http://doi.org/10.35784/iapgos.6173 *received: 30.04.2024 | revised: 10.09.2024 | accepted: 24.09.2024 | available online: 30.09.2024*

DETERMINING STUDENT'S ONLINE ACADEMIC PERFORMANCE USING MACHINE LEARNING TECHNIQUES

Atika Islam[1], Faisal Bukhari[2], Muhammad Awais Sattar[1], Ayesha Kashif[1]
[1]Riphah International University, Riphah School of Computing and Innovation, Lahore, Pakistan, [2]University of The Punjab, Faculty of Computing and Information Technology, Lahore, Pakistan

Abstract. *Predicting student's academic performance during online learning has been considered a major task during the pandemic period. During the online mode of learning, academic activities have been affected in such a way that the management of educational institutions has planned to design support systems for predicting the student's performance to reduce the dropout ratio of the students and bring improvement in academic activities. During COVID-19, the main challenge is maintaining student's grades by predicting their academic performance using different techniques such as Education Data Mining and Learning Analytics. Different features have been identified related to the teaching mechanisms in online learning, which have a great impact on the improvement of academic performance. A high-quality dataset helps us to generate productive results, which in turn helps us to make effective decisions for promoting high-quality education. In this research, five prediction models for predicting academic performance have been proposed by collecting an imbalanced dataset of 350 students from the same computer science domain. After applying pre-processing techniques for cleaning the data, machine learning models have been applied, including K-Nearest Neighbor Classifier, Decision Tree, Random Forest, Support Vector Classifier, and Gaussian Naive Bayes. Results have been predicted for an imbalanced and balanced dataset after feature selection. Support Vector classifier has produced the best results in a balanced dataset with selected features by giving an accuracy of 96.89%.*

Keywords: educational data mining, learning analytics, random forest, support vector classifier

OCENA WYDAJNOŚCI AKADEMICKIEJ STUDENTÓW W NAUCE ONLINE ZA POMOCĄ TECHNIK UCZENIA MASZYNOWEGO

Streszczenie. *Przewidywanie wyników akademickich studentów podczas nauki online było uważane za ważne zadanie w okresie pandemii. W trakcie nauki w trybie online działalność akademicka była zakłócana w taki sposób, że zarządy instytucji edukacyjnych planowały projektowanie systemów wsparcia do przewidywania wyników studentów w celu zmniejszenia wskaźnika rezygnacji ze studiów i poprawy działalności akademickiej. Podczas COVID-19 głównym wyzwaniem jest utrzymanie ocen studentów poprzez przewidywanie ich wyników akademickich za pomocą różnych technik, takich jak Edukacyjna Analiza Danych i Analityka Edukacyjna. Zidentyfikowano różne cechy związane z mechanizmami nauczania w nauce online, które mają duży wpływ na poprawę wyników akademickich. Wysokiej jakości zestaw danych pomaga generować produktywne wyniki, które z kolei pomagają podejmować skuteczne decyzje na rzecz promowania wysokiej jakości edukacji. W tym badaniu zaproponowano pięć modeli predykcyjnych do przewidywania wyników akademickich, zbierając niezrównoważony zestaw danych 350 studentów z tej samej dziedziny informatyki. Po zastosowaniu technik przetwarzania wstępnego do oczyszczania danych, zastosowano modele uczenia maszynowego, w tym klasyfikator K-Najbliższych Sąsiadów, Drzewo Decyzyjne, Las Losowy, Klasyfikator Wektorów Wspierających oraz Naiwny Klasyfikator Bayesa Gaussowskiego. Wyniki przewidziano dla niezrównoważonego i zrównoważonego zestawu danych po selekcji cech. Klasyfikator wektorów wspierających wyprodukował najlepsze wyniki w zrównoważonym zestawie danych z wybranymi cechami, osiągając dokładność 96,89%.*

Słowa kluczowe: edukacyjna eksploracja danych, analityka uczenia się, losowy las, klasyfikator wektora wsparcia

Introduction

The best education is thought to be crucial for a student's successful life. The success of students is the responsibility of all the academic institutions. To assess student's academic performance and ensure that it is up to the mark, specific actions should be taken by all educational institutions. The student's activities during their education have been analyzed for prediction. Students have encountered numerous issues with the online education system regarding their academic performance. Many researchers have performed their best in determining the grades of the students. On the data set, various machine learning approaches have been applied to examine the student's academic achievement. Other researchers have described a variety of ways. There is too much dataset related to student's academic performance, so this has become a challenge. The next crucial step is to list down all the critical factors required for forecasting the student's academic achievement. Those attributes include GPA, previous grades, academic progress, mental and psychological condition, and family educational background. Research has introduced two more significant attributes, including the effect of students' internet usage during their studies. As is well known, students are taking classes online throughout the COVID-19 era, and this online learning environment has increased internet usage. Students must enroll in many platforms to complete their coursework, have a reliable internet connection, and must be familiar with using social media platforms. Schools are facing a severe problem of student failure in their primary grades. To study every element that contributed to this failure is a time-consuming process [4]. Educational data Mining has been used to extract meaningful information from educational data that have been extracted using mining. Education has grown essential and is spreading to others via various avenues. Any internet source can be used to obtain information. In addition, if there are too many students in a class and they are having difficulties in listening to the Lecture, the innovations in this technology allow the instructors to transmit knowledge to the individual student [5]. Grades are used to measure the performance of the students during their course. Essential variables have been investigated that could influence academic success or activity most effectively. This study provides us with necessary information about the academic situation of students. As is well known, students are never accustomed to learning using online educational platforms during the online mode of the education system. This research is involved in finding the significant factors that could best determine academic activities performance.

The use of machine learning techniques to forecast student's academic achievement has already been studied. Particularly during the COVID-19 pandemic, the educational system changed from traditional to online. The success or failure of a student in online learning is evaluated based on their behavior in class and their academic interests during online class assessments [7]. Since students are not in direct contact with their lecturers during online learning, the ratio of dropout students can be predicted early on by observing their actions. Many forecasting techniques, including a machine learning model called support vector machines (SVM) have been applied, and the results have been quite good. Naive Bayes and Neural Networks have shown excellent results [25]. It is possible to assess academic performance using educational data mining (EDM). A very traditional method that can be used to arrive at the right judgments is data mining. Educational Data Mining can help us to design and introduce new techniques in the education system [15].

