Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 5, 2025

# PREDICTING ACADEMIC SUCCESS: A MACHINE LEARNING APPROACH USING DECISION TABLES AND RANDOM FORESTS ALGORITHMS

**Malak Roman*[1], Aftab Ullah[2], Muhammad Asad Ullah[3], Farzana Hussain[4], Sana Shaiza Shams[5], Aisha Bint-e-Meraj[6], Fardad Ali Shah[7], Sajad Ali[8]**

*[1,2,3,4,5,6,7]*Department of Computer Science, University of Chitral, KPK, Pakistan*
[8]*Department of Mathematics, University of Chitral, KPK, Pakistan*

[*1]malak_5116@uoch.edu.pk

DOI: https://doi.org/10.5281/zenodo.15380415

**Keywords**
*Machine Learning, Educational Data Mining- EDM, Best First Search, Decision Table, Random Forest*

**Article History**
*Received on 02 April 2025*
*Accepted on 02 May 2025*
*Published on 10 May 2025*


**Corresponding Author: * Malak Roman**

**Abstract**
*Machine Learning (ML) in educational data prediction refers to the use of AI-driven algorithms to analyses academic data (e.g., grades, attendance, engagement) and forecast student performance, identify at-risk learners, and recommend interventions. By processing historical and real-time data, machine learning (ML) models uncover hidden patterns that enable educators to optimize their teaching strategies and enhance learning outcomes. This research cones with data collected from 'UCI Machine Learning Repository' and the database has total of 33 attributes along 395 rows. The two classification classifiers used in this paper were Decision Table (DT") and Random Forest (RF). The best first search algorithm has been used as a preprocessing step with both classifier models. The distribution of these models is based on the analysis of the Mean Square Root Error between the predicted and actual values. The proposed decision table yields a better result as compared to the random forest algorithm with the blast 1.92 root mean squared error.*

## INTRODUCTION

Artificial Intelligence or AI known as Machine Learning, can be defined as making the computers acquire the ability to gain intelligence to boost the performance of process without coding. It is a much more flexible approach than a set of rules within which the choice must be made, normally, the ML algorithms learn about the sample data and proceed to make predictions based on information obtained thereafter. Machine learning (ML) plays a transformative role in analyzing and predicting students' academic performance by leveraging historical data, behavioral patterns, and learning trends [1]. ML has revolutionized the education sector by enabling data-driven insights into learners' academic enhancement and accurate grade prediction. By analyzing vast datasets, including past grades, attendance records, engagement metrics (e.g., LMS logins, assignment submissions) [2], and even socio-economic factors, ML algorithms can identify patterns that influence learning outcomes.

Early identification of such students is possible by using predictive models which means that the right type of intervention such as tutoring, course modification or counselling can be offered at the correct time. These insights are important to institutions in matters concerning the retention rates, curriculum enhancement as well as the methods of instructions [3]. In addition to predicting results and probabilities, modern intelligent tutoring systems (ITS) and leaning analytics dashboards

promote self- regulated learning through formative feedback. However, challenges such as data secrecy concerns, algorithmic prejudice, and model interpretability must be addressed to ensure ethical deployment. With advancements in natural language processing (NLP), ML can even assess essay-based responses or discussion forum participation to predict performance. As educational technology evolves, machine learning continues to bridge gaps between student potential and achievement, paving the way for a more personalized, equitable, and efficient academic environment [4]. Essentially, grade prediction and the evaluation of the probability of students dropping out necessitate a clean data set, since real-life data is imprecise, skewed or inconsistent [5]. In this research work, we employ Best-First Search as the preprocessing technique and a classifier, utilizing random forest and decision tables. Since BFS can select the most basis features in the dataset, this could be very crucial in cutting on processing time thereby reducing size of the data without affecting the analysis results.

