Proceedings of 9th IOE Graduate Conference
Peer Reviewed
ISSN: 2350-8914 (Online), 2350-8906 (Print)
Year: 2021 Month: March Volume: 9

Performance Analysis of Engineering Students in Academic using Ensemble Methods

Tek Bist Bithari [a], Sharan Thapa [b]

[a, b] *Department of Electronics and Computer Engineering, Paschimanchal Campus, IOE, Tribhuvan University, Nepal*
Corresponding Email: [a] tekbist54@gmail.com, [b] sharant@ioepas.edu.np

Abstract
The declining scenario of student pass rates in the engineering institutions in Nepal has led to the need for Educational Data Mining (EDM) in the engineering sector. The research aims to investigate the key factors influencing the academic success of undergraduate engineering students and apply data mining techniques (classification and clustering) to predict engineering student's academic outcomes. For this study, the data of 3201 engineering graduates from Paschimanchal Engineering Campus and Dhangadhi Engineering College were collected manually from the admission file of each student. On the included dataset, first, K-mean clustering was used to group students with similar academic performances, and classification was performed using three traditional algorithms; Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbors (KNN), and three ensemble methods; Bagging, XGBoost, Voting. The improvement in classification result was observed using ensemble methods where XGBoost outperforms the other models with an accuracy of 84.7% and F1-score 85.0%. The trend analysis of student's performance has revealed that engineering students in Nepal tend to achieve a better academic result in ending semesters than in commencing semesters. By using the Correlation-based feature selection method, we identified the most impactful core courses, challenging semesters, and study years for the respective six engineering programs that have a huge impact on their academic performance. The study can be used in the early identification of at-risk students so that proper steps can be taken to improve their performance and ultimately reduce college dropouts.

Keywords
Institute of Engineering, Ensemble Methods, Engineering Students, Nepal

1. Introduction

Educational Data Mining (EDM) focuses on analyzing educational data to extract useful information and knowledge by the application of data mining tools and techniques [1]. EDM can act as a mediator to bring change in the quality of education. In recent times, Universities have been using EDM in various educational sectors to make proper planning for dropout students, identification of weak learners, number of student's enrollment and make optimum use of available resources [1]. In the context of Nepal, EDM has not been used to an extent it must have been used. This is one of the reasons behind the lagging educational standard compared to European countries. The educational data generated from the educational institution in Nepal has not been properly stored and organized causing difficulty in extracting useful information. Institute of Engineering (IOE), the top engineering institution in Nepal known for its quality education has suffered from a reduction in semester pass percentage from 50 to 40 since 2009 AD [2]. Similarly, almost all the engineering colleges in Nepal are facing the common problem of lower pass rates. To counter this situation, we felt the need for EDM in the engineering education.

The study is a step towards utilizing educational data generated from the engineering institution in Nepal to extract useful information that can help concerned authorities to make proper planning. Due to unavailability of required data in respective colleges, at first, we prepared the dataset of Nepalese engineering graduates by collecting attributes from their admission files. The admission files were kept by the exam department of respective colleges that encloses a copy of each level transcript, form filled during admission, mark sheet of every semester, scholarship form, hostel form, and other related

information. The study analyses the various factors including academic, demographic, and socio-economic information of engineering students, and then identifies the factors that affect the most in the academic success. The inclusions of the attributes were based on the literature studies.

Another aim of the study is to predict engineering student's academic performance labeled into four categories; excellent, good, medium, and satisfactory. To accomplish this task, the data mining techniques such as classification and clustering were used. Classification is a supervised method that works by assigning each instance to a particular label. In contrast, Clustering is an unsupervised method that works by grouping the unlabeled instances with some similarities. For clustering, we used K-means clustering to group the students with similar performances. Mainly four types of students were observed in the cluster analysis and the respective labels were assigned to each student. The classification was then performed using both traditional classification methods (SVM, DT, KNN) and common ensemble methods (Bagging, XGBoost, Voting). Ensemble methods improve the model performance by combining the decisions of multiple classifiers. They combine the results of multiple models to develop an optimum predictive model. Previous reasearches have shown that combining results of multiple models reduces generalization error [3].

