A STUDY ON THE IMPACT OF INTERNET USAGE ON THE ACADEMIC PERFORMANCE OF THE STUDENTS

SHAJID HOSSAIN HEMAL

BACHELOR OF SCIENCE IN INFORMATION AND COMMUNICATION ENGINEERING

NOAKHALI SCIENCE AND TECHNOLOGY UNIVERSITY

A STUDY ON THE IMPACT OF INTERNET USAGE ON THE ACADEMIC PERFORMANCE OF THE STUDENTS

SHAJID HOSSAIN HEMAL

This report submitted in fulfillment of the requirements for the award of the degree of Bachelor of Science in Information and Communication Engineering

Department of Information and Communication Engineering

NOAKHALI SCIENCE AND TECHNOLOGY UNIVERSITY

SEPTEMBER 2021

DECLARATION

This thesis proposal is submitted to the Department of Information and Communication Engineering of Noakhali Science & Technology University in partial fulfillment of the requirements for the award of the degree of Bachelor of Science (B.Sc) Engineering in Information and Communication Engineering (ICE). So, I hereby declare that this report has not been submitted elsewhere for the requirement of any kind of degree, diploma or publication.

Shajid Hossain Hemal

Student of Information and Communication Engineering, NSTU

Roll No.: ASH-1711019M

Session: 2016-17

Date: 30. 09. 2021

ACCEPTANCE

This research proposal is submitted to the Department of Information & Communication Engineering, Noakhali Science & Technology University, Sonapur, Noakhali in partial fulfillment of the requirements for the Bachelor of Science (B.Sc) degree in Information & Communication Engineering.

Professor Dr. Md Ashikur Rahman Khan

Chairman

Department of Information and Communication Engineering

Noakhali Science & Technology University

Noakhali - 3814, Bangladesh

ABSTRACT

Education systems have greatly changed with the emergence of Internet. It has a significant impact on how students learn things and how educational institutes operate. But its impact can also be contradicting. Internet addiction can slowly poison the minds of our youths and stand in the way of pursuing their goals. Proper analysis of the effects this technology has on today's youth can help determine the factors that allow a student to succeed academically and those that do not. Also, predicting the academic performance beforehand can help determine the changes that must be incorporated to improve the educational system. So, this research attempts to accurately analyze the effects of internet usage on student's academic progress and also predict the class performance of the students by using different machine learning algorithms. About 700 student's data were collected for this study from an offline survey taken from different schools, colleges and universities, all situated in Noakhali district. These data will be analyzed with the help of statistical evaluation methods and machine learning algorithms such as Logistic Regression, Decision Trees, Random Forests and Naive Bayes Classifier. The results from these models will be analyzed using different performance metrics to determine the factors that contribute the most in their academic activities.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND STUDY

The academic achievement of students is an important benchmark in comparing the quality of education in a state or a country. Predicting the academic performance beforehand can help determine the changes that must be incorporated to improve the educational system. With the help of various statistical methods and the application of machine learning and artificial intelligence techniques, it's now possible to predict a students' academic performance based on several diverse factors. The goal of this research is to analyze the effects of the internet usage on students of this generation and predict their academic performance based on these factors. Several statistical evaluation methods and machine learning algorithms will be employed to perform this prediction.

Following the invention of the Internet, our education systems have greatly changed. It opened doorways to a wealth of information, knowledge and educational resources, increasing opportunities for learning in and beyond the classroom [13]. Because of this, the Internet has superseded libraries as a source for information gathering and research. Many teachers will now ask students to visit specific websites to study from home and online encyclopedias because they provide masses of knowledge on almost every topic imaginable. Teachers can also make use of the Internet by giving students extra resources and material from the Internet, such as interactive lessons and educational games. Many university and college courses use a "hybrid" model where many lessons are done online, requiring fewer in-class meetings. This saves students from having to commute to campus with their heavy textbooks every day. The Internet allows instantaneous connection to classmates and teachers. Improving communication between students and teachers allows teachers to assist students without having to stay after class. It also allows for students to have greater efficiency when working on projects with their peers when everyone cannot attend or asking for clarification when something is unclear. A number of universities, such as Harvard, Yale and Stanford, have opened up free courses on a variety of subjects that are accessible to anyone for free. These typically come in the form of lectures on video, but some also have notes attached. This means there is easy access to plenty of free lectures without emptying your bank account to pay tuition [14].

The internet also has a diverse effect on the young generation. Access to internet has made self-learning easily possible. Students can pursue their respective career paths without any assistance of institutional education through the internet. This is a revelation for students who do not have the financial capability to pay for schools and universities. But internet can also harm the youth. Social networking sites have become the need of the present world. Social networking is a medium through which people can broadcast their knowledge, experience, and views. These networking sites are changing the mindsets of the youths of this generation. Studies show depression and anxiety problems among teenagers and young adults can rise from too much use of social networking sites. Also, Internet addiction is a real source of worry, and even comparable to substance-related addictive disorders. Changes in brain chemistry and structure resulted from the Internet overuse are of particular concern to researchers and educators, because of their profound effects on young people's development [15].

The Internet proves to be a valuable tool for young people provided that it is used sensibly in close connection with teachers' guidance. For every person you find who thinks the internet has improved our access to information and made us self-sufficient learners you will find someone else who thinks we have become addicted to entertainment because of it. This research focuses on some of these factors influenced by Internet use that affect the students' academic performance.

1.2 LITERATURE REVIEW

A number of reviews pertaining to not only the diverse factors like personal, socio-economic, psychological and other environmental variables that influence the performance of students but also the models that have been used for the performance prediction are available in the literature and a few specific studies are listed below for reference.

Dorina Kabakchieva implemented well known data mining algorithms, including two rule learners, a decision tree classifier, two popular Bayes classifiers and a Nearest Neighbor classifiers to predict student data from the two databases provided by the technical staff of a Bulgarian university [1]. The WEKA software is used for the implementation of this study. The classifiers perform differently for the five classes. The decision tree classifier achieved an accuracy of 65.94% in the 10-fold cross-validation testing. The KNN classifier is not able to predict the classes which are less represented in the dataset. The Bayes classifiers are less accurate than the others. All tested classifiers are performing with an overall accuracy below 70% which means that the error rate is high and the predictions are not very reliable. Similar attempt was made by Zaidah Ibrahim, Daliela Rusli using three predictive models [2]. But they were developed using SAS Enterprise Miner to predict student's academic performance from 206 student's data that were collected for the study. Model selection is based on the square root of average squared error (RASE) for the validation dataset. Based on the validation dataset, the smallest RASE is produced by Artificial Neural Network model which is 1.714, followed by Decision Tree model which is 0.1769 and linear regression which is 0.1848. The result of this study shows that all of the three models produce more than 80% accuracy which is not bad. It was seen that, ANN model is better than the other two models in the first 10th deciles. Ali Daud, Naif Radi Aljohani, Rabeeh Ayaz Abbasi, Miltiadis D. Lytras, Farhat Abbas and Jalal S. Alowibdi used two types of classification models to predict the performance of the students [9]. Two generative and three discriminative model was used for this experimental analysis. About 3,000 student records from different universities of Pakistan were collected initially. After removing inconsistencies and duplications in the dataset, only 776 student instances remained for experiments. SVM performed the best for the proposed feature sets with an F1 score of 0.867. There were a few features that performed very well for the prediction. The features related to family expenditure such as natural gas, electricity, telephone, water, accommodation, miscellaneous expenditures, and most importantly family expenditure on education are found to be most effective in predicting academic performance.