The efforts made to employ data analysis for the improvement of the education system led to the development of EDM as a comprehensive subject. EDM can be defined as a method in the educational process to come up with a model based on educational data with the aim of describing students and enhancing educational effectiveness. Machine learning technology applications have experienced explosive growth throughout recent years. The educational data mining discipline offers educators and researchers educational metrics, including success indicators, failure indicators, and dropout metrics, to help them predict educational outcomes and simulate the learning process. The paper presents a student performance analysis enabled by data mining techniques.

**LITERATURE REVIEW:**

Burman, I., and Som, S. [6] employ a multi-classifier SVM to categorize students into high, average, and low categories based on their academic marks. It is implemented with linear kernels and radial basis kernels. It is observed that SVM with a radial basis kernel provides improved results as compared to a linear kernel.

A predictive study by H. Alamri et al. [7] focused on analyzing academic performance to enhance the effectiveness of educational organizations, leading to improved academic results among students. Support Vector Machines (SVMS) and Random Forests (RFS) were the primary algorithms and techniques used for classification in the study. Support Vector Machines and Random Forest work together for binary classification and regression applications. The experimental results demonstrate that the SVM and RF algorithms achieve their highest prediction accuracy rates of 93% for binary classification, with RF exhibiting the smallest root mean squared error (RMSE) of 1.13.

Naicker, N, et al [8] pursued a research goal to evaluate linear support vector machines alongside contemporary machine learning classical algorithms for determining the optimal predictive algorithm for student achievement. Student performance predictions demonstrated higher accuracy when linear support vector machines were evaluated against ten categorical machine learning activities. The existing research has shown that students' race, alongside their gender, influences their mathematics outcomes, but accessing lunch significantly impacts their reading and writing results.

Al-Shehri, H., et al. [9] conducted their research using the famous University of Minho dataset from Portugal, which contains mathematics performance data for 395 records. Future forecasting enables educators to take preventive measures and perform necessary actions, or select students based on their competency for suitable tasks. The research utilized the Support Vector Machine algorithm and the K-Nearest Neighbor algorithm on the dataset to forecast student grades and then measured the accuracy between the two algorithms. Results from empirical research confirmed that the Support Vector Machine achieved a superior performance with a correlation value of 0.96, while the K-Nearest Neighbor showed a correlation value of 0.95.

Hassan, C. A., et al. [10] explored the prediction of coronary heart disease using various machine learning classifiers. With a UCI dataset of 303 examples and 14 features, which was clean, he utilized eleven algorithms, including Gradient Boosted Tree, Random Forest, and Multilayer Perceptron. The Random Forest achieves a better

performance level in heart disease prediction, with an accuracy rate of 96%.

Aman, F., et al. [11] The selection of the exact academic program at the right time and the prediction of students' academic performance can save students and their parents effort, resources, and time. In the present work, initially, the subjectively identified academic and socioeconomic attributes are determined, based on which the prediction exhibit is developed. Then, a decision-tree-based approach, Logistic-Model Trees (LMT), is employed. The existing system is tested and trained on a real-life dataset of 1,021 records obtained from the University of Peshawar's examination database. The proposed system achieved a predictive accuracy of 83.48%, enabling parents, higher education institution management, and students themselves to determine whether to proceed further or withdraw from the admitted program.

Through Hussain, S., & Khan, M. Q [12] suggested 'Predicting students' academic performance at secondary and intermediate level using machine learning'. The data set for this work is directly sourced from the Board of Intermediate and Secondary Education (BISE), Peshawar, KPK. A genetic algorithm with a forward approach selects 30 optimal attributes out of 126 to train the k-nearest neighbor (KNN) and decision tree (DT) classifiers. The decision tree outperforms the KNN classifier with an accuracy of 96.64%, which is 6.72% higher than the KNN classifier.

Zacharias, N. Z. [13] The objective of the present work was to evaluate the capability of artificial neural networks to predict student success, using data gathered during students' online activities within a Web-Based blended learning environment. A multilayer perceptron network was trained using the backpropagation algorithm to forecast students' capacity to pass the course successfully. The accuracy rate for classifying students into the predicted success and failure categories was extremely high, at 98.3%.

