OPEN ACCESS

European Journal of Teaching and Education
ISSN 2669-0667

Comparative Analysis of Machine Learning Models on Student Performance Data: Insights from Test Scores and Survey Data

Sanjana Sundararaman[1*], and Maheen Hasib[2]

[1] BSc (Hons) Statistical Data Science, Heriot Watt University, Dubai, UAE
[2] Department of Mathematics and Computer Science, Heriot Watt University, Dubai, UAE

ARTICLE INFO

Keywords:
AI in Education,
Educational Data Mining,
Machine Learning in Education,
Predictive Modeling,
Test Scores

ABSTRACT

With the increasing use of digital learning platforms, large volumes of student data have become available for analysis. This paper investigates how machine learning, learning analytics, and educational data mining can be utilized to gain insights into student performance. Various predictive modeling techniques, including Random Forest (RF), K-Nearest Neighbor (KNN), and Decision Trees (DT), are evaluated for their ability to forecast student test scores. Clustering algorithms like K-means are employed to identify patterns within the data. The study integrates these predictive models with survey data collected from undergraduate students at Heriot-Watt University Dubai, aiming to identify factors that influence academic outcomes. The research uses comparative analysis across different machine learning models which is applied to both the survey data and Kaggle test score data. The analysis reveals that linear regression is the most effective model for the Kaggle test score dataset, while K-means clustering provides the best insights from the survey data. The survey model is determined to be more comprehensive due to its inclusion of more predictors. Key metrics, such as accuracy scores, precision, recall, F1 score, and mean squared error, were calculated for both datasets to provide a quantitative overview, enabling a comparative evaluation of model performance and predictor effectiveness for both the datasets. The findings contribute to understanding how data-driven approaches can support educational decisions and interventions while addressing ethical considerations and inclusivity in educational settings.

1. Advancing Learning Outcomes through Machine Learning and Predictive Modeling

Artificial intelligence (AI) has expanded significantly in the field of computer sciences (CS) and education. To better illustrate AI's direct impact on educational outcomes, it's crucial to highlight specific advancements that have tangibly enhanced teaching and learning. The evolution of AI has enhanced teaching, learning, and administrative processes, improving their effectiveness and efficiency (Chen et al., 2020). These enhancements include adaptive learning systems that customize content to meet individual student needs and real-time feedback mechanisms that help teachers adjust instructional strategies. This, in turn, has propelled the

* Corresponding author's E-mail address: sanjanasundar2020@gmail.com

Cite this article as:
Sundararaman, S., & Hasib, M. (2025). Comparative Analysis of Machine Learning Models on Student Performance Data: Insights from Test Scores and Survey Data. *European Journal of Teaching and Education, 7*(1): 61-76. https://doi.org/10.33422/ejte.v7i1.1459

development of machine learning (ML). Recent advancements in ML models have accelerated AI's ability to process data and provide significant insights into learning experiences (Sghir et al., 2023).

Many strategies for educational analysis, such as learning analytics (LA) and educational data mining (EDM), are derived from ML models. Supervised learning techniques, like Random Forest (RF) and Linear Regression (LR), are used for trend prediction, while unsupervised techniques, such as clustering, identify hidden patterns in data (Umer et al., 2017).

Meanwhile, LA represents a methodical approach to gather and analyse vast datasets from online sources, aimed at enhancing learning processes. A relatively new field in education, LA is closely interlinked with academic analytics, learning analytics, and educational data mining (Zilvinskis et al., 2017). LA and EDM offer analytical tools that enable educators to observe and predict student performance, identify risks, and implement timely interventions, fostering student success (Alalawi et al., 2024).While LA seeks to use existing methods and models to address difficulties impacting student learning and organizational learning systems, EDM aims to forge new paths in computational data analysis (Peña-Ayala, 2014).

One of the primary applications of LA is the observation and forecasting of learner performance, as well as the identification of potential problematic situations and students at risk (Khine, 2019).

These analytical tools, rooted in ML and PM, offer a proactive approach to education by enabling educators to anticipate challenges, provide timely interventions, and ultimately foster an environment conducive to student success.

Moreover, predictive modeling (PM) provides actionable insights by analyzing historical trends, enabling proactive management of student outcomes (Sghir et al., 2023). While LA employs existing methods to enhance student learning, EDM focuses on forging new paths in computational analysis to uncover complex data characteristics.

Despite these advancements, several gaps remain in literature. For instance, previous studies often relied on outdated datasets or had a narrow focus, such as dropout prediction or blended learning environment (Baek & Doleck, 2023; de Oliveira et al., 2021). This study aims to address these limitations by leveraging both synthetic and real-world datasets to compare the effectiveness of various ML models in predicting student performance.

Notably, these literature reviews exhibited varied scopes, with some providing a broader examination of LA and EDM and others having a more specific focus on particular learning issues. Additionally, certain reviews discussed the field from an educational perspective without delving into the technical aspects (Sghir et al., 2023). Such gaps underscore the need for comprehensive studies that bridge the divide between educational and technical perspectives.

Our research seeks to address these limitations by advancing predictive modeling techniques in education, leveraging both synthetic and real-world datasets. Specifically, this study incorporates LA, EDM and PM to enhance the accuracy and applicability of models in diverse educational contexts. This paper focuses on comparing and contrasting the distinct tendencies of student test scores using the Kaggle SPSS data set and the survey data collected from undergraduates of Heriot Watt University Dubai. It offers insight into how learning practices, as well as other aspects like sleep, eating habits, study schedule, and so on, affect academic performance. Additionally, the study explores how high school learning habits influence university-level outcomes, providing a holistic view of academic performance predictors.

