JCBI
Journal of Computing & Biomedical Informatics Volume 09 Issue 02
ISSN: 2710 - 1606 2025

Research Article
https://doi.org/10.56979/902/2025

A Comparative Study of Data Mining Techniques for Predicting Candidates' Performance Based on High School Records

M. Imran[1], Saad Ahmed[2], Raheela Asif[1], Saman Hina[3], and Hira Farman[2]

[1]Department of Computer Science and Information Technology, NED University, Karachi, Pakistan.
[2]Department of Computer Science, IQRA University, Karachi, Pakistan.
[3]Department of Computing, Imperial College London, England.
*Corresponding Author: Saad Ahmed. Email: saadahmed@iqra.edu.pk

Received: June 25, 2025 **Accepted:** August 04, 2025

Abstract: Pre-admission entry requirements vary between higher education institutions. The pre-test performance of applicants in the entrance exam of a public engineering university is examined in this article based on their high school grades. Additionally, this study looked into the relationship between the scores obtained on high school exams and the entrance exam. Creating a prediction model to aid admissions committees in their work was the study's main objective. The candidates' pre-admission entry exam scores as well as their high school scores are included in the dataset. To determine the association between these scores, a number of statistical analytic methods and machine learning algorithms, "Decision trees", "K-Nearest Neighbour, "Neural Networks," " and "Naive Bayes" were used. Accuracy metrics was applied to calculate the performance of the selected methods.. The 'Decision tree' and 'Neural network' models fared better than the other models, according to the results. The study concluded that the high school scores are significant predictors of pre-test performance and that machine learning models can be effective tools in predicting the performance of applicants in university entrance exams. The results of this study have applications for universities and admissions committees because they can be used to enhance selection procedures and spot potential high achievers.

Keywords:. Higher Education; Machine Learning; Predictive Model

1. Introduction

The data mining techniques are used to mine crucial information from enormous amounts of data. These techniques are used to extract patterns, association or relationship between data points. The extracted knowledge can then help to know the future trends of data to make proactive decisions. Numerous data mining applications have been successfully used in a variety of industries, including finance, manufacturing, telecommunications, healthcare, science, and fraud and risk analysis.

Predicting applicants' performance is a critical task in data mining. It helps to know not only that how the applicants learn but also how they will perform or underperform. Educational Data Mining (EDM) is a domain-specific field of study that focus on the development of new methods for mining the educational data. EDM methodologies are used to explore the deep understanding of students' learning using the distinct sort of data that originate from educational settings [1]. Moreover, hidden facts in data can be predicted to improve quality of education and learning process. [2] Proposed that analyzed the connection between pre-admission standards and pupil performance in medical courses at a Saudi public university. In their research, they found that the optimized weightings not only improved the student performance but also predicted the best applicants for medical programs. In another research [3]. Investigated the correlation between students' academic performance and their internal assessment marks

Prediction of students' performance in different professional disciplines has also been predicted using external independent testing and intermediate certifications/sessions [4]. Their predictive model was able to identify students who may struggle in certain subjects or topics and those who excel in particular areas. In another study [5], applied some criteria like high school grade, average, and score in test with data mining techniques to predict students' academic performance at the university with high accuracy. This can assist universities in identifying high-performing applicants and making more informed admissions decisions.

Researchers working in this domain have mostly investigated correlation between students' performance in internal assessments or during the course of study [6, 7]. However, this research focuses on more important investigation that is prior to admission in graduate degree programs. Other than the eligibility criteria, the mechanism focuses on the applicants' performance in pre-admission entry test.

This research is framed to assist the students in their entrance examination preparations. For this purpose, high school scores for the admission in public sector engineering university in Pakistan were analyzed for the prediction of their performance in pre-admission entry test scores. In addition to this study, correlation between high school scores and pre-admission entry test scores has also been evaluated. For this purpose, various data mining algorithms and correlation methods were employed and assessed. The applied algorithms comprised "Decision Trees (DT)", "Neural Networks (NN), K-Nearest Neighbor (KNN), and Naïve Bayes. For analysis and visualization of the results, an open-source data mining tool called "Rapid Miner" was utilized.

2. Literature Review

Data mining methods are used to identify and evaluate the variables or patterns in existing data that helps to enhance the learning process. EDM is a developing research field in the domain of education (Romero and Ventura 2010). Some researchers applied DT model to predict the students' division in their final year of the bachelor's program [8] while others predicted students' success either pass or fail using machine learning methods (Naive Baye, DT and NN methods) [9]. DT was also used to analyze students' performance using the factors like class attendance, tests, seminars and assignments [10]. The study's objectives were to locate dropouts and find students who required extra help. They used the Bayes classifier to analyze the factors (students' category, language, and background qualifications) that affect students' performance as either the student is performer or underperformer. Ayesha S. et al. ([11]) proposed clustering algorithm (k-means) to foresee learning activities of students. Al-Radaideh et al., [12] predicted the final grade of students using decision tree (ID3, C4.5) and Naïve Bayes model. Yadav et al. [13] used various decision tree algorithms, such as C4.5, ID3, and CART, to analyze the data of engineering students to predict their performance (pass or fail) on the final exam. A case study performed [14] on the data of post-graduate students to predict students' performance at the end of the semester using decision trees, neural networks, and clustering techniques. Singh et. Al [15] conducted a study using decision tree to predict their final grade.

