

**Development of Predictive Models for Students Learning Outcomes
based on Spatial Learning Analytics**



By

Hajrah Amir

(2021-NUST-MS-GIS-363752)

**A thesis submitted in partial fulfillment of the requirements for the
degree of Master of Science in Remote Sensing and GIS**

**Institute of Geographical Information Systems
School of Civil and Environmental Engineering
National University of Sciences & Technology
Islamabad, Pakistan**

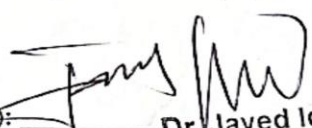
May 2024

THESIS ACCEPTANCE CERTIFICATE

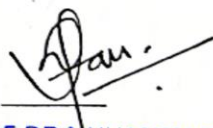
Certified that final copy of MS/MPhil thesis written by **Hajrah Amir (Registration No. MSRSGIS 00000363752), of Session 2021 (Institute of Geographical Information Systems)** has been vetted by undersigned, found complete in all respects as per NUST Statutes/Regulation, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.



Signature: _____
Name of Supervisor: Dr Ali Tahir
Date: 20-5-2024



Signature (HOD): _____
Date: 20-5-2024
Dr. Javed Iqbal
Professor & HOD IGIS, SCEE (NUST)
H-12, Islamabad



Signature (Principal & Dean SCEE): _____
Date: 21 MAY 2024
PROF DR MUHAMMAD IRFAN
Principal & Dean
SCEE, NUST

ACADEMIC THESIS: DECLARATION OF AUTHORSHIP

I, **Hajrah Amir**, declare that this thesis and the work presented in it are my own and have been generated by me as the result of my own original research.
Development of Predictive Models for Students Learning Outcomes based on Spatial Learning Analytics.

I confirm that:

1. This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text;
2. Wherever any part of this thesis has previously been submitted for a degree or any other qualification at this or any other institution, it has been clearly stated;
3. I have acknowledged all main sources of help;
4. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
5. None of this work has been published before submission.
6. This work is not plagiarized under the HEC plagiarism policy.

Signed: *Hajrah Amir*

Date: 22nd may, 2024

DEDICATION

Dedicated to my exceptional parents and adored siblings whose tremendous support and cooperation led me to this wonderful accomplishment.

ACKNOWLEDGEMENTS

I want to convey deep appreciation to my supervisor “Dr. Ali Tahir” for his unwavering support, expert guidance, insightful suggestions, and constant motivation during the entire journey of crafting this thesis. Additionally, I'd like to express my heartfelt thanks to the members of my GEC (Graduate Examination Committee) for their valuable assistance and significant contributions to this work.

I'd also like to acknowledge the Constructor research team (based in Singapore), whose continuous support played a crucial role throughout my research. Furthermore, my sincere gratitude goes to Assistant Director (IT) “Mr. Kamran Mir” and the Vice-Chancellor of Allama Iqbal Open University, “Dr. Nasir Mehmood”, for their invaluable assistance in facilitating the collection of data from their institution.

In addition, I want to extend my thanks to my colleagues and all the individuals who supported and encouraged me during the development of this thesis, even if I may have inadvertently omitted some names. Your collective support has been truly invaluable.

Hajrah Amir

TABLE OF CONTENTS

THESIS ACCEPTANCE CERTIFICATE.....	i
ACADEMIC THESIS: DECLARATION OF AUTHORSHIP	ii
DEDICATION.....	III
ACKNOWLEDGEMENTS.....	IV
TABLE OF CONTENTS.....	V
LIST OF FIGURES	VII
LIST OF TABLES	VIII
ABSTRACT	IX
INTRODUCTION.....	1
1.1 Background.....	1
1.2 Objectives.....	4
1.3 Literature review	5
MATERIALS AND METHODS.....	9
2.1 Study area	9
2.2 Data collection & preprocessing.....	10
2.3 Exploratory data analysis.....	12
2.3.1 Mapping student distribution using geographic data visualization.....	12
2.3.2 Correlation matrix	15
2.3.3 Spatial autocorrelation analysis	16
2.4 Feature engineering & selection	18
2.5 Predictive model development	19
2.5.1 Machine learning basic.....	19
2.5.2 Predictive analytics	20
2.6 Selected methods	21
2.6.1 Predictive model evaluation.....	21
RESULTS AND DISCUSSIONS	25