2.1 Research Aim

The aim of this research is to perform comparative analysis of ML models applied to student test score datasets from Kaggle and HWUD survey, with a focus on the most influential features (study hours, previous year scores, high school test scores). The study aims to evaluate the survey datasets to identify the impact of key predictors such as study habits for educational outcomes and to find patterns using educational analytics.

2.2 Objectives

The objectives of this research are as follows:

- Investigate and evaluate appropriate ML models for analyzing student test scores, emphasizing their impact on the learning process.
- Apply ML models to the survey data and Kaggle data to identify patterns and relationships comprehensively.
- Determine key characteristics for predicting student results by assessing feature importance within the selected ML models.
- Optimize ML models based on findings to improve accuracy and relevance in predicting student outcomes, fostering a deeper understanding of learning dynamics.

2. Literature Review

2.1 Background

AI has emerged as a transformational force in a variety of sectors, with a particularly major influence on education (Adıgüzel et al., 2023). In the realm of education, one of the most promising applications of AI is in the domain of LAS. The collecting, analysis, and interpretation of data from diverse educational sources to enhance learning outcomes, instructional practices, and student assistance. Educators and institutions may obtain deeper insights into the learning process by using the power of AI, allowing them to make data-driven decisions to improve the educational experience. While the concept of using AI in education (AIEd) has been explored for around 30 years, its practical implementation in educational settings has only gained traction in the past decade (Richter et al., 2019). This background sets the stage for investigating how specific AI tools and techniques directly contribute to improvements in student engagement and achievement, areas which our research specifically targets. The potential applications of ML in education have piqued the interest of researchers, educators, and policymakers.

ML is widely used in almost all fields in today's world. Due to its increase in popularity and varied applications it is used in education more wide from the early 2010's. It is also used in recommendation systems and personalized learning systems. Many studies have leveraged ML to predict student performance. In (Rahman & Abdullah, 2018), a fuzzy clustering technique combined with a decision tree, was employed to construct their system.

The primary learning category on which our analysis will be centre is supervised learning, which covers prediction and classification tasks. ML algorithms used are supervised learning and it is as follows: Naive Bayes, Random Forest, Linear Regression and K nearest neighbour, Support Vector Machine and Neural Networks (Umer et al., 2017). In addition, unsupervised learning techniques like K-means clustering have gained traction for grouping unlabeled data points into predefined clusters.

In supervised learning, input and output pattern pairs are the items associated with a certain notion. Unsupervised learning involves passively mapping or clustering data based on some order rules to comprehend or derive a succinct description of the data. Accordingly, the aim of unsupervised learning is to group or cluster similar things based on a similarity criterion and then deduce a concept that these objects have in common (Umer et al., 2017).

One of the key challenges of LA lies in its application within distinct learning environments while maintaining the flexibility to be implemented across various courses and institutions. As LA continues to gain prominence, it will progressively adopt advanced methodologies to support students, educators, administrators, and educational institutions in improving the learning experience (Chen & Cui, 2020).

In parallel, EDM has emerged as a field of study and practice that focuses on utilizing DM, ML, and statistical approaches in educational settings to analyze large datasets (Dutt et al., 2017).

As DM continues to grow in popularity, it is proving to be a powerful tool for enhancing the learning process and fostering mastery of knowledge. By recognizing patterns and extracting hidden insights, DM allows educators to make data-driven adjustments that can improve curricular development within the educational system (Chen et al., 2020). Together, LA and EDM represent complementary approaches, both working towards the shared goal of optimizing educational environments.

Figure 1 below gives a visual understanding on how EDM works in the educational system while also providing important findings and patterns for students and educators.

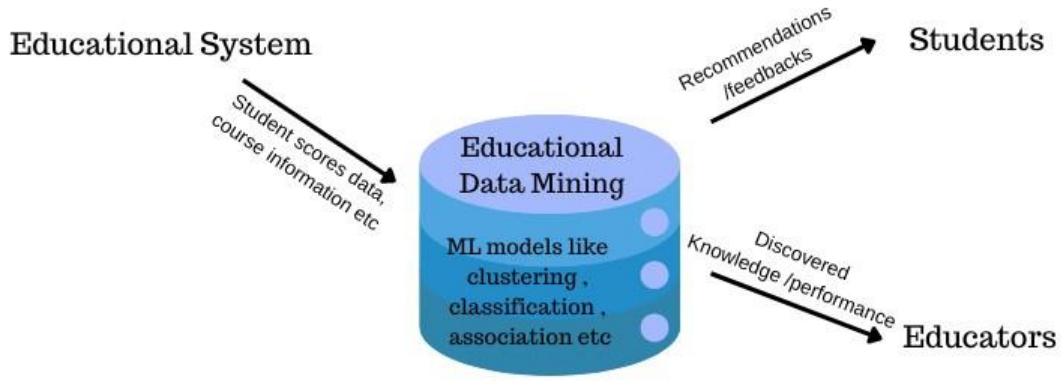

Figure 1. Applications of Educational Data Mining
Note. Created by author

This also illustrates the way EDM and ML are related and the manner in which they aid in predicting certain outcomes.

Table 1 illustrates a range of techniques used for predicting student performance through LA, EDM and Predictive Analytics (PA). This table serves as a vital reference, providing a nuanced understanding of the research methodologies employed in educational contexts. The table offers a comprehensive perspective, highlighting the intersections and distinctions among these analytical approaches. It emphasizes the diverse methodologies within LA, EDM, and PA, underscoring their collective significance in developing predictive models for education. By categorizing and comparing these techniques, researchers gain valuable insights into the strategies used to forecast student outcomes, guiding the application of learning algorithms to transform educational pathways.