In the same context, Hijazi S.T. et al. [16] conducted a study on a sample of 300 students (225 males and 75 females) by examining their attendance in class, daily time spent after college, family income, and mother's level of education. The study found that educational attainment of mother and family income highly impacted the academic performance of students.

In a research led by Khan [17] 400 students (200 boys and 200 girls) were investigated by different factors that include; demographic, personality and other variables that can influence students' performance in academia. The findings revealed that girls with having high socio-economic backgrounds do better in science while boys with lower socio-economic backgrounds perform better in general subjects. Another study conducted by Shah N.S. [18] involved a sample of 231 students out of 637. The objective was to predict the performance levels of students, categorizing them as best or poor, by analyzing multiple factors influencing their academic achievements. The students were divided into the following categories: fail, very good, good, average, satisfactory, and below satisfactory. Similarly, Walters et al. [19] studied; factors such as gender, grade level, school location, student type, and socioeconomic background have a substantial influence on the academic performance of students.

Romero et al. [20] using Moodle to classify students, different data mining techniques were evaluated for effectiveness and performance. Ahmedi et al. [21] compared real data with predicted data by choosing

different attributes of students and found that the knowledge produced using incomplete data is less trustworthy than using complete data. They discovered that if the data is complete, the C4.5 algorithm performed best for predicting students' success. [22] Applied various data mining algorithms to analyze the learning model. They classified the performance as average, below average, or above average, and the data set—which includes 3500 records with 99 attributes each—was taken from the learning model for classes 1 through 7. Pal et al. [10] performed a study to identify the weak students and lower the failure rate of the students by predicting the division of a newcomer based on the analysis of data from the previous year. Elaraby et al. [2] applied DT (ID3) to identify the students who required extra attention by forecasting the final grade of the class. Marquez et al. [19] applied 10 classification algorithm (5 induction algorithm and 5 decision trees algorithm) using WEKA (a data mining software) to predict the students' performance, specifically the students who might fail. Kabakchieva [23] found similarities in the data that might be utilized to forecast student performance based on their pre-university and personal variables using a variety of classification methods.

Alhazmi, E. et. al. [24] used data mining methods to evaluate the performance of students and pinpoint the effect of early-stage factors on students' GPA. The study employed both clustering and classification techniques using machine learning models to explore the relationship between admission scores, academic achievement tests, general aptitude tests, and GPA. The results showed that educational systems can reduce the risk of students' failures in the early stages.

In [25] a framework was developed for training and evaluating machine learning models for academic success based on different measures like credits, grade, and exams. The evaluation generates PDF reports, and the tool is designed for socially responsible AI with fairness checks for identifying potential discrimination in data and predictions.

The study proposed in [26] used orange technology for data mining to identify educational trends and forecast university students' academic performance at King Khalid University. K-Means clustering grouped students into three layers based on their scores, and Linear Regression was the most effective algorithm for predicting academic performance. Most courses' academic performance was predicted using activities and quarterly tests, except for one course where only semester tests were effective. SVM algorithm was found to be the least effective in predicting academic performance.

In the research study [27] developed a model to predict whether students will graduate on time using data on age, gender, test results, and final exam scores by implementing top predictive algorithm, PART, J48, Random Forest and Bayes Net. The J48 algorithm emerged as the most effective among the other employed algorithms.

Sateesh, N et. al. [28] proposed ensemble classifier based on rule mining for predicting students' performance in academia using by employing "weighted Rough Set Theory" method. The proposed technique achieved 92.77% and 94.87% accuracy and sensitivity rate respectively that demonstrate its superiority over traditional approaches in predicting student performance. Students' performance was also examined for online education [29] to analyze how students behave while watching online course videos, such as pausing, forwarding, and rewinding, can be used to predict their test performance. The study was conducted in two experiments, one with 22 university students and basic statistical methods, and another with 16 students and more advanced data mining techniques. The results showed that students who clicked more, rewound or paused more slowly had better test performance than those who clicked less or used fast-forward.

The aim of the authors [3] was to investigate the correlation between students' academic performance and their internal assessment marks. To accomplish this, a survey questionnaire and a qualitative data collection method "Focus Group Discussion (FGD)" were used. The results of the study revealed a modest correlation between internal assessment scores and final exam scores.

This work [30] proposed student achievement prediction system that uses educational data mining algorithms, with Qiannan Normal University for Nationalities in China as a case study. The system predicts student achievement by combining K-means, decision tree, box technology, and SVM algorithms. The system achieved an accuracy rate of 78.3% and meets the ISO/IEC 25010 software quality standard.