3.1 Predictive model comparison.....	25
3.1.1 Performance analysis of different models	25
3.2 Selection of the best performing model.....	32
3.3 Features importance graphs.....	33
CONCLUSIONS	36
4.1 Conclusions.....	36
LIMITATIONS AND FUTURE RESEARCH DIRECTIONS.....	37
5.1 Limitations.....	37
5.2 Future research directions	38
REFERENCES	39

LIST OF FIGURES

Figure 1: Students diversity regions of Allama Iqbal Open University (AIOU) Islamabad	9
Figure 2: Teams device usage detail of students.	11
Figure 3: User activity details of students.	11
Figure 4: Geographic distribution and density of students in various regions.	13
Figure 5: Geographic locations of students residing in highlighted areas.	13
Figure 6: Web map with hierarchical clustering for student information access.	14
Figure 7: Correlation matrix showing association between variables.	15
Figure 8: Moran's scatterplot with trend line.	17
Figure 9. Random Forest features importance graph.	34
Figure 10. Decision Tree features importance graph.	34
Figure 11. SVM features importance graph.	34
Figure 12. KNN features importance graph.	34
Figure14. Logistic Regression features importance graph.	35
Figure 13. Light GBM features importance graph.	35
Figure 16. Neural Network features importance graph.	35

LIST OF TABLES

Table 1: Spatial autocorrelation stats.	17
Table 2: Brief description of ML models.	22
Table 3: Example data about students who passed or failed in Quiz.	24
Table 4: Decision tree evaluation metrics.	30
Table 5: KNN evaluation metrics.	30
Table 6: Light GBM evaluation metrics.	30
Table 7: Logistic Regression evaluation metrics.	31
Table 8: Neural Network evaluation metrics.	31
Table 9: Random Forest evaluation metrics.	31
Table 10: SVM evaluation metrics.	31
Table 11: Evaluation metrics comparison of predictive models.	33

ABSTRACT

Learning analytics is a data-driven methodology that provides instructors with important insights regarding student interactions with course materials, allowing them to make informed decisions about content delivery and structure. We investigated the use of spatial learning analytics in the context of online and remote education, with a focus on fulfilling the United Nations Sustainable Development Goal (SDG) 4 - guaranteeing inclusive and high-quality education for everyone. By combining data from Moodle logs and device usage and user activity metrics from Microsoft Teams, this study aimed to predict students quiz performance and enhance personalized learning interventions. The primary objectives were to create accurate predictive models for student learning outcomes and evaluate their effectiveness. Additionally, mapping tools were utilized such as Leaflet and ArcGIS to craft interactive maps, enriching the data-driven learning journey. Study employed supervised machine learning techniques, including Random Forest, Decision Trees, Support Vector Machine, LightGBM, K-Nearest Neighbors, Logistic Regression, and Neural Networks. These predictive models were trained and tested on the preprocessed dataset to predict quiz scores using binary classification method technique with threshold of 10. The predictive models' performance was evaluated using metrics such as precision, recall, F1 score, training time, and prediction time. Results indicate that the Support Vector Machine (SVM) model achieved the highest recall (100%) and F1 score (85.19%). Logistic Regression and Neural Networks also performed well, with Logistic Regression showing a recall of 95.57% and an F1 score of 83.84%, and Neural Networks exhibiting a recall of 94.38% and an F1 score of 83.56%. This study contributes to learning analytics by demonstrating the potential of spatial data in predicting and improving student outcomes, aligning with United Nations Sustainable Development Goal 4 for accessible and equitable education. Future research can refine these models by incorporating additional data sources and advanced machine learning techniques. Enhancing geographic insights and addressing ethical considerations in data usage will be crucial.

