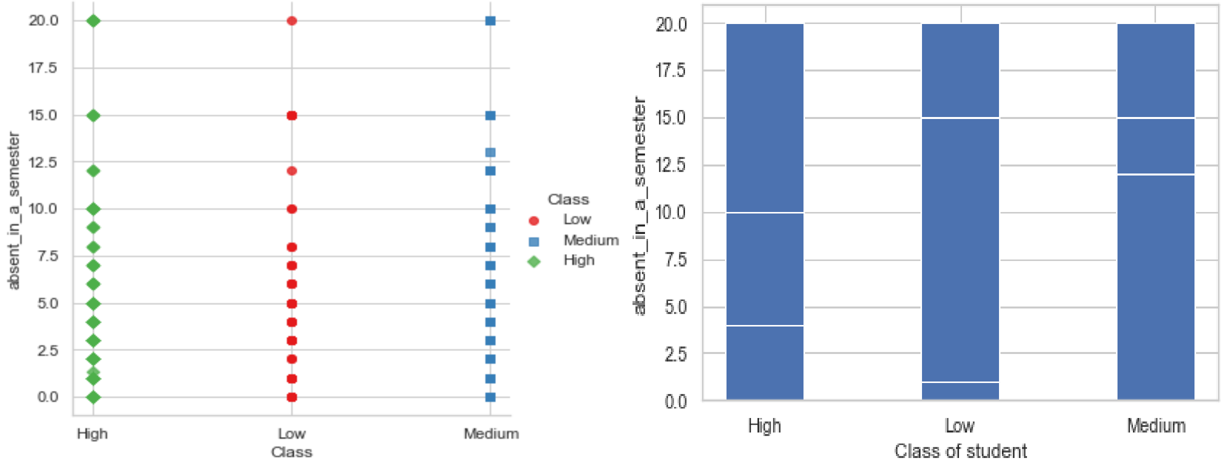


Figure 3.6: Absent in a Semester plot

Amount of drop semester: In this dataset, highest CGPA (CGPA 4.00) student holder didn't drop out the semester. It happened with them rarely like it's an exceptional case. There are 2 students are dropout which is cgpa range (3.5 to 4).

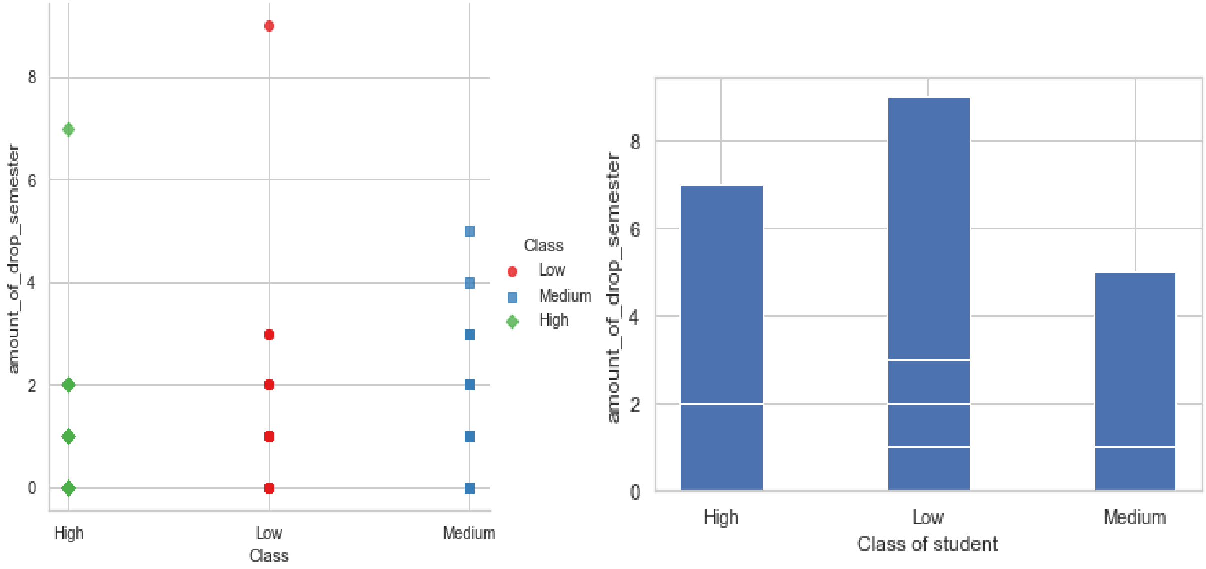


Figure 3.7: Amount of Drop Semester plot

Parents satisfied with result: (Figure 3.8), Shows that parent satisfied with student's result ratio are high when the student's CGPA is good. But when Student's result fall that time parents satisfaction result is low. So it's like a linear relation which is show that if one is increased than other variable will be increased and also it happens on opposite side.