Only developing a predictive model is not able to lift the quality of teaching procedure and student's academic performance so, we also performed data analysis to identify the most effective courses for the final outcome and challenging semesters that can be helpful for the students to update their study plan and ultimately achieve better results in semester exams.

2. Related Works

The rising interest in the field of Educational Data Mining has led to numerous researches in the field of education over the past few decades. Mostly, the studies were focused on the application of various data mining algorithms for predicting student's academic success and identifying the critical factors that have an impact on their academic performance. In [4], the authors reviewed and analyzed the research papers related to student's academic success of the past few years to determine the crucial factors influencing the student's academic success and find the algorithms that were mostly used for the studies so far. After reviewing almost 47 papers, they found that Support Vector Machine, Decision Tree, Neural Network, K-nearest Neighbors, and Logistic Regression were the most used classification algorithm whereas the academic, social, and personal attributes were included in most of the articles.

The paper [5] presents a framework for predicting the learning outcome of engineering students using various data mining algorithms. The authors discovered 12 highly ranked attributes among the total 24 attributes using different feature selection methods. It was found that CGPA, back exams, 12 marks, the medium of education, Engineering cutoff, Board type were the best predictor of the final outcome and the j48 algorithm produced the best result. Rifat et al. [6], performed a prediction of student's academic results, trend analysis and introduced a generalized structure for comprehensive educational analysis for the Business students from the marketing department of renowned Bangladesh University. The authors used real-world data collected from the transcript directly including grades in each course, Semester Grade Point Average (SGPA), states of students according to semester GPA and final GPA. The prediction was done using six different classification algorithms where Random Forest outperforms all the others. The study also suggested the most impactful course for the Business student's that can be helpful for the course advisors, college authority, students, and other concerned parties to improve the academic result.

Soobramoney et al.[7], reviewed the previous works on identifying weak students using a single traditional machine learning algorithm. The authors concluded that a single machine learning algorithm cannot perform better in every circumstance to predict students at risk. They suggested using ensemble techniques like bagging and boosting instead of the individual machine learning algorithm and including several influential factors to achieve a better result. The authors in [8] performed classification on educational data collected from Learning Management System (LMS) using four different classification algorithms individually. They achieved better results when ensemble methods like bagging, boosting, and voting were applied to each of the models. Further, the authors found improvement in the accuracy of the models when student's behavioral attributes were also included along with other

academic factors.

Bajracharya et al. [2], proposed restructuring the examinations system of IOE to improve the semester pass rate of the students and quality of education. The authors reviewed the data of admitted engineering students and their semester results where they found that only 40 percent of students pass as regular graduates each year and the remaining students go for back exams. After the complete revision of the data, they concluded that the traditional examination system, teaching process, syllabus were the major causes of the lower pass rate so they need to be updated to gain better results. The authors in [1] proposed a predictive model using a decision tree that can predict the final outcome of Information Technology (IT) students using their grades in mandatory courses and classification rules were generated using the j48 algorithm. They discovered the most impactful courses for the final outcome that can help students to update their study plan accordingly and improve their final grade. The authors recommended using other data mining algorithms and adding more relevant attributes as future works.

In Nepal, there aren't many studies regarding EDM. Although, there exists some studies related to quality improvement of engineering education in Nepal but most of them were done from organizational perspectives. They concluded updating the traditional examination system, admission process, teaching and management process [2, 9] can improve the existing situation of engineering education. Apart from these factors, there are some factors related to students themselves that have impact on their academic success but none of the studies had covered them yet. The study aims to discover those impactful attributes related to engineering students in Nepal. This study is an extended version of our previous work [10].