The approach proposed [6] aims to increase the precision of a model used to predict academic performance for higher studies. The results revealed that the Nave Bayes produced high accuracy as compare to other classification models. The study focuses on internal marks and sessional marks to

accomplish this task. Early prediction of students' academic performance can help identify factors that may lead to their failure in academia, thereby enabling timely intervention to address these issues.

The focused of the research in [7] is to analyze undergraduate students' performance employing data mining methods. The study identifies two groups of students: low and high achievers. By identifying key courses that indicate good or poor performance, the study suggests that targeted support can be provided to low and high performing students. Data mining techniques can help identify at-risk students early and provide the necessary interventions to improve academic performance

This paper [6] presented a new approach for predicting student performance by incorporating behavioral features related to their e-learning management system interactions. The model is evaluated using various classifiers and ensemble methods, with results showing up to 25.8% improvement in accuracy. The study found that behavioral features are strongly correlated with academic achievement, and testing on newcomer students resulted in an accuracy of over 80%.

Based on the reported studies, it is evident that entrance examination plays a vital role in getting admission in university. The objective of this proposed research is to predict the applicant's performance in entry test using their marks in the subjects, 'Physics', 'Chemistry' and 'Mathematics' secured in High School examinations. The student's score on the entrance exam determines whether or not they are qualified for admission. Predicting whether a student will clear the entrance exam or not is therefore crucial. Extra efforts may be made to address a student's weaknesses if the prediction indicates that the student tends to perform poorly on the entrance examination prior to the entrance examination.

3. Materials and Methods

3.1. Preprocessing of Dataset

The data used for this study was gathered over a five-year period from the admissions database of a public sector engineering university. There are thousands of records in this dataset. Table 1 lists the attributes that are selected for this study.

The primary task of data mining is data collection, which is a time-consuming process. The pre-processed and completeness of the data are key considerations when collecting data [21]. It is crucial to emphasize that knowledge produced from incomplete data is never as reliable as knowledge produced from complete data [22].

Table 1. Attributes of Selected Dataset

Attribute	Description	Possible Values
Applicant_Code	Unique ID assigned to each applicant	Integer (unique)
Admission_Year	Year of admission	{Year1, Year2, Year3, Year4, Year5}
HS_Total_Marks	Total marks obtained in High School	660 – 1100
HS_Math_Marks	Marks obtained in Mathematics in High School	120 – 200
HS_Physics_Marks	Marks obtained in Physics in High School	120 – 200
HS_Chemistry_Marks	Marks obtained in Chemistry in High School	120 – 200
PAT_Total_Marks	Total marks obtained in Entry Test	1 – 100
PAT_Mathematics	Marks obtained in Mathematics in Entry Test	1 – 25
PAT_Physics	Marks obtained in Physics in Entry Test	1 – 25
PAT_Chemistry	Marks obtained in Chemistry in Entry Test	1 – 25
PAT_Result	Final result based on Entry Test	{Pass, Fail}

In order to identify the necessary and valuable features from the selected data, prior knowledge of the data is more helpful. Reviewing previous practices, paperwork, and forms, as well as speaking with staff members who are in charge of keeping the data, were used to get the knowledge for this job regarding the chosen data. The admission database including the five years' worth of records is where the data for this

study was gathered. The necessary officials have given their prior consent for the use of the data in this study.

Table 2. Year-wise Pass and Fail Applicants

S.#	Year for EntryTest	Pass	Fail	Total
1.	Year1	2426	207	2633
2.	Year2	2353	331	2684
3.	Year3	2578	654	3232
4.	Year4	3624	899	4523
5.	Year5	3642	512	4154
	GRAND TOTAL	**14623**	**2603**	**17226**

The admission database was made up of multiple tables with distinct properties. To reduce the noise, all redundant and irrelevant attributes are deleted, and any necessary preprocessing is done to prepare the data for data mining. To obtain precise and trustworthy findings, only necessary attributes are chosen. For this investigation, we used a dataset of applicants who were members of a certain board. Table 2 provides information about this board's pass and fail candidates by year.

This dataset was further segmented into Male and Female candidates, with each grouping of information further broken down into Pass and Fail to identify the qualities' correlations. Details of the data is tabulated in Table 3 and Table 4.

Table 3. Year-wise Pass and Fail Female Applicants

S.#	Year for Entry Test	Pass	Fail	Total
1.	Year1	755	054	809
2.	Year2	692	071	763
3.	Year3	785	145	930
4.	Year4	1175	176	1351
5.	Year5	1109	058	1167
	GRAND TOTAL	**4516**	**504**	**5020**

4. Methodology

After the data had been preprocessed, data mining techniques were applied to the datasets to find patterns and correlations between the different values. The study's main objective was to predict how applicants would perform on the pre-admission entry test, which helps university officials identify applicants who would not pass the test and assesses how well they know the three main subjects (Maths, Physics, and Chemistry) as well as how difficult the test will be. The most popular classification techniques include, "DT", "NN," "K-NN," and "Naive Bayes" have been used to address the aforementioned research difficulties. The data was analyzed and the results were presented using the free source data mining application Rapid Miner.