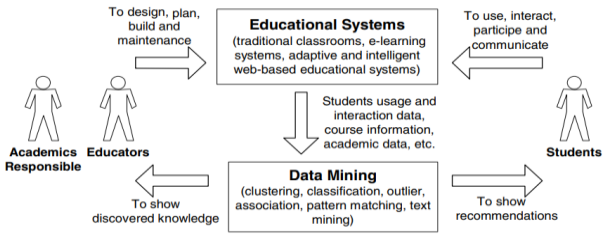


Fig. 1. Life cycle of Educational Data Mining

In the above Figure 1 from life cycle of educational data mining, it is clear that students and teachers use data mining techniques [20]. Students have to interact and communicate with their teachers to participate in their educational institutes. On the other hand, teachers and responsible of academics have to build or plan an academic structure for their educational institute. So, both educational systems and data mining techniques are directly dependent. During COVID-19, students are distracted while taking their classes because no one is observing them. They are free to use any other applications except the learning ones [1]. Three classifiers were utilized in this study to analyse the seven best features, with Naive Bayes having the highest accuracy at 85.7% [16]. It is challenging to predict students' academic achievement. Numerous research projects have already been conducted. The "Course Signals" early risk indicator was created by Purdue University. It is a remedy for detecting students' failures. It is initially created after making academic achievement predictions for kids. For teachers and students to communicate with one another and for students to share the issues they are having with the online educational system, many chatting technologies have been acknowledged. A few distinct factors make a prediction model accurate and effective. Artificial Intelligence (AI) has played a vital role in developing new tools and technologies for education. There are many fields of AI. One of the significant fields is Artificial Neural Networks (ANN). This field is helpful for the tutors to understand the learning situation of their students and is used for providing a comfortable and learnable environment for the students [11]. Techniques of AI are particularly beneficial for teachers to provide a suitable, hardworking, and attractive environment for learners. Coursera is an online learning platform for teachers and students to take complete and well-certified courses. Ultimately AI has become a source of improving student's academic performance [8]. While using the techniques of AI, the data set has been taken and then fed to different regressions and classification machine learning algorithms. A supervised neural network named Back Propagation performs best by giving an accuracy of 80.91% [22]. Predicting the students' academic performance is the key to improving the quality of education. Academic performance is affected by different factors like the financial condition of their families, learning abilities, and gender. Educational data mining results in forming those tools, which are the solutions to the problems the students identified and faced in online academic performance. Gathered data is assembled and labeled using the software SPSS and then used Decision Tree Algorithm [14]. Other researchers used three well-known algorithms, Bayesian, Neural Networks, and Decision trees, highlighting the critical factors for students to pass a course successfully. This will increase the passing ratio. After this research, educational planners and tutors know the students who need extra attention [6, 17].

Students' academic performance depends on their previous high school and college results. The student's success depends on the struggle of teachers as well as the students [21]. Academic performance is also affected by the facilities provided to the students. The selection of course content, teaching plans and methods, also management systems greatly impact academic performance. Social behavior is very important for students. When the teacher has to deliver the lectures online or in a physical environment. The tutors have to make it attractive for the students. They must make the topic easy for the students by making attractive prototypes to develop their interest. Using different designs gives students great visualization of their lectures, making it easy for them to remember. Learning through these tools is a very impressive and effective way of learning. Hard work is very important when you want to achieve a certain goal. Students, as well as teachers, have to show their concerns for the particular task. Rewarding students is also a good way to improve the student's academic performance. When you start appreciating the good qualities of the students and give them rewards, they also make their behavior good to take more and more appreciation. Quality of education should be improved so that the students get a suitable environment for studying and improving their skills. Learning Analytics has become very useful for helping students who are facing difficulties during their studies. Learning analytics proves to be an efficient method for maintaining the quality of education. This will lessen the failure rate of students [10].

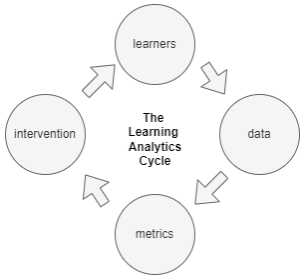


Fig. 2. Architecture of Decision Support System

As observed, educational plans have been made to improve student's performance, so accuracy is the major tool to help teachers learn and change the environment for students to improve their results. While predicting the results small-scale universities or the courses in which few students are enrolled, the data of their assignments and homeworks have been obtained, but the accuracy is not too good. To get better results from the model, preprocessing techniques have been applied to the dataset before applying machine learning algorithms [9]. When combined with learning data, behavioral data will generate better results in predicting academic performance. The teachers use a method to obtain weekly performance (high, medium, and low). This will increase performance. Clustering has been used here to make the classes of the students to which they belong according to their grades. Learning patterns and procedures have been recognized, and the learning systems have been built accordingly. Students have different styles of learning which their tutors should analyze. Sometimes different programs have been chosen during our education. It is very important to make proper decisions on time. This will lessen your efforts and time. Predicting academic performance on time is also very beneficial for parents and students. In this paper, an Intelligent Decision Support System (IDSS) has been introduced to give students the perfect directions to choose their program according to their learning abilities. This prediction helps in deciding whether a student can continue the program or quit the selected program [3]. Here Logistic Model Trees (LMT) have been used for predicting academic performance, and this figure explains the working of the Decision Support System.

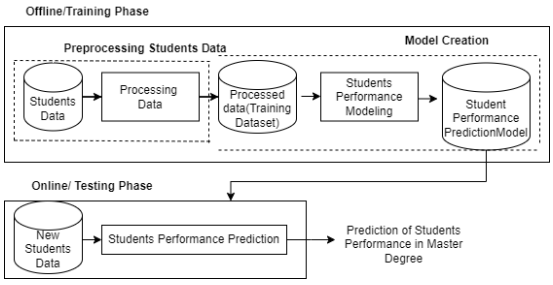


Fig. 3. Architecture of Decision Support System

Student analysis shows that Educational institutions have analyzed student's performance by considering the relationship between student achievement and attendance. They analyzed the characteristics of the toppers and the students being dropped out. Then the institution proposed a model to classify the students among the high and low achievers. After complete identification of the students who are performing well and can be high achievers shortly. The teachers have to arrange their training sessions in which they should be provided with complete guidelines. This will improve the student's results because the right training of the right student at the right time is necessary for improving academic behavior. By using data mining techniques, institutions have extracted the golden information from the present data, which results in the student's betterment. Many useful methods have been present to find the dropout ratio of the students. The method used for the prediction of dropout is the induction of rules. This can be done using IBM SPSS Answer Tree (AT) software. Different researchers have found major factors that affect the ratio of dropout students. Major factors involved are their economic status, family background, and grades in the previous class. A major problem during prediction is the unbalanced dataset. But when machine learning techniques are applied to this kind of dataset, it will give better results [24]. Another analysis has been performed which depicts that dropout ratio of the students depends upon the following factors:

- After matriculation, the proper guideline is not provided about the major subjects that have to be studied in the future. They don't know how to choose the major courses. This will become the reason for students being left out of the institute.
- Usually, the students drop out after choosing those courses that are not good for earning purposes shortly.
- In different countries, like India, they consider that a girl should be married after the age of 18. So, due to this culture, they have to be dropped out.
- The last reason for dropping out is the financial pressure of the courses the students are taking and the burden of studies on the students.

Another research has also been conducted using the dataset from Thailand, University of Technology. The dataset contains 15 features, and different feature selection algorithms have been compared, like Information Gain ratio (IG), Greedy, and Chi-Square [19]. This research shows that the Greedy Forward selection algorithm, when combined with Neural Networks, gives the best accuracy in predicting academic performance instead of predicting through Decision Tree and Naive Bayes. After doing this literature review, results have been analyzed from different journals, their dataset, and the techniques they have used in their prediction are described in the following table 1.

Table 1. Comparison of this Research with literature

Paper	Dataset	ML algorithms	Best algorithm	Accuracy
[16]	60	Naive Bayes (NB), NN, DT	NB	85.70%
[3]	1021	Logistic Model Tree (LMT), Random Forest and J48	LMT algorithm	83.09%
[14]	211	Decision Tree (DT) with nodes 2, 3, 4	DT with 4 nodes	65.52%
[2]	161	Artificial Neural Network, LR, NB, DT	ANN	77.04%
[18]	648	LR for Supervised Learning (SL), LR with Deep Learning and NN	LR for SL	Mean Average Error 3.26
[13]	323	Linear Regression, ANN, NN, SVM	SVM	86.6%
[12]	914	NB, DT, LR, SVM, ANN, KNN	SVM, KNN, NB	73.3%
This Research	350	SVC, KNN, RF, DT, GNB	SVC	96.89%

1. Materials and methods

All the individuals related to this research work lie in the Research Population. Education is now becoming one of the major sectors. Academic performance is significant for a student's progress as this research is related to the field of education. So, this includes the interests of parents, students, and teachers. After COVID-19, the online education system was followed by almost all educational institutions, so this research has been performed to predict the student's academic performance. This study was conducted between August 2021 and July 2022. It has been decided to gather information from students of Information Technology. Table 2 shows the data collected from both Universities.

Table 2. Participation of different Universities in collecting dataset

Universities	No of students
University of the Punjab (PU), Lahore	283
Information Technology University (ITU), Lahore	67

A questionnaire has been designed including all the questions regarding the student's current situations during their online learning. The questionnaire is designed using Google Forms. An online link to the questionnaire has been shared with the students from fall 2016 to fall 2022. This research is description-based, so a questionnaire is the best approach for collecting data. Many questionnaires have been analysed from different research papers and then designed a questionnaire named Academic Performance during Online Learning. The questionnaire consists of four different sections. It is based on analyzing the pros and cons of the online mode of learning. The questionnaire was divided into the following components. Predefined styles should also be used for bulleted and numbered lists, as in the examples:

- In the first portion, generic questions about the student's gender, the subjects they have studied online, their percentage of marks gained, etc., have been made.
- Different skills-related questions have been designed for the second segment, such as what challenges individuals encounter while taking online classes.
- Academic performance has been compared between the traditional classroom setting and online learning in the third section of the questionnaire. Some questions have been raised regarding the workload and homework offered to students during the online semester.
- Questions in the final portion are intended to gauge how satisfied the students are with this online learning environment. How much do the staff and teachers cooperate with their students in online lectures?

1.1. Input features

Completing research on determining academic performance, 38 essential elements have been identified that help us in determining the student's academic performance. Different input features have been used for this prediction. Few features are in binary format, and most are in string format. The following Table 3 shows all the features, options, and their format.

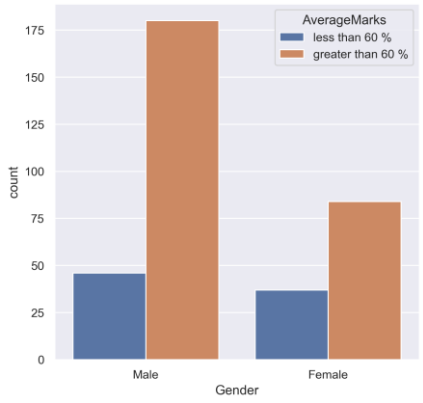
Table 3. Questions asked from the students to collect data

Sr.	Questions	Possible answers	Attribute type
1	What is your gender?	Male/Female	Binary
2	Which programming subject have you studied during the online semester?	PF/OOP/DSA/WE/EAD/AOA /MC/DS	String
3	In which online semester have you studied your mentioned subjects?	1st/2nd/3rd/4th/5th/6th/7th/8th	String