Table 1. Describes the different papers along with the type of algorithm used in them based on LA, EDM and PA

Reference	LA	EDM	PA	Student Performance Prediction
(Alshamaila et al., 2024)		yes	yes	Prediction of student performance using deep learning
(Shou et al., 2024)	yes		yes	A student performance prediction model based on multidimensional time-series data analysis
(Costa et al., 2017)		yes		A comparative study of EDM techniques to predict those students who are likely to fail in a programming course
(Ahmad et al., 2015)		yes		EDM techniques are used to predict academic performance of first-year students in a computer science course
(Ulfa & Fatawi, 2021)	yes			Predicting Factors that Influence Students' Learning Outcomes Using Learning Analytics in Online Learning Environment

Note. Created by author.

The identified literature evaluations done so far had several significant limitations; for example, some articles evaluated LA and EDM in a larger framework, whilst others had a narrow emphasis on a specific learning difficulty. Other evaluations explored the field from an educational standpoint while ignoring the technical side (Sghir et al., 2023). These studies often lacked diversity in datasets, focusing predominantly on dropout prediction or historical data analysis. LA focuses on real-time data to enhance the learning experience in the moment, while EDM derives insights from historical educational data. PA on the other hand, employs statistical algorithms to forecast future student performance based on current data patterns.

While EDM remains a widely adopted approach, the integration of LA and PA presents a promising direction for more nuanced and robust analyses in educational research. Key to advancing this field is emphasizing the importance of data pre-processing, understanding the impact on educational policy and practice, and promoting inclusivity and diversity in research methodologies (Al-Gerafi et al., 2024).

After a thorough literature review, four models—Random Forest (RF), Logistic Regression (LR), K-Nearest Neighbors (KNN), and Decision Trees (DT)—were selected for comparative analysis in predicting student performance based on test scores. These models were chosen for their proven effectiveness in handling complex relationships and their speed during training. Their diverse methodologies allow a thorough exploration of predictive capabilities, providing depth to the comparative evaluation. In addition, the inclusion of unsupervised learning through K-means clustering provides complementary insights into patterns within educational data.

In conclusion, the combination of LA, EDM, and PA, alongside the strategic use of machine learning models like RF, LR, KNN, and DT, offers a multifaceted approach to improving student performance prediction. Additionally, by comparing the results from these models with information collected through surveys, we can better predict and understand a broader range of factors that impact student success. This approach provides a fuller picture by considering not just test scores from kaggle dataset, but also insights gathered from surveys, which reflect various personal and contextual influences on academic performance. This integrated approach strengthens current educational practices while paving the way for advancements in predictive methodologies.

3. Methodology

The dataset, sourced from Kaggle and IBM SPSS, focuses on student test scores. It contains 11 distinct features and contains approximately 2000 observations, offering a comprehensive view of student test scores and associated variables. Features include school setting, test scores and other related indicators. This diversity of features provides a robust foundation for analyzing various influences on student performance across different educational environments. RF and LR are utilized for modelling, with DT as a benchmark. Additionally, KNN is employed to contrast and explore patterns. These methodologies are chosen to cover a broad spectrum of statistical learning approaches, from ensemble methods to nearest neighbors, enhancing the robustness of our predictions. They help forecast test scores, offering valuable insights into survey data gathered at Heriot Watt University Dubai (HWUD) from undergraduate students (UG).

The survey data comprises 10 key features and 125 observations. Some of the key features are previous year scores, high school scores, and study hours, among others. By focusing on these variables, we aim to explore the direct and contextual factors affecting academic success, linking educational achievements to lifestyle and study habits. This dataset is analyzed to determine which features significantly influence student performance. This serves as a critical benchmark for comparing Kaggle models with real-world survey data, validating our model's effectiveness in authentic educational settings. Specifically, the study investigates the influence of study hours on test performance, providing comprehensive analysis and predictive capabilities. This particular focus allows us to quantify the impact of study habits on academic outcomes, a key consideration for educational planners and policymakers.

Additionally, EDA was performed to determine the relevant features for the model. Exploratory Data Analysis (EDA) is crucial as it ensures that our modeling efforts are guided by a clear understanding of data trends and relationships, preventing model overfitting and enhancing predictive accuracy. Both datasets underwent preprocessing to ensure accuracy and consistency, followed by comparative analyses using machine learning algorithms and evaluation metrics.

3.1 Data Collection

Data collection is a critical component in the research process, necessary for obtaining insights, making educated decisions, and drawing relevant conclusions (Figure 2). It involves gathering information from various sources to address research questions, test hypotheses, or evaluate trends. The validity and reliability of study conclusions are directly influenced by the quality and relevance of the data gathered.

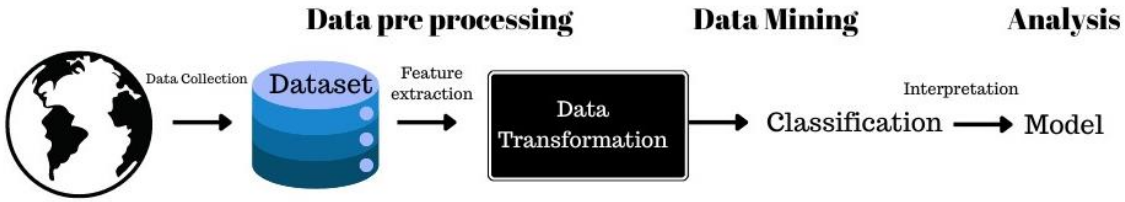

Figure 2. Method used for data collection, processing and analysis
Note. Created by author.