The proposed research includes two analyses; the first aims to expose whether a correlation exists between the high school marks and entrance test marks while the aim of second was to perform early prediction of applicant's performance in entrance test based on their high school marks.

Analysis-1: Correlation between total marks and HSC marks all Male and Female applicants

The correlation matrix operator in Rapid Miner is used to determine the correlation between the attributes. There was a measured correlation between the entry test scores and the high examination scores. We compared the overall marks attained in both the admission test and the high school final exams. Additionally, the relationship between success on various high school topics (such as Mathematics, Physics, and Chemistry) and entry test examinations was examined.

To find the correlation among various attributes, all the applicants applied for admission from the Year-1 to Year-5 were selected. These data were further divided into pass and fail, pass only, and fail only and categorized the as follows:

Table 4. Analysis table

Category-1	Category-2	Category-3
All Male and Female Applicants	**All Female Applicants**	**All Male Applicants**
•Pass & Fail •Pass only •Fail only	•Pass & Fail •Pass only •Fail only	•Pass & Fail •Pass only •Fail only

Four data mining algorithms—DT, NN, KNN, and Naive Bayes — were used to predict the performance of the applicants. For the experiment, two datasets were made.

Dataset I: Training Data Year1 and Year2 while the testing data Year3.

Dataset II: Training Data Year4 and the testing data Year5.

The classifier models are as follows:

As stated earlier in this section, correlations among different attributes of the selected data have been analyzed. Correlation, in general, describes the connection between two attributes or two data sets. Correlation coefficient values can be used to gauge how closely two attributes are related. A positive or negative correlation coefficient value is possible. If the value is negative, the relationship between the attributes is weak. The value of the correlation coefficient increases as the strength of the relationship between the attributes increases.

The datasets have been separated into three groups in order to determine the link between the various high school characteristics and the entrance exam: First, information on all applicants who passed or failed the entrance exam. Second, the information about all applicants who passed the entrance test, and third is information about all applicants who failed the entrance test. To conduct a thorough analysis of the association, the data were divided into many groups. The results of the correlation for each category of data are tabulated from Table 5-7. Table 5 displays the correlation between all applicants' qualities, while Table 6 solely displays the connection between candidates who passed. The outcomes of the applicants who failed the entry test are shown in Table 7. Table 5 clearly reveals that the total scores in entry tests for Years 1 and 2 are significantly correlated with the total scores in high school exams. The correlation between high school grades in math, physics, and chemistry and the results on entry exams for those subjects is, however, quite weak.

Table 5. Correlation between High School score and Test score (All Pass and Fail Applicants)

S. #	High School Attribute	Entry Test Attribute	Year1	Year2	Year3	Year4	Year5
1	HSC_Total_Marks	PAT_Total_Marks	0.595	0.586	0.458	0.440	0.327
2	HSC_Mathematics	PAT_Mathematics	0.450	0.406	0.320	0.359	0.157
3	HSC_Physics	PAT_Physics	0.478	0.475	0.308	0.308	0.215
4	HSC_Chemistry	PAT_Chemistry	0.422	0.436	0.350	0.239	0.169

Table 6. Correlation between High school scores and Test scores (all pass applicants)

S. #	High School Attribute	Entry Test Attribute	Year1	Year2	Year3	Year4	Year5

1	HSC_Total_Marks	PAT_Total_Marks	0.582	0.607	0.508	0.525	0.405
2	HSC_Mathematics	PAT_Mathematics	0.443	0.473	0.398	0.440	0.255
3	HSC_Physics	PAT_Physics	0.461	0.485	0.364	0.374	0.286
4	HSC_Chemistry	PAT_Chemistry	0.380	0.440	0.373	0.311	0.235

Table 6 demonstrates that while the correlation between these variables is significant in other years, there is a significant association between the total marks obtained in the high school examination and the total marks obtained in the entry test in Year2. Other variables' correlations are almost identical as of Table 5. Table 7 with the data only includes applicants who failed the entry test that reveals no correlation between the variables.

Table 7. Correlation between High school scores and Test scores (all fail applicants)

S. #	High School Attribute	Entry Test Attribute	Year1	Year2	Year3	Year4	Year5
1	HSC_Total_Marks	PAT_Total_Marks	-0.015	-0.114	-0.046	-0.114	0.031
2	HSC_Mathematics	PAT_Mathematics	0.060	-0.070	-0.002	-0.064	-0.031
3	HSC_Physics	PAT_Physics	0.026	-0.021	-0.102	-0.062	0.078
4	HSC_Chemistry	PAT_Chemistry	0.066	-0.041	-0.009	-0.059	0.060

Analysis-2: Correlation between total marks and HSC marks of Female applicants: We further divided our data into male and female applicants in order to determine the correlation between the variables. Each of these genders was then further divided into the following categories:

- Pass and Fail: Include all Male and Female applicants who cleared or not cleared the entry test.
- Pass only: Include all Male and Female applicants who cleared the entry test.
- Fail only: Include all Male and Female applicants who not cleared the entry test.