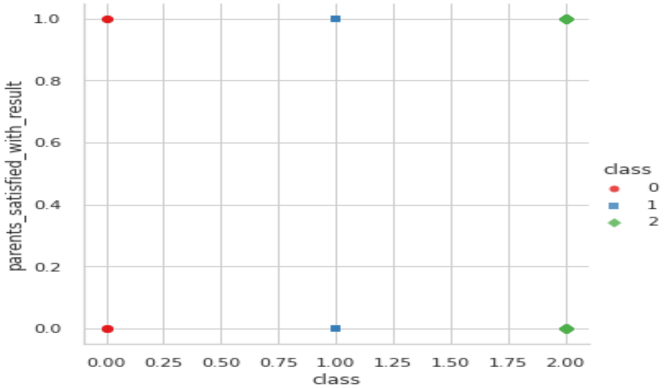


Figure 3.8: Parents satisfied with result plot

Current result CGPA: Chi2 square and Extra Tree classifier model feature importance shows that current result cgpa one of the importance feature for student classification. Those students are high categorical student which is CGPA is high and those students are low categorical which is CGPA is low.

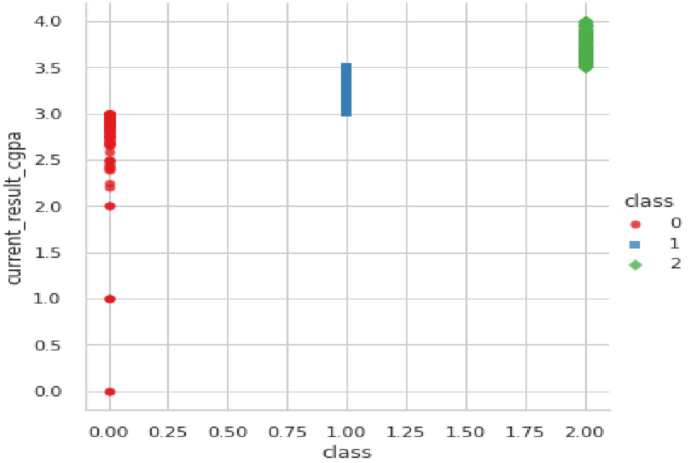


Figure 3.9: Current result CGPA plot

Drop Reason: (Figure 3.10), shows that financial drop reason is also there to all categorical students. Because of in Figure 3.7 shows that some High categorical students give drop out in any semester for any reason.

