Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 12, 2025

DATA MINING–DRIVEN STUDENT PERFORMANCE PREDICTION: A SYMMETRIC UNCERTAINTY-BASED COMPARATIVE ANALYSIS OF DECISION TREE AND SUPPORT VECTOR MACHINE MODELS

Muntazira[1], Sarwat Jahan[2], Malak Roman[*3], Sana Shaiza Shams[4], Husan Jamal[5]

[1,2,4]BS-Computer Science, Department of Computer Science, University of Chitral, KP-Pakistan.
*[*3]Department of Computer Science, University of Chitral, KP-Pakistan.*
[5]BS-Computer Science, Government Degree College Chitral, KP-Pakistan.

[*3]malak_5116@uoch.edu.pk

DOI: https://doi.org/10.5281/zenodo.17909413

Keywords
Machine Learning, Symmetric Uncertainty, Support Vector Machine, Decision Tree, Classification

Article History
Received: 11 October 2025
Accepted: 21 November 2025
Published: 12 December 2025

Copyright @Author

Corresponding Author: *
Malak Roman

Abstract
Student academic performance refers to the measurable educational outcomes achieved by students through their learning process, while education represents the systematic framework designed to facilitate knowledge transfer and skill development. The use of machine learning and data mining has brought a noticeable shift in how educational data is studied. These methods help researchers and institutions make sense of complex student information by turning it into insights that can actually be used. With these tools, it becomes easier to predict how students might perform, spot those who may need support at an early stage, and take steps that can genuinely improve learning outcomes as well as the overall performance of an institution. In this study, the results clearly show that machine learning can be very effective for educational prediction. Among the models tested, the Decision Tree performed the best, reaching an accuracy of 86.5%, which is considerably higher than the Support Vector Machine's 76.4% accuracy. The Decision Tree not only achieved perfect precision (1.00) for high-achieving students but also delivered stable results across all categories. Overall, these findings highlight how powerful data-driven approaches can be in helping educational systems respond to individual student needs and support learning in a more informed and timely way.

INTRODUCTION

Artificial Intelligence (AI), particularly Machine Learning (ML) can be defined as enabling computers to acquire intelligence and improve performance without explicit programming [1]. Unlike rigid rule-based systems, ML algorithms learn from sample data and generate predictions based on acquired knowledge. This flexibility has transformed the education sector by providing data-driven insights into learners' academic enhancement and enabling accurate grade prediction. There is a well-established link between a country's development and the educational attainment of its citizens, as demonstrated by global research [2]. Machine learning (ML) has become a powerful tool in this context, improving the analysis and prediction of student academic performance by drawing on historical records, behavioral patterns, and learning trends [3]. Through predictive analytics, ML is transforming educational systems by helping institutions recognize performance patterns using academic data, engagement statistics, and socioeconomic indicators. This enables timely, proactive support that enhances both student learning outcomes and institutional planning [4,5].

Recent progress in educational data mining emphasizes the value of integrating diverse data sources—including demographic, behavioral, and institutional factors to increase the accuracy of predictive models [6]. However, a key challenge remains: processing these multi-dimensional features effectively while ensuring that the resulting models remain interpretable and reliable [7]. As a result, educational institutions are increasingly turning to predictive analytics to identify at-risk students, design targeted interventions, and enhance overall academic success [8]. Machine learning techniques such as decision trees and neural networks have shown strong performance in classifying and predicting academic outcomes by uncovering subtle patterns across complex datasets [9,10]. Predicting student performance, therefore, becomes essential for institutions that aim to support learners more effectively and improve academic results.

In response to this need, the present study investigates how machine learning techniques can enhance the accuracy of predicting student academic performance using demographic and academic data. We propose an ML-based framework that integrates Symmetric Uncertainty (SU) for feature selection with Support Vector Machines (SVM) and Decision Tree algorithms for classification. SU is used to identify the most informative features from a dataset that includes academic records, demographic attributes, and engagement metrics collected from a local university [11]. These refined features are then used to train SVM and Decision Tree models leveraging SVM's strength in capturing complex patterns and Decision Trees' interpretability for practical, educational insights. By validating the framework using real-world data, this study aims to provide a scalable and interpretable solution that supports educational institutions in enhancing student outcomes through data-driven strategies.