The correlation among attributes of all female applicants who passed or unable to pass the entry test is depicted in Table 8. Table 9 represented correlation of all female applicants who cleared (pass) the entry test while Table 10 shows the results of female applicants who did not cleared the entry test.

Table 8 demonstrates that, from Year 1 through Year 4, there is a strong correlation be-tween the total marks achieved in the HSC examination and the total marks obtained in the entry test, with the exception of Year 5, where there is a weak correlation between these attributes. While there is a weak correlation between the individual marks in math, physics, and chemistry at the entry test and the marks at the high school examination, there is a correlation nonetheless.

Table 8. Correlation between High school scores and Test scores (Female pass and fail applicants)

S. #	High School Attribute	Entry Test Attribute	Year1	Year2	Year3	Year4	Year5
1	HSC_Total_Marks	PAT_Total_Marks	0.627	0.651	0.557	0.610	0.494
2	HSC_Mathematics	PAT_Mathematics	0.506	0.490	0.394	0.513	0.333
3	HSC_Physics	PAT_Physics	0.491	0.510	0.440	0.456	0.369
4	HSC_Chemistry	PAT_Chemistry	0.416	0.482	0.448	0.416	0.295

Table 9 demonstrates that there is a significant correlation between the HSC examination's overall score and the entry test's overall score for each year (Years 1 through 5). However, there exist no or weak correlation among individual subject marks secured high school examinations and entry test marks.

Table 9. Correlation between High school scores and Test scores (Female pass applicants)

S. #	High School Attribute	Entry Test Attribute	Year1	Year2	Year3	Year4	Year5
1	HSC_Total_Marks	PAT_Total_Marks	0.604	0.654	0.562	0.611	0.577
2	HSC_Mathematics	PAT_Mathematics	0.474	0.509	0.406	0.520	0.408
3	HSC_Physics	PAT_Physics	0.454	0.496	0.440	0.476	0.438
4	HSC_Chemistry	PAT_Chemistry	0.388	0.481	0.458	0.406	0.359

Table 10 demonstrates that none of the characteristics of the data set consisting of the female applicants who are still unable to pass the entry test are correlated.

Table 10. Correlation between High school marks and Test marks (Female fail applicants)

S. #	High School Attribute	Entry Test Attribute	Year1	Year2	Year3	Year4	Year5
1	HSC_Total_Marks	PAT_Total_Marks	0.021	-0.176	-0.238	-0.131	-0.021
2	HSC_Mathematics	PAT_Mathematics	0.332	0.002	0.094	0.037	-0.130
3	HSC_Physics	PAT_Physics	0.208	-0.090	-0.077	-0.146	0.129
4	HSC_Chemistry	PAT_Chemistry	0.035	-0.139	-0.152	0.022	-0.026

Analysis 3: Correlation between total marks and HSc marks of Male applicants: In experiment 3, we chose data from applicants who were male and were further broken down into those who passed the test but did not clear it, those who passed the test, and those who did not.

Table 11 demonstrates that there is only a marginal relationship between the total marks earned in the HSC examination and the total marks earned in the Year 1–Year 3 entry test. While there is no connection between the individual HSC exam scores in mathematics, physics, and chemistry and the entry test scores in those subjects, there is a correlation between the two.

Table 11. Correlation between High school scores and Test scores (Male pass and fail applicants)

S. #	High School Attribute	Entry Test Attribute	Year1	Year2	Year3	Year4	Year5
1	HSC_Total_Marks	PAT_Total_Marks	0.584	0.560	0.432	0.374	0.285
2	HSC_Mathematics	PAT_Mathematics	0.428	0.377	0.301	0.306	0.127
3	HSC_Physics	PAT_Physics	0.475	0.461	0.260	0.254	0.197
4	HSC_Chemistry	PAT_Chemistry	0.424	0.418	0.314	0.178	0.137

Table 12 reveals a weak correlation from Year1 to Year4, but no correlation exists in Year5. Additionally, there is no correlation between individual subject marks in mathematics, physics, and chemistry in high school examination and corresponding marks in the entry test.

Table 12. Correlation between High School Marks (Male pass applicants)

S. #	High School Attribute	Entry Test Attribute	Year1	Year2	Year3	Year4	Year5
1	HSC_Total_Marks	PAT_Total_Marks	0.577	0.591	0.486	0.478	0.345
2	HSC_Mathematics	PAT_Mathematics	0.430	0.458	0.393	0.397	0.213
3	HSC_Physics	PAT_Physics	0.466	0.480	0.331	0.314	0.251
4	HSC_Chemistry	PAT_Chemistry	0.378	0.425	0.338	0.258	0.192

Table 13 demonstrates that there is no correlation between any of the characteristics of the data set made up of male applicants who are still unable to pass the entry test.