The authors Alhazmi, E. and Sheneamer, A. [14] employed clustering, joined with classification approaches, to study the performance-stage influence on GPA data. The clustering technique employs T-SNE dimension reduction to analyses early-stage factors, including admission scores and first-level courses, as well as academic achievement tests (AAT) and general aptitude tests (GAT). For the classification technique, include XGBoost, Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Random Forest (RF).

Wilt, C., et al [15] Compare and contrast the connections between three types of greedy heuristic search: best-first, hill-climbing, and beam search. Discuss the following best-first searches: weighted A, greedy search, *A∗*, *window A* and multi-state commitment k-weighted A*. For the hill climbing procedures, utilize enforced hill climbing and LSS-LRTA*. BULB and beam-stack operations make up some of the multiple beam search techniques available. An empirical evaluation of the six standard benchmarks indicates that best-first search and beam search demonstrate very similar performance, excelling that of hill-climbing procedures in both solution quality and time to solution.

## RESEARCH METHODOLOGY:

Researchers obtained the dataset of 'Students' Academic Performance and Grade Prediction ' from the database of the UCI Machine Learning Repository. It contained 33 attributes and 395 instances. After data pre-processing, two datasets were made, one for training the machine and the second for model testing. In "Greedy heuristic search," the Best-First Search algorithm was implemented for optimal feature selection. The dataset was loaded in order to train the model, i.e. Decision Table (DT) and Random Forest (RF). Classification algorithms were applied individually to check and test the accuracy of the models. The model was trained first, and then test data was used to calculate the accuracy and root mean square error of the proposed model as shown in Figure 1.

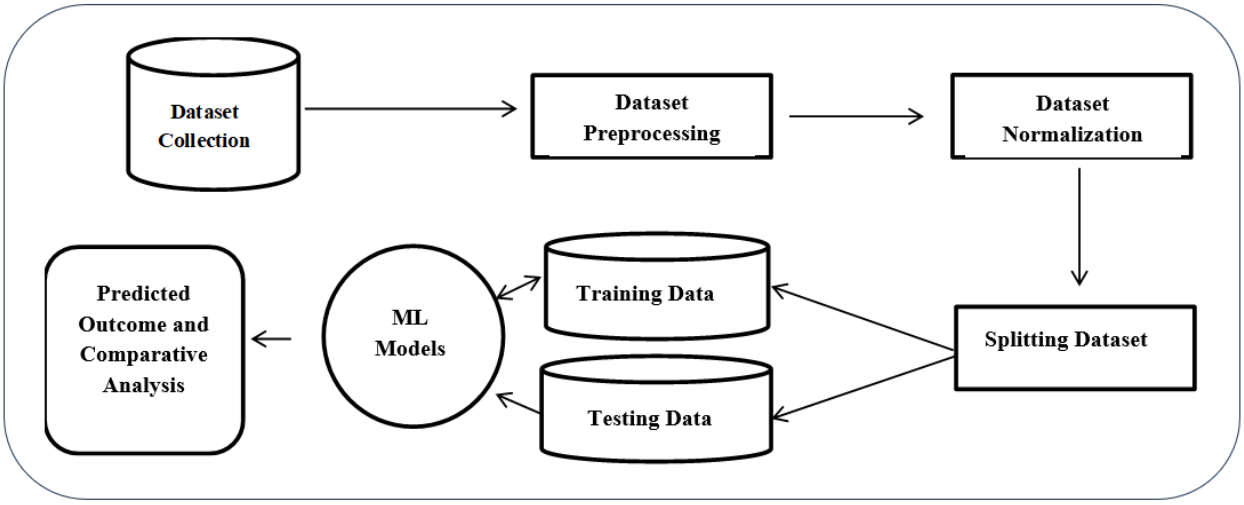


**Fig. 1** Research Methodology Flow Chart

**DATA SET COLLECTION:**

The Students' Academic Performance and Grade Prediction dataset consists of 33 attributes. The attribute values are explicated in Table 1.