For the survey data, UG students were selected through random sampling techniques from diverse academic disciplines to ensure a balanced representation across different majors. This helped the sample achieve an accurate representation of the UG student's population.

Challenges such as scheduling conflicts and low availability were mitigated by making the survey available online for an extended duration, accommodating students' busy schedules.

Ethical considerations are paramount, especially when human subjects are involved. At HWUD ethical guidelines were strictly followed throughout the data collection process to ensure the well-being and rights of participants. Informed consent was obtained, and measures were implemented to safeguard privacy and confidentiality.

3.2 Data Pre - Processing

Ensuring the reliability of raw data is crucial to its utility, necessitating consistency and cleanliness. Real-world data often presents challenges such as inconsistency, incompleteness, or inaccuracies. To tackle these issues, various techniques are employed in data pre-processing, enhancing the predictive capabilities of models as in Figure 2. In the data pre-processing stage, several steps were taken to ensure the quality and relevance of the data for analysis.

Data Cleaning:

Checking for duplicates in the dataset to prevent skewing of the analysis results and to maintain data integrity. Kaggle dataset and survey data set showed no duplicate values.

Handling of Null/Missing Values:

Checking for null or missing values in the dataset using python. Kaggle dataset and survey data set showed no missing values.

Feature Engineering:

Features were evaluated based on their relevance to the target variable (e.g., test scores in Kaggle and high school scores in survey data). Both the Kaggle data and survey data only relevant features were included and some features were combined for further analysis. The survey data uses techniques such as one – hot encoding to transform the categorical variables into numerical format suitable for modelling and combining multiple variables.

3.3 Assumptions

The following are the assumptions made for the HWUD survey dataset:

Response Validity: It is assumed that the responses provided by the students are accurate and truthful to the best of their knowledge.

Representativeness: It is assumed that the sample of UG students surveyed is representative of the larger student population at HWUD.

Independence of Responses: It is assumed that the responses provided by one student are independent of the responses provided by other students.

3.4 Exploratory Data Analysis (EDA)

EDA constitutes a critical initial phase within any research endeavor. Its primary function entails the examination of data distributions, outliers, and anomalies to guide subsequent hypothesis testing (Sahoo et al., 2019). The data collected through survey was preprocessed and the Kaggle data set was already cleaned before EDA was performed. This process helps evaluate the quality of the data for building models. To further explore the relationships between the features correlation heat map is used.

3.4.1 Correlation Heat Map and Feature Selection

A commonly used statistical visualization, the heat map, illustrates shared patterns among subsets of rows and columns in matrix-like data (Gu, 2022). This visualization provides insights into the relationships between different features in the dataset (Zhang et al., 2014).

A value close to 1 or -1 indicates a strong correlation, while a value close to 0 indicates no correlation. All variable have a strong positive correlation with themselves, reflected in a value of 1 on the diagonal. Strong correlations may indicate potential predictors or influential variables for further analysis (Zhang et al., 2014). These correlations guide feature selection, ensuring the most relevant predictors are included in the ML models. This systematic approach to feature selection ensures that our models are not only grounded in empirical data but are also efficient in terms of computational resources and less prone to overfitting.

In the heat map (Figure 3, Figure 4) darker red colors indicate a strong positive correlation between variables, while darker blue color indicate a strong negative correlation. The lighter colors represent weaker correlations, either positive (peach) or negative (light blue).

