05

A Data Mining Perspective on Academic Performance and Extracurricular Engagement

[1]*Paras Kacha,* [2]*Tanya Shruti, and* [3]*Rajeswari K*

Pimpri Chinchwad College of Engineering, Pune, India
[1]paras20pk@gmail.com, [2]tanya.shruti@pccoepune.org, [3]kannan.rajeswari@pccoepune.org

Abstract

There is still a lack of understanding about the relationship between extracurricular activities and academic performance, as educa- tional institutions often focus on test scores rather than sports, arts, or volunteer work contributions. This undervaluation hinders the devel- opment of education policies supporting academic success and holistic student growth. This study uses a dataset of 2,392 high school students aged 15–18 to explore how extracurricular activities impact students' academic performance and personal development by taking into account such attributes as age, gender, parental education, study time, absen- teeism, and participation in sports, music, and volunteering. The ma- chine learning models used were Naive Bayes, Logistic Regression, SVM, XGBoost, and Random Forest. Preprocessing of the data included impu- tation of missing values and feature standardization. Model evaluation metrics consisted of accuracy, precision, and recall, for which Random Forest attained the highest accuracy at 91%. Random Forest's ensemble approach aggregated multiple decision trees and selected key predictors like parental support, study hours, and extracurricular involvement. The findings point out that extracurricular activities significantly enhance academic and behavioral outcomes. Sports and music foster cognitive skills, discipline, and teamwork, while volunteer work promotes social awareness, indirectly improving academic success. This study underlines the need for balanced educational policies integrating academics with extracurricular activities to support holistic student development.

Keywords: *Extracurricular Activities.Academic Performance.Ran- dom Forest. Machine Learning Models · Holistic Student Development. Educational Policies.*

1. Introduction

1.1 Background

A student's performance in school is the most significant determinant of a stu- dent's academic and career future. Not only does it portray the acquisition of knowledge within the curriculum, but it also demonstrates the application of that knowledge to solve real problems and situations of everyday life. Researchers have traditionally perceived factors such as study habits, attendance, and parental involvement as contributors to academic success. They have dedicated much effort to studying these factors, leading to the creation of well-established interventions to enhance students' academic outcomes. Recent advances in data analytics and educational data mining have opened new avenues for understanding the influ- ences on academic success. For instance, researchers are now exploring social and behavioral factors that were previously ignored. However, despite innovations, the role of extracurricular activities such as sports, music, and volunteering has been underexplored. The predominant educational structure emphasizes measurable academic outcomes like test scores, grades, and attendance; it rarely evaluates the impacts of extracurricular activities on students. Institutional au- thorities underplay the valueadded outcome in developing a student's holistic development. Policies are mainly concerned with academics, as there is little space to implement extracurricular activities in curricula. This lacuna of study presents a significant challenge for educators and policymakers charged with bal- ancing academia with preparing a rounded individual. While activities outside of class contribute to fundamental human skills such as teamwork, discipline, and creativity, their direct link to better academic performance is far less docu- mented. Therefore, a balanced education that advances academic and extracur- ricular priorities is necessary. This study explores how extracurricular activities such as sports, music, and volunteering impact a student's academic performance in terms of GPA and grade classification.

1.2 Problem Statement

While the positive effects of extracurricular activities are well documented, their direct implications on academic performance are poorly understood. Educational institutions focus more on assessment-based measures of success, such as test scores and attendance than on the effects of non-academic engagement on student learning and development. This focus creates a significant gap in educa- tional strategies, as there is a lack of comprehensive, data-driven insights into how extracurricular activities contribute to academic outcomes. Available studies suggest correlations between extracurricular involvement and academic perfor- mance, yet these findings have not been sufficiently adapted to shape educational policy. Furthermore, traditional statistical methods are often in a position to ad- dress the complexity of nonlinear relationships between extracurricular participation and academic achievement. Consequently, educators, administrators, and policymakers are unsure about the true impact of extracurricular activities on students' academic performance. This study addresses these challenges by con- ducting a comprehensive, data-driven analysis of how extracurricular activities

influence academic performance. The research will identify hidden patterns and gain insights that traditional methods may overlook through the use of machine learning techniques. This study aims to identify concrete proof for incorporating a more balanced educational policy that places value on the stimulation of stu- dent development through academic and extracurricular activities. Results will lead to the recommendation of utilizing extracurricular activities in educational interventions to improve student performance.