LITERATURE REVIEW:

Naicker, N. et al., (2020) [12] investigated the effectiveness of linear Support Vector Machines (SVM) compared to traditional machine learning algorithms to identify the most suitable model for predicting student achievement. Their results indicated that linear SVMs achieved higher accuracy across ten categorical machine learning tasks. Additionally, prior studies have shown that factors such as students' race and gender can influence mathematics performance, while access to lunch significantly affects reading and writing outcomes. Hussain, S. & Khan, M. Q. (2023) [13] developed a machine learning model to predict student performance at both secondary and intermediate levels. The dataset was sourced from BISE, Peshawar, and a genetic algorithm was applied to select the 30 most relevant features from an initial set of 126. These selected features were then used to train K-Nearest Neighbors (KNN) and Decision Tree classifiers. The Decision Tree model outperformed KNN, achieving an accuracy of 96.64%, which was 6.72% higher.

Yadav, D. et al., (2012) [14] applied Decision Tree algorithms to classify student grades, achieving moderate accuracy. While Decision Trees are valued for their simplicity and interpretability, they also face limitations such as overfitting and difficulty in modeling complex data patterns. Dutta, S. et al., (2024) [15] explored the diagnosis of chronic kidney disease using Logistic Regression, Decision Trees, and Random Forests. Their dataset contained 400 patient records, including 250 CKD-positive and 150 CKD-negative cases. Logistic Regression provided the best classification results with an accuracy of 1.0, highlighting its reliability for early diagnosis and clinical application. Shah, M. B. et al., (2019) [16] developed a "Student Performance Assessment and Prediction System" using machine learning. Their dataset from the UCI Machine Learning Repository included extensive details on students from two Portuguese secondary schools, covering grades, social activities, and demographics. Among the evaluated algorithms, Gradient Boosting achieved the highest accuracy at 93.8%, followed by Random Forest (91.7%) and SVM (89.8%). Decision Tree and XGBoost performed similarly (88.2% and 88.21%), while Logistic Regression had the lowest accuracy at 59.8%.

Agyemang, E. F. et al., (2024) [17] conducted research work on predicting students' academic performance. The dataset was obtained from XAPI-

Edu Kaggle, consisting of 16 attributes. Preprocessing techniques included hyperparameter tuning and standard scalar. The following machine learning algorithms were applying namely vector machine (SVM), Decision tree (DT), K- Nearest neighbors (KNN) , Random forest (RF) and logistic regression (LR). Among the tested ML algorithms the best one was Random Forest (RF)85%, and the worst was Logistic Regression (LR) 81% among all. El Jihaoui, M. et al. (2025) [18] performed research work on predicting and interpreting student academic performance using a Deep Learning and Shapley Additive Explanations approach. The data was sourced from the Moroccan Ministry of Education's integrated information system and the High Commission for Planning (HCP), consisting of 17 attributes. Preprocessing techniques included Removing outliers, Handling Missing values, Encoding and Scaling, and Feature Engineering. Only one algorithm was used, the Deep Learning Neural Network Model Mean Absolute Error (MAE) 0.392%. Yang, Y. et al., (2025) [19] performed research work on student academic performance prediction via Hypergran and TabNet. The dataset was taken from a University Information System. Preprocessing techniques included Anomaly removal, Missing value handling, Sample imbalance correction, Feature engineering, One-hot coding, and Feature selection. Several machine learning algorithm were employed TabNet ,K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression (LR), Adaboost, Decision Tree (DT), Random Forest (RF), Hypergraph Neural Network (HGNN) Grap convolutional network (GCN) andHyper Tab .Relational-aware knowledge tracking self-attention modelThe best algorithm was Hyper Tab 93%, and the worst was Decision Tree (DT) 47%.