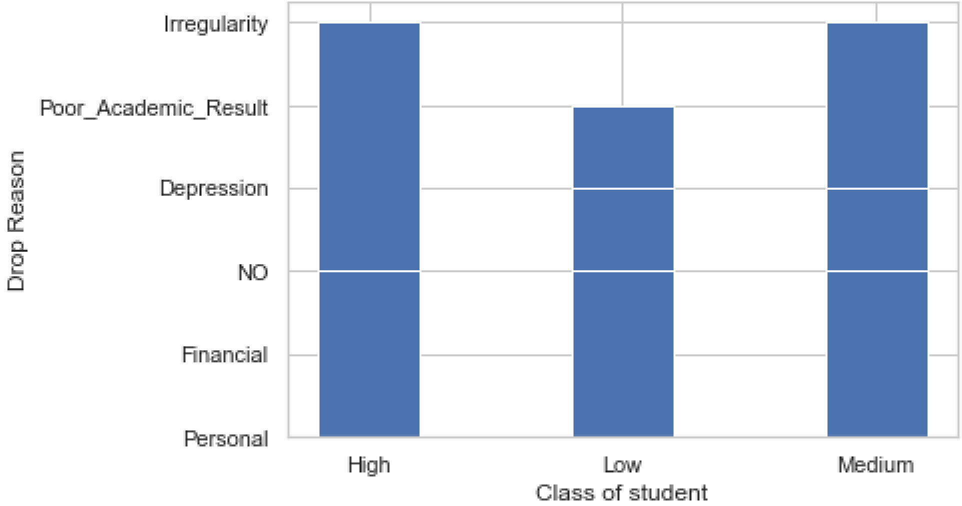


Figure 3.10: Drop Reason plot

3.3 Optimum Feature Stand point

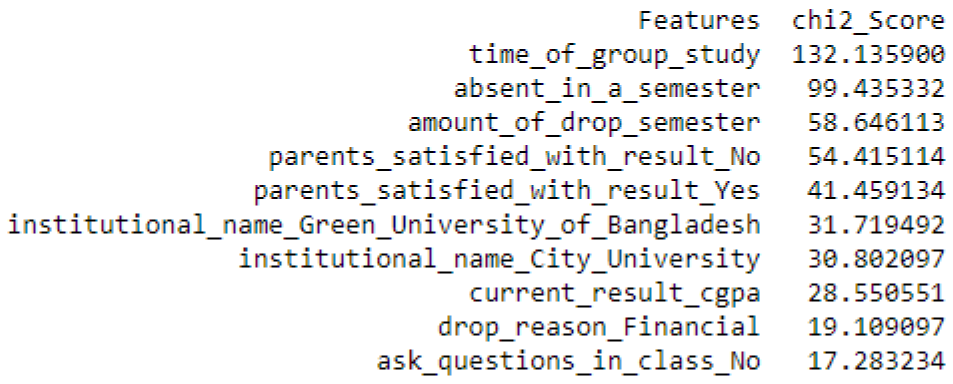
We were applied 4 different data pre-processing steps that represent in Section 3.1 and Table 3.2, the left column describes the optimum features list.

3.4 Feature Importance

Feature Importance describe a score for each feature of dataset. The highest score feature more important or relevant feature towards with output feature. Feature importance reduce overfitting and less opportunity to make decision based on noise. It reduces algorithm complexity and make faster to train model. Thus why, model accuracy improves and less misleading data.

Chi-Square

When the feature is categorical Chi-Square is to be used. Chi-Square show the feature importance ratio and measures the degree of association between two categorical variables [15].

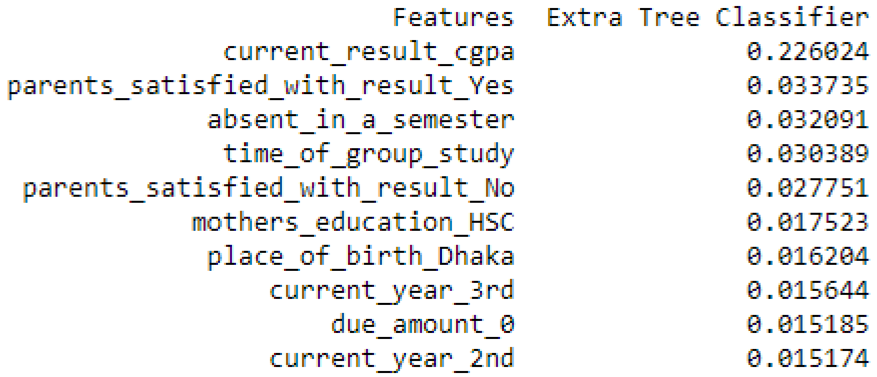


```
                                   Features   chi2_Score
                         time_of_group_study  132.135900
                        absent_in_a_semester   99.435332
                      amount_of_drop_semester   58.646113
             parents_satisfied_with_result_No   54.415114
            parents_satisfied_with_result_Yes   41.459134
institutional_name_Green_University_of_Bangladesh   31.719492
          institutional_name_City_University   30.802097
                          current_result_cgpa   28.550551
                        drop_reason_Financial   19.109097
                     ask_questions_in_class_No   17.283234
```

Figure 3.10: Features score of Chi-Square

Extra Tree Classifier

Extra Tree Classifier is an ensemble machine learning algorithm. It’s very closer to RF and only differs from RF to the construct of the DT in the forest. In feature selection, it performs using the forest structure and normalized total reduction in the mathematical criteria. The most important variable score determines with the output label.



```
                         Features  Extra Tree Classifier
              current_result_cgpa               0.226024
parents_satisfied_with_result_Yes               0.033735
             absent_in_a_semester               0.032091
              time_of_group_study               0.030389
 parents_satisfied_with_result_No               0.027751
            mothers_education_HSC               0.017523
             place_of_birth_Dhaka               0.016204
                current_year_3rd               0.015644
                    due_amount_0               0.015185
                current_year_2nd               0.015174
```

Figure 3.11: Features score of Extra Tree Classifier

3.5 Predictive Algorithms

3.5.1 XGBoost

Extreme Gradient Boosting (XGBoost) is an implementation of Gradient Boosting Decision Tree. It is an additive tree model means it add new trees that complement the already build ones. It is highly scalable end to end tree boosting framework (Chen, et al., 2016). Compared to other gradient boosted machines, it uses a more regularized- model formalization to control over-fitting, which gives it better performance. It supports various objective functions, including classification, regression and ranking. Feature selection is not necessary when we use XGBoost because of it can andle both data characteristics and missing values. [16] During training period of XGBoost, good features would be chosen as node in trees, which means features not used are abandoned.