1.3 Objectives

This paper examines how participation in extracurricular activities is related to student's academic outcomes, significantly whether sports, music, and vol- unteerism activities predict students' GPAs and grade levels. Though schools historically value test scores and attendance as evidence of school success, this paper explores the unheralded contributors toward a student's academic success - non-academic activities. The other demographic and behavioral factors that this study will examine include age, gender, ethnicity, parental involvement, and study habits to evaluate which variables significantly impact academic success. By focusing on the broader picture of student development, this study expands on the traditional scope of test scores in illuminating the influence of extracur- ricular engagement on student outcomes. The study will also include multimodal data such as clickstream activity and peer interactions in addition to traditional data about demographics and extracurricular participation to better illustrate students' learning behaviors and academic trajectories. Such an approach would provide a better understanding of the complexities of determining academic per- formance. The study compares the performance of five machine learning models: Naive Bayes, Logistic Regression, Support Vector Machines (SVM), XGBoost, and Random Forest on predicting student success based on the involvement level. The work shall identify specific patterns that might guide educational approaches and policies; such knowledge helps design appropriate interventions that provide academics with challenging conditions and support integration with engaging activities.

1.4 Scope of the Research

The research study focuses on high school students aged 15-18 years, a pe- riod crucial for teenagers' academic and personal development. The dataset collects demographic data (age, gender, and ethnicity), academic metrics, es- pecially GPA, grade classifications, and behavioral factors such as study habits, absenteeism, and involvement of parents. Additionally, the dataset captures data from various extracurricular activities such as sports, music, and volunteering. The influence of demographic factors, such as gender and socioeconomic status, on educational outcomes is well documented. This study will carefully consider these factors to ensure a nuanced understanding of the results. However, the study has several limitations. It does not account for extrinsic factors, such as socioeconomic background, peer influence, or schoolspecific variables, such as teacher quality and curriculum differences, which can also affect academic perfor- mance. However, the dataset included only one age group and location, limiting the ability to generalize to a broader population. Nonetheless, this study's find- ings can open the way for more studies to find out how different extracurricular activities operate across various age groups, geographic locations, and educa- tional systems. These findings can be a stepping stone toward more holistic educational policies that

can balance academic learning with nurturing personal growth through extracurricular engagement, thus encouraging a more holistic approach to student success. Despite all this, this research will provide the basis for future studies into the role of extracurricular activities in academic success. It is expected that findings from this research will inform succeeding research and guide policy development that fosters a more balanced and holistic educational approach whereby academic achievement and personal growth through engagement with extracurricular activities are placed on the same pedestal.

2. Literature Review

2.1 Related Work

Several studies have examined the relationship between extracurricular activities and academic performance using machine learning models. The research explored the impact of time spent on extracurricular activities on academic outcomes us- ing Logistic Regression and K-Nearest Neighbors, finding Logistic Regression to be more accurate but based on a smaller dataset of 395 responses and limited predictors. Similarly, The research employed Random Forest, Decision Tree, and KNN with six predictors, including time spent on extracurricular activities, study time, and absenteeism, concluding that Random Forest performed best with an 89% accuracy, though the dataset size was limited to 390 students. Unlike these studies, which were confined to small datasets and fewer predictors, your research uses a larger dataset of 2,392 students and advanced algorithms like Random Forest to identify the broader impact of extracurricular activities on academic and personal development. The influence of extracurricular activities (EAs) on students' academic performance has been a significant area of educational research. Studies reveal that participation in EAs fosters improved academic outcomes by enhancing students' motivation, emotional wellbeing, and stress management. Gutierrez et al. (2024) demonstrated that students involved in extracurricular activities often report lower levels of academic stress and burnout while maintaining higher academic performance in STEM electives. Their study highlights the importance of balancing academic pursuits with extracurricular involvement to mitigate depression and tension. Furthermore, research by Yaacob et al. (2019) emphasizes the value of using predictive models to analyze students' performance in academics. Employing supervised data min- ing techniques, including Naïve Bayes and Decision Trees, this study identifies critical factors contributing to academic success. The inclusion of cocurricular attributes such as attendance and engagement in educational activities underpins the importance of non-academic influences on academic outcomes. Another perspective is provided by Santiago et al. (2024), who found a negative correlation between excessive extracurricular activity hours and academic performance. This suggests the need for a balanced approach to avoid academic fatigue while leveraging the motivational benefits of EAs. Their findings also stress the role of institutional support and tailored strategies to help students achieve optimal academic and emotional outcomes.