Yan, C. (2022) [20] performed research work on student academic performance prediction methods. The dataset was collected from a school data platform combined with field research and literature statistics. Preprocessing techniques included Feature Encoding Categorization, Assumption, and Train test split.The Linear regression and Random forest were used is a machine learning algorithm. Randa forest head the lower RMSE (less than 2) meaning it provided the best fit and highest prediction accuracy among the models tested and linear regression had a higher RMSE indicating less accurate prediction. Junejo, N. U. R. et al., (2024) [21] performed a research work on SAPPNet (students' academic performance prediction) during COVID-19 using a neural network. The dataset was taken from a Jordanian university via a questionnaire and contained 46 attributes. Preprocessing techniques used were Numerical conversion, One-hot encoding, Label encoding, and Data cleaning. The following machine learning are used in this research paper,SAPPNet,Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), Decision Tree (DT),Random Forest (RF),Artificial Neural Network (ANN), Convolutional Neural Network (CNN). The best algorithm was SAPPNet 93 %, and the worst was Decision Tree.

RESEARCH METHODOLOGY:

The groundwork phase involves the collection of data, analyzing and validating. The Symmetric Uncertainty algorithm is used to refine and select significant features, enhancing the model's capability to recognize the most important attributes. Support Vector Machine and Decision tree classification algorithms are trained and tested. Figure 1 demonstrates our research methodology.

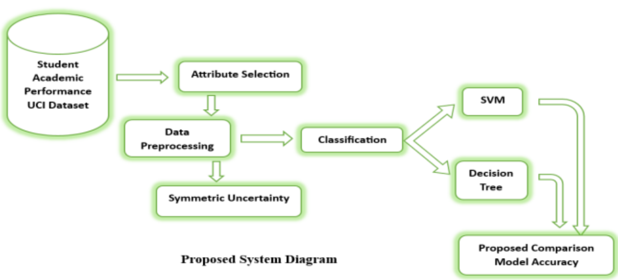

Fig.1: Research Flow Chart Diagram

DATA COLLECTION:

The dataset is downloaded from UCI repository used in this study comprises 395 records (student instances) and 33 attributes, which include both categorical and numerical variables related to student demographics, family background, social behaviour, academic performance, and other personal attributes. These features provide a comprehensive overview of each student's profile, making the dataset suitable for predictive modelling and feature evaluation.

Table 1: UCI-Repository Dataset Attribute Description

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,3:1hrs, or 4>1 hrs)'	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'			

PREPROCESSING AND FEATURE SELECTION:

Feature selection, a key step in data preprocessing, has demonstrated its effectiveness in preparing data particularly high-dimensional datasets for diverse data mining and machine-learning tasks [22]. Its primary goals are to create simpler, more interpretable models, enhance data-mining performance, and deliver clean, well-structured data for analysis [23]. Prior to feature selection, several preprocessing steps were carried out using Python (pandas and scikit-learn libraries).

Encoding Categorical Variables: All categorical features were transformed into numerical form using label encoding, ensuring that the dataset could be processed effectively by the machine-learning algorithms. This step allowed each category to be represented in a consistent and interpretable way for further analysis.

```
PS C:\Users\muhib> python -u "d:\progm folder\.vscode\SUpreprocessing.py"
Top 19 Most Important Features:
G2: 1.4228
G1: 0.9143
absences: 0.5783
failures: 0.1500
age: 0.1137
Mjob: 0.1120
Walc: 0.1101
Medu: 0.0973
goout: 0.0964
Dalc: 0.0950
Fjob: 0.0948
health: 0.0858
Fedu: 0.0830
freetime: 0.0776
studytime: 0.0737
famrel: 0.0713
reason: 0.0688
traveltime: 0.0498
schoolsup: 0.0462

Data ready! Using 19 features to predict 3 grade categories.
Low: 73, Medium: 130, High: 192
```

Fig.2: Symmetric Uncertainty Attributes Ranking Score

Entropy and Symmetric Uncertainty (SU) Calculation: To assess how strongly each feature is related to the target variable (G3), we applied entropy-based measures and computed the Symmetric Uncertainty for every feature target pair. This method helps quantify the mutual dependence between variables while adjusting for their individual levels of entropy. As a result, it provides a balanced and reliable indication of which features contribute the most to predicting student performance.