3. Problem Statements

Among the total admitted students in IOE, 60% of students go for back exams [2]. This scenario raises the question mark against the quality of engineering education of Nepal. Some of the students are forced to leave engineering due to lots of back subjects. The engineering students are not aware of the impactful courses that needs more time for study. Similarly, they are unaware of the difficult semesters where they have to labor hard. There is a lack of proper study plan and fear of examination among Nepalese engineering students. The management, course advisor, and teachers are unknown of the trend of student's performances that has created difficulty in proper planning. The identification of the weak learner can help the concerned authorities to update their plan accordingly. The prediction of the final outcome can reduce frustration among students and encourage them to achieve the better academic result. Moreover, the model can be used as a decision support system for choosing the right engineering program.

4. Methodology

To, achieve the objectives of this study, we followed the sequential steps shown in Figure 1.

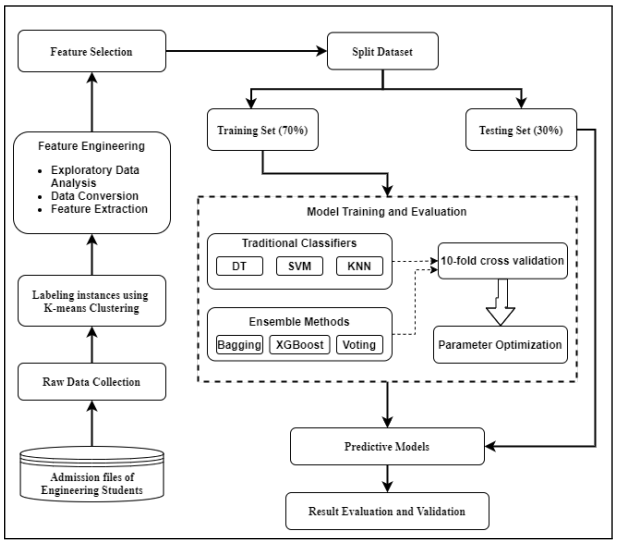

Figure 1: Steps followed for the study

4.1 Data Collection

The past studies have revealed that demographic and socio-economic factors of the students also had an impact on their academic success along with their academic backgrounds [11]. Therefore, we included 36 attributes covering academic, demographic, and socio-economic information of an engineering student in our dataset. The dataset with required attributes was not available in the respective colleges so we had to collect these data manually from the admission files of each engineering student that were kept by the respective exam department of campus administration. The total data samples of 3201 engineering graduates of batch 2061 to 2072 from Paschimanchal Engineering Campus, Pokhara (TU affiliated), and Dhangadhi Engineering College, Dhangadhi (PU

affiliated) of batch 2062 to 2072 were collected for the study. Since, Pokhara University (PU) follows the grading system so, we converted the final CGPA and grades of the included subjects into the cumulative scale of 0 to 100 based on the criteria followed by the grading system of PU. The data collection was one of the challenging steps in our study that took us around three months for its completion. Table 1 show the descriptions of included attributes.

Table 1: Descriptions of included attributes

Attributes	Descriptions/Denotations
Gender	Male or Female (M, F)
Scholarship	Has got scholarship or not (Yes, No)
Entrance Rank	Rank of student in entrance examination (under_500, 500-1000, 1000-1500, 1500+)
Interested program	Priority number given by student to admitted program (1, 2, 3, 4..)
Age	Age at the time of admission
Plus two percentage	Aggregate percentage scored in plus two
College location	Province where plus two college is located (province1,.....,province7)
Gap	Year gap between plus two and bachelor (0, 1, 2, 3..)
Major Subject	Major subject taken by student in class 12 (Math, Biology, Other)
Ten Class Percentage	Percentage secured by student in class 10
School Location	Province where school is located (province1,.....province7)
School Type	Governmental or Private (G, P)
Ethnic Group	Ethnic category of student (brahmin, chhetri, janajati, madesi, others)
Batch	Year of admission
Father Occupation	Occupation in which father is engaged (agriculture, teaching, government, other)
Program	Engineering program the student is admitted (BCE, BEX, BCT, BME, BEL, BGE)
University	affiliated University of college (TU, PU)
BE percentage	Aggregate percentage secured in bachelor
Failures	Total number of back exams attended before pass out
Semester Percentage	Percentage of student in each semester (sem1,.....,sem8)
Core subjects marks	Marks in core subjects of each semesters for respective program (subj1,.....,subj7)
Back subject counts	Count of back subjects in each year (back1,.....,back4)

4.2 Labeling Instances

The supervised machine learning algorithm must-have label for each instance therefore, we had to label each instance before feeding them to the machine learning algorithm. K- mean clustering was performed on the student's data based on the two attributes; bachelor aggregate percentage and the total number of failures before graduation. The elbow method was used to visualize the optimum numbers of clusters where 4 clusters were found to be optimum as shown in Figure 2. The inertia starts converging after (k=4), therefore it was chosen as the optimum value for clustering.