Early prediction of student performance has become a crucial task for educational institutions to enable timely interventions and improve academic outcomes. This survey explores data mining techniques, such as Decision Trees, Neural Networks, and SVMs, for identifying factors that influence performance, including socioeconomic and

psychometric variables. Lastly, research from Samdrup Jongkhar schools (2021) links structured EA programs with better discipline, leadership, and academic achievements among secondary students. This underscores the broader developmental benefits, suggesting that schools with wellintegrated EA programs witness holistic student growth. Extracurricular activities (ECAs) are structured, nonacademic engagements that include sports, arts, and community services. These activities play a significant role in complementing formal education by fostering academic performance, social skills, and career readiness. Numerous studies have documented the positive relationship between ECA participation and improved academic outcomes, such as better grades, enhanced test scores, and higher attendance rates. For instance, students involved in sports and music often develop critical thinking, discipline, and time management skills that contribute to their academic success. The impact of ECAs varies based on the types of activities students engage in. Research highlights that different ECA profiles, such as "sports-only," "community clubs," or "highly involved" groups, yield distinct outcomes. While participation in sports and diverse ECAs generally correlates with better academic performance, some profiles, such as those involving community clubs, have occasionally been associated with undesirable behaviors like delinquency. This nuanced understanding underscores the need to analyze the specific types of ECAs students participate in to gauge their full impact. Beyond academics, ECAs contribute significantly to students' personal and social development. They help build leadership, collaboration, and emotional intelligence, enabling students to connect better with their peers and community. Students who engage in ECAs often report increased confidence and reduced negative behaviors, such as drug use and disciplinary issues. These activities also cultivate skills like teamwork, communication, and problemsolving, which are essential for future career success. However, balancing ECAs with academic commitments remains a challenge, as they demand significant time and energy, potentially adding stress to students' lives. Methodologies used in the reviewed studies include surveys, longitudinal data collection, and statistical analyses. These studies have examined diverse populations, such as a survey of 27,121 students in British Columbia, Canada. Metrics like grade point averages (GPAs), attendance records, and behavioural assessments were employed to measure the impact of ECAs. The findings consistently suggest that while ECAs have predominantly positive effects, the outcomes can vary de- pending on the type and intensity of participation. To maximize the benefits of ECAs, schools and policymakers should ensure their integration into regular programs, backed by adequate funding and resources. Programs should align with students' interests and abilities to enhance engagement and outcomes. Furthermore, parental awareness and involvement are crucial in encouraging participation and overcoming barriers. By addressing these considerations, ECAs can continue to play a transformative role in students' academic and personal development. Students who participate in extracurricular activities (ECA) experience better academic performance, as highlighted in the study. The involvement in ECA not only enhances their overall university experience but also contributes to the development of critical skills such as time management and teamwork. This improvement in academic performance is linked to greater student satisfaction and a stronger academic reputation for institutions that support these activities, ultimately attracting high-achieving students and fostering their overall development.

3. System Methodology

It's designed to predict grade classifications based on student's GPA as well as the influencing factors, which include extracurricular participation, parental involvement, and study habits. A data-driven pipeline of machine learning is employed in this. Each stage of the pipeline ensures that raw data transforms into actionable insights through structured steps; hence, accuracy and reliability are maintained.

There are four main stages in the system architecture. The first is Input Stage, which is called Data Ingestion and refers to the loading of the dataset. This en- compasses demographic attributes such as age and gender; academic metrics like GPA and grade classification; and behavioral variables such as study time and absenteeism. This stage typically involves an initial inspection of data to understand its structure; identify missing or inconsistent values in the data; and detect some potential outliers. These represent the foundation for robust preprocessing. The second stage involved is Preprocessing and Feature Engineering, which transforms the original raw dataset into a form clean and structured enough to be modeled. Key steps will involve handling missing data by minimizing information loss, categorical attributes such as "Gender" and "Participation in Sports" encoded into a numerical value, normalization of the numerical features like "GPA" and "Weekly Study Time" to make scales uniform, and further feature engineering, in this case, a composite score for parental involvement, to make complex relationships between variables more effective. Such cleaned data will be used in model training and evaluation to train several selected machine learning models optimized based on hyperparameter tuning to raise their performance.

3.1 Data Collection

This research derives its dataset from Kaggle. It provides a basis for assess- ing factors influencing academic outcomes in high school students. The data include 2,392 records from students aged between 15–18 years old, comprising details on demographic attributes, academic metrics, behavioral patterns, and extracurricular activities. Key features are binned into groups: demographic at- tributes include age (a numeric attribute), gender (a binary categorical variable: Male → 1, Female → 0), and ethnicity, which is diverse; academic measures in- clude GPA, which is the most important numeric measure of performance, and grade categories, the target variable which classifies levels of performance like High, Medium, Low. The behavioral factors comprise the following: study time in weeks (number of hours per week), absenteeism (days absent), and parental involvement-a composite measure of education levels and support. Participation records in extracurricular activities include the variables for sports, music, and volunteering, coded as Yes/No.