3.5.2 Decision Tree

Decision Tree classifier is one of the predictive modeling approaches used in Classification (Leo Breiman et al 1984). Decision tree used simple decision rules for learning. It's a supervised learning algorithm and solved classification problem. It starts from the root of the tree for predicting a class label for a record. It uses multiple algorithm to decide to split a node into two or more sub-nodes. It detects the feature which is more relevant to classify the label or target variable and make node of the tree with that feature variable. Then it splits the node on all available variables and select the split sub nodes. Thus it continues can't further classify the nodes and called the final node as a leaf node. DT want less requirement of data cleaning compared to other algorithms [17].

3.5.3 Artificial Neural Network

Artificial Neural Network is formed from Hum Biological Neural Networks that develop the structure of a human brain. ANN develop to follow the concept of How human brain's neurons are interconnected to one another. ANN also have neurons that are interconnected to another and have various layers. In ANN, neurons are known as nodes and dendrites represent inputs and axon represents outputs. ANN is the best to represented as a weighted directed graph. For using ANN, we need to use Keras API [18]. ANN receives the input signal from the external source like dataset. Then inputs are then mathematically assigned by the notations x(n) for every n number of inputs which are multiplied by its corresponding weights. Afterward, if the weighted is equal to zero then bias is added to make output non-zero. Then the total of weighted inputs is passed through the activation function. In Ann, using the final linkage weight score the activation rate of the output nodes. And each epochs its reduce loss function with increase accuracy score. Thus why, it gives the best accuracy compare to other classification algorithm.

3.6 Performance Evaluation Metrics

Performance evaluation metrics shows the evaluation of the model accuracy. So, performance evaluation evaluates after doing data pre-processing and implementing several models. Getting output after implementing several models forms of a probability or a class. The next step is to detect how effective is the model based on some metric using test datasets. To evaluate performance of different Machine Learning models with different performance metrics. There are a large number of evaluation metrics are available to evaluate and compare the performance of Supervised Machine Learning models. In this study, we used the different classification based evaluation metrics to evaluate performance of the model such as Accuracy, Precision, Recall, F1, MCC.

3.6.1 Accuracy

Accuracy score is a faster way to evaluate a set of predictions on a classification problem. It is describing the ratio of classification accuracy with the number of correct prediction out of all predictions that were made. The formula of calculating a classification problem shown below:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

3.6.2 Precision

Precision is a useful in cases where False Positive is a higher concern than False negatives. It describes how many of the correctly predicted cases actually turned out to be positive.

$$Precision = \frac{TP}{TP + FP}$$

3.6.3 Recall

Recall is a useful metric in cases where False Negative to False Positive. It describes how many of the actual positive cases we were able to predict correctly with our model.

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

3.6.4 F1-Score

F1 score is a harmonic mean of Precision and Recall thus why it gives a combined score about these two metrics. When Precision is equal to Recall, it is maximum.

$$F1 - score = \frac{2}{\frac{1}{Recall} + \frac{1}{Precision}}$$

3.6.5 MCC

Matthews Correlation Coefficient get true and false positive and negative and there have a range between -1 to 1 where -1 defines a complete wrong binary classifier and 1 defines a completely correct binary classifier. The formula of MCC:

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

CHAPTER 4

RESULTS AND DISCUSSION

In this study, we applied total 4 experiments with 4 different standpoints. In this section. We describe all the experiments results and describe briefly. Experiment 1and 2 are based on sequentially Chi Square and Extra Tree classifier highest score based features. Then we experiment 3 & 4 are based on optimum features, reduced features respectively.

4.1 Experiment 1

Experiment 1, is based on all feature in dataset which get after data pre-processing (Section 3.1). Here, we inquire all features dataset and make 3 different models: XGBoost, DT, ANN. After making those models where calculate the performance through several evaluation metrics describe in Section 3.6.

Evaluation Metrics	**ANN**	**DT**	**XGBoost**
Accuracy	0.55	0.53	0.57
Precision	0.48	0.36	0.55
Recall	0.48	0.41	0.48
F1 Score	0.47	0.38	0.48
MCC	0.26	0.17	0.27

4.2 Experiment 2

Experiment 2, is based on optimum feature which are select by Chi-Square Feature Selection Technique. We select the top 10 features which are highest scorer at chi-square feature selection technique. Here, we conduct optimum dataset listed in Table 3.2 and make 3 different models: XGBoost, DT, ANN. After making those models where calculate the performance through several evaluation metrics describe in Section 3.6.