Cluster plot in Figure 3 show four clusters of students with different performances. It was concluded that cluster 1 (with a high percentage and very low failure counts) belong to students of an excellent class, cluster 3 belongs to a good class, cluster 2 belongs to a medium class, and cluster 4(with very low percentage and high failure counts) belongs to a satisfactory class. A new column named 'Class' was added that include label of students based on the cluster they fall. Among the total 3201 students, 344 attained 'Excellent', 642 attained 'Good', 1286 attained 'Medium' and 949 attained 'Satisfactory' class.

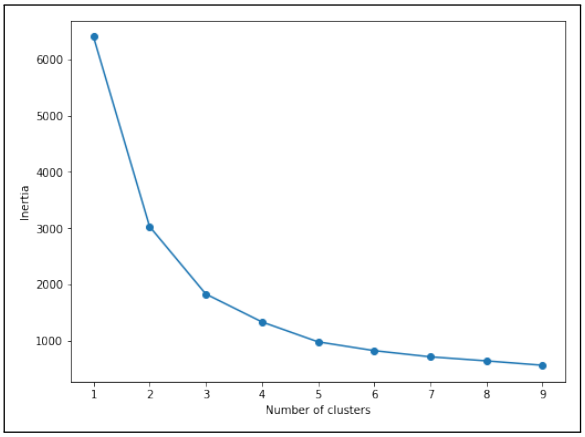

Figure 2: Elbow method to find optimum number of clusters

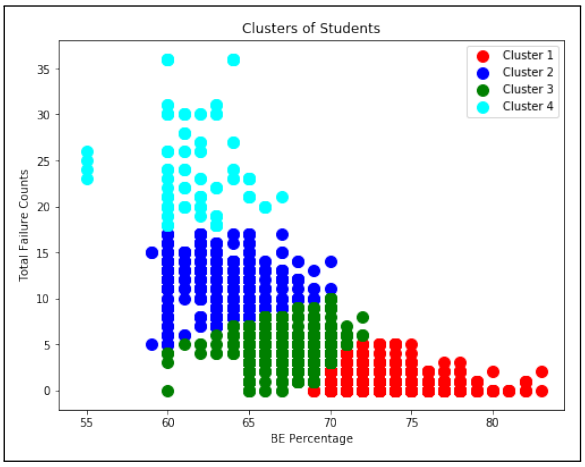

Figure 3: Cluster visualization

4.3 Feature Engineering

The main goal of feature engineering is to prepare input dataset and making it compatible with the machine learning algorithm requirement. It is a very

important step as the performance of a model entirely depends on the standard of the input data. The following techniques were applied for feature engineering:

Exploratory Data Analysis: Basically, the real-world data are inconsistent, prone to error and noise, missing value, and outliers so we used different methods like scatter plot diagram, box plot diagram to visualize data. The missing data were handled by imputing a constant value.

Data Conversion: The data gathered by us consists of both numeric and categorical data. To feed them to the machine learning algorithm all of them need to be converted into numerical form. In our collected data, there were two types of categorical data; ordinal and nominal. For preserving the order of ordinal categorical data we used replace method for data conversion and for the nominal categorical data where there is no natural ordering we used one hot encoding method which is a common data conversion method for nominal categorical data.

Feature Extraction: Feature Engineering deals with discovering new features from the existing ones. For performing the trend analysis of student's academic performance, a new feature set was created that indicates the state of students after each semester examination. The four terminologies had been used to indicate the states as shown in Table 2.