The dataset was preprocessed in order to have intact and reliable data. The missing values did not pose significant problems, so mean imputation for numerical variables and mode imputation for categorical variables were carried out to keep the completeness of the dataset. Distribution of grade categories was looked at for any imbalance, with plans to over-sample or under-sample if that would be the case. These steps ensure the robustness of the dataset to be analyzed. The selected features capture not only direct academic efforts, such as GPA and study time, but also emphasize the indirect contributions from extracurricular activities that enhance cognitive skills, teamwork, and discipline that shape students into being academically successful.

3.2 Tools/Technologies Used

A variety of tools and technologies were used for data collection, analysis, modeling, and evaluation. The primary programming language used for this research was Python (v3.9), chosen for its versatility and extensive ecosystem of libraries for data manipulation, machine learning, and visualization. Key libraries included Pandas, which was used for efficient data handling and preprocessing, and NumPy, which provided numerical computation capabilities for handling large datasets. To support the visualization of data, both Matplotlib and Seaborn were used to display data trends and model performance visually with infor- mative charts and plots. In this research, diverse machine learning algorithms were tested with regards to the prediction of student academic performance. During that process, Scikit-learn was used when building and testing models. Random Forest, XGBoost, SVM, Naive Bayes, and Logistic Regression all followed through this process. These models were evaluated in terms of metrics such as accuracy, precision, recall, and F1-score. For the research to be reproducible and for efficient execution of computationally intensive models, the project was developed on Google Colab, a cloudbased platform, allowing for access to computational resources without taxing local machines. The initial data exploration and organization was carried out using Microsoft Excel in inspecting and arranging the CSV data files, ensuring the suitability of the dataset for further analysis. These tools streamlined a workflow from data ingestion to model evaluation, enabling an allaround analysis of how extracurricular activities influence students' academic performance.

3.3 Algorithms

This research used a few machine learning algorithms to predict the academic performance of students. It aimed at choosing the best model to predict grade classifications based on various input features, such as extracurricular activities, parental involvement, and study habits.

Naive Bayes Classifier: Naive Bayes is a probabilistic classifier that depends on Bayes' Theorem and assumes features to be independent given the class label. It is known as "naive" because of its assumption of independence. Although Naive Bayes does reduce the realworld complexities to some extent, it performs quite well for tasks such as text classification and smaller datasets. Naive Bayes was used for student classification on the basis of GPA and cocurricular activities, which were helpful in handling categorical data and checking the efficiency. Its simplicity, efficiency with a large feature set, and applicability to categorical data make it a valuable model.

Logistic Regression: Logistic Regression is a linear model used for classification purposes. It computes the probability of a categorical dependent variable using a logistic function, which is sigmoid, and then predicts based on the highest probability. In this research, Logistic Regression categorized students' academic performance into different classes by analyzing factors like GPA, extracurricular activities, and study habits. Its main benefits are that it can generate probabil- ities for predictions, supports both binary and multi-class classification, and is easy to interpret and apply, which makes it a very useful resource for this study. Support Vector Machine (SVM): SVM is a supervised learning model that constructs a hyperplane to separate classes in a feature space. They are very effective when the data cannot be linearly separated. In

this research, SVM was applied to classify the students' grade categories based on demographic and behavioral attributes. SVM performs well with high-dimensional data and classifies non-linear patterns by using the kernel trick, making it appropriate for identifying complex relationships within the data.

Random Forest: Random Forest is an ensemble learning method that con- structs multiple decision trees and aggregates their predictions to improve ac- curacy and reduce overfitting. It was the most effective model for this research, achieving the highest accuracy of 91%. Random Forest worked exceptionally well because of its ability to handle mixed data types, including both numerical and categorical features, and its robustness against overfitting through ensemble averaging. Further, it gave me feature importance scores that helped to under- stand which variables would affect the academic outcome. These would include parental involvement, study habits, and participation in extracurricular activities. The nature of the data, with such variables as demographic, and behavioral measures, necessitated a model able to capture higher-order interactions among the features. And Random Forest performed well. Its interpretability and strong performance with datasets containing nonlinear relationships further validated its suitability as the best choice for this study.