Entropy is a key concept in information theory, measuring the uncertainty present in the distribution of a variable. For a discrete random variable X, entropy is defined as:

$$H(X) = -\sum_i P(x_i)log_2(P(x_i)) \quad (1)$$

where p(x) is the probability mass function of X. After observing another variable Y, the conditional entropy of X given Y is:

$$H(X|Y) = -\sum_j P(y_j)\sum_i P(x_i|y_j)log_2(P(x_i|y_j \quad (2)$$

Here, p(y) is the prior probability of Y, and p(x|y) represents the posterior probability of X given Y. The information gain (**IG**) between two variables X and Y quantifies the reduction in the uncertainty of X after observing Y, and is defined as:

$$IG(X|Y) = H(X) - H(X|Y) \quad (3)$$

Information gain is a symmetric measure, meaning that IG (X; Y) =IG (Y; X), and it reflects the shared information between the two variables [24]. If IG (X; Y)>IG (X; Z), it implies that variable Y contributes more to the reduction in uncertainty of X than variable Z does.

To normalize the information gain and eliminate bias toward features with many distinct values, Symmetric Uncertainty (**SU**) is used. It is defined as:

$$SU(X, Y) = \frac{2 \times IG(X|Y)}{H(X) + H(Y)} \quad (4)$$

SU values lie in the range [0,1], where a value of 1 indicates complete dependence (knowing one variable fully predicts the other), and a value of 0 implies independence [25]. Because SU is symmetric and normalized, it is widely used for evaluating feature relevance and redundancy in feature selection tasks [26]. For continuous variables, a discretization step must be applied prior to the computation of entropy, IG, and SU to ensure accurate results [27].

Feature Ranking and Selection: Features were ranked based on their SU scores, and the top 19 were selected as the most relevant predictors for the three grade categories Low, Medium, and High. This selection process helped reduce the dimensionality of the dataset while preserving the variables that carried the most meaningful information for predicting student performance.

Attribute Description			
S.No	Attributes	S.No	Attributes
1	G2	11	Family relation
2	G1	12	Go out
3	Absence	13	Health
4	Failures	14	Father’s education
5	Mother’s Job	15	Study time
6	Student’s Age	16	Free time
7	Weekend Alcohol Consumption	17	School up
8	Workday Alcohol Consumption	18	Reasons
9	Father’s Job	19	Travel time
10	Mother’s education		

Table 2: Best Featured Attributes Selected by Symmetric Uncertainty

As a result of this process, the original 33 features were reduced to 19 features that demonstrated strong associations with the target variable.

CLASIFICATION MODEL:

Classification refers to the process of building a predictive model that can categorize a given set of items into predefined classes. This model is later employed to assign class labels to previously unseen or unlabelled instances based on their attributes [28].

In this research SVM and decision tree are being used as a classifier.

DECISION TREE:

A decision tree is a tree-structured flowchart where internal nodes represent attribute validations, branches indicate the outcomes of these validations, and leaf nodes correspond to class labels. The topmost node is called the root node [29]. To classify an unknown instance, the process starts at the root and traverses down the tree, following the branches based on the

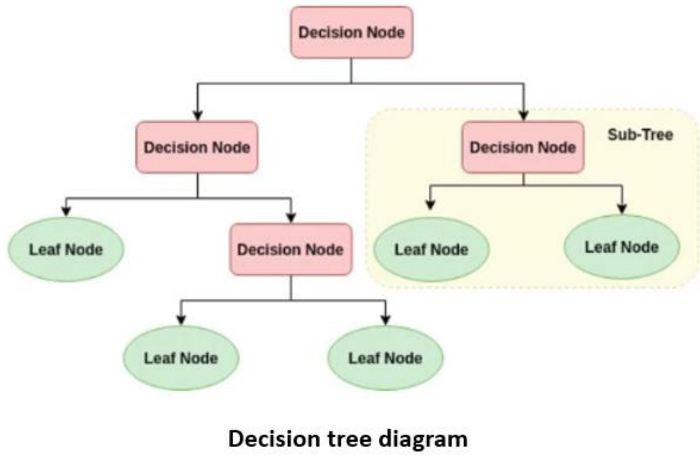

Fig.3 Decision Tree Classification Flow Diagram

instance's attribute values. Once a leaf node is reached, the instance is assigned to the class label associated with that node [30].