V. Ramesh, P. Parkavi and K. Ramar used WEKA's implementation of Naive Bayes, Multi-Layer Perceptron, SMO, J48, REPTree algorithms were used for predicting student performance [3]. The dataset was collected from randomly selected nine schools from Kancheepuram district. A sample of 900 students was taken from the group of schools. Results from each algorithm are surprisingly very different. Naive Bayes had an accuracy of 49.5%, Multi-Layer Perceptron had 72.38%. From the results, it was seen that Multi-Layer Perceptron classifier is most appropriate for predicting student performance. Results obtained from hypothesis also testing revealed some interesting observations. The type of school does not influence student performance. On the other hand, parent's occupation plays a major role in predicting grades. As a result, educational institutions would be able to identify students at risk early and provide better additional training for the weak students. But it has not collected additional features like psychological factors which disturb the students, motivational efforts taken by the teachers and e-learning materials available to the students. Another study done by E. T. Lau, L. Sun and Q. Yang collected a sample data of about a total of 1000 students consisting of 275 female and 810 male students selected from the undergraduate students from one Chinese University [5]. Conventional statistics and neural networks were used to identify the factors that likely to affect the student's performance and predict their performance. The neural network is modelled with 11 input variables, two layers of hidden neurons, and one output layer. Levenberg-Marquardt algorithm is employed as the backpropagation training rule. This neural network model had a good prediction accuracy of 84.8%. Some observations that was detected using statistical evaluations were: there is significant difference of students' yearly CGPA performance, with notable improvements in their performance for next two academic years. The female students score better than male students. Also, the students CGPA are significantly affected by mothers' occupation. Fathers' motivations are not sufficiently strong enough to improve academic performance among students. ANN performed poorly in classifications of students according to their gender, as high False Negative (FN) rates are obtained in the results. This is likely due to high imbalance ratio of two different types of sample data (students' gender) applied.

A unique approach was taken by Nguyen Thai-Nghe, Lucas Drumond, Artus Krohn-Grimberghe and Lars Schmidt-Thieme in predicting student performance [6]. They implemented recommender systems for this purpose which is very unconventional. An open-source framework named

MyMedia was used to implement two Collaborative Filtering and Regularized Matrix Factorization recommender systems on several data combinations. They used two datasets from the Knowledge Discovery and Data Mining (KDD) challenge 2010. Results from these recommender systems were then compared with traditional machine learning algorithms like logistic regression. Experimental results show that the proposed method, which employs recommender system techniques to predict the student performance, outperforms the traditional regression methods. But stronger methods for dealing with the cold-start problem when using matrix factorization is needed for better performance. Also, an ensemble model on different combinations of "items" is required.

Bogdan Oancea, Raluca Dragoescu, Stefan Ciucu implemented a multilayer perceptron with one input layer, two hidden layer and one output layer and it was trained using a version of the resilient backpropagation algorithm called iRPEOP+ [4]. The neural network is implemented with the Encog framework (Heaton, 2011) using Java programming language. For this study, a sample of 1000 students is used from the last three graduate generations from "Nicolae Titulescu" University from Bucharest. This network achieved an accuracy of 86% with an MSE of 1.7%. The test data had an MSE of 1.91%. This model has accurately predicted 86.6% of the poor results, 94.2% of the medium results and 85.7% of the good results. But it failed to identify those parameters from the input data that influence in a greater extent the students' decision to leave education. Similarly, V.O. Oladokun, A.T. Adebanjo and O.E. Charles-Owaba aimed to predict students' academic performance by employing Multi-layer Perceptron topology as the ANN model [7]. Five generations of graduates from the pre-admission record forms of an engineering department of University of Ibadan., Nigeria were collected as the dataset. Results show that, this model achieved an accuracy of over 74% which is a fair performance. One limitation of this model stems from the fact not all the relevant performance influencing factors are obtainable from the pre-admission record forms filled by the students. Also, the extension of this research to non-engineering departments is recommended. An efficient approach for predicting student graduation outcomes was proposed by Stamos T. Karamouzis and Andreas Vrettos using Artificial Neural Networks [8]. A sample of 1,407 profiles of students was collected for this study. The sample represented students at Waubonsee College and it was divided into two sets. The first set of 1,100 profiles was used for training and the 307 profiles were used for testing. A three-layered Perceptron Network was trained using the backpropagation principles for this purpose. Training the network within 6000 epochs, the mean square error achieved was 0.18 and the network was able to correctly classify 148 out of 172 successful graduates (86.04%) and 633 out of 928 unsuccessful graduates (68.21%). The average predictability rate for the training and test sets were 72% and 68%, respectively. Expanding the repertoire of parameters included in each student profile we believe that the network will achieve even higher levels of predictability. By eliminating the parameters that are the least contributory to successful predictability we can reduce input noise and thus optimize the network's performance. Another attempt in predicting the performance of student's taken by O. C. Asogwa, A. V. Oladugba collected a total of 420 students records were used in the analysis [10]. About 58.8% of the total data were used as the training set, 31.0% as the testing set and 10.2% was used for cross validation as each network was run 100 epochs. The neural network gives an accuracy of 100% for promoted students, 62.5% for repeated students, and 93.5% for the demoted students. The Mean Correct Classification Rate or accuracy of about 97.07% for the Artificial Neural Network model architecture developed which shows a good performance.

All these papers reviewed here represent different studies very close to our research topic. But almost all of these papers tried to predict the students' performance based on the usual factors contained in the databases of the educational institutions or survey data containing factors like age, total score from previous examinations, university admittance year, admittance exam and achieved score, university specialty/direction, total university scores on their academic result etc. [1]. The sample data of students' socio-economic background in [5]. In another study, the primary data was collected using a questionnaire which includes questions related to several personal, socio-economic, psychological and school related variables that were expected to affect student performance [3]. But our goal is to analyze the effect of Internet usage on the academic performances of the students. So most of our dataset features are about the academic and non-academic usage of the internet. This data was collected from an offline survey.

The following table summarizes each reviewed paper and compares with our research objectives:

Serial No.	Paper's Title [Reference Number]	Methodology	Contribution/Advantages	Limitations	Difference To Us
1	Predicting Student Performance by Using Data Mining Methods for Classification [1]	Used two rule learners, a decision tree classifier, two popular Bayes classifiers and a nearest neighbor classifier. The WEKA software is used for the study.	WEKA's implementation of Decision tree classifier achieves 65.94 % accuracy, Bayesian classifiers is about 60%, and the k-NN classifier accuracy is about 60 %.	The error rate is high and the predictions are not very reliable. The classifiers are not able to predict the classes which are less represented in the dataset.	The dataset only contained factors related to the university admission campaigns. The WEKA software is used for the study.
2	Predicting Students' Academic Performance: Comparing Artificial Neural Network, Decision Tree And Linear Regression [2]	Three predictive models have been developed using SAS Enterprise Miner: Artificial Neural Network, Decision Tree and Linear Regression.	The result of this study shows that all of the three models produce more than 80% accuracy. The best model in predicting the final CGPA of the students upon graduation is ANN. Probably due to smaller dataset.	Only 206 student's data were obtained for this study.	Research was done in SAS Enterprise Miner. Also, Factors considered for this is substantially different from us.
3	Predicting Student Performance: A Statistical and Data	A sample of 900 students was taken from a group of schools. For this study WEKA's	From the results, it was proven that Multi-Layer Perception (MLP) classifier is most appropriate for predicting student	This study has not collected additional features like psychological factors,	Research was performed using WEKA. Also, Factors considered for this is

	Mining Approach [3]	implementation of Naïve Bayes, Multi-Layer Perception algorithms were used.	performance. MLP gives 72.38% prediction which is relatively higher than other algorithms.	motivational efforts taken by the teachers and e-learning materials available to the students.	substantially different from us. Our study mainly focuses on effects of internet usage.
4	Predicting students' results in higher education using neural networks [4]	A sample of 1000 students is used collected for this study. The implemented neural network was a multilayer perceptron trained using a version of the resilient backpropagation algorithm.	86.6% of the poor results, 94.2% of the Medium and 85.7% of the Good results students were predicted by the network. This network achieved an accuracy of over 86%.	There were parameters in the dataset that influence in a greater extent the students' decision to leave education. These parameter are only noise.	The neural network was implemented with the Encog framework using Java programming language.
5	Modelling, prediction and classification of student academic performance using artificial neural networks [5]	A total of 1,000 sample data were collected. The neural network is modelled with Levenberg–Marquardt algorithm is employed as the backpropagation training rule.	Overall, the neural network model has achieved a good prediction accuracy of 84.8%, along with limitations.	ANN performs poorly in classifications of students according to their gender, as high False Negative (FN) rates are obtained as results.	Our study includes the application of only machine learning algorithms. Factors considered for this study is substantially different from us.
6	Recommender System for Predicting Student Performance [6]	User/item average Collaborative Filtering and Regularized Matrix Factorization on several data combinations was used for this research.	Experimental results show that the proposed approach can improve the prediction results. Using all of the combinations together yields better results for both matrix factorization and the collaborative filtering baseline.	Stronger methods for dealing with the cold-start problem when using matrix factorization is needed for better performance.	The algorithms implemented in MyMedia open source framework. Our study mainly focuses on effects of internet usage which this study does not.
7	Predicting Students' Academic Performance using Artificial	The model was based on the Multilayer Perceptron Topology.	It achieved an accuracy of over 74% which shows the potential efficacy of Artificial Neural Network as a prediction tool and a	Not all the relevant performance influencing factors ore obtained. Also,	Factors considered for this is substantially different from us. Our study