Evaluation Metrics	**ANN**	**DT**	**XGBoost**
Accuracy	0.47	0.52	0.54
Precision	0.39	0.44	0.58
Recall	0.40	0.42	0.47
F1 Score	0.39	0.40	0.48
MCC	0.12	0.17	0.22

4.3 Experiment 3

Experiment 3, is based on optimum feature which are select by Extra Tree Classifier Feature Selection Technique. We select the top 10 features which are highest scorer at Extra Tree Classifier feature selection technique. Here, we conduct optimum dataset listed in Table 3.2 and make 3 different models: XGBoost, DT, ANN. After making those models where calculate the performance through several evaluation metrics describe in Section 3.6.

Evaluation Metrics	**ANN**	**DT**	**XGBoost**
Accuracy	0.52	0.55	0.53
Precision	0.47	0.37	0.47
Recall	0.43	0.43	0.45
F1 Score	0.44	0.40	0.45
MCC	0.17	0.22	0.21

4.4 Experiment 4

Experiment 4, is based on reduced feature and selected common features with those techniques. Here, we conduct optimum dataset listed in Table 3.2 and make 3 different models: XGBoost, DT, ANN. After making those models where calculate the performance through several evaluation metrics describe in Section 3.6.

Evaluation Metrics	**ANN**	**DT**	**XGBoost**
Accuracy	0.57	0.53	0.58
Precision	0.57	0.36	0.61
Recall	0.49	0.41	0.47
F1 Score	0.50	0.38	0.47
MCC	0.26	0.17	0.27

4.4.1 Accuracy & Loss Graph of ANN

ANN model show the best accuracy than the other experiment of ANN model in Experiment 4. Accuracy Graph of ANN present that train accuracy increase with increasing iteration but test accuracy decrease with increasing iteration. It means that there have overfitting, because ANN model fits with the train dataset but has poor fit with test dataset and does much better on the train dataset than on the test dataset. Therefore, find out the overfitting reason and improve the model.

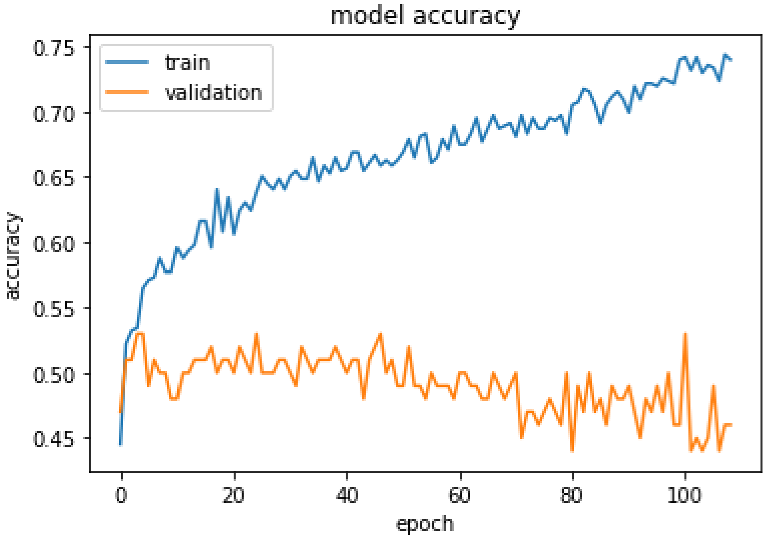


Figure 4.4.1: Accuracy Graph of ANN

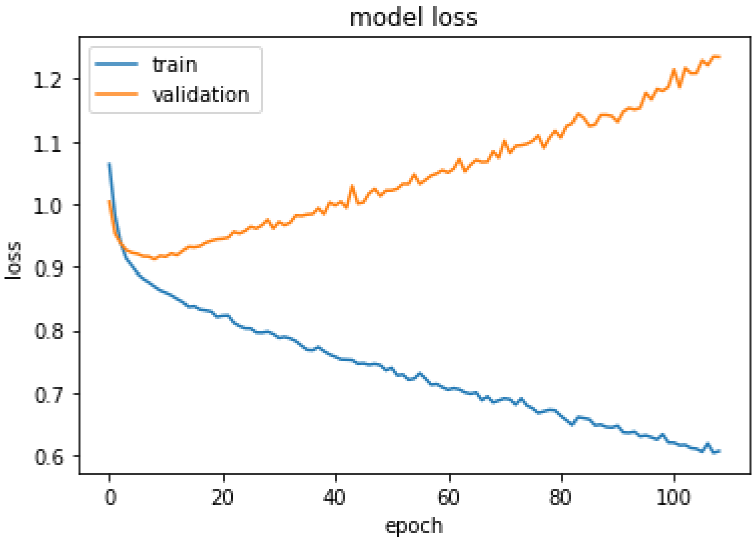


Figure 4.4.2: Loss Graph of ANN

To conclude, we can say ANN model perform best with common (both Extra Tree Classifier & Chi-Square) features but model test loss value increase with increasing iteration and accuracy decreases with increasing iteration. In ANN & Decision Tree give prediction accuracy most similar but Decision Tree model give the best accuracy in Experiment 3. Another side, XGBoost gives the best accuracy to compare to other models with 3 experiments e.g. All features, Common features, and Chi-Square features.