Table 2: Percentage range of every defined states

Percentage	Terminology
Above 80	High
70-80	Good
60-70	Average
Below 60	Weak

4.4 Feature Selection

The inclusion of irrelevant features leads a model towards the curse of dimensionality so, we applied Correlation based Filter Selection (CFS) method for feature selection. The performance of CFS was found to be better than other feature selection algorithms [12]. We calculated the correlation of features with each other and included only one if they were highly correlated (correlation coefficient>0.9) among themselves. After applying this process, 4 features were excluded.

4.5 Model Training and Evaluation

The preprocessed dataset was spitted into training (70%) and testing set (30%). The training set was used to build predictive models and a testing set was used to validate them. For model optimization, 10-fold cross-validation was conducted after balancing the training data using SMOTE (Synthetic Minority Oversampling Technique). The popular library function of the 'sklearns' package 'GridSearchCV' was used for tuning the hyperparameters. It selects the best parameter that fits the estimator on the training dataset. The predictive models were built using traditional classification algorithms; DT, SVM, KNN, and common ensemble methods; Bagging, XGBoost, Voting. The results acquired from the models were validated and comparative analysis was performed in the next stage.

5. Evaluation Measures

The accuracy of the model cannot be a measure for model evaluation in case of imbalanced classes. Therefore, the model evaluation was performed using following evaluation metrics:

Accuracy is defined as ratio of correctly classified instances and the total number of instances

$$\text{accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

Precision is defined as the ratio of correctly classified positives cases and the positively predicted cases

$$\text{precision} = \frac{TP}{TP+FP} \tag{2}$$

Recall is defined as the ratio of correctly classified positive cases and the total positive cases

$$\text{recall} = \frac{TP}{TP+FN} \tag{3}$$

F1-score is the harmonic mean of precision and recall

$$\text{F1-score} = \frac{2*precision*recall}{precision+recall} \tag{4}$$

6. Result and Analysis

6.1 Trend Analysis

To, determine the trend of engineering student's performance, a feature set using four terminologies;

'High', 'Good', 'Average', and 'Weak' was created previously that indicates the performance state of students based on the semester percentage. From Figure 4, it can be observed that the trend line for 'High' and 'Good' states has inclined whereas the trend line for 'Average' and 'Weak' states has declined with ongoing semesters. This indicates that the number of students who attained 'High' and 'Good' states has increased with semester's completion and the number of students who attained 'Average' and 'Weak' states has decreased. From this analysis, we can conclude that engineering students tend to perform better in ending semesters than in commencing semesters.

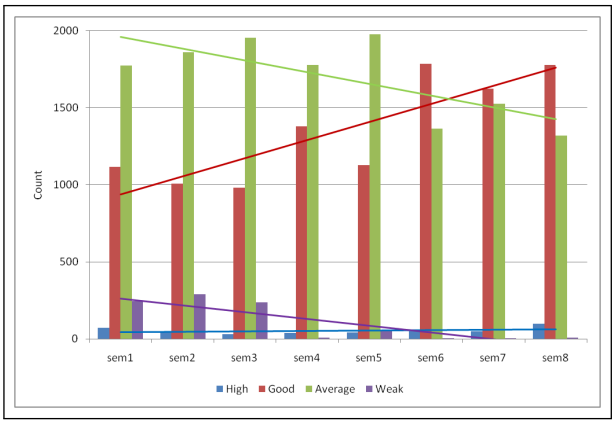

Figure 4: Trend analysis of student's performance

6.2 Feature Importance Analysis

The total data was separated into two parts for proper correlation analysis; Transcript data and Background data. Transcript data consists of those data that were available from a student's bachelor transcript and it may differ based on the engineering program. Background data were the data of the students collected from outside the transcript. We applied Correlation based feature selection method where the attributes are ranked based on Pearson's correlation coefficient.

6.2.1 For Background Data

Figure 5 and Figure 6 show the features (among background data) having a strong positive and negative correlation with the student's academic performance. The figure show that 'plus two percentage', 'Entrance Rank', 'Class Ten percentage' has strong positive correlation and the 'Age' have a negative correlation with the success of students in engineering study. Apart from them, 'Program', 'University', 'Plus two college location', 'School location', 'Father Occupation', 'Ethnicity' also impacts the student's success.