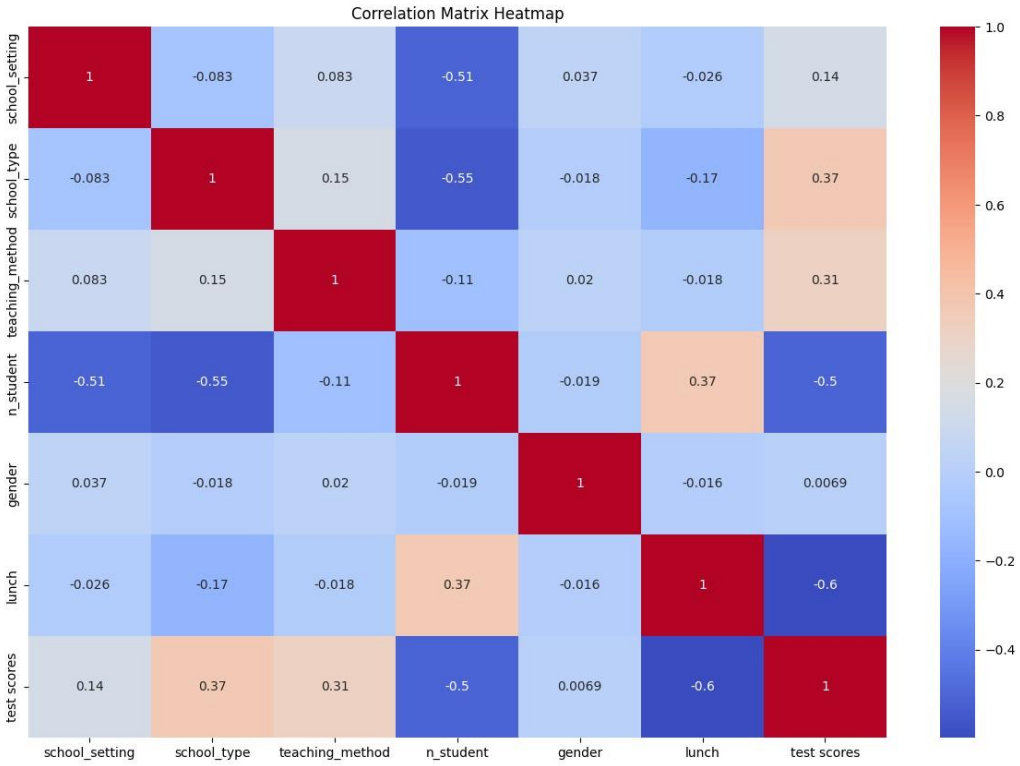

Figure 3. Correlation Matrix Heat Map for Kaggle dataset
Source: Results obtained from Python

Feature selection is a crucial step in ML tasks, particularly when dealing with high dimensional datasets. It aims to identify a subset of relevant features that contribute most significantly to the target variable, while discarding redundant or irrelevant ones. This process improves model performance by reducing overfitting, enhancing interpretability, and potentially reducing computational cost (Cherrington et al., 2019). Based on the correlation matrix heat map in Figure 3, the following features are selected with respect to the target variable (test scores):

The column “n students” (number of students) has a moderately strong negative correlation(-0.5). The column “school type” and “teaching method” have relatively low correlations, ranging from 0.31 to 0.37. While these features may still have some predictive power, their

impact on test scores appears to be moderate, suggesting that they may play a supportive rather than a primary role in the models.

This section focuses on exploring the relationships within the survey dataset and analysis of a correlation heat map.

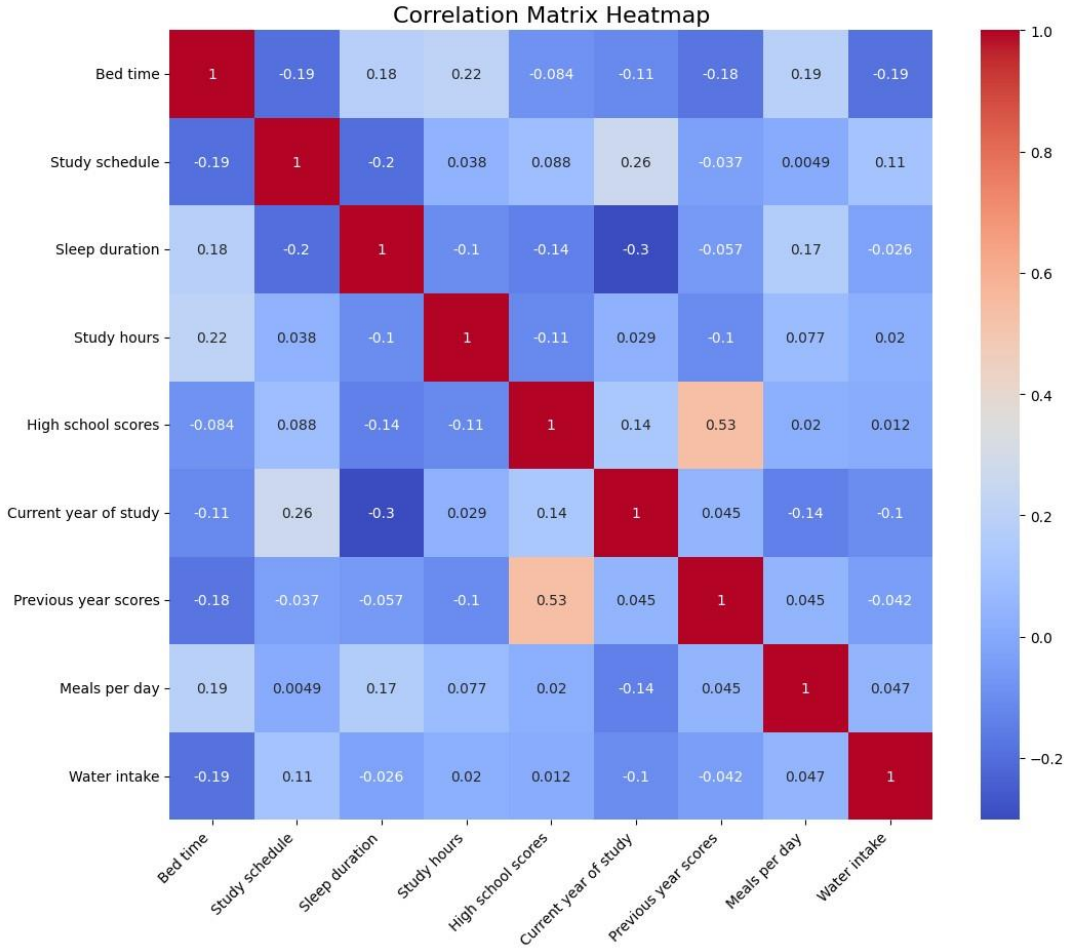

Figure 4. Correlation Matrix Heat Map for HWUD Survey Dataset
Source: Results obtained from Python.

Based on the correlation heat map in Figure 4, here are the features that have a significant correlation with high school scores:

- Previous year scores have a moderately strong positive correlation (0.53).
- Study hours have a relatively low positive correlation (0.14).
- Current year of study has a relatively low negative correlation (-0.11).

3.5 Datasets Used for Model Building

Kaggle Dataset:

Dataset 1: This includes all the features from the data set which are school type, teaching method, number of student, gender and lunch (except school setting).