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

5.1 Findings and Contributions

In this study, we have presented 4 different supervied learning based machine learning technique on a real dataset of students and start our work with an imbalance, noisy dataset which we have gotten from survey questionnaires answer. Then we applied 4 different data pre-processing techniques and make the dataset clean and tidy and getting optimum feature from dataset. We have gotten some side result which are giving information without predict. Such as, time_of_group_study feature displays that some of students didn't study with group any day, therefore we can say that, those students aren't qualified for team work. Another side, several students didn't drop semester without any reason so it will be their personal or financial reason. We clustered the data based on current_result_cgpa feature with the range and category e.g. High, Medium, Low. We applied different feature importance techniques to select best features which are more highly correlated with the label or target variable. We discover true key factors and reasons of student classification. Four experiment was leaded in order to find the best model with highest accuracy. We presented a comparison analysis in order to select which classification algorithm is more efficient for a 19 dimensional dataset from Experiment 2 and 3. Evaluate model performance with different evaluation matrices, we represent result in section 3. In Experiment 1,2,3 &4, XGBoost and Decision Tree model gives the highest results among all other models with optimum features. But those model are find out the best node for labeled the data based on information or given data. Also, those model are classified target variable based on one features samples which is Current Result CGPA. On the other side, DT model perform 55% with Extra Tree Classifier features (best 10 features) and show best result with testing set. Because it is not take specific one feature. XGBoost model perform best on of the three experiments e.g. All features, Common Features (Chi Square-Extra Tree Classifier), Chi-Square Features. Therefore, we can say that XGBoost model will be performed best on real world to decrease the dropout rate and it will be train on stored data which model already predicted on real-time data. So, model accuracy will be increase after train with store data a period of time.

5.2 Recommendations for Future Works

In this dataset, we collect a student’s academic information with their parent information. Therefore, we can call this dataset real-world dataset, and thus why we can predict or work on the different side of student academic performance with this dataset. In this dataset, there have all features to need to predict the dropout rate and classify the drop reason. So, in the future we will be worked on drop out problem and dropout reason problem can be predicting with this real-world dataset which also helps to reduce dropout rate and improve education quality.

References

[1] "Dropout rates | National Assessment of Education Progress (NAEP)," [Online]. Available: https://nces.ed.gov/fastfacts/display.asp?id=16. [Accessed 30 05 2021].

[2] "Higher Secondary dropout rates | Bangladesh Bureau of Education Information and Statistics (BanBEIS).," [Online]. Available: http://data.banbeis.gov.bd/images/4ce.pdf. [Accessed 30 05 2021].

[3] H. L. E. A. Nasir Bhanpuri, "Who, When, and Why: A Machine Learning Approach to Prioritizing Students at Risk of not Graduating High School on Time," in *Fifth International Conference on Learning Analytics and Knowledge*, 2015.

[4] M. S. K. M. K. I. A. K. D. K. M. M. M. A. Md. Haider Ali, "A Reduced Feature Based Neural Network Approach to Classify the Category of Students," in *ICIAI '18: Proceedings of the 2nd International Conference on Innovation in Artificial Intelligence*, 2018.

[5] J. M. ,. J. K.M. Mohiuddin, "Artificial neural networks: a tutorial," 1996.

[6] M. A. F. Hendri Murfi, "The Accuracy of XGBoost for Insurance," *International Journal of Advances in Soft Computing and its Applications 10(2):159-171,* 2018.

[7] M. K. A. Souhaib Aammou, "Enhancing the prediction of student performance based on the machine learning XGBoost algorithm," *Interactive Learning Environments,* 2021.

[8] D. S. K. Jithin Kumar.K.P, "ANALYSIS OF STUDENT PERFORMANCE BASED ON CLASSIFICATION AND MAPREDUCE," *International Journal of Pure and Applied Mathematics,* 2018.

[9] S. T. H. S.M.M. Raza Naqv, "FACTORS AFFECTING STUDENTS' PERFORMANC," 2016.

[10] R. G. B. M. S. A. L. Kecia L. Addison, "A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcomes," 2015.

[11] A. V. d. B. C. P. Cuijpers, "Predicting student performance in a blended MOOC," 2018.

[12] J. X. Kyeong Ho Moon, "A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs," 2017.

[13] T. H. ,. I. A. Elaf Abu Amrieh, "Mining Educational Data to Predict Student's academic Performance using Ensemble Methods," 2016.