Entropy(S) = $-\sum p_i \log_2(p_i)$, *Where p_i is the probability of each class.*

Information Gain (S, Feature) = Entropy(S) - $\sum (|S_j| / |S|) * \text{Entropy}(S_j)$

Decision Trees played an important role in analyzing student academic performance by presenting complex information in a clear and structured manner. The model operates like a flowchart, beginning with the most influential attribute such as study hours and then moving through factors like attendance and parental education level. The model may highlight that students with limited study time combined with frequent absences are at a higher risk of poor performance. In this way, Decision Trees function as an explainable AI technique, turning detailed student data into practical, actionable insights and supporting the development of more targeted and effective educational strategies.

SUPPORT VECTOR MACHINE:

Support Vector Machine (SVM) is a robust supervised learning algorithm based on Structural Risk Minimization (SRM) from statistical learning theory. It excels in classification, regression, and other tasks by finding an optimal hyperplane that maximizes the margin between classes, even in non-linearly separable data as shown in figure 4 [31].

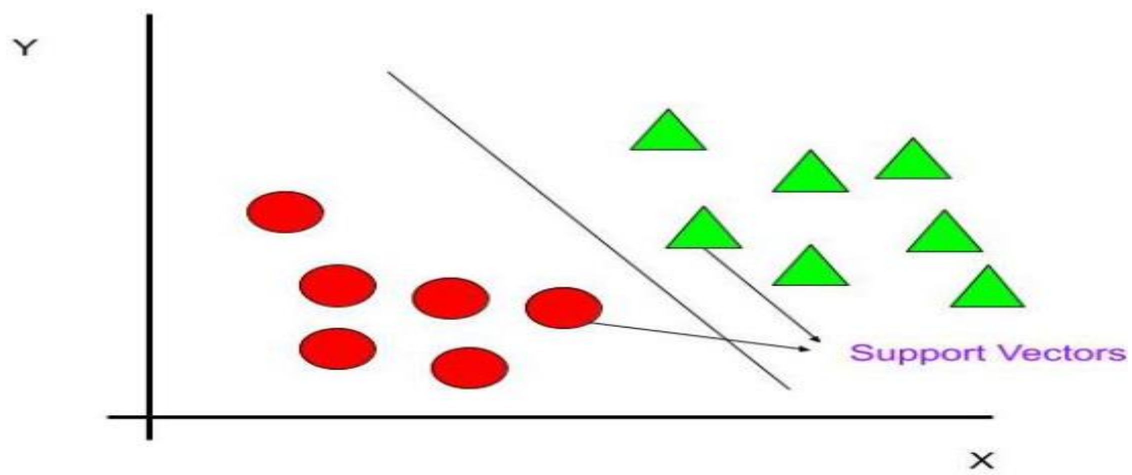

Fig.4: Support Vector Machine Working with Hyperplane

Support Vector Machine (SVM) operates by using kernel functions to transform input data into a higher-dimensional space, where it becomes easier to separate classes. The algorithm emphasizes the most critical training samples, known as support vectors (SVs), which determine the decision boundary and ensure both efficient and generalizable performance [32]. This capability makes SVM especially effective for complex datasets where traditional linear methods may fail [33].

In the context of predicting student academic performance, SVM plays a key role by analyzing multiple student attributes and classifying them into performance categories. The algorithm identifies the most influential patterns that distinguish high-performing students from those who may be at risk. By considering factors such as study habits, attendance, and demographic information, SVM builds a reliable prediction model. This allows educational institutions to implement early interventions for struggling students, supporting better learning outcomes. In essence, SVM serves as an intelligent tool that enables educators to make informed, data-driven decisions, thereby optimizing educational support system.