	Neural Network: A Case Study of an Engineering Course [7]	Source data spans five generations of graduates of a university.	selection criterion for candidates seeking admission into university.	this research only considers engineering departments of a university.	mainly focuses on effects of internet usage which this study does not.
8	An Artificial Neural Network for Predicting Student Graduation Outcomes [8]	A sample of 1,407 profiles of students was used. The network was developed as a three-layered perceptron and was trained using the backpropagation principles.	It turned out that after 6000 training epochs the network produced its best predictive power. When tested with the testing set it produced a mean square error of 0.22 on both successful and unsuccessful graduates.	Small number of profiles in the training set. Also, there were a lot of parameters that did not contribute to the predictions.	Artificial Neural Networks are not implemented in our study. Also, factors considered for this is substantially different from us.
9	Predicting Student Performance using Advanced Learning Analytics [9]	About 3000 student records were collected. Implemented models were, Support Vector Machine (SVM), Classification and Regression Tree (CART), Naïve Bayes (NB).	From the comparative analysis that our proposed features are important predictors and achieved F1-score of 86% on real life undergraduate students' data.	Only the features related to family expenditure, most importantly family expenditure on education are found to be most effective in predicting academic performance.	Our study mainly focuses on effects of internet usage which this study does not. Here, factors considered for this is substantially different from us.
10	Of Students Academic Performance Rates Using Artificial Neural Networks (ANNs) [10]	A total of 420 students records were used in the analysis. Artificial Neural Networks were implemented for this study.	This indicates a Mean Correct Classification Rate or accuracy of about 97.07% for the Artificial Neural Network model architecture developed which shows a good performance	The model was developed based on some selected input variables from the pre admission data contained in the student's records.	Artificial Neural Networks are not implemented in our study. We utilize machine learning algorithms.
11	Student Performance Prediction Using Support Vector Machine and K-Nearest	Implemented algorithms are Support Vector Machine and K-Nearest Neighbor to 395 datasets	Support Vector Machine achieved slightly better results with correlation coefficient of 0.96, while the K-Nearest	The dataset only includes 395 instances which is very small.	The data only contains standard academic information's. But our study mainly focuses

	Neighbor [27]	collected from University.	Neighbor achieved correlation coefficient of 0.95.		on the effects of internet usage.
12	Academic performance, adaptation and mental health of nursing students: A cross-sectional study [28]	A multiple linear regression, through the backward method was used for predicting depressive symptoms. This analysis was performed in the R statistical program.	Multiple linear regression produced a prediction model with 74.6% accuracy which is not that good.	Only one university is included in the study. So, the dataset is not diverse.	This study predicts the depressive symptoms (mental health, dropout rates).This is not the focus of our study.
13	Data survey on the factors affecting students' satisfaction and academic performance among private universities in Vietnam [29]	The Cronbach's alpha, exploratory factor analysis (EFA), and confirmatory factor analysis (CFA) were utilized to examine the internal reliability and validity of each scales.	Results indicate that quality of academic staff has the strongest effect on students' academic performance, followed by service accessibility, university, and education program. However, students' academic performance is not related to training environment.	There were limited datasets of primary data which is available to explore. This study did not use machine learning or deep learning algorithms, which tend to be more reliable.	Factors are very different than our dataset. Also, data is processed using SPSS 23.0 and AMOS 23.0.
14	Impact of online education on fear of academic delay and psychological distress among university students following one year of COVID-19 outbreak in Bangladesh [30]	Chi-square ($\chi 2$) test is used for searching the association of various variables. A binary logistic regression was applied to determine the potential variables associated with FAD and psychological stress after adjusting other factors.	Near 60% of the current students exerted extreme FAD and 78.1% of students were severely psychologically stressed. Logistic regression analyses revealed that the students of the female gender, rural area, lower-income families, and who suffered from the highest FAD were more significantly stressed than their reference groups.	Internet-based responses may not be considered as representative data for those having no online facilities. Also, the current research did not investigate the association of all the critical parameters of online education with FAD.	All the analyses have been conducted using software IBM SPSS (version 20). We are using Python based framework. Online data collection was conducted in this study. We conducted offline data collection which more reliable.
15	Impacts of urban landscapes	Statistical methods used in this study are	Students in the Atlanta CBSA performed best in math	The relationships we observed	The factors that are included in this study are

	on students' academic performance [31]	Pearson correlation tests, Ordinary Least Squares (OLS) regressions, Variance inflation factor (VIF), Simultaneous autoregressive models (SAR) etc.	for both measures, followed by science, then social science, and ELA. Results suggest that urban planning for increased canopy cover and aquatic environments may improve academic performance in English Language Arts, mathematics and the sciences.	between gender, school enrollment, and academic success were not statistically significant.	very different from our research goal. Our study mainly focuses on the effects of internet usage.
16	Predictors of nursing students' academic performance in the National Autonomous University of Mexico, 2010–2019: A retrospective study [32]	Multiple-regression models is used in order to identify predictors of school performance.	Changes were noticed regarding the sex distribution and the enrolment age. School performance was higher among female students, among students in their last semesters, and among students who had been granted a scholarship. In contrast, students who had previously failed a subject performed lower.	Even though the estimated multiple regression model of this study showed acceptable goodness of fit, other predictors recognized in other studies could have an important influence well.	Psychological variables like emotional conditions, stress, and alcohol and drug use a d workplace variables like employment status and work schedules are not included in our study.
17	The impact of social media usage and lifestyle habits on academic achievement: Insights from a developing country context [33]	The study employed linear regression to test the underlying hypotheses and logistic regression to predict good academic performance.	The overall accuracy for logistic regression was 70.1% indicating an acceptable discrimination.	The convenience sampling method used might not guarantee the generalizability of findings. Also, the subjective nature of the self-reporting.	The factors that are included in this study are very different from our research goal. Our study mainly focuses on the effects of internet usage.
18	Identifying the K-12 classrooms' indoor air quality factors that affect student academic	A multivariate linear regression analysis was used to investigate the associations between IAQ factors and student scores using	The results revealed associations between student scores and ventilation system type, ventilation rates, fine particle counts, and O3 and CO concentrations.	Classroom characteristics (such as floor and ceiling types) were not used in the statistical model, because the research team did not	This paper predicts academic performance of students based on the type of classrooms and indoor air quality. This study very

	performance [34]	demographic data as controls.		find enough variations in the current sample to warrant further study.	different from our objective.
19	Impact of intrapersonal and interpersonal emotional intelligence and self-directed learning on academic performance among pre-university Science students [35]	Statistical methods like Chi-square test (χ2), Comparative Fit Index (CFI), Goodness of Fit Index (GFI) etc. were used. Also, the hierarchical regression analysis was used to predict student academic performance	The hierarchical regression analysis reveals that student Academic performance was positively predicted by perceived Interpersonal and Intrapersonal EI, whereas self-directed learning has an inconsistent predictive impact at different steps in the model.	The sample size is limited to only the successful students in the program and they comprised primarily of female students accounting for more than 62% of the sample size.	The analysis was performed using SPSS 26 and AMOS 24. We perform our study using Python based framework.
20	Predicting Secondary School Students' Performance Utilizing a Semi-supervised Learning Approach [36]	Classification algorithms were selected for this study. These classification algorithm utilized two wrapper based semi-supervised learning techniques: (i) Self-training and (ii) YATSI (Yet Another Two-Stage Idea)	The classification performance with Self-training: 1. SMO (sequential minimum optimization) and C4.5 was slightly improved from 0.2% to 3.2%. 2. For MLP, the accuracy with self-training is improved up to 6.9% over the accuracy of MLP alone.	YATSI does not always improve the performance of naive Bayes when unlabeled data are used together with labeled data.	The implementation code was written in JAVA, using WEKA Machine Learning Toolkit. Our study implements Python based framework.

Table 1.1: Summary of the literature review.