[14] Ş. Aydoğdu, "Predicting student final performance using artificial," 2019.

[15] "Sklearn.feature selection.chi2. (n.d).," [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.chi2.html. [Accessed Retrieved 10 October 2020].

[16] "XGBoost Documentation," [Online]. Available: https://xgboost.readthedocs.io/en/latest/. [Accessed Retrives 2020].

[17] P. P. Vivek Kumar Sharma, "A Decision Tree Algorithm Pertaining to the Student".

[18] "Keras," [Online]. Available: https://keras.io/.

Appendix – A

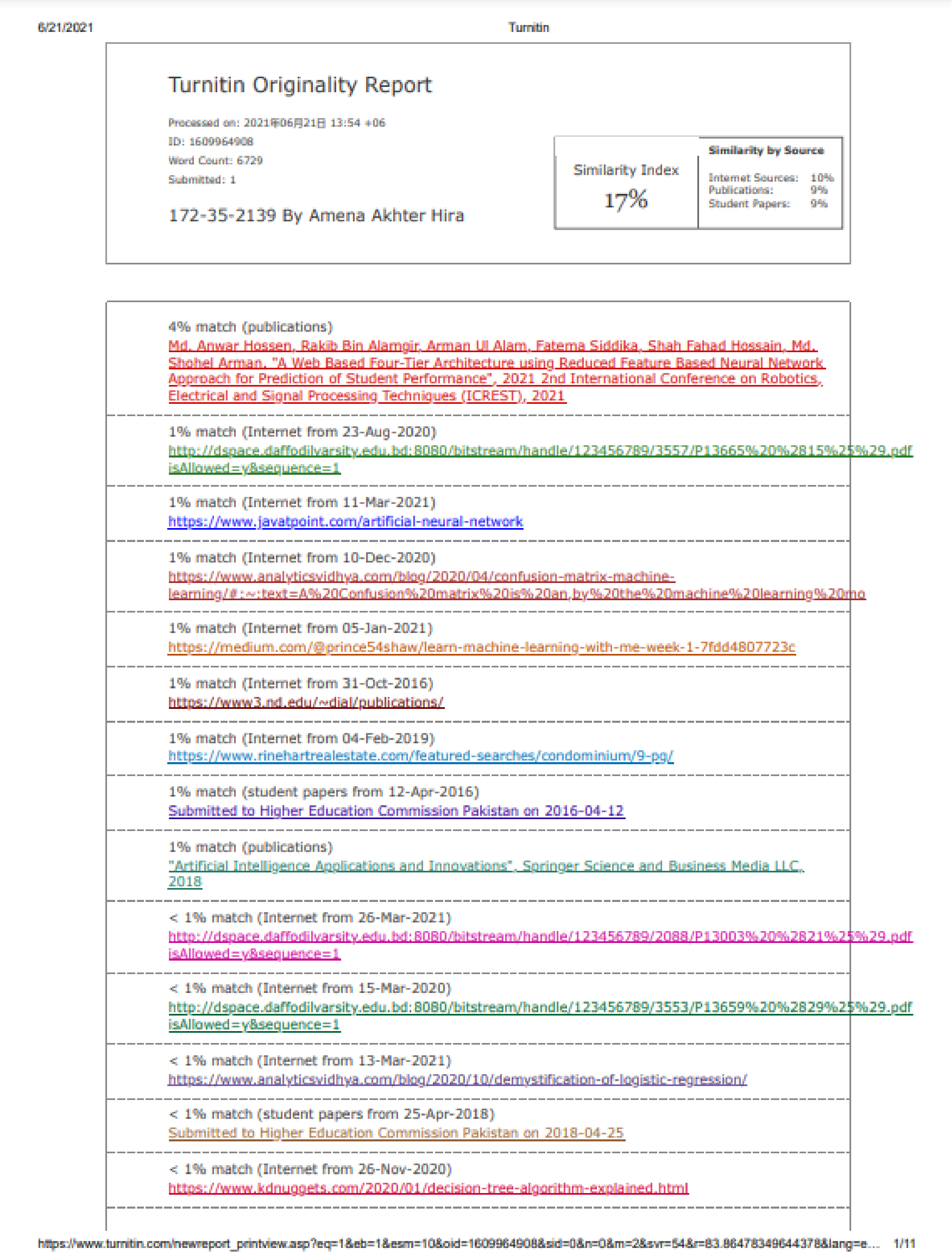
Name	**Description**
gender	Categorical Value, (Female, Male)
nationality	Categorical Value, (Bangladesh, Somalia)
place_of_birth	Categorical Value, where student birth.
department	Categorical Value, where student study at the University
current_year	Categorical Value, Student’s study period
time_of_group_study	Numerical Value
absent_in_a_semester	”
parents_satisfied_with_results	Categorical Value(Yes=0, No = 1)
amount_of_drop_semester	Numerical Value
drop_reason	Categorical Value, why student drop the semester
meet_with_advisor	Categorical Value(Yes=0, No = 1)
ask_questions_in_class	Categorical Value(Yes=0, No = 1)
use_additional_course_material	Categorical Value(Yes=0, No = 1)
due_amount	Categorical value, how amount due of tuition fee
institutional_name	Categorical Value, where student study
fathers_education	Categorical Value
mothers_education	”