XGBoost: XGBoost (Extreme Gradient Boosting) is an optimized gradient boosting algorithm renowned for its speed and performance. It handles large datasets effectively and excels in classification problems. For this research, XG- Boost was used to build a robust model for predicting academic performance based on multiple features. Its ability to deal with missing values and to detail feature importance analysis made it a strong contender. XGBoost is a strong contender since it has good efficiency in large datasets, with internal mechanisms that manage missing data, and with detailed insights into the feature importance of what could make a student more successful in academic pursuits. Even though Naive Bayes, Logistic Regression, SVM, Random Forest, and XGBoost performed exceptionally well in producing valuable insights regarding the prediction of academic performance, Random Forest provided the best model accuracy of 91%. This model can deal with categorical and numerical fea- tures; overfitting is also diminished by averaging different ensembles, and it can of- fer feature importance scores. The nature of the dataset, containing a mix of demographic, behavioral, and academic features, required that the model could capture complex higherorder interactions. The robustness of Random Forest in dealing with such interaction and its capacity to generalize well was a crucial fac- tor in its better performance. Additionally, the interpretability of Random For- est and its strong performance with datasets containing nonlinear relationships further validated its suitability as the best choice for this study.

3.4 Machine Learning Model Development Workflow

They studied it rigorously and followed scientific research techniques for adequate accuracy and consistency for the resultant report. The first step was collecting data, wherein a complete dataset was gathered from Kaggle with 2,392 high school students aged 15-18 years. This dataset encompasses a wide range of features such as demographic details like age, gender, and parental education; academic information regarding GPA and grade classification; and behavioral factors related to study time, absenteeism, and involvement in extracurricular activities. Once the data was

collected, preprocessing was performed to prepare the raw data for analysis. These include handling missing values by imputation, normalizing numerical features to ensure uniformity, and encoding categorical variables like extracurricular participation into a numerical format suitable for modeling. Feature engineering was then done by extracting meaningful attributes, which could enhance the model's performance. This included identifying correlations between features and creating new interaction terms to provide better accuracy. After developing and training multiple machine learning models, including Random Forest, SVM, Logistic Regression, and XGBoost, the next step was to optimize the developed models using hyperparameter optimization techniques for performance maximization. These models were evaluated based on primary metrics, such as accuracy, precision, and recall, and relied on crossvalidation techniques to prevent overfitting, thereby ensuring robust results. Finally, the results were visualized through graphs, correlation matrices, and model comparison plots. Such visualization helped the stakeholders to understand the trends and infer actionable insights from the data. In this process, from collecting data to evaluating models, utmost care was taken for reliability in results while establishing a relationship between extracurricular activities and academic performance. Figure 1 depicts the model's workflow, explaining each of these key steps in the process.

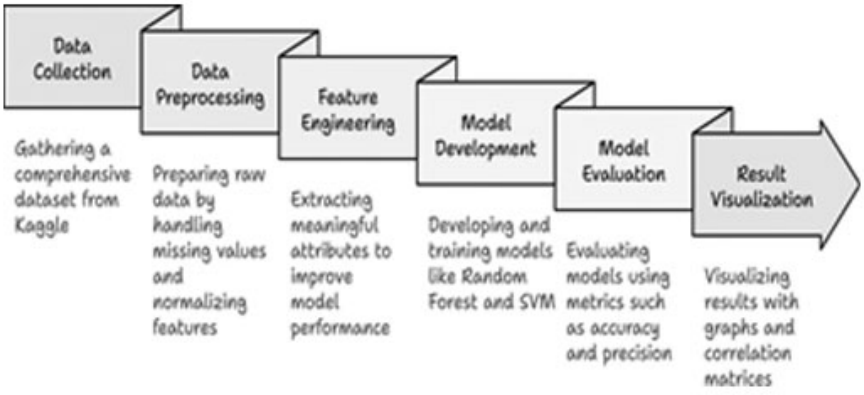

Fig. 1. Machine Learning Model Development Workflow

4. Implementation

The implementation of the research was divided into a number of welldefined steps, starting with the acquisition and preprocessing of the data, followed by feature engineering, model selection, training, evaluation, and final result interpretation. Each stage was done with high precision to ensure accurate insight into the relationship between extracurricular activities and academic performance by students.

4.1 Data Acquisition and Exploration

The dataset for the study had included comprehensive information regarding academic performance, involvement in extracurricular activities, demographic information, and so on, that influence learning. Initially, data exploration was done in Python to load into the environment where an initial exploration led to identifying some patterns,

an assessment of the quality of data and distribution of major features. In this stage, descriptive statistics and visualizations were used to understand the structure of the data set and its aptness for the analysis. Initial data inspection to ensure the absence of inconsistencies or errors in imported data was also carried out using tools such as Microsoft Excel.