INTRODUCTION

1.1 Background

In recent years, learning analytics has emerged as a powerful field that utilizes data and analysis to gain insights into students' interactions with educational materials. As a result of ICT technology improvement, the discipline of learning analytics, also known as data analytics in education or educational data mining, is quickly gaining footing in education management, government, and industry. The desire for technology developments in support of learning delivery is driven by the constant demand for knowledge and knowledge management (Romero & Ventura, 2007). Learning analytics is defined as “the measurement, collection, analysis, and reporting of data about learners and their contexts, for the purpose of understanding and optimizing learning and the environments in which it occurs” (Long & Siemens, 2011). In short Learning Analytics is understanding, analysing, and converting the educational data into useful actions.

Learning analytics has potential for predicting and enhancing student success and retention, in part because it enables teachers, institutions, and students to make data-driven decisions on student success and retention (Olmos & Corrin, 2012; Smith et al., 2012). Along with other benefits, learning analytics holds the promise of more "personalised learning" that would, among other things, help students learn more effectively (Greller & Drachsler, 2012). This personalised learning experience is crucial in overcoming the "efficient learning hypothesis," as Siemens (2010) refers to the belief and practice of many course designers that learners begin the course at the same stage and progress through it at the same rate. Most

LMSs automatically collect data, which can be used by teachers to influence how students go through a course. For instance, (Smith et al., 2012) discovered that students' participation in the material, frequency with which they logged into their LMS, pace, and assignment grades all effectively predicted how well they would do in the course.

One specific area of interest within learning analytics is spatial learning analytics, which explores the role of spatial data in online and distance education settings. Current approaches to learning analytics provide insights on assessment performance and student interactions, but do not clarify the roles that geography may play. Becker (2013) identifies location as one of three main data types required to support learning analytics: timing, location, and population. The use of location data in learning systems is also noted in talks about the spectrum of ethical and privacy issues raised by advances in learning analytics (Pardo & Siemens, 2014). This implies that using such data may violate students' privacy rights, needing careful consideration of problems such as informed consent, data security, and the appropriate handling of sensitive information. These ethical and privacy concerns are crucial to arguments about the future of learning analytics. Recognizing geography's impact on learner engagement enables educators to design courses that consider cultural nuances, time zone differences, and language variety. This understanding allows for the creation of inclusive, high-quality learning experiences that cater to students' diverse geographical backgrounds, enabling increased engagement and higher learning results in a global setting. The fundamental goal of this study is to develop prediction models that can anticipate students' quiz score performance based on their involvement with the course learning material. The method used intends to provide educators with a tool for early detection and support of students who may require further assistance by merging varied datasets such as

Moodle logs data and Microsoft teams' data. Predicting a student's performance based on previous academic data is one of the most prominent applications of educational data mining, and as such, it is a useful source of information that can be utilised to improve students' performance (Buenaño-Fernández et al., 2019). Predicting student performance assists educational institutions in improving learning and teaching approaches by identifying instructional methods that suit students based on a variety of background information (Belachew & Gobena, 2017).

Exams, assignments, quizzes are often used as course assessments to check students' understanding as well as progress. Analysing student performance is a difficult undertaking due to the large amount of educational data that must be examined Pojon (2017). Predicting students' performance accurately based on their ongoing academic records is critical for carrying out appropriate educational interventions to ensure students' on-time and satisfactory course completion (Belachew & Gobena, 2017). To achieve these objectives, a large amount of student data must be examined and predicted using multiple machine learning models. Thus, the predictive models have the potential to significantly improve learning outcomes in online education contexts. By leveraging geographical insights from latitude and longitude data, spatial learning analytics provides inclusive and high-quality education for all. This strategy improves course design, develops student global connectivity, and accommodates time zone variances, ensuring that education is accessible and equitable for learners globally. Furthermore, by promoting cross-cultural awareness and tailoring support based on geographic data, we align our efforts with UN Sustainable Development Goal 4, which calls for the creation of an online learning environment that embodies excellence, inclusivity, and accessibility for all students, regardless of geographic location.