Appendix – B

List of Abbreviation

ANN	Artificial Neural Network
DT	Decision Tree Classifier
FP	False Positive
FN	False Negative
LR	Logistic Regression
MCC	Matthews Correlation Coefficient
ETC	Extra Tree Classifier
TP	True Positive
TN	True Negative
XGBoost	Extreme Gradient Boosting

Plagiarism Report



6/21/2021 Turnitin

Turnitin Originality Report

Processed on: 2021年06月21日 13:54 +06
ID: 1609964908
Word Count: 6729
Submitted: 1

172-35-2139 By Amena Akhter Hira

Similarity Index	Similarity by Source
17%	Internet Sources: 10% Publications: 9% Student Papers: 9%

4% match (publications)
Md. Anwar Hossen, Rakib Bin Alamgir, Arman Ul Alam, Fatema Siddika, Shah Fahad Hossain, Md. Shohel Arman. "A Web Based Four-Tier Architecture using Reduced Feature Based Neural Network Approach for Prediction of Student Performance", 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), 2021

1% match (Internet from 23-Aug-2020)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/3557/P13665%20%2815%25%29.pdf
isAllowed=y&sequence=1

1% match (Internet from 11-Mar-2021)
https://www.javatpoint.com/artificial-neural-network

1% match (Internet from 10-Dec-2020)
https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/#:~:text=A%20Confusion%20matrix%20is%20an,by%20the%20machine%20learning%20mo

1% match (Internet from 05-Jan-2021)
https://medium.com/@prince54shaw/learn-machine-learning-with-me-week-1-7fdd4807723c

1% match (Internet from 31-Oct-2016)
https://www3.nd.edu/~dial/publications/

1% match (Internet from 04-Feb-2019)
https://www.rinehartrealestate.com/featured-searches/condominium/9-pg/

1% match (student papers from 12-Apr-2016)
Submitted to Higher Education Commission Pakistan on 2016-04-12

1% match (publications)
"Artificial Intelligence Applications and Innovations", Springer Science and Business Media LLC, 2018

< 1% match (Internet from 26-Mar-2021)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/2088/P13003%20%2821%25%29.pdf
isAllowed=y&sequence=1

< 1% match (Internet from 15-Mar-2020)
http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/3553/P13659%20%2829%25%29.pdf
isAllowed=y&sequence=1

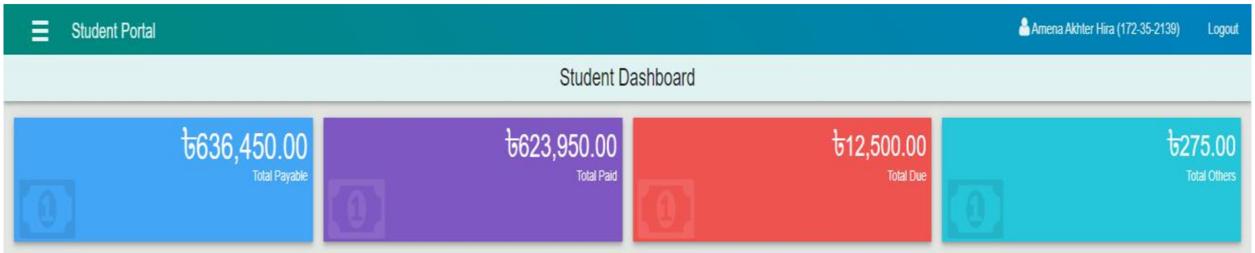
< 1% match (Internet from 13-Mar-2021)
https://www.analyticsvidhya.com/blog/2020/10/demystification-of-logistic-regression/

< 1% match (student papers from 25-Apr-2018)
Submitted to Higher Education Commission Pakistan on 2018-04-25

< 1% match (Internet from 26-Nov-2020)
https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

https://www.turnitin.com/newreport_printview.asp?eq=1&eb=1&esm=10&oid=1609964908&sid=0&n=0&m=2&svr=54&r=83.86478349644378&lang=e... 1/11

Clearance



Total Payable	Total Paid	Total Due	Total Others
৳636,450.00	৳623,950.00	৳12,500.00	৳275.00