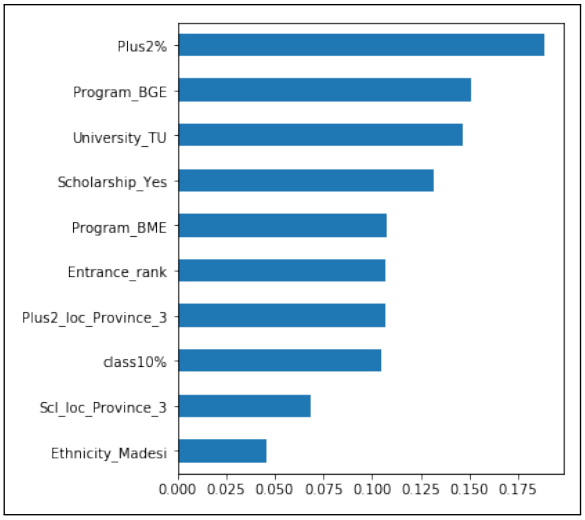

Figure 5: Features with positive correlation

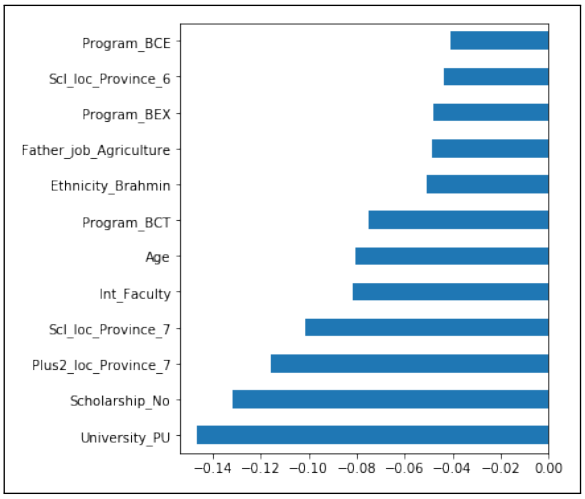

Figure 6: Features with negative correlation

6.2.2 For Transcript Data

The Semester Percentage, Each semester core subject mark (7 subjects), and Count of back subjects each year for respective six engineering programs were included in the dataset with the aim to identify impactful courses, difficult semesters and difficult study years. Table 3 show the top two impactful courses, semesters and study years for respective engineering program using correlation based feature ranking technique. For pattern visualization, we have constructed a rule set for core subjects of the BEX and BCT program using a decision tree as shown in Figure 7 and Figure 8.

Table 3: Summary of impactful courses, semesters, and study years for respective engineering programs

Programs	Impactful Courses	Impactful Semesters	Impactful Study Years
BEX	Microprocessor, Electromagnetism	Sem3, Sem5	Second, Third
BCT	C programming, Object Oriented Programming	Sem5, Sem3	First, Second
BCE	Design of Steel and Timber Structure, Foundation Engineering	Sem4, Sem6	Third, Second
BEL	Electrical Circuit Theory, Electrical Machine Design	Sem5, Sem6	Third, Second
BME	Machine Design II, Fluid Mechanics	Sem6, Sem4	Third, Second
BGE	Design and Implementation of GIS, Fundamental of Geodesy	Sem7, Sem6	Third, Second