| Attribute Description | | | | | |
|---|---|---|---|---|---|
| **No** | **Attribute Name** | **Value/Description** | **No** | **Attribute Name** | **Value/Description** |
| 1 | 'School' | 'student's school' | 18 | 'paid: extra classes' | 'binary: yes or no' |
| 2 | 'Gender' | '(binary: "F"- female or "M" – male)' | 19 | 'activities: extra-curricular activities.' | 'binary: yes or no' |
| 3 | 'Age' | 'Numeric' | 20 | 'nursery: school' | 'binary: yes or no' |
| 4 | 'Address' | '(binary: "U"- urban or "R"- rural)' | 21 | 'higher: interested in higher edu' | 'binary: yes or no' |
| 5 | 'famsize' (family size) | '(binary: LE3<=3 or GT3>3)' | 22 | 'internet - Internet access at home' | 'binary: yes or no' |
| 6 | 'pstatus' (parent status) | '(binary: "T": living together or "A": apart)' | 23 | 'romantic: relation' | 'binary: yes or no' |
| 7 | 'Medu mother's education' | '(numeric: 0=none, 1=primary, 2=5th to 9th grade, 3=secondary, 4=higher edu)' | 24 | 'Fedu father education' | '(numeric: 0=none, 1=primary, 2=5th to 9th grade, 3=secondary, 4=higher edu)' |
| 8 | 'free time - free time' | '(numeric: 1-5 (low to very high)' | 25 | 'famrel: family relationships' | 'numeric: 1-5 (bad to excellent' |
| 9 | 'Mjob mother's job' | 'nominal' | 26 | 'go out – going out with friends' | 'numeric: 1-5 (low to very high' |
| 10 | 'Fjob father's job' | 'nominal' | 27 | 'Dalc: use of alcohol' | 'numeric: 1-5 (low to |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | very high' |
| 11 | 'reason to choose this school' | '(Nominal: close to "home", "reputation", "course" preference or "other")' | 28 | 'Walc: weekend alcohol' | 'numeric: 1-5 (low to very high' |
| 12 | 'guardian' | '(nominal: "mother", "father" or "other")' | 29 | 'health: current health status' | 'numeric: 1-5 (bad to very good' |
| 13 | 'Travel time' | '(numeric: 1<15 min., 2 15 to 30 min., 3: 1 hrs, or 4>1hrs)' | 30 | 'Studytime: weekly study time' | '(numeric: hrs 1<2, 2: 2 to 5, 3: 5 to 10, or 4: >10)' |
| 14 | 'absences' | 'numeric: from 0 to 93' | 31 | 'G1: 1st period grd' | 'numeric: 0 to 20' |
| 15 | 'Failures: past classes' | 'numeric: n if 1<=n<3, else 4' | 32 | 'G2: 2nd period' | 'numeric: 0 to 20' |
| 16 | 'schoolsup: extra educ. support' | 'binary: yes or no' | 33 | 'G3: final grade' | 'numeric: 0 to 20' |
| 17 | 'famsup - family educ. support' | 'binary: yes or no' | | | |

**Table 1:** Dataset Attribute Description

**Greedy Heuristic Search: Best-First Search:**

Best First Search is a heuristic search strategy where the most promising node to be expanded is determined by the evaluation function. The two versions of BFS are A* (Best-First Search) and Greedy Best-First Search. Greedy BFS utilizes the Heuristic function search and enables us to leverage both the algorithms [16]. Greedy BFS utilizes the heuristic to rank nodes within the search space and estimate their potential. It assumes that the most promising node at each iteration is the one that will reach the goal state efficiently and is found to work very well for optimization problems [17]. The pseudocode for finding the solution using the Greedy Best-First Search algorithm is presented in Figure 2.

In our research work, the best first search algorithm selects the best 11 attributes out of 33. The selected attributes are gender, age, medu, reason, failures, higher, romantic, famrel, G2, and G3.

**CLASSIFICATION MODELS:**

The classifiers Random Forest and Decision Table were applied separately on the best features selected by GBFS techniques.