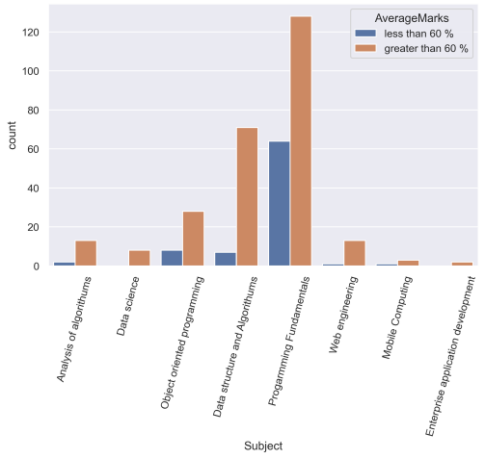
Sr.	Questions	Possible answers	Attribute type
4	Do the long-term use of digital devices affect your studies?	Yes/No	Binary
5	Staying a long time in the house makes you lazy during online lectures?	Yes/No	Binary
6	Is it hard for you to use mobiles and laptops for taking online lectures?	Yes/No	Binary
7	Distraction from surroundings lessens your attention during class.	Yes/No	Binary
8	Do you have a quiet place to study?	Often/Rarely/Always/Never/Sometimes	String
9	Do you have the required software and programs?	Often/Rarely/Always/Never/Sometimes	String
10	Do you have headphones and microphones?	Often/Rarely/Always/Never/Sometimes	String
11	Do you have a Webcam?	Often/Rarely/Always/Never/Sometimes	String
12	Do you have a strong internet connection?	Often/Rarely/Always/Never/Sometimes	String
13	Do you have a Computer/Laptop?	Often/Rarely/Always/Never/Sometimes	String
14	Do you have course study material?	Often/Rarely/Always/Never/Sometimes	String
15	Are you familiar with browsing information and sharing digital content?	Often/Rarely/Always/Never/Sometimes	String
16	Focusing during the online Lecture is more difficult for me than in physical lectures.	Strongly disagree/ Disagree/Neutral/Agree/Strongly agree	String
17	During physical classes, my academic performance has been improved	Strongly disagree/ Disagree/Neutral/Agree/Strongly agree	String
18	During online classes, my performance has worsen	Strongly disagree/ Disagree/Neutral/Agree/Strongly agree	String
19	During online classes, I am interested in listening to the Lecture	Strongly disagree/ Disagree/Neutral/Agree/Strongly agree	String
20	How much I became a master in the skills taught during online classes	Strongly disagree/ Disagree/Neutral/Agree/Strongly agree	String
21	Mental stress and depression during COVID 19 affect your learning abilities	Strongly disagree/ Disagree/Neutral/Agree/Strongly agree	String
22	Familiar with online learning platform (Google meet)	Strongly disagree/ Disagree/Neutral/Agree/Strongly agree	String
23	Familiar with online learning platforms (Zoom)	Strongly disagree/ Disagree/Neutral/Agree/Strongly agree	String
24	One assignment during a week	Strongly disagree/ Disagree/Neutral/Agree/Strongly agree	String
25	Providing feedback on my assignments	Strongly disagree/ Disagree/Neutral/Agree/Strongly agree	String
26	Give responses to my questions promptly	Strongly disagree/ Disagree/Neutral/Agree/Strongly agree	String
27	Have taken student's suggestions	Strongly disagree/ Disagree/Neutral/Agree/Strongly agree	String
28	Have informed us about online exam patterns	Strongly disagree/ Disagree/Neutral/Agree/Strongly agree	String
29	Lecture timings	Very dissatisfied/Dissatisfied/Neutral/Satisfied/Very satisfied	String
30	Supervisions	Very dissatisfied/Dissatisfied/Neutral/Satisfied/Very satisfied	String
31	Way of teaching online	Very dissatisfied/Dissatisfied/Neutral/Satisfied/Very satisfied	String
32	Career counseling services for students	Very dissatisfied/Dissatisfied/Neutral/Satisfied/Very satisfied	String
33	Online in real-time (videoconference)	Very dissatisfied/Dissatisfied/Neutral/Satisfied/Very satisfied	String
34	Online with a video recording (not in real-time)	Very dissatisfied/Dissatisfied/Neutral/Satisfied/Very satisfied	String
35	Online with an audio recording (not in real-time)	Very dissatisfied/Dissatisfied/Neutral/Satisfied/Very satisfied	String
36	Online by sending presentations to students	Very dissatisfied/Dissatisfied/Neutral/Satisfied/Very satisfied	String
37	Written communication (forums, chat, etc.)	Very dissatisfied/Dissatisfied/Neutral/Satisfied/Very satisfied	String
38	Average marks	less than 60% /greater than 60	Binary

1.2. Data analysis and pre-processing

After collecting data from 350 students, the second step is to analyze our data. When the dataset has been collected, all the responses are gathered in a Comma Separated Values (CSV) file. Major factors are identified that help us in predicting academic performance. Jupiter notebook is used for implementation in Python. Then used pandas' library to read the CSV file and convert it into a data frame. These figures will help you check the relationship between different features which are helpful to predict the student's academic performance. These figures are designed using a library named Seaborn in Python. To create a better visualization using a counting chart. Eight different features have been presented here, which are in binary and string format. Each figure corresponds to a single feature related to the outcome, known as average marks. Each feature has been visualized according to two possible labels of average marks (Less than 60% and greater than 60%).



a)



b)

Fig. 4 (I). The distribution of eight features with respect to Average marks (a) Gender, (b) Subject, (c) With respect to Students laziness (d) Affect of surroundings on studies, (e) With respect to Improved performance during online learning, (f) With respect to Worse performance during online learning, (g) Using Google Meet for online lectures, (h) With respect to Zoom platform for online lectures