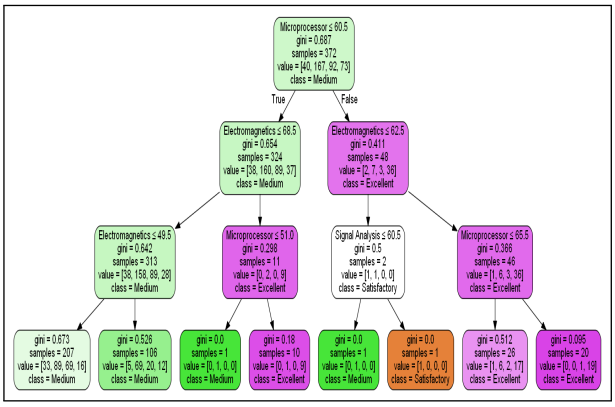

Figure 7: Ruleset using DT for BEX program

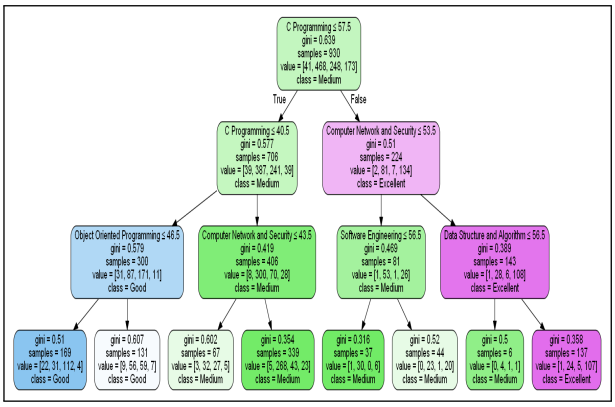

Figure 8: Ruleset using DT for BCT program

6.3 Classification Results

In this study, we developed classification models using three traditional classifiers and three ensemble methods. The models were trained using 70% training data with optimized parameters. For validation, the unseen test data (30%) were fed to each trained model, and evaluation results were compared based on four evaluation metrics. Figure 9 show the predictive accuracy of each classifier. It can be observed that XGBoost performed best with an accuracy of 84.7%. Similarly, from Table 4, it can be observed that ensemble methods has shown better results compared to traditional classifiers in terms of precision, recall, and F1-score.

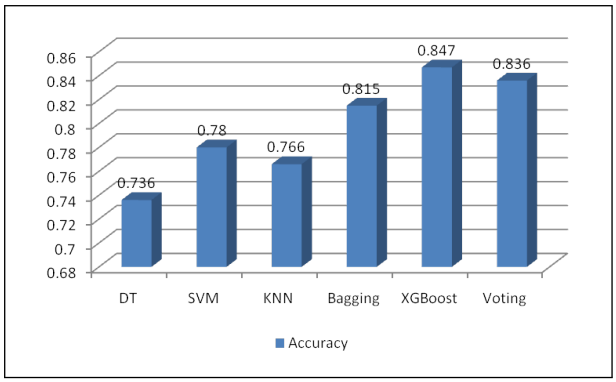

Figure 9: Comparison of Predictive accuracy of all Classifiers

Table 4: Classification results of each Classifier

Algorithm (Weighted Avg.)	Precision	Recall	F1-Score
Decision Tree	0.74	0.74	0.74
Support Vector Machine	0.78	0.78	0.78
K-Nearest Neighbors	0.77	0.77	0.76
Bagging	0.83	0.82	0.81
Ensemble Voting	0.84	0.84	0.84
XGBoost	0.85	0.85	0.85

The learning curve for the predictive model developed by XGBoost is shown in Figure 10. Figure 10 show an acceptable gap between the training and validation accuracy curve indicating no sign of over-fitting and under-fitting.