1.3 PROBLEM STATEMENT

Bangladesh currently ranks 128th in global literacy with 72.8% of its population aged 15 or older being literate. This is very poor compared to 86% average worldwide [17]. So, more work must be done on this field to get us out of this predicament. If we analyze the factors that affect students' performance and try to predict how their performance can be nudged into the positive direction by applying changes to these factors, this illiteracy problem can then be solved. So, students' performance assessment is not separate from the learning process of a student. Actually, it is a continuous and an integral part of learning processes.

Predicting the performance of a student has been an important field of study and many researchers gave numerous contributions to find solutions to improve students' academic performance. Correctly assessing students' performance and weeding out the reasons for their failure we can ensure a higher success rate of students. Early prediction of students' success could help teachers to identify students that have potential for advanced courses and also students who need the additional care in order to improve their knowledge.

The Internet has already become a vital part of our education systems. We may fall behind with the rest of the world if we don't manage to utilize this powerful tool properly. So, it is imperative to check first how much influence the Internet has on our students and how it has affected their academic performance. As we already know that, Internet also has a side that distracts today's youth and sends them towards emotional despair, we have to find a way to navigate our youth so that they can use this technology to its fullest potential. Recent studies have greatly depended on the data that covers the effect of socio-economic factors in the performance of a student. But Internet is also an important factor that can significantly affect the academic performance of a student.

Student performance has long been a key issue for all kinds of educational organizations. This research attempts to discover how academic performance of the students can be affected by different factors relating to the Internet and build a reliable model that predicts their performance based on these factors. Predicting student performance will help in identifying which student possibly will succeed. If they do not succeed, we can figure out what are the main reasons for it. We will use data collected from different schools, colleges and one university situated in Noakhali for this study. About 702 data samples have been collected with no null values. We will then analyze this dataset with the help of different statistical techniques like correlation, machine learning, etc.

1.4 RESEARCH HYPOTHESIS

This study focuses on how the use of internet can impact the academic performance of students by the application of statistical methods and machine learning algorithms. For predicting performance, first and most challenging task is to collect real-world data. Here we use 702 student's data for the research which is taken from various school, college and university. 302 data samples were collected from Noakhali Science and Technology University, 200 data samples from two colleges situated at Maijdee, Noakhali, and the last 200 from data samples were collected from three schools situated at Sonapur, Noakhali. These data were collected through an offline survey by handing out questionnaires. Each data sample contained about 20 factors. These factors contained information about academic, non-academic uses of internet, popular websites that were frequently visited by each student, how much time is spent on surfing the internet, how much time on academic. Other than that, the survey inquired about the family, living area, academic result, social media interaction of each student. But it mostly focuses on the interaction with internet. Some categorical attributes like Total Internet Usage, Time Spent in Academics, Years of Internet Use etc. have feature engineered into numerical attributes to make analysis easier. After completing these steps, we start analyzing the data.

There is assumed to be some impact of total time spent in academics and total time of internet usage in the academic performance of the students. If more time is spent in academic studies, the student's should perform well in exams. Also which websites a student visits while on the internet might relate to how well a student performs on exams. If the student mainly uses internet for online video games, social-media, Netflix, YouTube and other entertainment based websites, that student may get poor scores in his/her exams. But on the other hand, if taking online courses and surfing educational blogs are prioritized while surfing the internet, it can have positive effect on the academic performance. Also the total time spent on the internet can have a significant impact on the students' academic results. Too much time spent on internet means less time for studies.

Online classes and exams have become a regular thing in the midst of pandemic that is raging through the world right now. So, the availability of the devices that can access internet has a significant effect on students' academic career. Finally, the numerous obstacles that some rural parts of our country face like unavailability of internet connection, bad service, overpriced etc. can become a hindrance in a student's academic activities, which in turn affects their performance.

Through this study, we will analyze all of the factors mentioned and check their validity with real world data. Then with the help machine learning algorithms, we will try to predict students' academic performance based on these factors.

1.5 OBJECTIVES OF THE STUDY

The purpose of this study is to analyze the effects of internet usage in the academic performance of students and to predict their performance with the help of statistical methods and machine learning algorithms. To achieve this, the following sub-objectives will have to materialize:

I. To analyze various research works done on students' performance prediction using various machine learning and deep learning techniques.

II. To identify and study the attributes that contribute the most in predicting their performance.

III. To enhance the student's learning capabilities by seeking ways to make the process more effective, efficient and reliable.

IV. To better understand the effectiveness of the application of machine learning algorithms in the educational systems.

V. To obtain a clear idea of machine learning classification algorithms and see how each of them perform.

CHAPTER 2

METHODOLOGY

2.1 INTRODUCTION

This section involves the working procedures of how the whole study is conducted. In this research Logistic Regression, Stochastic Gradient Descent, Decision Tree, Random Forest and Naïve Bayes classifiers are used for analysis. These algorithms are also introduced in this section. These algorithms are used to predict the performance of the students. Then the models are evaluated using proper evaluation metrics. Various performance evaluation metrics are also applied for data assessment e.g., confusion metrics, correlation matrix, accuracy, precision, recall etc. which helps our research consequently. In the following section 2.2 provides overview of the methodology is described, the overall process is shown in a graph figure and then step by step processing is also showed along with algorithmic view. Then section 2.4 explains the whole data collection and preprocessing process that include collecting, preprocessing, encoding and standardizing of the dataset. In section 2.7 provides a description of the selected algorithms. At last, in section 2.8 the performance evaluation metrics are described with their educational form such as confusion matrix, accuracy, precision, recall etc.

2.2 RESEARCH DESIGN

The design of the system is shown in Figure 2.1. This figure paints a simplistic picture of the overall system. It provides a clear view of how the prediction process is conducted. First of all, we collect data from educational institutions. After collecting the data, we need to remove any unclear, duplicate or missing values. That's why, the next process is to preprocess the data. We preprocess the data using data mining tools. If there are other features that is needed for this study, we can engineer them in this step. Also, we have to scale the features so that the training the model can converge faster. We'll use standardization technique for feature scaling. After that we can start designing the models.

We will design a supervised machine learning model which trains the already preprocessed data and using the trained data predict students' academic performance. Different machine learning algorithms are employed for this experimentation. The last step is to calculate the accuracy of these models by using numerous performance measuring techniques. The best accuracy from these models will be selected by comparing them. Each part of our propose system is described in the following subsections.