4.2 Data Preprocessing

A considerable part of the implementation was performed in preprocessing raw data. It consisted of several stages for preparing the given dataset in order to be ready for machine learning tasks. Missing values in some variables are replaced appropriately by means or medians in numeric variables and mode for categorical variables. Outliers will also be spotted using some statistical methods; either correction or omission of such outliers was made to avoid changing the outcome. Other categories such as extracurricular participation and demographic details were encoded as numerical representations through one-hot encoding and label encoding. The reason for doing so is that the model required all those values to be of compatible forms for computation. Continuous variables such as GPA and study time were standardized or normalized, making the features of all equal scale, which helped the model perform better and converge at a faster rate.

4.3 Feature Engineering and Selection

The next step would be feature engineering, enhancing the ability of the dataset to make predictions. Correlation analysis has been done in order to identify the relationship between the variables. Features with high correlation values with the target variable were given first priority, such as extracurricular participation and study habits. Features that seem redundant or irrelevant were deleted to reduce noise and to decrease computational complexity.

New features were also generated by combining or transforming existing ones, such as the total time spent on extracurricular activities or deriving interaction terms between parental support and study habits. Feature importance analysis using tree-based models like Random Forest was also performed to ensure that only the most relevant features were retained for training the models.

4.4 Model Selection and Training

Implementation of various machine learning algorithms ensued for classification purposes. These include Naïve Bayes, KNN, SVM, Logistic Regression, XG- Boost, and Random Forest. The algorithms were chosen since they have diverse strengths in their respective approaches, especially when it comes to dealing with a classification task and the kind of data distribution.

Each model was trained on the prepared dataset. Grid search and crossvalidation techniques have been used for tuning each model's hyperparameters. For instance, the number of trees and the maximum depth in the case of Random Forest, and learning rate and tree depth for XGBoost, are adjusted iteratively. That way, no overfitting to the training data or underfitting should occur, making the models' generalization capabilities better.

4.5 Model Evaluation and Comparison

In the evaluation, comparison was made between the model performance with respect to standards of metrics like accuracy, precision, recall, and receiver operating

characteristic curves. Such detailed insights were obtained as regards strengths and weaknesses from every model. Random Forest and XGBoost ranked among the best models regarding the level of accuracy attained; the Random Forest model stood ahead a little in aspects like generalization and feature in- terpretation.

	train_accuracy	test_accuracy	train_precision	test_precision	train_recall	test_recall
0	0.792115	0.767409	0.797607	0.759596	0.792115	0.767409
1	0.781959	0.710306	0.771689	0.706637	0.781959	0.710306
2	0.873357	0.846797	0.836549	0.811006	0.873357	0.846797
3	0.807646	0.753482	0.795335	0.728864	0.807646	0.753482
4	1.0	0.9039	1.0	0.902156	1.0	0.9039
5	0.943847	0.916435	0.94463	0.914396	0.943847	0.916435

Fig. 2. Model Evaluation and Comparison

Visualization tools were extensively used during this phase to illustrate model performance. For instance, bar charts compared the accuracy of all models, while confusion matrices highlighted classification errors. Feature importance plots from tree-based models provided insights into the most influential factors affecting students' performance.

5. Results and Discussion

Figure 3 illustrates the analysis of the correlation matrix produced useful relationships of attributes influencing student's academic performance. Heatmap reveals that support by parents correlates at a moderate level positively with GPA; in other words, higher parental support influences good performance of the student, whereas study time per week has shown to have positive associations with GPA; in this respect, good steady habits while studying matter significantly. On the other hand, attributes like absences show a strong negative correlation with GPA, indicating that frequent absences heavily impact academic success.

Extracurricular activities showed a moderate positive correlation with GPA, signifying their role in enhancing the overall development and performance of the students. The visualization provided a very comprehensive understanding of how different factors interplay in influencing students' academic outcomes.