Table 13. Correlation between High school scores and Test scores (Male fail applicants)

S. #	High School Attribute	Entry Test Attribute	Year1	Year2	Year3	Year4	Year5
1	HSC_Total_Marks	PAT_Total_Marks	-0.026	-0.094	0.014	-0.088	0.039
2	HSC_Mathematics	PAT_Mathematics	-0.006	-0.081	-0.012	-0.061	0.008
3	HSC_Physics	PAT_Physics	-0.039	0.001	0.006	0.005	0.007
4	HSC_Chemistry	PAT_Chemistry	0.065	-0.003	0.047	-0.023	0.074

We implemented various machine learning models, DT, NN, KNN and Naïve Bayes algorithms with different parameters settings to assess its impact on the model's accuracy. It is observed that the decision tree algorithm with the parameters, leaf size 30 and max depth 10 produced best results with an accuracy of 79.98% among all other algorithms on dataset-I. However, the neural network model is found produced best results with an ac-curacy of 89.25% on dataset-II. It is also noted that other classifiers also showed an acceptable accuracy.

Table 14. Prediction results on dataset-I

Classifier	Dataset I				
	Accuracy	Pass		Fail	
		TP	FP	TP	FP
Decision Tree	79.98%	2549	619	35	28
Neural Networks	79.76%	2577	654	00	00
Naïve Bayes	76.26%	2273	463	191	304
KNN	76.42%	2334	519	135	243

Table 15. Prediction result on dataset-II

Classifier	Dataset II				
	Accuracy	Pass		Fail	
		TP	FP	TP	FP
Decision Tree	87.44%	4207	443	97	175
Neural Networks	89.25%	4300	447	93	82
Naïve Bayes	80.46%	3869	449	91	513

KNN	77.71%	3675	390	150	707

It was observed that the decision tree model for data set-I (Figure 1) demonstrates that the majority of applicants who received 768 or higher are likely to have passed the entry test. If an applicant's scores fall below 768, the outcome will be determined by comparing the applicant's high school cumulative, HSC Chemistry, and HSC_Mathematics scores. Additionally, it should be noted that HSC Physics does not factor in marks earned.

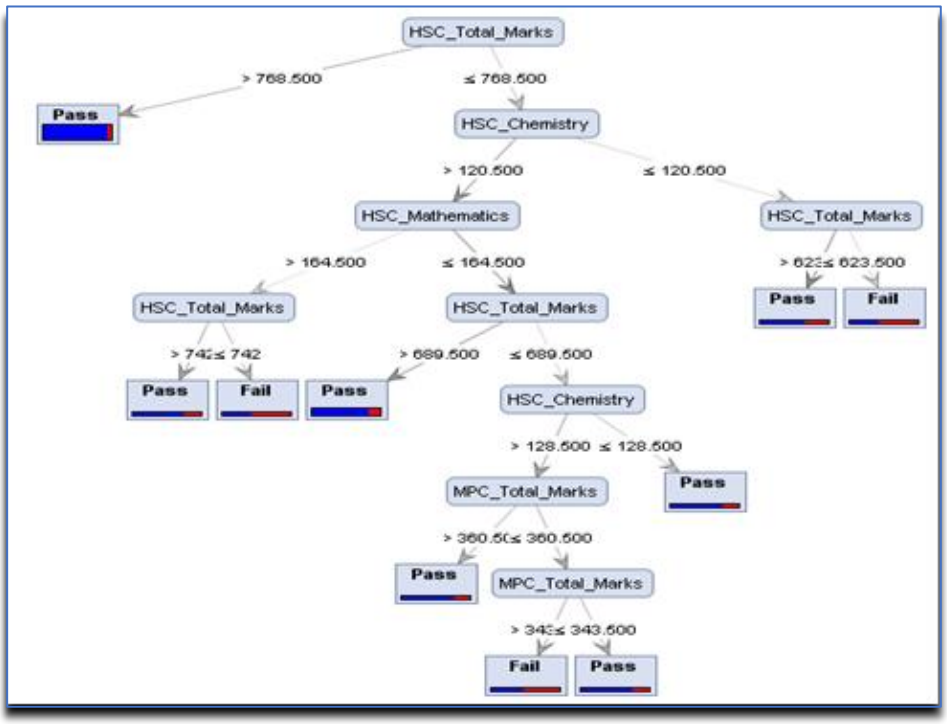

Figure 1. Decision Tree (Dataset-I)