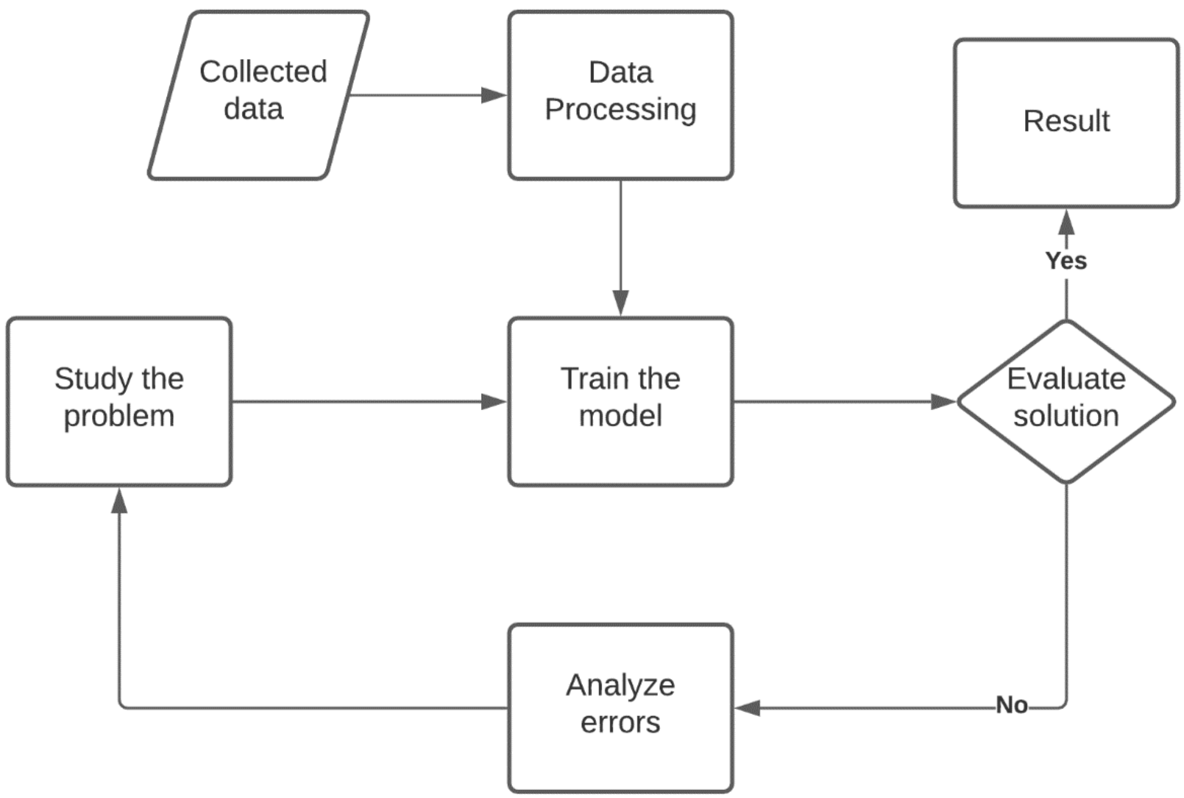

Figure 2.1: System Overview

2.3 REQUIRED TOOLS

Different tools will be used for this study. All of them are free and open source. A short description of these tools is given below,

1. **Python:** Python is a high-level general programming language.

2. **Virtual Environment:** We will use Jupyter Notebook as the environment to conduct our experiment. Anaconda framework provides Jupyter Notebook along with other tools that provide assistance in conducting such experiments. We will create a virtual environment inside Anaconda and install all the required tools.

3. **Scikit-learn:** Scikit-learn is widely used popular library for machine learning.

4. **NumPy:** NumPy is a very powerful package which enables us for scientific computing.

5. **Matplotlib:** Matplotlib is a plotting library for the Python and its numerical mathematics extension NumPy.

6. **Pandas:** Pandas is an open-source BSD licensed software specially written for python.

7. **Seaborn:** Seaborn is a library used for data visualization and is created by using python.

8. **SciPy:** SciPy is a collection of fundamental mathematical functions and is built on the top of Scikit-learn.

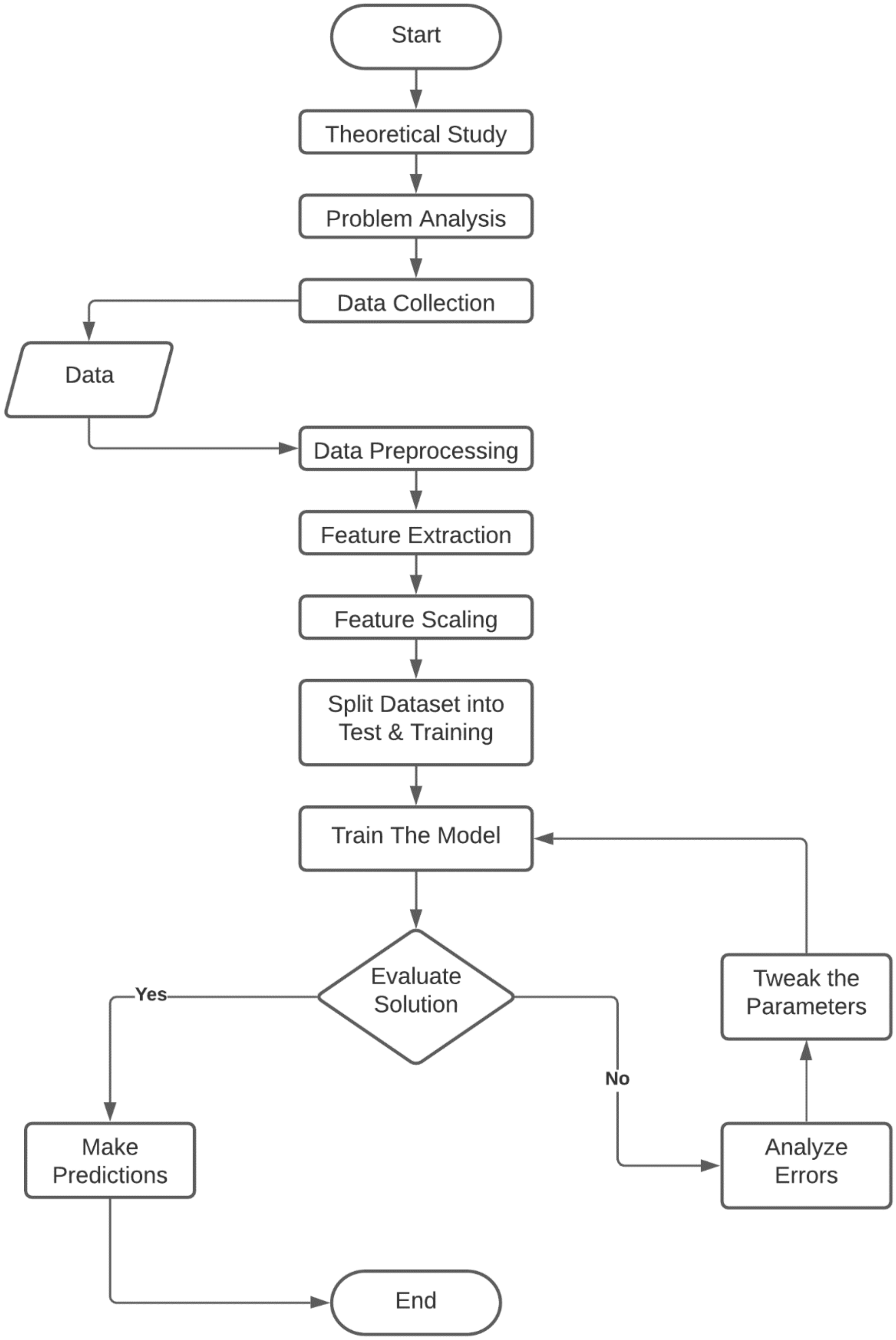

Figure 2.2: Workflow Diagram

2.4 DATA COLLECTION

2.4.1 PARAMETERS

For predicting the student's performance at first, we have to collect relevant data from students. Here we use 700 student's data for the research which is taken from various school, college and university. We collect these data from the students by offline survey by handing out questionnaires. About 20 factors are considered in this survey. In these factors, we consider not only academic data but also others personal information such as family, living area, social media interaction and mostly internet usage etc. Factors that are considered in this questionnaire is given below.

Variable Name	Type	Value
Gender	Nominal	Male, Female
Age	Numeric	14-27
Most Frequently Visited Website	Nominal	Google, Youtube, Whatsapp, Messenger, Instagram, Twitter, Gmail and others
Proficiency	Nominal	Good, Average, Better, Not at All
Medium	Nominal	Laptop, Mobile, Desktop, Laptop and Mobile
Location	Nominal	Home, Library, University, Cyber Cafe, School, College
Household Internet Facilities	Nominal	Connected, Not Connected
Browsing Time	Nominal	6am-12pm, 12pm-6pm, 6pm-12am, 12am-6am
Browsing Status	Nominal	Daily, Weekly, Monthly
Residence	Nominal	Village, Town, Remote
Total Duration of Internet Use	Nominal	Rarely, 0-5 hours, 5-10 hours, 10-15 hours, 15-20 hours, 20-24 hours
Time Spent in Academic Use	Nominal	Rarely, 0-5 hours, 5-10 hours, 10-15 hours, 15-20 hours, 20-24 hours
Purpose of Use	Nominal	Social Sites, File Sharing News, Group Study, Music, Shopping

Years of Internet Use	Nominal	Greater than a year, 1 to 5 years, 5 to 10 years, less than 10 years
Browsing Purpose	Nominal	Academic, Non-Academic
Enrollments in Online Courses	Nominal	Yes, No
Attendance in Webinars	Nominal	Yes, No
Academic Performance	Nominal	Good, Average, Satisfactory, Not Satisfactory
Obstacles	Nominal	Slow Internet, High Price, Power Failure, Poor Computer Skills, Lack of Necessary Hardware's, Cost of Computer Equipment
Graded Result	Numeric	from 2.50 to 4.00 (5.00 in case of schools and colleges)

Table 2.1: Dataset Description

2.4.2 DATA COLLECTION PROCESS

As discussed above, the required data was collected through offline surveys by handing out questionnaires. The factors considered in the questionnaires are already discussed in the above sub-section. 702 student's data were collected for this research which are taken from various school, college and university. 302 data samples were collected from Noakhali Science and Technology University, 200 data samples from two colleges situated at Maijdee, Noakhali, and the last 200 from data samples were collected from three schools situated at Sonapur, Noakhali. These data were collected through an offline survey by handing out questionnaires. Each data sample contained about 20 factors.