This research focuses on supervised learning, specifically predictive analytics (Nyce & Cpcu, 2007), using machine learning to anticipate future outcomes. We seamlessly integrate EDM and LA techniques in this study. EDM includes several machine learning approaches, whereas LA is concerned with optimizing learning environments. This collaboration allows us to use data to improve online education while adhering to the study's aims and methods. In the field of EDM, a variety of machine learning methods, such as kNN, Random Forests, Decision Tree Classifiers and others have been utilised with varying degrees of success (Romero & Ventura, 2010). These algorithms will be trained on training set, with the remaining data utilised to assess the models' performance. There are widely used indicators for evaluating the effectiveness of algorithms, such as precision, recall and F-measure Powers (2011). Precision, recall, and F-measure are popular metrics for assessing the effectiveness of machine learning models Powers (2011). Algorithms are compared in terms of indicator values to see which algorithm produces the best outcomes. This classification is based on data gathered from Moodle Learning Management System (LMS) and MS Teams. Interactive map provides educators with insights into spatial patterns, assisting comprehension of the link between spatial interactions and academic success, and improving online and distance learning outcomes. Prediction models and data analysis in learning analytics enable proactive personalized support for students, enhancing achievement, retention, and global education accessibility, contributing to an inclusive and successful academic environment.

1.2 Objectives

The objectives of research are:

1. To develop accurate predictive models for student learning outcomes by combining Moodle Logs data and Microsoft Teams data.
2. To evaluate the effectiveness of the predictive models.

1.3 Literature review

In education, student retention is a critical issue. While intervention programs can increase retention rates, they require prior knowledge of student performance (Yadav et al., 2012). This is when performance prediction comes into play. Various data mining approaches and strategies for prediction have been implemented in research instances. This study's research falls under the category of EDM. This is done to better understand how students learn, to research educational issues, and to improve the effectiveness of teaching and learning activities (Papamitsiou & Economides, 2014). This is accomplished by converting raw data into information that has a direct impact on educational practice and research (Romero & Ventura, 2010).

The quantity of research publications devoted to EDM in its many forms has increased dramatically in recent years (Peña-Ayala, 2014). This has been connected to an increase in the availability of educational data as well as the broad availability of low-cost computing power and easily accessible digital technologies (Johnson & Samora, 2016). Quinn and Gray (2019) used Moodle data analysis to predict student academic progress, allowing them to intervene proactively with at-risk students. Dondorf (2022) investigates the secure deployment of Moodle's learning analytics in higher education, focusing on data protection, ethics, informed permission, and privacy measures. With so much high-quality data available and the potential for important educational insights, educational institutions, governments, and researchers are increasingly seeking for ways to put these techniques to use.

Machine learning predicts student achievement by analyzing demographics, grades, and involvement for timely assistance (Xu et al., 2017). Study uses machine learning in MOOCs for accurate student achievement prediction using demographics, engagement, and course data. (Al-Shabandar et al., 2017). (Robinson et al., 2020) study geovisualization's impact on spatial learning, improving pattern identification and comprehension. Spatial learning analytics have been undertaken using students' demographic data to measure their residence locations and performance, using visualizations using maps and spatial autocorrelation. In addition to spatial autocorrelation, correlation analysis has been performed to examine the relationships between features and the target variable.

The primary goal of this research is to investigate the applicability and usefulness of specific machine learning algorithms for determining whether a student needs academic help in a specific course based on an analysis of their quiz scores. (Agrawal et al., 2017) used data mining classifiers to predict undergraduate academic success and identify factors for targeted interventions. A comparison of machine learning algorithms was performed to predict success or failure in an Intelligent Tutoring Systems course (Hämäläinen & Vinni, 2006). Other comparisons of different data mining techniques are done to predict students' final marks based on Moodle usage data (Romero et al., 2008), predict student final grade based on features derived from logged data (Minaei et al., 2003), and predict university students' academic achievement (Ibrahim & Rusli, 2007). The goal of these studies' research differs. The goal of several of them is to discover the best prediction system. In others, the goal is merely to see if machine learning is a realistic method for predicting student performance. We employed different classifiers to perform predictions on the data to have a baseline to compare the performance of the machine learning techniques. These algorithms have all