Results and Discussions:

The classifiers were assessed using standard evaluation metrics including accuracy, precision, recall (sensitivity), false positive rate (FPR), and false negative rate (FNR) [34]. These metrics were computed using the following formulas:

$$Accuracy = \frac{No\ of\ correct\ predictions}{Total\ no\ of\ predictions} \quad (equ.3)$$

$$Precision = \frac{True\ Positive}{True\ positive + False\ Positive} \quad (equ.4)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (equ.5)$$

$$F1_{Score} = 2 \times \frac{Precision\ \times Recall}{Precision + Recall} \quad (equ.6)[35]$$

A visual comparison of model performance was generated using a Seaborn (python library) bar plot, which clearly illustrated that the Decision Tree outperformed SVM in terms of accuracy. This comparative analysis highlights the importance of choosing an appropriate classifier for the dataset and problem type. While SVM provided a strong baseline, the Decision Tree demonstrated superior capability, making it a more effective choice for multi-class student performance prediction in this study.

IEER
Institute for Excellence in Education & Research

Matric	Accuracy	Best performer
Support Vector Machine	76.4%	**Decision Tree**
Decision Tree	86.5%	

Table. 3: SVM and DT Classifiers Comparison Table in terms of Accuracy.

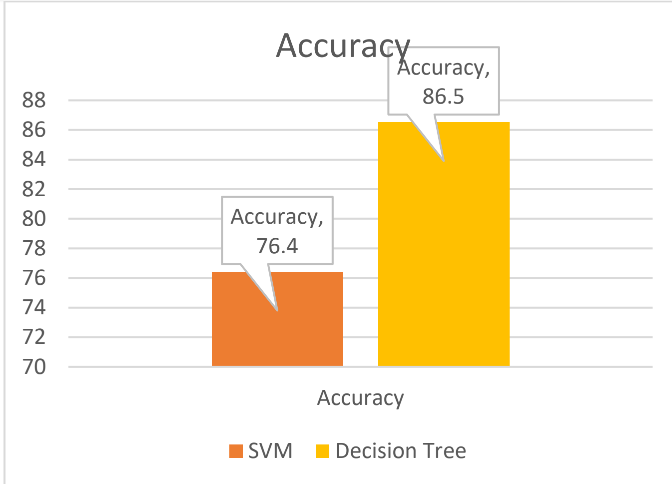

Fig.5: Classifier Performance Evaluation in terms of Accuracy

The SVM model achieved an overall accuracy of 76.47%, reflecting a satisfactory level of predictive performance. Class-wise results showed that the model performed reasonably well across all categories, with balanced precision, recall, and F1-scores. However, a slightly lower recall for the High category indicated that some instances were misclassified. This limitation may be linked to the linear decision boundary of SVM and its sensitivity to variations in class distribution.

```
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.3,random_state=42
)
```

The target variable was divided into three categories Low, Medium, and High based on the final grade (G3), allowing the problem to be approached as a multi-class classification task. To evaluate the model's ability to generalize, the dataset was split into training and testing subsets using a 70:30 ratio. This ensured that the model was trained on one portion of the data and then assessed on unseen samples, providing a reliable measure of its real-world performance.

In contrast, the Decision Tree classifier demonstrated notably stronger performance, achieving an overall accuracy of 86.55%. It delivered excellent results across all categories, including perfect precision (1.00) for the High class and a high recall (0.97) for the Low class. The macro-averaged and weighted F1-scores of 0.87 further highlight its stability and reliability as shown in figure 6. This improvement can be attributed to the Decision Tree's ability to capture complex, non-linear relationships and interactions among features capabilities that SVM handles less effectively in this context.