2.5 DATA PREPROCESSING

This is one of the most important states of the implementation module. We may collect a large amount of data from several sources so some of the data might contain garbage values, some may be null. Mostly, there can be unclear, duplicate and missing values in the dataset. Also, some data may contain large dimensions that can be bothersome because it can significantly increase the training time. That's why, data preprocessing is a vital part of this process. Data preprocessing includes procedures to handle duplicate, missing values, dimension reduction etc. The following Figure 2.2 displayed below indicates the steps involved data preprocessing.

In the case of our dataset, there were no duplicate or missing values. Dataset dimension is not that large so, it will not have much effect on training time. But, we will transform some categorical features like Total Duration of Internet Use, Time Spent in Academic Use into numerical features by taking the average of the ranges described in the survey question. Numerical data performs better in generalizing the models than categorical ones.

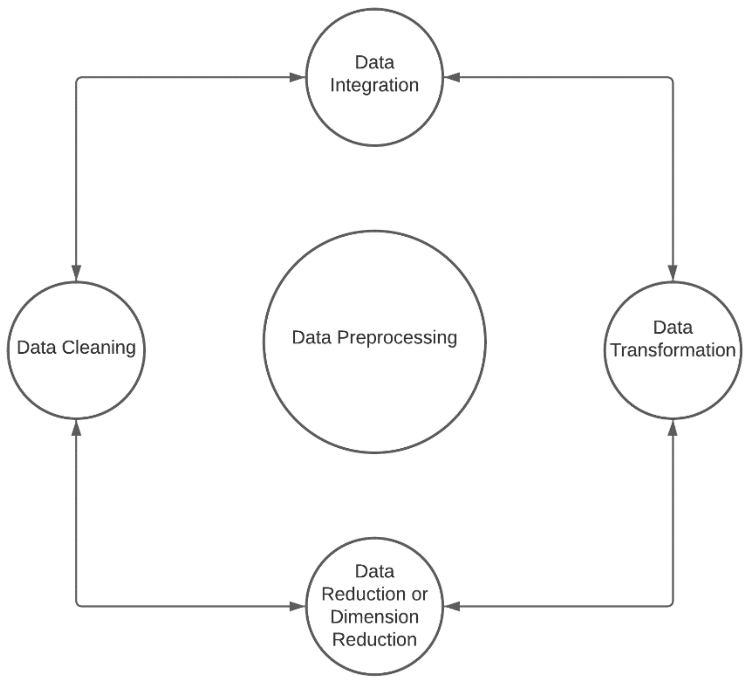

Figure 2.3: Tasks Involved in Data Preprocessing

Unfortunately, there are some unclear values existing in the dataset. We will remove the entries that contain unclear values because it can confuse the model while training it. Other than that, there are no problems with the dataset. We can move on to the next step.

2.6 FEATURE SCALING

One of the most important transformations that is needed to apply to the data is feature scaling. With few exceptions, Machine Learning algorithms don't perform well when the numerical attributes have very different scales. This is the case for our data too. Note that the numerical attributes Total Duration of Internet Use, Time Spent in Academic Use, Age and Graded Result do not have the same scales. So, it is imperative that we scale our data. We will use Standardization to complete this process.

Standardization process is actually quite simple: first it subtracts the mean value with each of the values of the attribute (so standardized values always have a zero mean), and then it divides by the variance so that the resulting distribution has unit variance. This is an ideal technique because standardization is much less affected by outliers.

$$X_{Stand} = \frac{X - \overline{X}}{\sigma}$$

2.7 SPLITTING THE DATASET

All of the dataset cannot be used at once. It is standard practice to split the dataset into two sets. Separating data into training and testing sets is an important part of evaluating data mining models. Typically, a dataset is separated into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing. Analysis Services randomly samples the data to help ensure that the testing and training sets are similar. By using similar data for training and testing, the effects of data discrepancies can be minimized and better understanding on the characteristics of the model can be gained.

The training set will contain the majority of the dataset. We will use 80% of the dataset for training the models. Remaining 20% will be kept in store for testing purposes. After a model has been processed by using the training set, the model is tested by making predictions against the test set. Because the data in the testing set already contains known values for the attribute that is to be predicted, it is easy to determine whether the model's guesses are correct. Testing comes at the last step of the whole process. Testing set only checks if the trained models generalizes well with the dataset.

2.8 MODEL DEVELOPMENT

This is the most important phase which includes building the models that will aid in the prediction. We will implement five machine learning algorithms for prediction. These algorithms include Logistic Regression, Decision Tree, Naive Bayes, Random Forest and Stochastic Gradient Descent. All of these algorithms will be discussed in the following sub-sections.

2.8.1 LOGISTIC REGRESSION

Logistic regression is a machine learning algorithm for classification. This algorithm is commonly used to estimate the probability that an instance belongs to a particular class. If the estimated probability is greater than 50%, then the model predicts that the instance belongs to that class, or else it predicts that it does not. This makes it a binary classifier [12]. Normally, to predict which class a data belongs, a threshold is set. Based upon this threshold, the obtained estimated probability is classified into classes. Decision boundary can be linear or non-linear. Polynomial order can be increased to get complex decision boundary. Usually sigmoid function is used as the activation function for this algorithm. Logistic regression is designed to solve classification problems, and is most useful for understanding the influence of several independent variables on a single outcome variable [19].

Unfortunately, this algorithm works only when the predicted variable is binary, assumes all predictors are independent of each other and assumes data is free of missing values. But it is possible to perform multi-classification tasks using one-vs-all method.

2.8.2 DECISION TREES

Decision Trees are versatile Machine Learning algorithms that can perform both classification and regression tasks, and even multi-output tasks. It is a very powerful algorithm, capable of fitting complex datasets. A Decision Tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. An instance is classified by starting at the root node of the tree, testing the attribute specified by this node and then moving down the tree branch corresponding to the value of the attribute. This process is then repeated for the sub tree rooted at the new node [19].

Decision Tree is simple to understand and visualize, requires little data preparation, and can handle both numerical and categorical data. But sometimes, Decision Tree can create complex trees that do not generalize well, and decision trees can be unstable because small variations in the data might result in a completely different tree being generated.

2.8.3 NAIVE BAYES

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e., every pair of features being classified is independent of each other [21].

When assumption of independence holds, a Naive Bayes classifier performs better compare to other models like logistic regression and you need less training data. It is easy and fast to predict class of test data set. It also performs well in multi class prediction. It performs well in case of categorical input variables compared to numerical variable(s). A Naive Bayes is usually used in real time prediction, multi class prediction, text classification, spam filtering, sentiment analysis, recommendation systems.

2.8.4 RANDOM FORESTS

Random forest classifier is a meta-estimator that fits a number of decision trees on various sub-samples of datasets and uses average to improve the predictive accuracy of the model and controls over-fitting. It is an ensemble of Decision Trees, generally trained via the bagging method. The Random Forest algorithm introduces extra randomness when splitting a node, it searches for the best feature among a random subset of features, this results in a greater tree diversity, which trades a higher bias for a lower variance, generally yielding an overall better model [12]. So the prerequisites for random forest to perform well are: There needs to be some actual signal in our features so that models built using those features do better than random guessing. Also, the predictions (and therefore the errors) made by the individual trees need to have low correlations with each other.

Reduction in over-fitting and random forest classifier is more accurate than decision trees in most cases. Slow real time prediction, difficult to implement, and complex algorithm.

2.8.5 STOCHASTIC GRADIENT DESCENT

Stochastic Gradient Descent (SGD) is a simple yet very efficient approach to fitting linear classifiers and regressors under convex loss functions such as (linear) Support Vector Machines and Logistic Regression [22]. Even though SGD has been around in the machine learning community for a long time, it has received a considerable amount of attention just recently in the context of large-scale learning.