received widespread application in EDM in recent years (Romero & Ventura, 2010). Anderson and Anderson (2017) determined SVM's superiority in predicting student grades, surpassing a basic average strategy, through an experiment on 683 students at California State University's Craig School of Business. SVM achieved 86.3% accuracy on a sample of 395 students (Cortez & Silva, 2008) where five labels had to be predicted, and 86.26% accuracy on a sample of 15150 students (Jayaprakash et al., 2014). With an accuracy of 68.2% (Stapel et al., 2016), logistic regression is commonly utilized in different data mining areas (Witten et al., 2016) and is thus included in our analysis. (Jayaprakash et al., 2014) predicted a pass or fail with the Decision Tree and achieved a prediction accuracy of 85.92% on a sample of 15150 students. The Random Forest classifier, which is an ensemble technique that uses a set of Decision Trees to make a prediction, is a variation on the Decision Tree classifier.

In the same way we have used Light GBM and neural network algorithm. So far, studies employing Neural Networks have only used modest sample sizes, with the largest encountered being 649 students in research by (Cortez et al., 2008). Although the performance revealed in these Neural Network experiments is excellent, the portability of their findings is difficult to assess. Brooks and Thompson (2022) emphasize predictive modelling in education for timely interventions and improved teaching strategies. Therefore, we have employed these algorithms in our dataset to predict students' performance. (Bujang et al., 2021) develops precise cross-class student grade prediction model using algorithm evaluation, highlighting educational data mining potential. We have used various evaluation metrics to evaluate and compare the performance of models.

In most of these studies, researchers follow a similar process. They use different algorithms to create models that make predictions. However, what's missing in these studies is a more in-depth comparison of different methods. In our research, we take a closer look at the data through exploratory analysis. We go beyond the standard metrics and consider things like how long it takes to train and make predictions with the models. Additionally, we investigate the importance of each feature in the models, which helps us understand which factors are most critical for predicting the outcome. This is the part where our research introduces a new approach. By comparing the effectiveness of different processes used in machine learning, this research can provide insight into more efficient methods for enhancing predictions of student performance.

The backdrop for the study is established in Chapter 1, which provides an overview of past work in the topic. The contents and methodology are covered in Chapter 2, which covers data preprocessing, exploratory data analysis, feature engineering, and predictive model creation utilizing various machine learning approaches. The results and debates are highlighted in Chapter 3, which includes a comparison of predictive models and the selection of the best performing one. Finally, Chapter 4 contains insightful comments and conclusions that summarize the findings and their consequences. Chapter 5 goes into future work and offers insights into potential breakthroughs and areas of investigation that can expand on existing research.

MATERIALS AND METHODS

2.1 Study area

Allama Iqbal Open University (AIOU) in Islamabad is the largest university in Asia dedicated to distance education. It offers a comprehensive range of academic programs from Matriculation to PhD, catering to students who are unable to attend traditional on-campus classes. Currently, AIOU serves approximately 1 million students. The university's diverse student body spans various regions, with significant representation from urban areas as well as rural communities, reflecting the institution's wide-reaching impact and commitment to accessible education. For a visual representation, Figure 1 map highlighting regions with the greatest diversity of AIOU students, emphasizing areas such as Islamabad, Punjab, Sindh, Khyber Pakhtunkhwa, and Balochistan, where the student population is particularly diverse and dense.