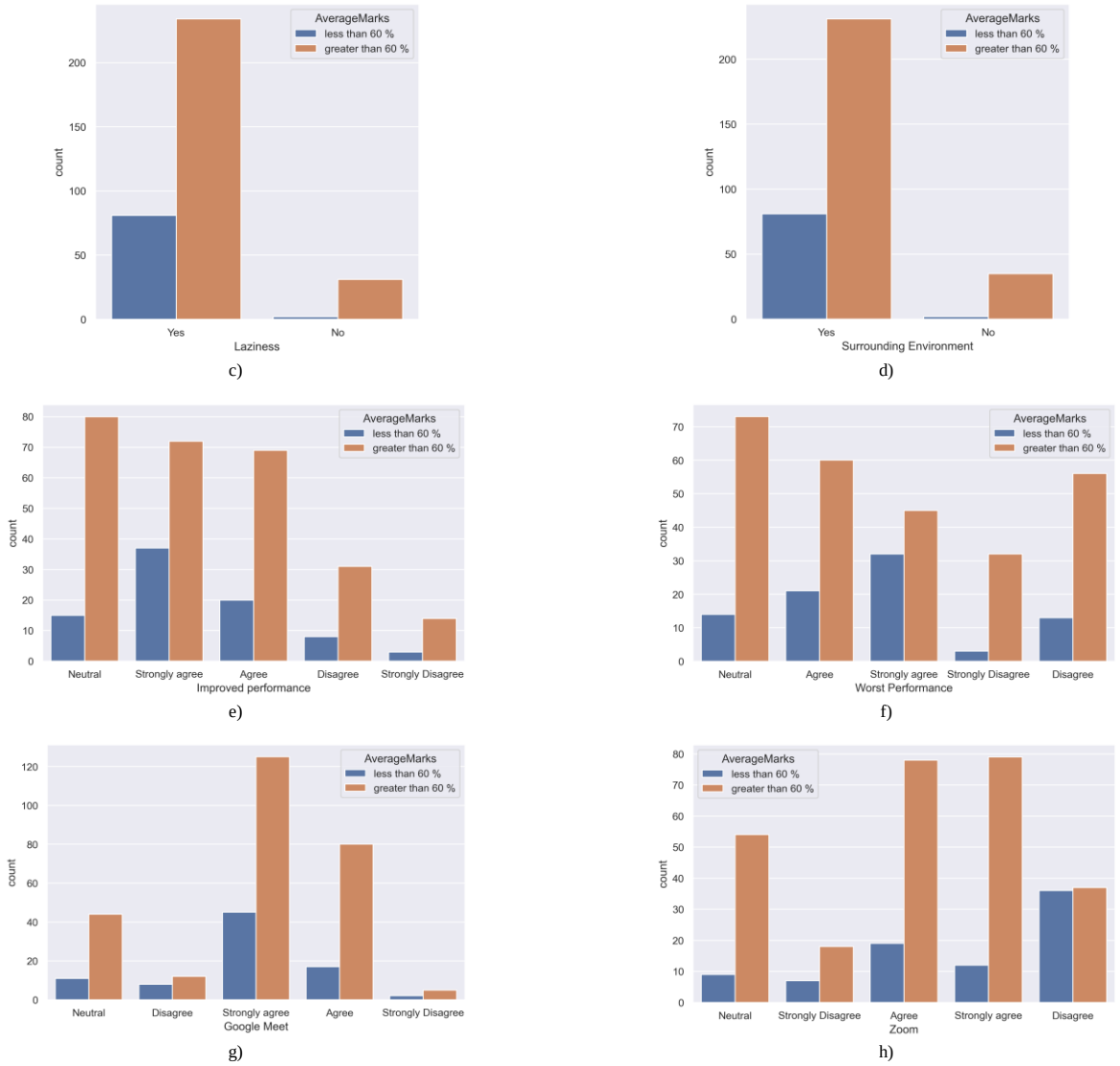


Fig. 4 (II). The distribution of eight features with respect to Average marks (a) Gender, (b) Subject, (c) With respect to Students laziness (d) Affect of surroundings on studies, (e) With respect to Improved performance during online learning, (f) With respect to Worse performance during online learning, (g) Using Google Meet for online lectures, (h) With respect to Zoom platform for online lectures

- Figure 4a will illustrate the count of male and female students who has participated in this data collection process related to average marks. Males who have greater than 60 percent marks have contributed more than females.
- Figure 4b will tell each subject's count related to the average marks. Students having marks greater than 60 percent have taken the course Programming Fundamentals.
- Figure 4c will show how much laziness the students feel during online learning. Mostly the students who have performed well claimed that they became lazy while taking online classes.
- Figure 4d will illustrate that the students with marks greater than 60% have claimed that the surrounding environment affects their studies.
- Figure 4e will explain the improvement in the student's performance during online learning. Most students consider this neutral and have performed well.
- Figure 4f will tell you about their point of view about online education. This illustrates that students are giving a neutral response to bad performance.
- Figure 4g will tell you that most students are more comfortable using the platform Google Meet.
- Figure 4h shows that some students also agree with Zoom for taking online lectures, but students are more convenient with Google Meet.

1.3. Data formatting

Data formatting can be done after data analysis. Before computing results, data is formatted into numeric labels.

All the missing values have been find out in our data frame. The the whole data frame has been checked and derived the count of NaN values in each column or feature using the IsNull function. After applying this function, it has been observed that there are five features in which there are missing values.

- First, all the missing values have been identified in our data frame. Complete data frame is checked and derived the count of NaN values in each column or feature using the IsNull function. After applying this function, it has been observed that there are five features in which there are missing values.
- As the task is to determine the student's academic performance during online education, different machine learning algorithms have been applied to the dataset. Still at first, converting these string labels into numeric labels so that the best results can be achieved after applying Machine Learning algorithms. Label Encoder can be applied to the structured dataset.
- Normalization is the second method, also known as rescaling, which have been applied to the encoded data. Normalization converts all the numeric data from 0 to 1 or -1 to 1. This range depends on the type of data values. Normalization helps us to get more accurate results because it increases the consistency

among features. This method will convert the values into a standard scale. Min Max Scaler function has been used to normalize the values. The mathematical formula for min-max scaler is as written below

$$y' = y - \frac{\min(y)}{\max(y)} - \min(y)$$

In this formula, y is the original data point, and y' is the normalized value of that data point.

2. Proposed methodology and experimental results

Significant steps have been observed in determining academic performance. Figure 5 contains a flowchart that is presenting our methodology.

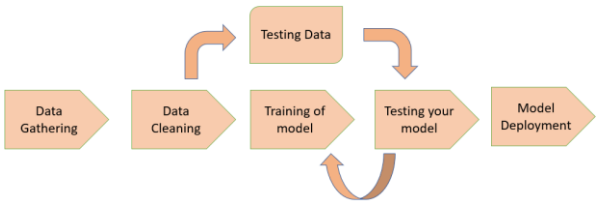


Fig. 5. Process of Machine Learning

1) Import the required libraries.
2) Data analysis and pre-processing techniques have been applied to the data to convert it into a structured form.
3) Division of data into two sets of training and testing sets. The prediction has been performed which includes removing all the null and empty data values, using Label Encoder to encode the labels with specific data values and Normalization has been performed using Min Max Scaler
4) At the end, accuracy, precision, and recall values are identified using every machine learning algorithm, and after comparison, results are generated.

2.1. Models of prediction

In this research, to determine the academic performance of the students, Five different machine learning models have been used [23].

K-Nearest Neighbor Classifier (KNN): The K-nearest neighbors (KNN) algorithm is a supervised machine learning algorithm. This classifier can predict the target variable by considering all the available independent variables. To perform classification, KNN utilizes a straightforward approach. In the training phase, it uses the entire dataset. Using the distance formula, it looks through the whole training dataset for k-most similar instances. The data with the most similar instance is finally returned as the prediction for an unseen data instance.