**RANDOM FOREST:**

Ensemble classification is a data mining technique that leverages multiple classifiers working in tandem to determine the class label of new, unlabeled data points. Among these methods, the random forest algorithm stands out by integrating several randomly generated decision trees and averaging their outputs. This method has garnered significant interest within the research community due to its notable accuracy and effectiveness, which have contributed to enhanced overall performance [19]. In essence, a random forest classifier aggregates the decisions of multiple individual classifiers, each casting a vote, and assigns the most commonly predicted class to the input vector (see Equation 1).

$$\llbracket (x), C\_rf\hat{}B = majority\ vote\ \{C\_b\ (x)\}\rrbracket\ \_1\hat{}B, \quad \text{(Equation 1) [20]}$$

where $C_b(x)$ represents the class prediction made by the $b^{th}$ tree within the random forest. Random forests enhance the diversity among individual trees by training each one on different subsets of the original data, generated through a technique known as bagging, or bootstrap aggregating [20].

**Algorithm 1** Best First Search
**Input** Initial state $I$, goal states $G$
**Output** A solution plan

```
1:  open.insert(I, 0, h(I))
2:  while open ≠ ∅ do
3:     n ← open.remove_min() {The open list is sorted differently for different
       algorithms.}
4:     if n ∈ G then
5:        return plan from I to n
6:     end if
7:     closed.insert(n)
8:     for each v ∈ successors(n) do
9:        if v ∉ closed then
10:          open.insert(v, g(n) + cost(n, v), h(v))
11:       else
12:          if g(n) + cost(n, v) < g(v) then
13:             closed.remove(v)
14:             open.insert(v, g(n) + cost(n, v), h(v))
15:          end if
16:       end if
17:    end for
18: end while
```

**Fig 2.** Pseudocode Greedy Best First Search [18]

**DECISION TABLE:**

A decision table, a classification model, is a decision-making tool that contains all logical conditions and outcomes, defined based on specific attributes, with each potential scenario represented across different columns in the table [21]. Decision Tables streamline complex decision-making by organizing information into a clear, structured layout that's easy to interpret. This approach is particularly popular in machine learning and data mining applications, where it's often employed for tasks like predicting student performance.

The decision table employs a hierarchical structure, where entries from upper-level tables expand into another table based on pairs of additional attributes entered into the system. The design follows the dimensional stacking approach, as described in [22]. The algorithm receives *the TrainD training data, along with minsup and minconf thresholds, as inputs, as depicted in Figure* 3. The *BestSplitAttr* function generates the best splitting attribute through a genetic search process. The input parameters for *CandidateDTable* include training data alongside the designated splitting attribute, while it produces candidate decision tables. The candidate decision table becomes the decision table through the execution of *PrunDTable*, which utilises *minsup* and *minconf* parameters [23].

```
begin
   bestSplitAttr := BestSplitAttr (TrainD);
   candDTable := CandidateDTable (TrainD,
                    bestSplitAttr);
   decisionTable := PrunDTable (candDTable,
                    minsup, minconf);
end.
```

**Fig 3.** The main algorithm that generates the decision table for a given data set

**RESULTS AND DISCUSSIONS:**

This section presents a comprehensive analysis of the experimental results obtained during this research, along with a detailed discussion of the findings about the research objectives. The empirical results demonstrate that individual classifiers offer distinct

advantages for effectively detecting student performance evaluation and prediction. The performance of two classification algorithms is evaluated using the best and optimal feature selection strategies. The effectiveness of a classification algorithm can be determined by assessing its correlation coefficient, root mean squared errors, and mean absolute errors, particularly when using the best attribute assessment technique. This evaluation is conducted on both the training and test data sets. To make it easier to understand for the performance of the presented model, the root mean square error (RMSE) is used as the measurement for the accuracy of the model in the experiment. Thus, if the value of RMSE is smaller, the value is nearer to true value and hence the model is more accurate. In the next section, I will construct the formula for the Root Mean Square Error, abbreviated as RMSE. [24].