<https://www.arcgis.com/apps/instant/basic/index.html?appid=c14242ed830643d892e91b8c87ff8c4f>

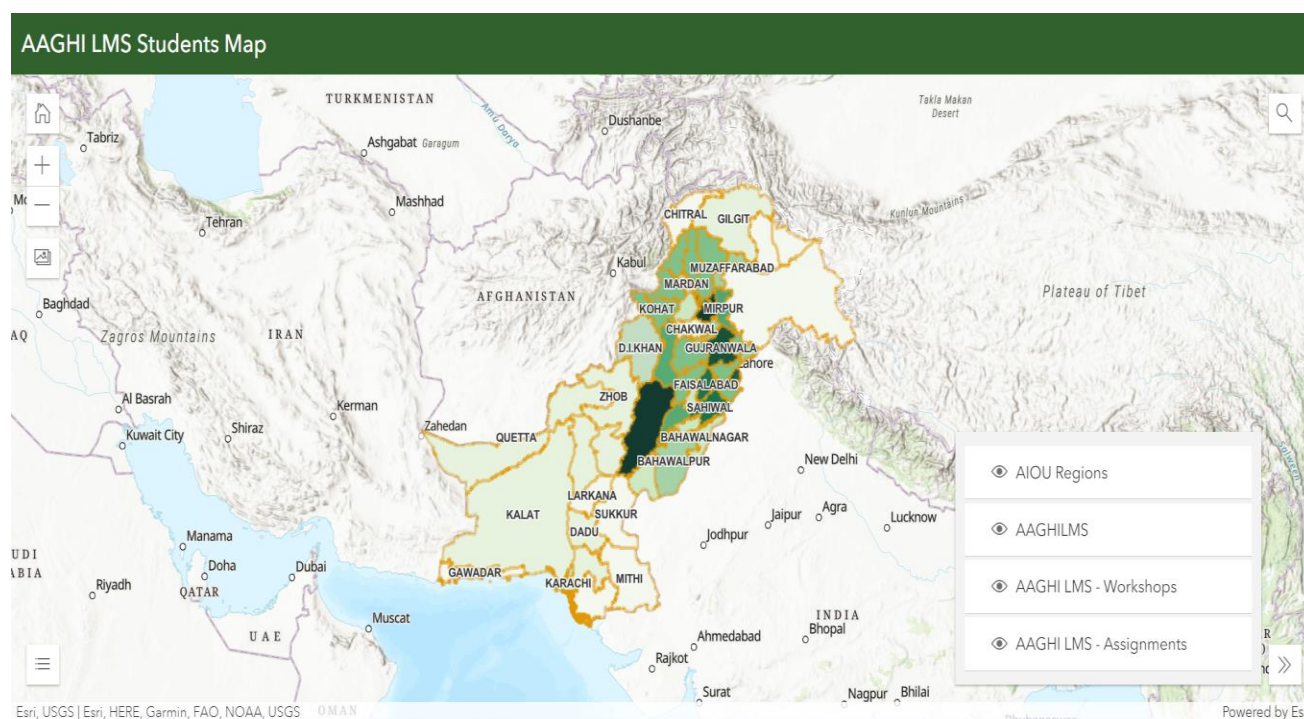


Figure 1: Students diversity regions of Allama Iqbal Open University (AIOU) Islamabad

2.2 Data collection & preprocessing

In this study data collected from MS Teams involves two main data which includes MS Team's device usage dataset and MS Team's user activity dataset. Figure 2 shows MS Teams device usage dataset includes information about how users' access MS Teams, categorizing their usage by device type.

The MS Team's device usage detail dataset contains 100,956 records. This shows most of the students used android phone, few of them accessed the course material by using windows, web, iOS, and none of them used mac, chrome OS and Linux.

MS Teams' user activity dataset provides specific metrics for each user's activity in MS Teams. It consists of 260,606 user-specific records. Figure 3 displays user activity details, including meetings attended, audio usage duration, video usage duration, and screen sharing duration.

In addition to the insights derived from the MS Teams device usage and user activity datasets, the study incorporated an analysis of user engagement patterns over time. By correlating these datasets, we can find out how students interacted with the course material and engaged with the platform. Moreover, the analysis revealed that students predominantly utilized audio features during virtual meetings, while screen sharing duration exhibited fluctuations depending on the nature of the content being shared.