Support Vector Classifier (SVC): The objective of this Support Vector Classifier is to classify the data of different classes using a hyperplane. This algorithm works on "best fit". You have to give your dataset to the hyperplanes for prediction. SVM helps us in the classification and detect all the outliers. It works best and is efficient because it consumes less memory for training data. There are two main goals of SVM, which are as follows:

- Support Vector Machine helps us classify by making different hyperplanes iteratively.
- In the second step, it will consider the hyperplane, which can separate classes more accurately.

There are support vectors which are the points closest to the hyperplane. These support vectors help us best identify lines separating between two classes. The following algorithm will explain the implementation of the Support Vector Classifier.

Decision Tree Classifier (DTC): Decision Tree Classifier is a tree-based structure algorithm. DT used for both classification and regression are very versatile. It uses a series of decisions to determine the class label it operates with the conduct of "If this, then that". DT is easy to decipher, quick, and appropriate for enormous datasets. DT provides an optimum solution for each step without the last stage, determining the optimum solution. The topmost node is the root, decision rules demonstrate branches, and the output signals the leaf node. The tree is recursively partitioned. A Decision Tree is just like a white box whose training time is faster than other algorithms, which are black boxes like Neural networks. A decision tree has the power to work with data having more dimensions.

Random Forest (RF): Random Forest is an algorithm that works for both classification and regression. In a random forest, decision trees are formed randomly using data points. Then predictions are made using every tree, and the voting algorithm is identified as the best solution. This algorithm has various advantages. It can generate the best results because many decision trees have been made. The collection of Decision Trees are random forests (RF), also called random selection timber. It is used for the choice of functions, clustering, and statistical inference. Categorical and Numerical records are used for these forests. The term woodland indicates that trees are accrued in certain areas, similar to the random woodland set of rules. When new records are obtained for classification, several trees are created, as every tree classifies the point according to choice-making guidelines. As an outcome, a new point is allotted to the class with the most elevated number of three votes. Python's sci-kit-learn library has been utilized to apply the previously mentioned machine learning algorithms as the library gives enhanced executions of a machine learning algorithm.

Gaussian Naive Bayes Classifier (GNB): Naive Bayes is a very efficient machine learning model that helps us predict. It uses a theorem known as Bayes Theorem. This theorem is defined as P(A|B) = P(B|A).P(A)/P(B)

Here, in this formula A and B are the two events. P(A|B) is the probability of event A, provided that event B has already happened. P(B|A) is the probability of event B provided that event A has already happened. P(A) is the independent probability of an event A. P(B) is the independent probability of an event B. In the Naive Bayes algorithm, it has been assumed that all the predictors play an equal role in predicting a class. All the predictors can individually and independently work for classification. Gaussian naive Bayes is one of the types of Naive Bayes. It is used when the data values are in continuous form.

2.2. Experiments and results

In this prediction, dataset is divided as the ratio of 70% in the training and 30% in the testing parts. All the features except average marks are in the X part, and the outcome average marks are in the Y part. Then split the dataset accordingly.

Now while training the data, Cross-Validation technique has been applied by dividing the data into five subsets. This technique is generally used for getting information about how the model is adequate.

As it has been observed, the results are not too satisfying due to the problem of unbalanced classes, as having two classes in the dataset. In the feature of average marks, the class of more significant than 60% has more records whose count is 267, and in comparison to that, there are fewer records of class less than 60% of average marks whose count is just 83. So, the results of classification can be improved by balancing the classes. Oversampling has been applied on our dataset to randomly increase the records of the minority class. For applying oversampling, the function named Random Over Sampler is applied using the strategy of minority sampling, so here having samples of 267 of the majority class and, after balancing the minority class, also have samples of 267 sample points.

Without balancing classes, K-nearest neighbor has given the highest accuracy value of 82.85% but now, after increasing the sample points of the class of average marks less than 60%.

Support Vector Classifier and Random Forest algorithms have performed very well by generating accuracy of 91.92% and 91.30%, respectively.

After balancing classes, Feature Selection is the second method for improving our results. There is a total of 38 different features in the dataset. Then technique of feature Selection has been applied by dropping some correlated features. Dimensionality reduction will reduce the number of input features. To obtain the correlation between all the features, the function named as corr has been used on the training set. Visualization can also be done on the correlation matrix using Heatmap. The threshold value of 0.5 is adjusted, which means if the two features are in 50% correlation, one of those features can be removed. Pearson correlation method has identified three features which are in 50% correlation. Supervisions, Written Communication, and sending presentations have been removed from the training and testing set.

Among all the algorithms KNN, SVC, RF, DT, and GNB, after doing all these experiments on data, it has been observed that the Support Vector Classifier has outperformed by giving an accuracy of 96%. This accuracy is observed using the Confusion matrix.

2.3. Confusion matrix

Confusion matrix is used in machine learning to evaluate the performance of a machine learning model. It is like a matrix with N*N order, where N is the number of classes. Here binary classification is used between two classes. So, the size of the matrix is 2 by 2.

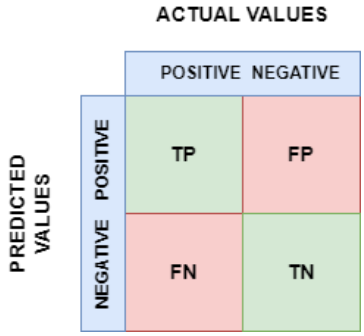


Fig. 6. Confusion matrix of size two by 2

- True Positive:
  When the model predicts the value correctly it is positive, and the model also predicts it as positive.
- True Negative:
  When the model predicts the value correctly it is negative, and the model also predicts it as negative.
- False Positive:
  When the model can not predict the value correctly it is negative, but the model predicts it as positive.
- False Negative:
  When the model can not predict the value correctly it is positive, but the model predicts it as negative.

The following are the confusion matrices that show the accuracy of each machine learning algorithm after balancing the classes and reducing the dimensions.