Dataset 2: This includes only the relevant featured noticed in the correlation matrix from Figure 3. The features are school type, teaching method and number of student.

HWUD Survey Dataset:

Dataset 1: This includes all the features from the data set which Bed time, Study schedule, Sleep duration, Study hours, High school scores, Current year of study, Previous year scores, Meals per day and Water intake.

Dataset 2: This includes only the relevant featured noticed in the correlation matrix from Figure 4. The features are Study hours, Current year of study and Previous year scores.

Dataset 3: This includes only the relevant featured noticed in the correlation matrix from Figure 4. The features are Study hours and Previous year scores.

Both datasets were evaluated using ML models to identify the optimal model based on evaluation metrics. Post-EDA feature selection ensured that the most relevant features were incorporated, improving model performance and directly influencing the accuracy and utility of the predictive analytics applied.

4. Result and Discussion

4.1 Performance Evaluation

RF and KNN are widely used for regression and classification. Whereas LR is used for regression and DT is used for classification.

The Kaggle data consists of both numerical and categorical data, making it suitable to use different ML models to predict (regression tasks) and to find patterns in the data (classification tasks). The comparison between the different ML models provides a better understanding of the patterns in the data based on its performance evaluation like accuracy, f1 score, recall and precision. The above mentioned ML models were applied to the dataset using python programming language, satisfying objective 2 of the paper.

The supervised learning models were assessed and compared using a confusion matrix. A confusion matrix is a widely used table for evaluating the performance of classification models in supervised learning. It compares the predicted outcomes with the actual values from the test dataset, providing insights into model performance. Key metrics such as accuracy, precision, recall, and F1 score are derived from the confusion matrix (Wankhade et al., 2022).

The linear regression model was evaluated using mean squared error (MSE), mean absolute error (MAE), and the R^2 score. For the unsupervised learning models, such as K-means clustering, the silhouette score was used to measure performance (Ahmed et al., 2020).

The evaluation metrics for the Kaggle and HWUD survey datasets are presented in Table 2 and Table 3, respectively.

Table 2. Kaggle data set and its evaluation metrics

Kaggle Data set						
Data set /Models		KNN	Decision Tree	Random Forest		Linear Regression
Data with all features (except school setting)	Accuracy	0.07	0.083	0.105	Mean Absolute Error	2.559
	Precision	0.066	0.062	0.112	Mean Squared Error	10.484
	Recall	0.07	0.083	0.105	R^2 Score	0.946
	F1 score	0.062	0.063	0.083		
Data with relevant features	Accuracy	0.0703	0.086	0.081	Mean Absolute Error	2.684
	Precision	0.058	0.058	0.047	Mean Squared Error	11.682
	Recall	0.0703	0.086	0.081	R^2 Score	0.94
	F1 score	0.058	0.058	0.049		

Source: Results obtained from Python.

For the Kaggle dataset with all features:

- LR: MAE is 2.559, and the MSE is 10.484, indicating a relatively high error rate. The R^2 score of 0.946 suggests a good fit for linear regression. Other models show relatively lower accuracy.

For the Kaggle dataset with relevant features:

- The accuracy and recall, were slightly higher for DT compared to using all features.
- For KNN, the performance metrics were similar to using all features. For RF the performance was significantly lower compared to the previous case.
- LR: The MAE and MSE were slightly higher compared to using all features, but the overall fit was still good.

Overall, including all features in the Kaggle dataset seemed to provide better performance for most models, except for KNN, where relevant features performed slightly better. LR performed the best for this dataset.

4.1.1 HWUD Survey Dataset

The HWUD survey data consist of categorical data appropriate ML models were used after encoding them numerically. The models were evaluated using supervised learning and clustering algorithms to satisfy objective 1 of the paper.

Table 3. Survey data set and its evaluation metrics

HWUD Survey data								
Data set /Models		**KNN**	**Decision Tree**	**Random Forest**		**Linear Regression**		**K means**
Data with all features	**Accuracy**	0.6	0.44	0.52	**Mean Absolute Error**	0.548	**Silhouette Score**	0.165
	Precision (Micro)	0.6	0.44	0.52	**Mean Squared Error**	0.539		
	Precision (Macro)	0.217	0.153	0.206	**R^2 Score**	-0.326		
	Recall (Micro)	0.6	0.44	0.52				
	Recall (Macro)	0.3125	0.172	0.271				
	F1 score (Micro)	0.6	0.44	0.52				
	F1 score (Macro)	0.256	0.162	0.234				
Data with relevant features ('Study hours','Current year of study', 'Previous year scores')	**Accuracy**	0.6	0.6	0.6	**Mean Absolute Error**	0.55	**Silhouette Score**	0.466
	Precision (Micro)	0.6	0.6	0.6	**Mean Squared Error**	0.5		
	Precision (Macro)	0.163	0.163	0.163	**R^2 Score**	-0.223		
	Recall (Micro)	0.6	0.6	0.6				
	Recall (Macro)	0.234	0.234	0.234				
	F1 score (Micro)	0.6	0.6	0.6				
	F1 score (Macro)	0.192	0.192	0.192				
Data with study hours , previous year score wrt high score scores	**Accuracy**	0.56	0.6	0.6	**Mean Absolute Error**	0.548	**Silhouette Score**	0.92
	Precision (Micro)	0.56	0.6	0.6	**Mean Squared Error**	0.51		
	Precision (Macro)	0.234	0.163	0.163	**R^2 Score**	-0.254		
	Recall (Micro)	0.56	0.6	0.6				
	Recall (Macro)	0.239	0.234	0.234				
	F1 score (Micro)	0.56	0.6	0.6				
	F1 score (Macro)	0.231	0.192	0.192				

Source: Results obtained from Python.