SGD has been successfully applied to large-scale and sparse machine learning problems often encountered in text classification and natural language processing. Given that the data is sparse, the classifiers in this module easily scale to problems with more than 10^5 training examples and more than 10^5 features.

2.9 PERFORMANCE ANALYSIS

Performance analysis involves systematic observations to enhance performance that improves decision making of data and gives feedback to our system. It is the process of studying or evaluating the performance of a particular scenario in comparison of the objective which to be achieved. For analyses we use confusion matrix, accuracy, precision, recall and f1 score.

2.9.1 CONFUSION MATRIX

A Confusion matrix is an N × N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. This gives us a holistic view of how well our classification model is performing and what kinds of errors it is making. For a binary classification problem, we would have a 2×2 matrix as shown below with 4 values:

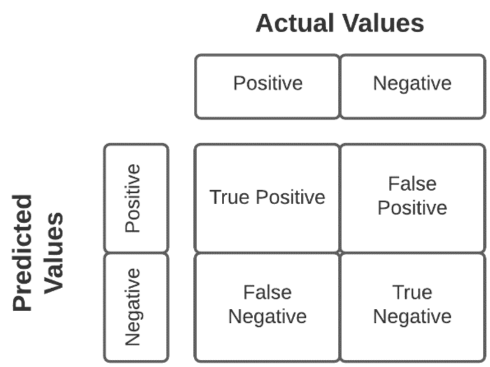

Figure 2.4: Confusion Matrix

The target variable in a binary classification has two values: Positive or Negative. The columns represent the actual values of the target variable. The rows represent the predicted values of the target variable.

1. **True Positive (TP):** The predicted value matches the actual value. The actual value was positive and the model predicted a positive value.
2. **True Negative (TN):** The predicted value matches the actual value. That means the actual value was negative and the model predicted a negative value
3. **False Positive (FP) – Type 1 error:** The predicted value was falsely predicted. That means the actual value was negative but the model predicted a positive value. Also known as the Type 1 error.
4. **False Negative (FN) – Type 2 error**: The predicted value was falsely predicted**.** That means the actual value was positive but the model predicted a negative value**.** Also known as the Type 2 error.

2.9.2 ACCURACY

In confusion matrix the accuracy is the ratio of correct prediction to the total predictions made. It is often presented as a percentage by multiplying the result by 100. The numerical value of accuracy represents the proportion of true positive results (both true positive and true negative) in the selected population. The formula shown as below

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

2.9.3 PRECISION

Precision tells us how many of the correctly predicted cases actually turned out to be positive. This would determine whether our model is reliable or not. Precision is a useful metric in cases where False Positive is a higher concern than False Negatives.

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

2.9.4 RECALL

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

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

2.9.5 F1-SCORE

It is often convenient to combine precision and recall into a single metric called the F1 score, in particular when we need a simple way to compare two classifiers. The F1 score is the harmonic mean of precision and recall. Whereas the regular mean treats all values equally, the harmonic mean gives much more weight to low values. As a result, the classifier will only get a high F1 score if both recall and precision are high,

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

CHAPTER 3

EXPECTED RESULTS

Several classification algorithms are employed in this study to predict the academic performance of the students. Each and every technique has its own advantages and disadvantages. So, some of the algorithms might perform better and some might not generalize quite that well. That's why, all of the classification techniques performance will be thoroughly analyzed and their accuracy will be compared to determine which algorithm performs the best for the data collected for this study. We expect that every algorithm has a satisfying performance range. Predictions depend on the factors that are included in the dataset. Some factors might help students to perform better academically and some might hinder it. We will also determine these factors that have the highest contributions in the prediction and which ones have the lowest.

From the statistical analysis, we can expect there will be a positive correlation between the total time spent in academic activities and academic performance. The more time a student spends in his/her studies, the better should be the grades. Again, there is expected to be a negative correlation between the total internet usage and their academic performance. But since the internet has become a vital part of our education system, if the internet usage is kept to a minimum it might have a negative effect in the academic performance of the students.

Total years of internet usage might not have any affect with how students perform academically. But proficiency might have some effects. How effectively a student can use the internet can indicate how much benefit the student can attain from this technology.

Frequently used websites should have a significant effect on the students psyche. Demotivated students usually spend most of their time on playing online games, watching videos on YouTube, Netflix or on social media etc. Hence their academic performance is expected to be unremarkable. Motivated students do not usually spend much time on entertainment. Although it is expected that most of the students use internet for non-academic purposes. It can be expected that most of the university students enrolled in online courses and attended webinars, but only a few of the college and high school students are involved in such extra-curricular activities.

The machine learning algorithms tend to perform differently for different type of datasets. So, it's quite hard to estimate which algorithm will perform the best among the implemented. An accurate example of this has to be Artificial Neural Networks. The Multilayer Perceptron Network in [3] had the accuracy of 72.38%, but this algorithm achieved an accuracy of 86% in [4]. This is quite a significant difference in performance. ANN algorithm in [10] achieved an accuracy of 97.07% which is quite remarkable.

Some algorithms like decision tree always tend to generalize in a similar manner. From the literature review we saw that, the decision tree classifier in [1] had 65.94% accuracy for the 10-fold cross validation testing but the REPTREE implementation had the accuracy of 60.13% in [3]. So, we expect that the decision tree classifier might not generalize that well with the dataset. The random forest classifier might perform better than decision tree classifier since it fits a number of decision trees to perform classifier. But this algorithm is prone to overfitting easily, so we might have to tweak the parameters and the number of epochs to avoid overfitting. The Naïve Bayes classifier might not have a great accuracy score because this algorithm performs best for small amount of training data. But our dataset has 702 students' data which is quite large. If we reduce the dataset to better fit the classifier, the model might not generalize well. Also, Naive Bayes implicitly assumes that all the attributes are mutually independent. In real life, it is almost impossible that we get a set of predictors which are completely independent [16]. The Logistic Regression algorithm is an efficient algorithm and least likely to overfit. So, this algorithm is expected to perform the best among the others.

CHAPTER 4

CONCLUSION

This study has proposed a research methodology for predicting the academic performance of student's using machine learning algorithms and analyzing the factors that affect the most. This has long been a key issue for all kinds of educational organizations. With this research, we attempt to discover how academic performance of the students can be affected by different factors relating to the Internet and build a reliable model that predicts their performance based on these factors. Predicting student performance will help in identifying which student possibly will succeed. If they do not succeed, we can figure out what are the main reasons for it. Identifying the factors that has an influence in motivating or demotivating a student can help teachers find an effective way to help students achieve their goals. The model required for this study will be developed based on the data collected from different from different educational institutes situated in Noakhali. About 702 data samples have been collected with no null values. The methodologies for the whole process have been explained with details in the previous sections. To validate this approach, we will compare the performance of each selected machine learning algorithms with each other. After conducting the research, experimental results will show if the proposed approach can actually predict students' performance accurately.

RESEARCH SCHEDULE

This section provides a table containing the time schedule of the whole research study from literature survey to the completion of the whole research study. Although this is not an accurate time schedule. It only provides an approximation of the amount of time each steps in the research took in completing the study.

The following Gantt-chart visualizes the project schedule. The objectives represents the major steps that are required to complete the research. The duration represents the progress of each steps in months.

Objectives	**Major Activities**	**Duration In Months**						
		Mar	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**
Study related and existing works to review them	Literature Review	■						
Design and develop the necessary statistical models	Design and Development		■	■				
Collecting required Data	Data Collection			■	■			
Setup the entire system and perform the experimentation	Experimental Setup					■		
Analyze the obtained result	Problem Analysis					■	■	
Prepare the research paper	Preparing the paper							■

Table 3: Gantt-Chart Representing Research Schedule

REFERENCES

[1] Kabakchieva, Dorina. (2013). Predicting Student Performance by Using Data Mining Methods for Classification. Cybernetics and Information Technologies. 13. 10.2478/cait-2013-0006.

[2] Ibrahim, Zaidah & Rusli, Daliela. (2007). Predicting Students' Academic Performance: Comparing Artificial Neural Network, Decision Tree and Linear Regression. 21st Annual SAS Malaysia Forum.

[3] Ramesh, Vamanan & P. Parkavi, & Ramar, K.. (2013). Predicting Student Performance: A Statistical and Data Mining Approach. INTERNATIONAL JOURNAL OF COMPUTER APPLICATIONS. 63. 975-8887.