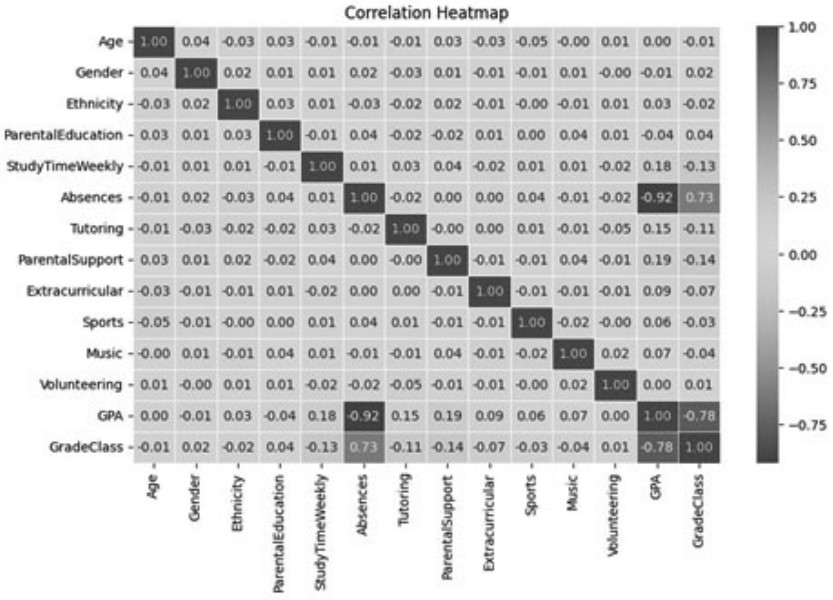

Fig. 3. Correlation Matrix

Figure 4 illustrates the analysis elaborated on the effect of after-school activities on GPA as presented in the bar chart. The results were evident, and there was a higher average GPA for students engaged in after-school activities compared to those who were not. This indicates that after-school activities positively impact academic performances through various lessons regarding discipline and time management while affecting wholesome development. This bar chart clearly presented the significant gap between GPA while emphasizing the role of extracurricular activities in educational institutions. Figure 5 illustrates the paper that assessed different machine learning models for predicting academic performance, namely 1 Naive Bayes, 2 SVM, 3 KNN, 4 Logistic

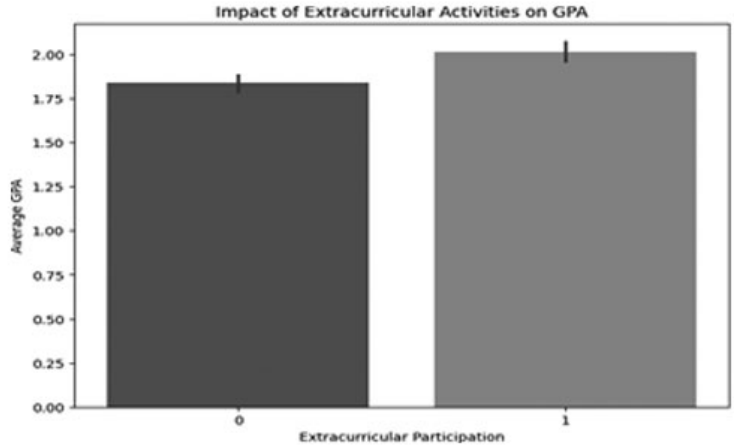

Fig. 4. Impact of Extra Curricular activites on students performance

Regression, 5 XGBoost, and 6 Random Forest. Results indicated that Random Forest (6) and XGBoost (5) have the highest accuracy and can be very effective with mixed data and complex patterns. On the other hand, Naive Bayes (1) and Logistic Regression (4) models were not very effective as they were not robust to nonlinear data. This demonstrates that Random Forest and XGBoost are strong predictive task models.