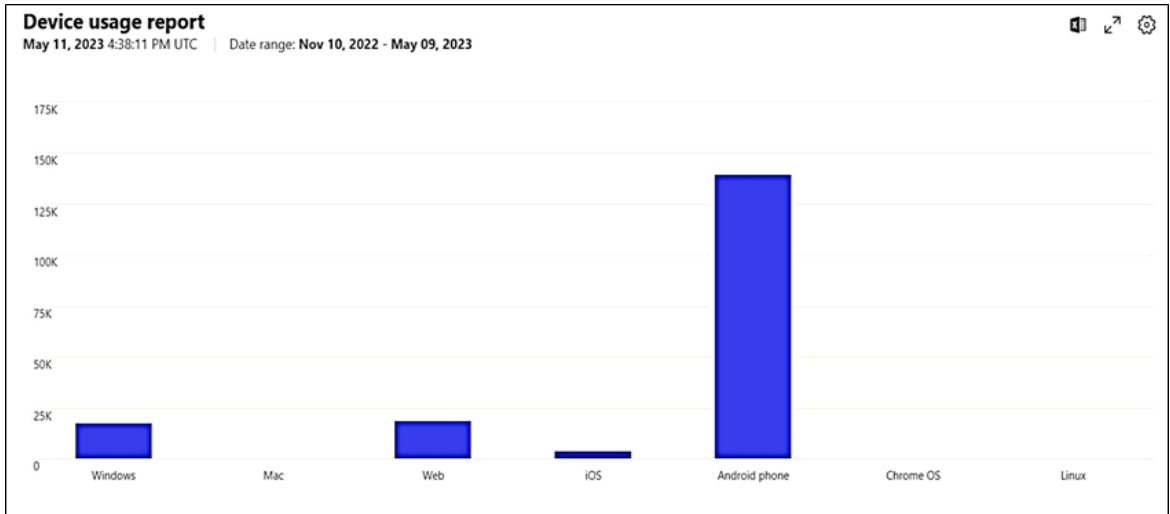


Figure 2: Teams device usage detail of students.

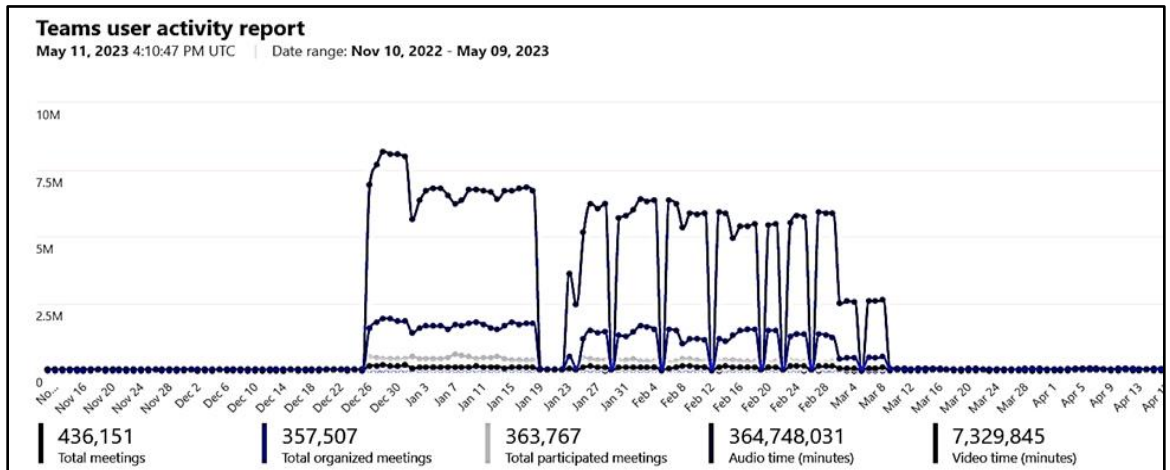


Figure 3: User activity details of students.

The total number of meetings indicates students' active participation in virtual meetings or online collaboration sessions on the MS Teams platform. The total duration of audio usage measures time spent using audio capabilities, while the total duration of video usage reflects time spent using video features. The total duration of screen sharing indicates screen sharing time with other participants during MS Team's sessions.