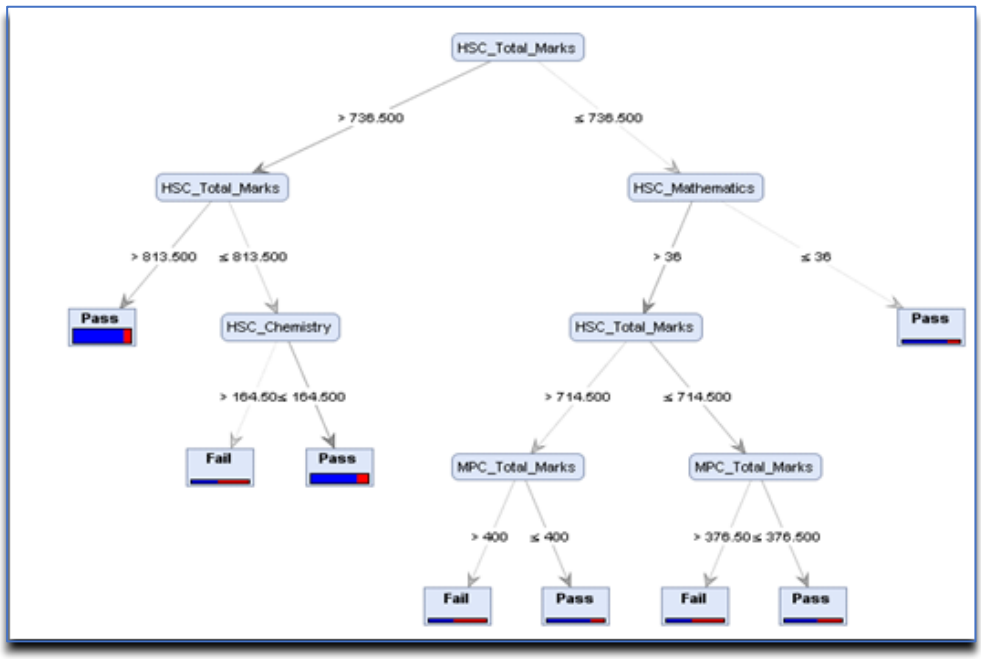

Figure 2. Decision Trees (Using Dataset II)

It was observed that the decision tree model for data set-II (Figure 2) demonstrates that the majority of applicants who received 813 or higher are likely to have passed the entry test. If an applicant's scores fall below 813, the outcome will be determined by comparing the applicant's high school cumulative, HSC Chemistry scores. Similarly, if an applicant's scores fall below 736, the outcome will be determined by

comparing the applicant's high school cumulative, and HSC_Mathematics scores. Additionally, it should be noted that HSC Physics does not factor in marks earned.

5. Conclusions

The objective of this study was to determine the correlation (if any) among the various at-tributes of high school examinations with the attributes of the Pre-Admission Entry Test (PAT) in the data sets, as well as to forecast applicants' success in the entry test. To find the correlation, we chose data from the Karachi board candidates and classified them as all male and female applicants, female applicants alone, and male applicants solely. All of this information was classified as successful and not successful, successful only, and the students who did get the success in the entrance examinations only. Viewing the results, we found that there was no consistency among the applicants. This means that an applicant who scored highly on mathematics, physics, and chemistry in high school does not necessarily need to score similarly highly on those same subjects on the entry test, and vice versa for an applicant who scored poorly on those same subjects in high school.

This analysis can be broadened to incorporate additional data sets, such as applicant’s information from other boards, in order to ascertain the pattern of scoring well on each of the aforementioned criteria as well as the relationships between them. In addition to the ones used in this study, other data mining techniques can be used to produce more accurate results.