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

(equation 2)

where $n$ denotes the number of samples, $y_1$ represents the true value, and $\hat{y}_i$ represents the estimated value.

Table 2. Shows the summarized results of the decision table and the random forest algorithm classifiers with attribute evaluation techniques. The classifiers are compared based on correlation coefficient, root mean squared error, and mean absolute error.

| ALGORITHM | CORRELATION COEFFICIENT | ROOT MEAN SQUARED ERROR | MEAN ABSOLUTE ERROR |
|---|---|---|---|
| DECISION TABLE | 0.92 | 1.92 | 1.23 |
| RANDOM FOREST | 0.90 | 1.99 | 1.38 |

**Table 2:** Comparison of Decision Table and Random Forest classifiers

Figure 4 presents a graphical comparison of the two models in terms of correlation coefficient, mean squared error, and mean absolute error, which shows that the Decision Table outperforms in terms of RMS error.

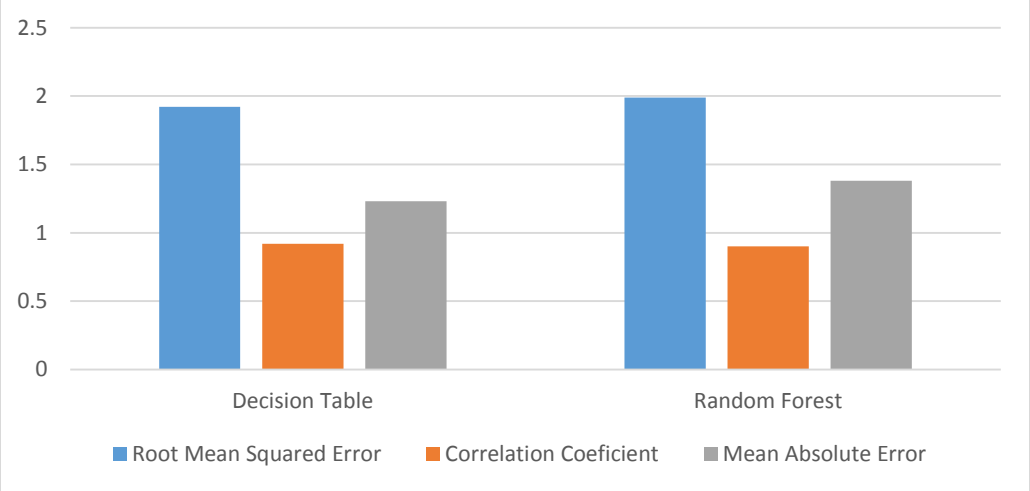


**Fig 4** Diagrammatic Comparison of ML Model

**FUTURE SCOPE:**

Data mining is used for classifying large datasets, allowing various assumptions to be made. In the classification process, attributes are responsible for generating rules. Our proposed solution offers an effective malware detection system with enhanced accuracy and reduced execution time, leveraging its parallel processing capabilities, feature normalization, and selection methods. In further research, selection of attributes has been enhanced based on different artificial intelligence approaches that can be used as fitness factors for evaluation of attributes.

**REFERENCE:**

Roman, M., Naz, I., Luqman, M. A., Ali, J., Jan, M. S., & Nawab, H. U. (2024). Stroke Disease Prediction Using K-Nearest Neighbor and Decision Tree Algorithms with Machine Learning Pre-Processing Techniques. *Migration Letters*, *21*(S4), 2015-2027.

Halde, R. R. (2016, September). Application of Machine Learning algorithms for betterment in education system. In *2016 international conference on automatic control and dynamic optimization techniques (ICACDOT)* (pp. 1110-1114). IEEE.

Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance.

Kumar, P. (2025). Evaluating Machine Learning Algorithms for Enhanced Prediction of Student Academic Performance.

Roman, M., Nawab, H. U., Ahmad, S., & Khan, I. A. (2022). K-Nearest Neighbor And Fuzzy K-Nearest Neighbor Algorithm Performance Analysis For Heart Disease Classification. *Webology (ISSN: 1735-188X)*, *19*(1).