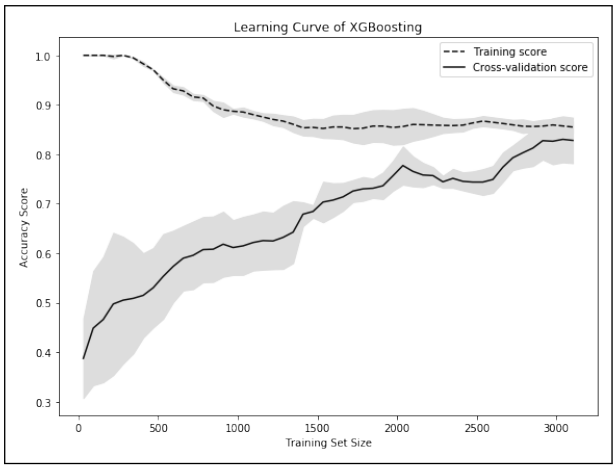

Figure 10: Learning Curve of XGBoost model

7. Conclusion and Future Works

The study was intended to improve the student's pass rates of the engineering institutions in Nepal. Data mining techniques like classification and clustering were utilized to discover useful information for the students, teachers, course advisors, and management. The dataset of 3201 Nepalese engineering graduates from two engineering colleges in Nepal with 36 attributes was used for the study. Predictive models were developed using traditional classification methods as well as ensemble methods with parameter optimization using 10-fold cross-validation. Ensemble methods (Bagging, XGBoost, Voting) demonstrated better classification results compared to traditional methods. Among all the predictive models, a model developed by using XGBoost outperforms (with an F1-score of 85%) others in terms of all evaluation metrics. The study has identified the top two impactful core courses, semesters and year of study for respective six engineering programs (BEX, BCT, BCE, BME, BEL, BGE) using Correlation-based feature selection method. Similarly, visualization of the trend analysis of student's performances show that engineering students in Nepal tend to perform better result in ending period than commencing period. The study has tried to perform EDM using engineering data that can be a step towards enhancing the quality of engineering education in Nepal. The study can be extended by including marks of all the courses and other behavioral attributes.

Acknowledgments

The authors acknowledge the Campus Chiefs of Paschimanchal Engineering Campus and Dhangadhi Engineering College for granting permission to collect data of students. The authors also Er. Hari Baral, Er. Nabin Lamichanne, Er. Ramesh Thapa for their invaluable guidance, comments, and suggestions.

References

[1] Mashael A Al-Barrak and Muna Al-Razgan. Predicting students final gpa using decision trees: a case study. *International journal of information and education technology*, 6(7):528, 2016.

[2] Tri Ratna Bajracharya, Babu Ram Dawadi, and Ram Chandra Sapkota. Restructuring examination system of institute of engineering for establishing center of excellence in engineering education. *Journal of the Institute of Engineering*, 14(1):75–81, 2018.

[3] Mrinal Pandey and S Taruna. A comparative study of ensemble methods for students' performance modeling. *International Journal of Computer Applications*, 103(8), 2014.

[4] Mukesh Kumar and Yass Khudheir Salal. Systematic review of predicting student's performance in academics. *Int. J. of Engineering and Advanced Technology*, 8(3):54–61, 2019.

[5] Sadiq Hussain, Neama Abdulaziz Dahan, Fadl Mutaher Ba-Alwib, and Najoua Ribata. Educational data mining and analysis of students' academic performance using weka. *Indonesian Journal of Electrical Engineering and Computer Science*, 9(2):447–459, 2018.

[6] Md Rifatul Islam Rifat, Abdullah Al Imran, and ASM Badrudduza. Educational performance analytics of undergraduate business students. *International Journal of Modern Education and Computer Science*, 11(7):44, 2019.

[7] R. Soobramoney and A. Singh. Identifying students at-risk with an ensemble of machine learning algorithms. In *2019 Conference on Information Communications Technology and Society (ICTAS)*, pages 1–6, 2019.

[8] Pooja Kumari, Praphula Kumar Jain, and Rajendra Pamula. An efficient use of ensemble methods to predict students academic performance. In *2018 4th International Conference on Recent Advances in Information Technology (RAIT)*, pages 1–6. IEEE, 2018.

[9] Babu Dawadi and Daya Baral. Towards automation in the admission process as a tool to enhance quality of engineering education at tribhuvan university. *Journal of The Institute of Engineering*, 13, 05 2017.

[10] Tek Bithari, Sharan Thapa, and Hari K.C. Predicting academic performance of engineering students using ensemble method. *Technical Journal*, 2(1):89–98, Nov. 2020.

[11] Fazal Aman, Azhar Rauf, Rahman Ali, Farkhund Iqbal, and Asad Masood Khattak. A predictive model for predicting students academic performance. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–4. IEEE, 2019.

[12] T Velmurugan and C Anuradha. Performance evaluation of feature selection algorithms in educational data mining. *Performance Evaluation*, 5(02), 2016.