[4] Oancea, Bogdan & Dragoescu, Raluca & Ciucu, Stefan. (2013). Predicting students' results in higher education using neural networks.

[5] Lau, E. & Sun, L. & Yang, Qingping. (2019). Modelling, prediction and classification of student academic performance using artificial neural networks. SN Applied Sciences. 1. 10.1007/s42452-019-0884-7.

[6] Thai-Nghe, N., Drumond, L., Krohn-Grimberghe, A. and Schmidt-Thieme, L., 2010, Recommender System for Predicting Student Performance, Elsevier B.V., 10.1016/j.procs.2010.08.006.

[7] Oladokun, V.O., Adebanjo, A.T., and Charles-Owaba, O.E., 2008. Predicting Students' Academic Performance using Artificial Neural Network: A Case Study of an Engineering Course, The Pacific Journal of Science and Technology, Volume 9, Number 1, pp. 72-79.

[8] Karamouzis, Stamos & Andreas, Vrettos. (2008). An Artificial Neural Network for Predicting Student Graduation Outcomes. Lecture Notes in Engineering and Computer Science. 2173.

[9] Daud, Ali & Aljohani, Naif & Abbasi, Rabeeh & Lytras, Miltiadis & Abbas, Farhat & Alowibdi, Jalal. (2017). Predicting Student Performance using Advanced Learning Analytics. 10.1145/3041021.3054164.

[10] Oluchukwu C, Asogwa. (2015). Of Students Academic Performance Rates Using Artificial Neural Networks (ANNs). American Journal of Applied Sciences. 3. 10.12691/ajams-3-4-3.

[11] H. Al-Shehri et al., "Student performance prediction using Support Vector Machine and K-Nearest Neighbor," 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), 2017, pp. 1-4, doi: 10.1109/CCECE.2017.7946847.

[12] Géron A., 2018. Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow, Second edition, O'Reilly Media, Inc.

[13] Internet Society, 2017, Internet Access and Education: Key considerations for policy makers, https://www.internetsociety.org/resources/doc/2017/internet-access-and-education/. (Accessed September 22, 2021).

[14] Wright, J., 2019. Importance of Internet to Education. https://itstillworks.com/importance-internet-education-1232.html. (Accessed September 22, 2021).

[15] Pham, Linh. (2017). Effects of the Internet on young people: A view from neuroscience. VNU Journal of Science: Education Research. 33. 10.25073/2588-1159/vnuer.4068.

[16] Kumar, N. (2019). Advantages and Disadvantages of Naïve Bayes in Machine Learning. http://theprofessionalspoint.blogspot.com/2019/03/advantages-and-disadvantages-of-naive.html. (Accessed September 26, 2021)

[17] Roy, M. June 11, 2018. Digital Education Improves Literacy in Bangladesh (online). https://borgenproject.org/tag/literacy-in-bangladesh/. (Accessed September 27, 2021)

[18] Swaminathan, S., March 15, 2018. Logistic Regression – Detailed Overview (online). https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc. (Accessed September 27, 2021)

[19] Gupta, S., 22 Jun, 2021. Decision Tree (online), https://www.geeksforgeeks.org/decision-tree/. (Accessed September 27, 2021).

[20] Yiu, T., June 12, 2019. Understanding Random Forest – How the Algorithm Works and Why it is so Effective (online). https://towardsdatascience.com/understanding-random-forest-58381e0602d2. (Accessed September 27, 2021)

[21] Kumar, N., 15 May, 2020. Naïve Bayes Classifiers, www.geeksforgeeks.org (online). https://www.geeksforgeeks.org/naive-bayes-classifiers/. (Accessed September 27, 2021)

[22] Scikit-learn developers, 2021. Stochastic Gradient Descent, www.scikit-learn.org (online). https://scikit-learn.org/stable/modules/sgd.html/. (Accessed September 27, 2021)

[23] Hossain, S. 2021. Figure 2.1: System Overview. PNG formatted figure was made using Lucidchart, https://lucid.app/lucidchart/b6ce8fad-8661-422f-89ca a416b7202c87/edit?page=0_0 &invitationId=inv_511d1bb5-dd0e-469b-b973-bc79c8e55520#. (Accessed September 22, 2021)

[24] Hossain, S. 2021. Figure 2.2: Workflow Diagram. PNG formatted figure was made using Lucidchart, https://lucid.app/lucidchart/7b72a86e-ad15-4340-9b7f-35afd0bea482/edit?page= 0_0#. (Accessed September 23, 2021)

[25] Hossain, S. 2021. Figure 2.3: Tasks Involved in Data Preprocessing. PNG formatted figure was made using Lucidchart, https://lucid.app/lucidchart/5276cd37-fc0d-4e0e-b923-11c213617 be8/edit?page=0_0#. (Accessed September 22, 2021)

[26] Hossain, S. 2021. Figure 2.4: Confusion Matrix. PNG formatted figure was made using Lucidchart, https://lucid.app/lucidchart/58821ded-bf50-40d9-8b3a-d2d4e3bc4a0f/edit?Page =0_0#. (Accessed September 22, 2021)

[27] H. Al-Shehri et al., "Student performance prediction using Support Vector Machine and K-Nearest Neighbor," 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), 2017, pp. 1-4, doi: 10.1109/CCECE.2017.7946847.

[28] Silva G. O., Angelo Aredes D. N., Galdino-Júnior H., "Academic performance, adaptation and mental health of nursing students: A cross-sectional study, Nurse Education in Practice", Volume 55, 2021, 103145, ISSN 1471-5953, https://doi.org/10.1016/j.nepr.2021.103145. (https://www.sciencedirect.com/science/article/pii/S1471595321001815)

[29] Van T. D., Nguyen Thi K. C., Tran Thi H. P., "Data survey on the factors affecting students' satisfaction and academic performance among private universities in Vietnam", Data in Brief, Volume 33, 2020, 106357, ISSN 2352-3409, https://doi.org/10.1016/j.dib.2020.106357. (https://www.sciencedirect.com/science/article/pii/S2352340920312506)

[30] Hossain J., Ahmmed F., Rahman S.M.A., Sanam S., Bin Emran T., Mitra S., "Impact of online education on fear of academic delay and psychological distress among university students following one year of COVID-19 outbreak in Bangladesh", Heliyon, Volume 7, Issue 6, 2021, e07388, ISSN 2405-8440, https://doi.org/10.1016/j.heliyon.2021.e07388.(https://www. sciencedirect.com/science/article/pii/S2405844021014912).

[31] Lin, Meimei & Van Stan, John. (2020). Impacts of urban landscapes on students' academic performance. Landscape and Urban Planning. 201. 103840. 10.1016/j.landurbplan.2020.103840.

[32] Zárate-Grajales, Rosa & Ostiguín-Meléndez, Rosa & Aristizabal, Patricia & Serván-Mori, Edson & Nigenda, Gustavo. (2021). Predictors of nursing students' academic performance in the National Autonomous University of Mexico, 2010–2019: A retrospective study. Nurse Education Today. 100. 104790. 10.1016/j.nedt.2021.104790.

[33] Bou--Hamad, Imad. (2020). The Impact of Social Media Usage and Lifestyle Habits on Academic Achievement: Insights from a Developing Country Context. Children and Youth Services Review. 118. 105425. 10.1016/j.childyouth.2020.105425.

[34] Kabirikopaei A., Lau J., Nord J., Bovaird J., Identifying the K-12 classrooms' indoor air quality factors that affect student academic performance, Science of The Total Environment, Volume 786, 2021, 147498, ISSN 0048-9697, https://doi.org/10.1016/j.scitotenv.2021.147498. (https://www.sciencedirect.com/science/article/pii/S0048969721025699)

[35] Okwuduba, Emmanuel & Nwosu, Kingsley & Okigbo, Ebele & Samuel, Naomi & J., Achugbu. (2021). Impact of intrapersonal and interpersonal emotional intelligence and self-directed learning on academic performance among pre-university science students. Heliyon. 7. e06611. 10.1016/j.heliyon.2021.e06611.

[36] Livieris, Ioannis & Drakopoulou, Konstantina & Tampakas, V. & Mikropoulos, Tassos & Pintelas, P.. (2017). Predicting Secondary School Students' Performance Utilizing a Semi-supervised Learning Approach. Journal of Educational Computing Research. 57. 10.1177/0735633117752614.