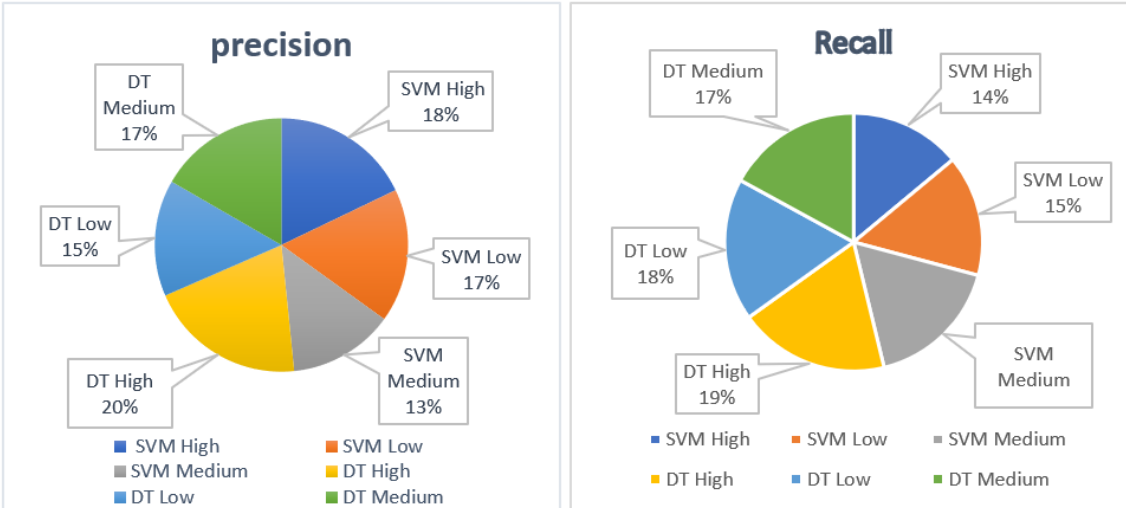

Fig .6: Classifiers Comparisons in terms of Precision and Recall

In summary, several key steps contributed to the success of the classification process: label encoding for categorical variables, feature scaling for SVM, appropriate train test splitting for unbiased evaluation, and the use of robust evaluation metrics such as accuracy and F1-score. Based on the comparative results, the Decision Tree classifier emerges as the most suitable model for this task, offering stronger predictive accuracy and more consistent performance across all grade categories.

REFERENCES:

Rahman, M. A. U., Roman, M., Ahmad, S., Jan, M. S., & Nawab, H. U. (2021). Optimizing Collaboration: Insights into Inter-Team Coordination and Self-Management in Distributed Agile Software Development. Webology (ISSN: 1735-188X), 18(5).

Timbal, M. A. (2019). Analysis of Student-at-Risk of Dropping out (SARDO) Using decision tree: An Intelligent predictive model for reduction. International Journal of Machine Learning and Computing, 9(3), 273-278

Roman, M., Ullah, A., Ullah, M. A., Hussain, F., Shams, S. S., Bint-e-Meraj, A., & Ali, S. (2025). PREDICTING ACADEMIC SUCCESS: A MACHINE LEARNING APPROACH USING DECISION TABLES AND RANDOM FORESTS ALGORITHMS. Spectrum of Engineering Sciences, 3(5), 205-213.

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.

Silva, A. (2008). Using data mining to predict secondary school student performance.

Syed Mustapha, S. M. F. D. (2023). Predictive analysis of students' learning performance using data mining techniques: A comparative study of feature selection methods. Applied System Innovation, 6(5), 86.

Ngulube, P. (2025). Predicting Academic Success and Identifying At-Risk Students: A Systematic Review of Data Analytics and Machine Learning Approaches in Higher Education Institutions. Educational Administration: Theory and Practice, 31(1), 117-134.

Shaun, R., De Baker, J., & Inventado, P. S. (2014). Chapter 4: Educational data mining and learning analytics. Computational Science, 7, 1-16.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? The journal of machine learning research, 15(1), 3133-3181.

IEER Institute for Excellence in Education & Research

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.

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).

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.

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

Yadav, D., Pal, S., & Thakur, P. (2012). Comparative study of data mining algorithms for predicting academic performance. International Journal of Computer Applications, 52(11), 43-48. doi:10.5120/8231-276.

Dutta, S., Sikder, R., Islam, M. R., Al Mukaddim, A., Hider, M. A., & Nasiruddin, M. (2024). Comparing the Effectiveness of Machine Learning Algorithms in Early Chronic Kidney Disease Detection. Journal of Computer Science and Technology Studies, 6(4), 77-91.

Shah, M. B., Kaistha, M., & Gupta, Y. (2019, November). Student performance assessment and prediction system using machine learning. In 2019 4th international conference on information systems and computer networks (ISCON) (pp. 386-390). IEEE.