For the survey dataset with all features:

- KNN: The accuracy was 0.6 and all the other evaluation parameters were the highest compared to other supervised learning models.
- LR: The MAE is 0.548, and the MSE is 0.539, indicating a moderate error rate. The R^2 score is -0.326, suggesting a poor fit for linear regression.
- K-means: The silhouette score is 0.165, indicating a moderate clustering quality.

For the survey dataset with relevant features:

- The accuracy for KNN, DT, and RF was slightly improved compared to using all features. The precision, recall, and F1-scores (micro) were generally higher for these models compared to using all features (except for KNN).
- LR: The MAE and MSE are slightly lower, but the R^2 score of -0.223 still suggests a poor fit.
- K-means: The silhouette score is 0.466, indicating a better clustering quality compared to using all features.

For the survey dataset with study hours, previous year scores, and high school scores:

- The accuracy for DT and RF was comparable to using relevant features. KNN performed worse compared to the previous cases.
- LR: The MAE and MSE were slightly lower, but the R^2 score of -0.254 still suggests a poor fit.
- K-means: The silhouette score is 0.92, indicating an excellent clustering quality.

Overall, for the HWUD survey dataset, using relevant features or a subset of highly relevant features (study hours, previous year scores, and high school scores) consistently improved the performance of most models, particularly DT, RF, and K-means clustering. However, LR showed a poor fit for this dataset, regardless of the features used.

5. Discussion

This part of the report provides a comprehensive comparison between the two data sets (i.e. Kaggle and HWUD survey data) and a brief on the way the objectives are satisfied for this paper.

Feature Selection and Relevance:

- For the Kaggle dataset, including all features generally led to better performance across most models, except for KNN. This highlights the collective contribution of features in synthetic datasets, even if individual correlations with the target variable are weak.
- In contrast, for the HWUD survey dataset, utilizing relevant features or a subset of highly relevant features (study hours, previous year scores, and high school scores) consistently improved the performance of models like KNN, DT, RF and K-means clustering. This highlights the importance of careful feature selection and the inclusion of features that directly impact the target variable. This also shows the importance/relevance of the survey data compared to the featured in the kaggle data. This supports Objective 3 by showcasing how specific features directly impact predictive accuracy.

Model Performance and Suitability:

- LR performed well on the Kaggle dataset but poorly on the HWUD dataset, emphasizing its limitations in handling non-linear relationships. Alternative models like KNN, DT, and RF were better suited for the complex nature of real-world data.
- Unsupervised techniques like K-means clustering further highlighted distinct patterns in survey data, validating their use for exploratory analysis.

Clustering Analysis:

- K-means clustering revealed the strongest clusters when only highly relevant features were used, achieving a silhouette score of 0.92. This result underscores the value of dimensionality reduction in clustering tasks.

Real-world vs. Synthetic Data:

- The HWUD survey dataset, with its real-world complexity, presented greater challenges but yielded richer insights compared to the Kaggle dataset. This reinforces the necessity of diverse and realistic data for building robust predictive models.

6. Conclusion

The primary aim of this paper was to perform a comparative analysis of ML models on student test scores using the Kaggle dataset and survey data collected from UG students at HWUD. The study successfully highlighted key differences between synthetic and real-world datasets, showcasing the importance of relevant features in model performance.

Key Findings:

- The HWUD survey dataset proved superior for predicting student outcomes due to its diverse features (e.g., study habits and previous scores), despite its complexity.
- K-means clustering emerged as a valuable unsupervised technique for identifying patterns in real-world data.
- LR performed well on the Kaggle dataset but was unsuitable for the HWUD dataset,

emphasizing the need for alternative, non-linear models in real-world scenarios.

Implications:

- Educational institutions can use these findings to refine admission criteria, focusing on factors like study habits and socioeconomic background.
- Policymakers may leverage these insights to design targeted interventions for at-risk students.
- Real-time data from digital learning platforms could further enhance predictive models, enabling timely interventions to improve academic outcomes.

These recommendations underscore the practical applications of our findings and suggest a framework for their implementation in educational settings.

Ultimately, this research underscores the value of data quality and feature relevance in educational predictive modeling, paving the way for more inclusive, effective, and equitable learning systems. Additionally, stakeholders in education could prioritize the implementation of personalized learning approaches tailored to individual student needs, supported by predictive models developed in this study. This research also highlights the potential of integrating real-time data from digital learning platforms to capture dynamic learning behaviors and improve predictive accuracy. Such real-time insights could allow educators to design timely interventions, enhancing student engagement and outcomes.

Furthermore, this research underscores the importance of data quality and relevance in predictive modeling for educational purposes. The superiority of the HWUD survey data in predicting student test scores emphasizes the value of incorporating diverse and impactful features. A detailed understanding of student learning patterns and influencing factors enables educators to create personalized learning experiences and foster more inclusive classrooms.

Ultimately, these efforts contribute to improved academic outcomes, student well-being, and a more equitable, inclusive, and effective educational system. Further studies could also investigate the long-term impacts of these interventions on educational equity and access, ensuring that advancements in educational technology contribute to broader societal goals.