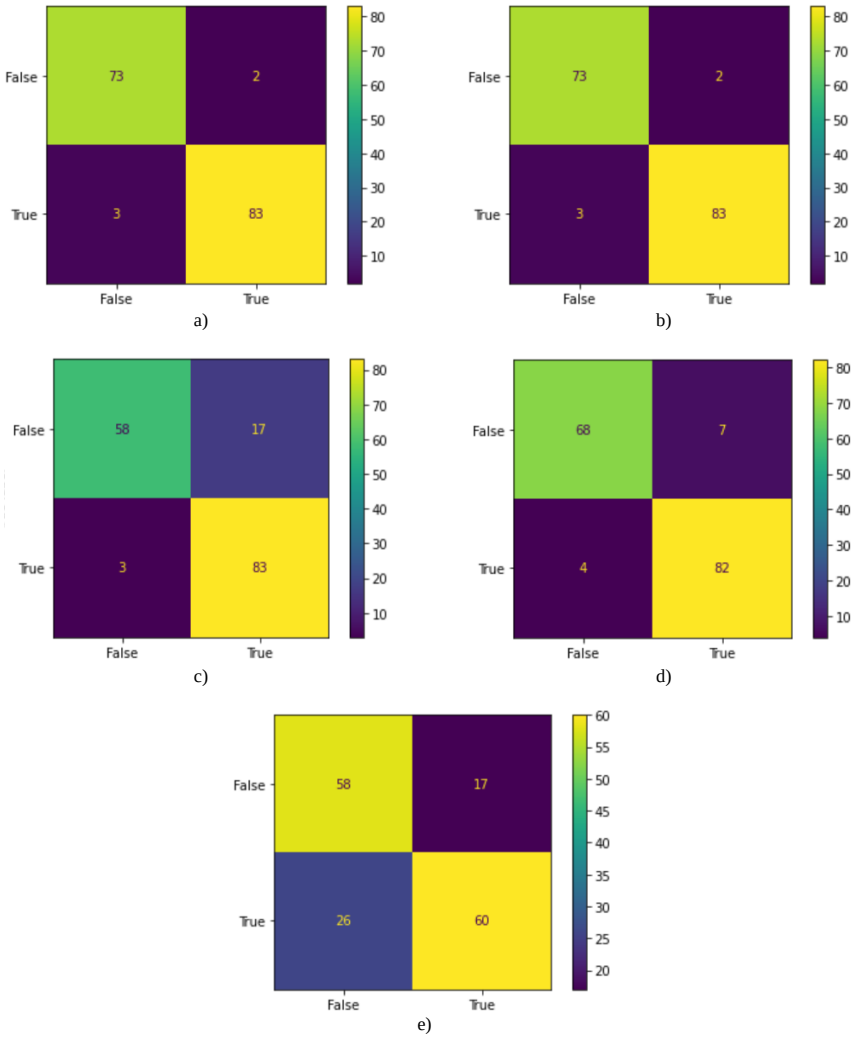


Fig. 7. Confusion matrices of five different training models after Feature Reduction (a), K-Nearest Neighbors (b), Support Vector Classifier (c), Decision Tree Classifier (d), Random Forest Classifier \ref{fi, Gaussian Naive Bayes

- Figure 7a shows that K-Nearest Neighbors model predicts 67 values as truly positive and 84 as truly negative. But it predicts 8 values as falsely positive and 2 as falsely negative.
- Figure 7b shows that the Support Vector Classifier predicts 73 values as truly positive and 83 as truly negative. But it predicts 2.
- Figure 7c shows that the Decision Tree model predicts 58 values as truly positive and 83 as truly negative. But it predicts 17 values as falsely positive and 3 as falsely negative.
- Figure 7d shows that the Random Forest Classifier predicts 60 values as truly positive and 82 as truly negative. But it predicts 7 values as falsely positive and 4 as falsely negative.
- Figure 7e shows that Gaussian Naive Bayes predicts 58 values as truly positive and 60 as truly negative. But it predicts 17 values as falsely positive and 26 as falsely negative.
- This is how all the models behave after performing the above experiments.
- Now, the following comparison Table 4 shows the accuracy of all the machine learning algorithms. When we compute the accuracy of unbalanced classes, then KNN performs best by giving an accuracy of 82.85%. After balancing the classes, the SVC performs best by giving an accuracy of 91.92%. Then, accuracy has been computed after feature reduction and gives an accuracy of 96.89%.

Table 4. Results of Different Machine Learning Algorithms

Machine learning algorithms	Accuracy of unbalanced classes	Accuracy with balanced classes	Accuracy after feature reduction
K-Nearest Neighbors(KNN)	82.85%	88.81%	93.78%
Support Vector Classifier(SVC)	81.90%	91.92%	96.89%
Decision Tree Classifier	81.90%	83.85%	87.57%
Random Forest	80.95%	91.30%	93.16%
Gaussian Naive Bayes	78.09%	68.94%	73.29%

3. Conclusion and future work

This research concludes that in the future, students' ultimate success depends on their academic grades and performance.

This prediction has been performed here using different algorithms of machine learning. In an unbalanced dataset, the Support Vector Classifier and Decision Tree have an accuracy of 81.90%. Then after balancing the dataset, SVM has the best accuracy of 91.92%. After that, again, to improve our result, Feature Selection technique has been applied, and then again, SVM outstands among all algorithms and gives an accuracy of 96.89%. This research shows all the features which directly affect academic performance and helps us predict students' average marks while observing those features. If a student is getting disturbed by the surroundings, doesn't have an internet connection, is not familiar with the usage of digital devices can't attend the online lectures hence the chance of failure increases. So, these few major factors should be noticed while predicting the performance of the students. This research only helps us to predict the average marks of the students. Student's marks can't be precisely predicted but after analyzing the features described in this research, instructors can predict students' performance. Due to limited time, data has been collected from 350 students from two universities and only considered the Department of Information Technology. In the future, more data can be collected to make our research more robust and precise. Due to the pandemic period, many educational institutions have shifted their education system from conventional mode to online mode of learning. So, without physical discussion between teachers and students during lectures, it is difficult for the instructors to predict the students' academic behavior. During online learning, it isn't easy to analyze students' interest in class activities, class discussions, and quizzes. To evaluate the performance, the management of educational institutions has planned their learning schedule using techniques of educational data mining (EDM).

References

[1] Akour I. et al.: Using machine learning algorithms to predict people's intention to use mobile learning platforms during the COVID-19 pandemic: machine learning approach. JMIR Medical Education 7, 2021, e24032.

[2] Altabrawee H., Ali O. A. J., Ajmi S. Q.: Predicting students' performance using machine learning techniques. Journal of University of Babylon for pure and applied sciences 27, 2019, 194–205.