References

1. C. Romero and S. Ventura, "Educational data mining: a review of the state of the art," IEEE Trans. Syst., Man, Cybern., vol. 40, no. 6, pp. 601–618, 2010.
2. A. Abdelmagid and A. Qahmash, "Utilizing the Educational Data Mining Techniques 'Orange Technology' for Detecting Patterns and Predicting Academic Performance of University Stu-dents," Information Sciences Letters, vol. 12, no. 3, pp. 1415–1431, 2023. [Online]. Available: http://dx.doi.org/10.18576/isl/120330
3. D. M. Thapa and S. Shakya, "Academic Performance Prediction Based on Internal Assessment Using Educational Data Mining Techniques: Shifting the Paradigm," in Intelligent Computing & Optimization, Springer, 2022. https://doi.org/10.1007/978-3-031-19958-5_49
4. O. Pronina and O. Piatykop, "Predicting Students' Academic Performance Based on the Cluster Analysis Method," in ICTERI 2021 Workshops, Springer, 2022.
5. H. A. Mengash, "Using data mining techniques to predict student performance to support decision making in university admission systems," IEEE Access, vol. 8, pp. 55462–55470, 2020.
6. R. Asad, S. Arooj, and S. U. Rehman, "Study of Educational Data Mining Approaches for Student Performance Analysis," Technical Journal, vol. 27, no. 1, pp. 68–81, 2022.
7. R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' perfor-mance using educational data mining," Computers & Education, vol. 113, pp. 177–194, 2017. http://doi.org/10.1016/j.compedu.2017.05.007
8. R. Asif, A. Merceron, and M. K. Pathan, "Mining student's admission data and predicting stu-dent's performance using decision trees," in ICERI2012 Proceedings, IATED, 2012.
9. E. Osmanbegovic and M. Suljic, "Data mining approach for predicting student performance," Econ. Rev. J. Econ. Bus., vol. 10, no. 1, pp. 3–12, 2012.
10. U. K. Pandey and S. Pal, "Data Mining: A prediction of performer or underperformer using classification," arXiv preprint, arXiv:1104.4163, 2011. https://doi.org/10.48550/arXiv.1104.4163
11. S. Ayesha, T. Mustafa, A. R. Sattar, and M. I. Khan, "Data mining model for higher education system," Eur. J. Sci. Res., vol. 43, no. 1, pp. 24–29, 2010.
12. Q. A. Al-Radaideh, E. M. Al-Shawakfa, and M. I. Al-Najjar, "Mining student data using decision trees," in Proc. Int. Arab Conf. Inf. Technol. (ACIT'2006), Yarmouk University, Jordan, 2006.
13. S. K. Yadav and S. Pal, "Data mining: A prediction for performance improvement of engi-neering students using classification," arXiv preprint, arXiv:1203.3832, 2012.
14. E. R. Chuchra, "Use of data mining techniques for the evaluation of student performance: a case study," Int. J. Comput. Sci. Manag. Res., vol. 1, no. 3, 2012.
15. S. Samrat and K. Vikesh, "Classification of Student's Data Using Data Mining Techniques for Training & Placement Department in Technical Education," Int. J. Comput. Sci. Netw., vol. 1, no. 4, pp. 1–4, 2012.
16. S. T. Hijazi and S. Naqvi, "Factors Affecting Students' Performance," Bangladesh e-Journal of So-ciology, vol. 3, no. 1, 2006.
17. Z. N. Khan, "Scholastic Achievement of Higher Secondary Students in Science Stream," Online Submission, vol. 1, no. 2, pp. 84–87, 2005.
18. [18] N. S. Shah, "Predicting factors that affect students' academic performance by using data mining techniques," Pakistan Business Review, vol. 13, no. 4, pp. 631–638, 2012.
19. Y. Beaumont-Walters and K. Soyibo, "An analysis of high school students' performance on five integrated science process skills," Res. Sci. Technol. Educ., vol. 19, no. 2, pp. 133–145, 2001. [Online]. Available: http://doi.org/10.1080/02635140120087687
20. C. Romero, S. Ventura, P. G. Espejo, and C. Hervás, "Data mining algorithms to classify students," in Educational Data Mining 2008, 2008.
21. L. B. Ahmedi, E. Rexha, and V. Raca, "Applying data mining to compare predicted and real success of secondary school students," Advances in Applied Information Science, pp. 178–181, 2012.
22. P. Srimani and A. S. Kamath, "Data Mining Techniques for the Performance Analysis of a Learning Model-A Case Study," Int. J. Comput. Appl., vol. 53, no. 5, 2012.
23. D. Kabakchieva, K. Stefanova, and V. Kisimov, "Analyzing university data for determining student profiles and predicting performance," in Educational Data Mining 2011, 2010.
24. E. Alhazmi and A. Sheneamer, "Early Predicting of Students Performance in Higher Education," IEEE Access, vol. 11, pp. 27579–27589, 2023. [Online]. Available: http://doi.org/10.1109/access.2023.3250702

25. M. K. Duong, J. Dunkelau, J. A. Cordova, and S. Conrad, "RAPP: A Responsible Academic Performance Prediction Tool for Decision-Making in Educational Institutes," in BTW 2023. [Online]. Available: https://doi.org/10.18420/BTW2023-29
26. A. Qahmash, N. Ahmad, and A. Algarni, "Investigating Students; Pre-University Admission Requirements and Their Correlation with Academic Performance for Medical Students: An Educational Data Mining Approach," Brain Sciences, vol. 13, no. 3, p. 456, 2023. [Online]. Avail-able: https://doi.org/10.3390/brainsci13030456
27. S. C. Mwape and D. Kunda, "Using data mining techniques to predict university student's ability to graduate on schedule," Int. J. Innov. Educ., vol. 8, no. 1, pp. 40–62, 2023. [Online]. Available: https://doi.org/10.1504/ijiie.2023.128470
28. N. Sateesh, P. S. Rao, and D. R. Lakshmi, "Optimized ensemble learning-based student's performance prediction with weighted rough set theory enabled feature mining," Concurrency Comput. Pract. Exp., vol. 35, no. 7, p. e7601, 2023. [Online]. Available: https://doi.org/10.1002/cpe.7601
29. O. R. Yürüm, T. Taşkaya-Temizel, and S. Yıldırım, "The use of video clickstream data to predict university students' test performance: A comprehensive educational data mining approach," Educ. Inf. Technol., vol. 28, no. 5, pp. 1–32, 2022. [Online]. Available: https://doi.org/10.1007/s10639-022-11403-y
30. Y. Wu, M. V. N. Gumabay, and J. Wang, "Student Achievement Predictive Analytics Based on Educational Data Mining," in Big Data Management and Analysis for Cyber Physical Systems, Springer, 2023. https://doi.org/10.1007/978-3-031-17548-0_6.

Abbreviations

The following abbreviations are used in this manuscript:

PAT	Pre-Admission Test
ML	Machine Learning
DT	Decision Tree
KNN	K-Nearest Neighbors
NN	Neural Network
EDM	Educational Data Mining