Agyemang, E. F., Mensah, J. A., Ampomah, O. A., Agyekum, L., Akuoko-Frimpong, J., Quansah, A., & Akinlosotu, O. M. (2024). Predicting Students' Academic Performance Via Machine Learning Algorithms: An Empirical Review and Practical Application. Computer Engineering and Intelligent Systems, 15(1), 86.

El Jihaoui, M., Abra, O. E. K., & Mansouri, K. (2025). Predicting and Interpreting Student Academic Performance: A Deep Learning and Shapley Additive Explanations Approach. In SHS Web of Conferences (Vol. 214, p. 01001). EDP Sciences

Yang, Y., Zhao, S., An, S., Li, X., & Zhang, Y. (2025). Student academic performance prediction via hypergraph and TabNet. Journal of Big Data, 12(1), 119

Yan, C. (2022). Research on Student Academic Performance Prediction Methods. Highlights in Science, Engineering and Technology, 24(1), 257-263.

Junejo, N. U. R., Huang, Q., Dong, X., Wang, C., Zeb, A., Humayoo, M., & Zheng, G. (2024). SAPPNet: students' academic performance prediction during COVID-19 using neural network. Scientific Reports, 14(1), 24605.

Shah, F. A., Meraj, A. B. E., Roman, M., Anwar, M., Zaib, A., Shams, S. S., & Hussain, F. (2025). Enhancing Weather Forecasting Accuracy: A Machine Learning Approach Using Genetic Algorithm and Random Forest. Global Research Journal of Natural Science and Technology.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. ACM computing surveys (CSUR), 50(6), 1-45. Mining, M. D. Data Mining: Concepts and Techniques (2nd.edition).

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on pattern analysis and machine intelligence, 27(8), 1226-1238.

Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., & Data, M. (2017, December). Practical machine learning tools and techniques. In Data mining (Vol. 2, No. 4, pp. 403-413). Amsterdam, The Netherlands: Elsevier Publishers.

Vergara, J. R., & Estévez, P. A. (2014). A review of feature selection methods based on mutual information. Neural computing and applications, 24(1), 175-186.

Song, L., Langfelder, P., & Horvath, S. (2012). Comparison of co-expression measures: mutual information, correlation, and model based indices. BMC bioinformatics, 13(1), 328.

Ullah, M. A., Ullah, A., Roman, M., Anwar, M., Siddiqi, M. U. M., Jaffar, M. H., & Ali, J. (2025). A Comparative Study of Machine Learning Algorithms for Cardiovascular Risk Prediction: Support Vector Machine, Gradient Boosting, And Rotation Forest. Spectrum of Engineering Sciences, 481-491.

Timbal, M. A. (2019). Analysis of Student-at-Risk of Dropping out (SARDO) Using decision tree: An Intelligent predictive model for reduction. International Journal of Machine Learning and Computing, 9(3), 273-278.

I. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, San Francisco: Morgan Kaufmann Publishers, 2000.

Islam, M. A., Majumder, M. Z. H., & Hussein, M. A. (2023). Chronic kidney disease prediction based on machine learning algorithms. Journal of pathology informatics, 14, 100189.

V. N. Vapnik, Statistical Learning Theory, John Wiley & Sons, Chichester, ISBN: 0-471-03003-1, 1998.

Xia, T. (2016). Support vector machine based educational resources classification. International Journal of Information and Education Technology, 6(11), 880.

Hussain, F., Shams, S. S., Roman, M., Anwar, M., Shah, F. A., Meraj, A. B. E., & Uddin, M. A. MACHINE LEARNING IN HEALTHCARE: PREDICTING CHRONIC KIDNEY DISEASE THROUGH FEATURE-DRIVEN HEURISTIC MODELS.

Anwar, M., Rahman, T., & Roman, M. (2025). Voice-Activated Smart Environments: Deep Learning Approach for Pashto Speech Command Processing. Spectrum of Engineering Sciences, 3(5), 541-550.

IEER
Institute for Excellence in Education & Research