References

Adıgüzel, T., Kaya, M. H., & Cansu, F. K. (2023). Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. *Contemporary Educational Technology*. https://doi.org/10.30935/cedtech/13152

Ahmad, F., Ismail, N. H., & Aziz, A. A. (2015). The prediction of students' academic performance using classification data mining techniques. *Applied mathematical sciences*, *9*(129), 6415-6426. https://doi.org/10.12988/ams.2015.53289

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, *9*(8), 1295. https://doi.org/10.3390/electronics9081295

Al-Gerafi, M. A. M., Goswami, S. S., Sahoo, S. K., Kumar, R., Simic, V., Bacanin, N., Naveed, Q. N., & Lasisi, A. (2024). Promoting inclusivity in education amid the post-COVID-19 challenges: An interval-valued fuzzy model for pedagogy method selection. *The International Journal of Management Education*, *22*(3), 101018. https://doi.org/10.1016/j.ijme.2024.101018

Alalawi, K., Athauda, R., Chiong, R., & Renner, I. (2024). Evaluating the student performance prediction and action framework through a learning analytics intervention study. *Education and Information Technologies*, 1-30. https://doi.org/10.1007/s10639-024-12923-5

Alshamaila, Y., Alsawalqah, H., Aljarah, I., Habib, M., Faris, H., Alshraideh, M., & Salih, B. A. (2024). An automatic prediction of students' performance to support the university education system: a deep learning approach. *Multimedia Tools and Applications*, *83*(15), 46369-46396. https://doi.org/10.1007/s11042-024-18262-4

Baek, C., & Doleck, T. (2023). Educational data mining versus learning analytics: A review of publications from 2015 to 2019. *Interactive Learning Environments*, *31*(6), 3828-3850. https://doi.org/10.1080/10494820.2021.1943689

Chen, F., & Cui, Y. (2020). Utilizing Student Time Series Behaviour in Learning Management Systems for Early Prediction of Course Performance. *Journal of Learning Analytics*, *7*(2), 1-17. https://doi.org/10.18608/jla.2020.72.1

Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *Ieee Access*, *8*, 75264-75278. https://doi.org/10.1109/access.2020.2988510

Cherrington, M., Thabtah, F., Lu, J., & Xu, Q. (2019). Feature selection: filter methods performance challenges. https://doi.org/10.1109/iccisci.2019.8716478

Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in human behavior*, *73*, 247-256. https://doi.org/10.1016/j.chb.2017.01.047

de Oliveira, C. F., Sobral, S. R., Ferreira, M. J., & Moreira, F. (2021). How does learning analytics contribute to prevent students' dropout in higher education: a systematic literature review. *Big Data and Cognitive Computing*, *5*(4), 64. https://doi.org/10.3390/bdcc5040064

Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. *Ieee Access*, *5*, 15991-16005. https://doi.org/10.1109/access.2017.2654247

Gu, Z. (2022). Complex heatmap visualization. *Imeta*, *1*(3), e43. https://doi.org/10.1002/imt2.43

Khine, M. S. (2019). Emerging Trends in Learning Analytics: Leveraging the Power of Education Data. https://doi.org/10.1163/9789004399273

Peña-Ayala, A. (2014). Educational data mining. *Studies in Computational Intelligence*, *524*. https://doi.org/10.1007/978-3-319-02738-8

Rahman, M. M., & Abdullah, N. A. (2018). A personalized group-based recommendation approach for Web search in E-learning. *Ieee Access*, *6*, 34166-34178. https://doi.org/10.1109/access.2018.2850376

Richter, O. Z., Juarros, V. I. M., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education: where are the educators? *International Journal of Educational Technology in Higher Education*(16), 6. https://doi.org/10.1186/s41239-019-0171-0

Sahoo, K., Samal, A. K., Pramanik, J., & Pani, S. K. (2019). Exploratory data analysis using Python. *International Journal of Innovative Technology and Exploring Engineering*, *8*(12), 4727-4735. https://doi.org/10.35940/ijitee.l3591.1081219

Sghir, N., Adadi, A., & Lahmer, M. (2023). Recent advances in Predictive Learning Analytics: A decade systematic review (2012–2022). *Education and Information Technologies*, *28*(7), 8299-8333. https://doi.org/10.1007/s10639-022-11536-0

Shou, Z., Xie, M., Mo, J., & Zhang, H. (2024). Predicting Student Performance in Online Learning: A Multidimensional Time-Series Data Analysis Approach. *Applied Sciences, 14*(6), 2522. https://doi.org/10.3390/app14062522

Ulfa, S., & Fatawi, I. (2021). Predicting factors that influence students' learning outcomes using learning analytics in online learning environment. *International Journal of Emerging Technologies in Learning (iJET), 16*(1), 4-17. https://doi.org/10.3991/ijet.v16i01.16325

Umer, R., Susnjak, T., Mathrani, A., & Suriadi, S. (2017). On predicting academic performance with process mining in learning analytics. *Journal of Research in Innovative Teaching & Learning, 10*(2), 160-176. https://doi.org/10.1108/jrit-09-2017-0022

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review, 55*(7), 5731-5780. https://doi.org/10.1007/s10462-022-10144-1

Zhang, Z., McDonnell, K. T., Zadok, E., & Mueller, K. (2014). Visual correlation analysis of numerical and categorical data on the correlation map. *IEEE transactions on visualization and computer graphics, 21*(2), 289-303. https://doi.org/10.1109/tvcg.2014.2350494

Zilvinskis, J., Willis Iii, J., & Borden, V. M. H. (2017). An overview of learning analytics. *New Directions for Higher Education, 2017*(179), 9-17. https://doi.org/10.1002/he.20239