Burman, I., & Som, S. (2019). Predicting students academic performance using support vector machine. In *2019 Amity international conference on artificial intelligence (AICAI)* (pp. 756-759). IEEE.

H. Alamri, L., S. Almuslim, R., S. Alotibi, M., K. Alkadi, D., Ullah Khan, I., & Aslam, N. (2020, December). Predicting student academic performance using support vector machine and random forest. In *Proceedings of the 2020 3rd international conference on education technology management* (pp. 100-107).

Naicker, N., Adeliyi, T., & Wing, J. (2020). Linear support vector machines for the prediction of student performance in school- based education. *Mathematical Problems in Engineering*, *2020*(1), 4761468.

Al-Shehri, H., Al-Qarni, A., Al-Saati, L., Batoaq, A., Badukhen, H., Alrashed, S., ... & Olatunji, S. O. (2017). Student performance prediction using support vector machine and k-nearest neighbor. In *2017 IEEE 30th canadian conference on electrical and computer engineering (CCECE)* (pp. 1-4). IEEE.

Hassan, C. A. U., Iqbal, J., Irfan, R., Hussain, S., Algarni, A. D., Bukhari, S. S. H., & Ullah, S. S. (2022). Effectively predicting the presence of coronary heart disease using machine learning classifiers. *Sensors*, *22*(19), 7227.

Aman, F., Rauf, A., Ali, R., Iqbal, F., & Khattak, A. M. (2019). A predictive model for predicting students academic performance. In *2019 10th International conference on information, intelligence, systems and applications (IISA)* (pp. 1-4). IEEE.

Hussain, S., & Khan, M. Q. (2023). Student-performulator: Predicting students' academic performance at secondary and intermediate level using machine learning. *Annals of data science*, *10*(3), 637-655.

Zacharis, N. Z. (2016). Predicting student academic performance in blended learning using artificial neural networks. *International Journal of Artificial Intelligence and Applications*, *7*(5), 17-29.

Alhazmi, E., & Sheneamer, A. (2023). Early predicting of student's performance in higher education. *Ieee Access*, *11*, 27579-27589.

Wilt, C., Thayer, J., & Ruml, W. (2010). A comparison of greedy search algorithms. In *proceedings of the international symposium on combinatorial search* (Vol. 1, No. 1, pp. 129-136).

Heusner, Manuel. "Search behavior of greedy best-first search." PhD diss., University_of_Basel, 2019.

Wilt, C., & Ruml, W. (2015). Building a heuristic for greedy search. In *Proceedings of the International Symposium on Combinatorial Search* (Vol. 6, No. 1, pp. 131-140).

Xie, F. (2016). Exploration in Greedy Best-First Search for Satisficing Planning.

Parmar, A., Katariya, R., & Patel, V. (2018, August). A review on random forest: An ensemble classifier. In *International conference on intelligent data communication technologies and internet of things* (pp. 758-763). Cham: Springer International Publishing.

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS journal of photogrammetry and remote sensing*, *67*, 93-104.

Göl, M., Aktürk, C., Talan, T., Vural, M. S., & Türkbeyler, İ. H. (2025). Predicting malnutrition- based anemia in geriatric patients using machine learning methods. *Journal of Evaluation in Clinical Practice*, *31*(2), e14142.

Becker, B. G. (1998). Visualizing decision table classifiers. In *Proceedings IEEE symposium on information visualization (Cat. No. 98TB100258)* (pp. 102-105). IEEE.

Lu, H., & Liu, H. (2000). Decision tables: Scalable classification exploring RDBMS capabilities. In *Proceedings of the 26th International Conference on Very Large Data Bases, VLDB'00* (p. 373).

Liu, X., Tang, Z., & Wei, J. (2025). Multi-Layer Perceptron Model Integrating Multi-Head Attention and Gating Mechanism for Global Navigation Satellite System Positioning Error Estimation. *Remote Sensing*, *17*(2), 301.