[3] Aman F. et al.: A predictive model for predicting students academic performance. 10th International Conference on Information, Intelligence, Systems and Applications – IISA. IEEE, 2019, 1–4.

[4] Arnold K. E., Pistilli M. D.: Course signals at Purdue: Using learning analytics to increase student success. 2nd International Conference on Learning Analytics and Knowledge, 2012, 267–270.

[5] Baraniuk R.: Open education: New opportunities for signal processing. IEEE International Conference on Acoustics, Speech and Signal Processing – ICASSP, 2015.

[6] Bhardwaj B. K., Pal S.: Data Mining: A prediction for performance improvement using classify cation. arXiv preprint arXiv:1201.3418, 2012.

[7] Bhutto E. S. et al.: Predicting students' academic performance through supervised machine learning. International Conference on Information Science and Communication Technology – ICISCT. IEEE, 2020, 1–6.

[8] Borge N.: Artificial intelligence to improve education/learning challenges. International Journal of Advanced Enginering & Innovative Technology – IJAEIT 2, 2016, 10–13.

[9] Chaudhury P. et al.: Enhancing the capabilities of student result prediction system. Second International Conference on Information and Communication Technology for Competitive Strategies, 2016, 1–6.

[10] Clow D.: An overview of learning analytics. Teaching in Higher Education 2013, 18, 683–695.

[11] Ever Y. K., Dimililer K.: The effectiveness of a new classification system in higher education as a new e-learning tool. Quality & Quantity 52, 2018, 573–582.

[12] Gray G., McGuinness C., Owende P.: An application of classification models to predict learner progression in tertiary education. IEEE International Advance Computing Conference – IACC. IEEE, 2014, 549–554.

[13] Huang S., Fang N.: Work in progress: Early prediction of students' academic performance in an introductory engineering course through different mathematical modeling techniques. Frontiers in Education Conference Proceedings. IEEE, 2012, 1–2.

[14] Kolo D. K., Adepoju S. A., Alhassan J. K.: A decision tree approach for predicting students academic performance. I.J. Education and Management Engineering 5, 2015, 12–19.

[15] Kotsiantis S. B.: Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades. Artificial Intelligence Review 37, 2012, 331–344.

[16] Mueen A., Zafar B., Manzoor U.: Modeling and Predicting Students' Academic Performance Using Data Mining Techniques. International Journal of Modern Education & Computer Science 8, 2016.

[17] Osmanbegovic E., Suljic M.: Data mining approach for predicting student performance. Economic Review: Journal of Economics and Business 10, 2012, 3–12.

[18] Oyedeji A. O. et al.: Analysis and prediction of student academic performance using machine learning. JITCE (Journal of Information Technology and Computer Engineering) 4, 2020, 10–15.

[19] Rachburee N., Punlumjeak W.: A comparison of feature selection approach between greedy, IG-ratio, Chi-square, and mRMR in educational mining. 7th International Conference on Information Technology and Electrical Engineering – ICITEE. IEEE, 2015, 420–424.

[20] Romero C., Ventura S.: Educational data mining: A survey from 1995 to 2005. Expert systems with applications 33, 2007, 135–146.

[21] Said M. A., Idris M., Hussain S.: Relationship between Social Behaviour and Academic Performance of Students at Secondary Level in Khyber Pakhtunkhwa. Pakistan Journal of Distance and Online Learning 4, 2018, 153–170.

[22] Sekeroglu B., Dimililer K., Tuncal K.: Student performance prediction and classification using machine learning algorithms. 8th International Conference on Educational and Information Technology, 2019, 7–11.

[23] Singh A., Halgamuge M. N., Lakshmiganthan R.: Impact of different data types on classifier performance of random forest, naive bayes, and k-nearest neighbors algorithms. International Journal of Advanced Computer Science and Applications 8, 2017.

[24] Thammasiri D. et al.: A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. Expert Systems with Applications 41, 2014, 321–330.

[25] Wolff A. et al.: Developing predictive models for early detection of at-risk students on distance learning modules. LAK Workshops, 2014.

M.Sc. Atika Islam
e-mail: atika.islam@riphah.edu.pk

Born in Pakistan in 1998. Currently working as Lecturer at Riphah International University, Lahore. Completed her Masters in Computer Science in 2022 and Bachelors in Software Engineering in 2020 from the University of the Punjab, Lahore.
Research areas are machine learning and artificial intelligence.

https://orcid.org/0009-0002-5400-7841

Dr. Syed Faisal Bukhari
e-mail: faisal.bukhari@pucit.edu.pk

Dr. Syed Faisal Bukhari is a distinguished academic with an extensive background in computer science and statistics. Holding a Ph.D. and an M.Sc. in Computer Science from the Asian Institute of Technology (AIT) in Thailand. Dr. Bukhari also possesses M.Sc. degrees in Computer Science and Statistics from Punjab University. His research and teaching interests are deeply rooted in computer vision, image processing, machine learning, and statistics. Dr. Bukhari's academic journey and professional dedication are reflected in his contributions to both theoretical and applied aspects of his fields of expertise.

https://orcid.org/0000-0002-7703-9742

Dr. Muhammad Awais Sattar
e-mail: awais.sattar@riphah.edu.pk

Earned his Ph.D. degree in Technical Computer Science and Telecommunication from Lodz University of Technology, Lodz, Poland, in 2022. Prior to that, he obtained an M.Sc. degree in Electrical Engineering from Rochester Institute of Technology, Dubai, UAE, in 2017, and a B.Sc. degree in Electrical (Computer) Engineering from Comsats University, Lahore, Pakistan, in 2014. He is currently serving as an assistant professor of Computing at Riphah International University, Lahore Campus, Pakistan. His research interests encompass machine learning, computer vision, and process tomography.

https://orcid.org/0000-0002-2431-8182

Dr. Ayesha Kashif
e-mail: ayesha.kashif@riphah.edu.pk

Dr. Ayesha Kashif completed her Ph.D. from University of Grenoble, France. Her research areas are Energy Management, modelling occupants' behaviour and their impact on the usage of appliances, building envelope and the interoperability of these models. She worked with multiple research laboratories (G-SCOP, GIPSA, LIG, University of Grenoble, France) and the Electricity Department of France (EDF). She is currently working at Riphah School of Computing and Innovation, Riphah International University Lahore as assistant professor and teaching courses including, artificial intelligence, data science, data mining and data warehousing.

https://orcid.org/0009-0007-7169-1610