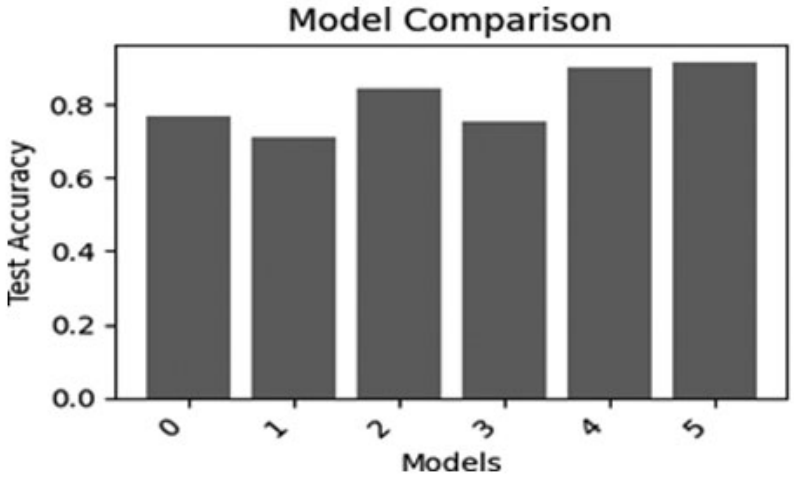

Fig. 5. Model Evaluation Bar Chart

6. Conclusion

The study on the impact of extracurricular activities on students' performance successfully demonstrated a positive correlation between student engagement in extracurricular activities and their academic outcomes. Using a heterogeneous dataset encompassing demographic, academic, and extracurricular attributes, the analysis revealed that students participating in such activities tended to achieve higher GPAs, attributed to the development of critical skills like time management, discipline, and teamwork. Correlation analysis highlighted other significant factors influencing academic performance, including parental support, persistence in studying, and regular attendance, while absenteeism emerged as a primary negative factor. Machine learning models, including Random Forest, XGBoost, and SVM, were employed to predict academic performance, with ensemble models like Random Forest and XGBoost delivering the highest accuracy by effectively capturing complex relationships in the data. This result would resonate with arguments for well-balanced education by incorporating academic knowledge and extra-academic activity to allow a holistic student's develop- ment. Despite this, however, there existed limitations that have been noticed and included: first, small-sized, therefore also biased sample and socio-economic aspects were left ignored. For improvement, in subsequent studies, large diverse data sets could be applied, together with other variables to strengthen the re- search conclusion. Future studies would focus on whether such changes are positive or negative among the students participating in extracurricular activities. On the positive side of the coin, increased selfconfidence, improvement of social skills, and

becoming better at emotive regulation would be achieved, thus boosting one's personal development. On the other hand, levels of stress, burnout, or times management where these matches create conflicts would characterize the negative approach to change. Understanding how these behavioral changes affect students' cognitive, emotional, and social growth will provide valuable in- sights into the comprehensive development of students. Ultimately, this research underscores the importance of extracurricular activities for academic excellence and personal growth, providing a foundation for educators and policymakers to design more inclusive and balanced educational frameworks. Thus, promoting education as a whole-which does not only mean a focus on extracurricular activities, but also one on academic rigorwill best prepare students to achieve success both academically and personally. Such balanced engagement in environments we can foster for our students will be sure to give them the necessary skills to succeed in such a diverse and dynamic world.

♣ References

1. Mushtaq Ahmad and Md Fashiur Rahman. Extracurricular activities and student's academic performance. *Journal of Armed Forces Medical College, Bangladesh*, 11(2):1–2, 2015.
2. Nasrulla Ahmed, Visama Hassan, and Khaulath Saeed. Effects of extracurricular activities on academic performance of secondary students in male. *International Journal of Scientific Research and Management (IJSRM)*, 12(06):3452–3464, 2024.
3. Hibah Qasem Alatawi and Shili Hechmi. A survey of data mining methods for early prediction of students' performance. In *2022 2nd International Conference on Computing and Information Technology (ICCIT)*, pages 171–174, 2022.
4. Shabiha Anjum. Impact of extracurricular activities on academic performance of students at secondary level. *International Journal of Applied Guidance and Counseling*, 2(2):7–14, 2021.
5. Santiago Gutierrez, Andrés Acero, Isaac Olivas, Jorge Alberto González-Mendívil, and Eduardo Caballero-Montes. The academic and emotional impact of extracur- ricular activities on college students. In *2024 IEEE World Engineering Education Conference (EDUNINE)*, pages 1–5. IEEE, 2024.
6. Farid Nassar, Ahmed Abbas, Hassan Al-Saify, and Omar Ali. The impact of ex- tracurricular activities on developing academic standing, student satisfaction, per- formance, and bolstering the academic reputation of higher education institutions from the perspective of sdgs. *Journal of Lifestyle and SDGs Review*, 5:e02758, 10 2024.
7. Tasnia Noshin. The role of extracurricular activities in child developmental out- comes. *The Child Health Interdisciplinary Literature and Discovery Journal*, 2(1), 2023.
8. Ugyen Penjor and Thinley Dorji. Impact of extra-curricular activities on students' academic performance at secondary schools in samdrup jongkhar. *Asian Journal of Education and Social Studies*, 32(3):8–19, Aug. 2022.

9. Shaikh Rezwan Rahman, Md Asfiul Islam, Pritidhrita Paul Akash, Masuma Parvin, Nazmun Nessa Moon, and Fernaz Narin Nur. Effects of co-curricular activities on student's academic performance by machine learning. *Current Research in Behavioral Sciences*, 2:100057, 2021.
10. Neeta Sharma, Mk Sharma, and Umang Garg. Predicting academic performance of students using machine learning models. In *2023 International Conference on Artificial Intelligence and Smart Communication (AISC)*, pages 1058–1063, 2023.
11. Neeta Sharma and Manoj Yadav. A comparative analysis of students' academic performance using prediction algorithms based on their time spent on extra-curricular activities. In *2022 Third International Conference on Intelligent Com- puting Instrumentation and Control Technologies (ICICICT)*, pages 745–750, 2022.
12. Wan Fairos Wan Yaacob, Syerina Azlin Md Nasir, Wan Faizah Wan Yaacob, and Norafefah Mohd Sobri. Supervised data mining approach for predicting student performance. *Indones. J. Electr. Eng. Comput. Sci*, 16(3):1584–1592, 2019.