Course content with the code of "5403-Basics of ICT" was delivered at undergraduate to open and distance learning students was selected and its quiz score data is included in the Moodle (LMS) logs data. It comprises forum logs (4601 count) that reflect user involvement, h5p logs (5666 count) that show interactions with interactive content, resource logs (3785 count) describing material access, and URL view logs (5308 count) that track external resource interactions.

Data from MS Teams and Moodle is pre-processed and analyzed to meet research objectives. Data security measures have been put in place. A dataset including student quiz results is cleaned and combined with MS Teams data for course 5403, yielding 4557 instances. Alignment is by registration numbers, enriched with location data, and influential features are selected for the prediction models.

2.3 Exploratory data analysis

2.3.1 Mapping student distribution using geographic data visualization

The research emphasizes the use of interactive maps for exploratory data analysis, revealing regional trends in student performance. Geographically mapping student locations and quiz scores provides insights into academic achievements, allowing for more focused interventions through improved data visualization.

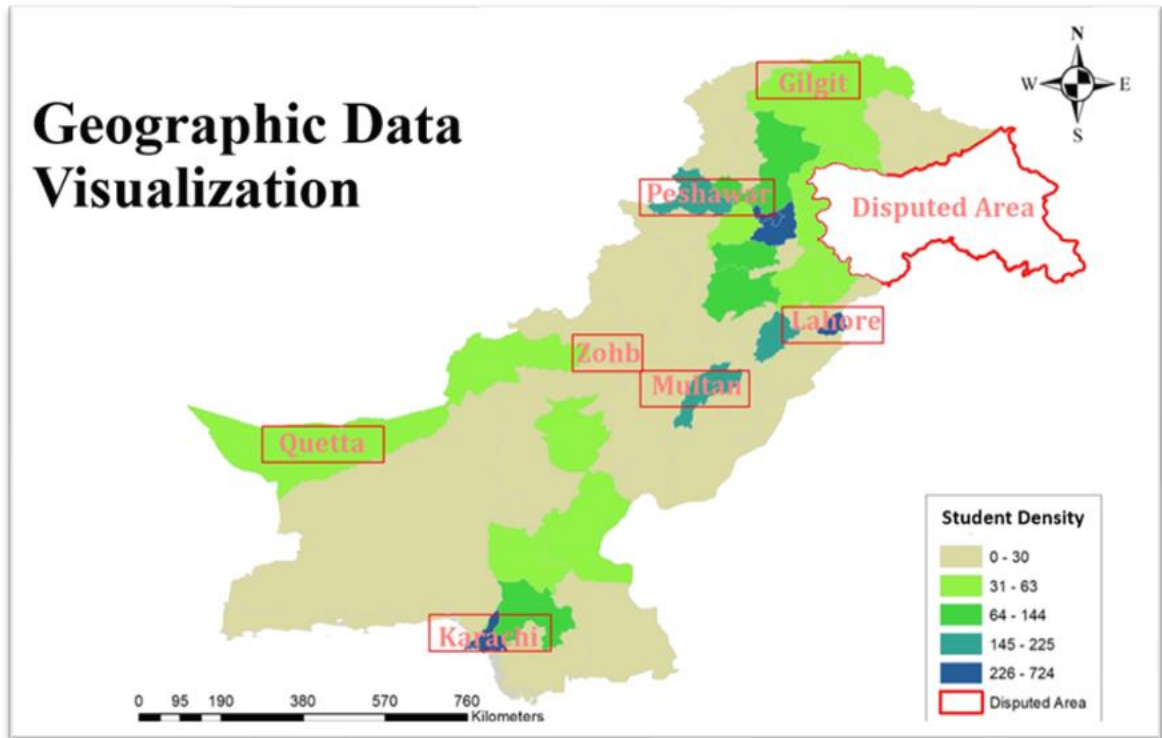


Figure 4: Geographic distribution and density of students in various regions.

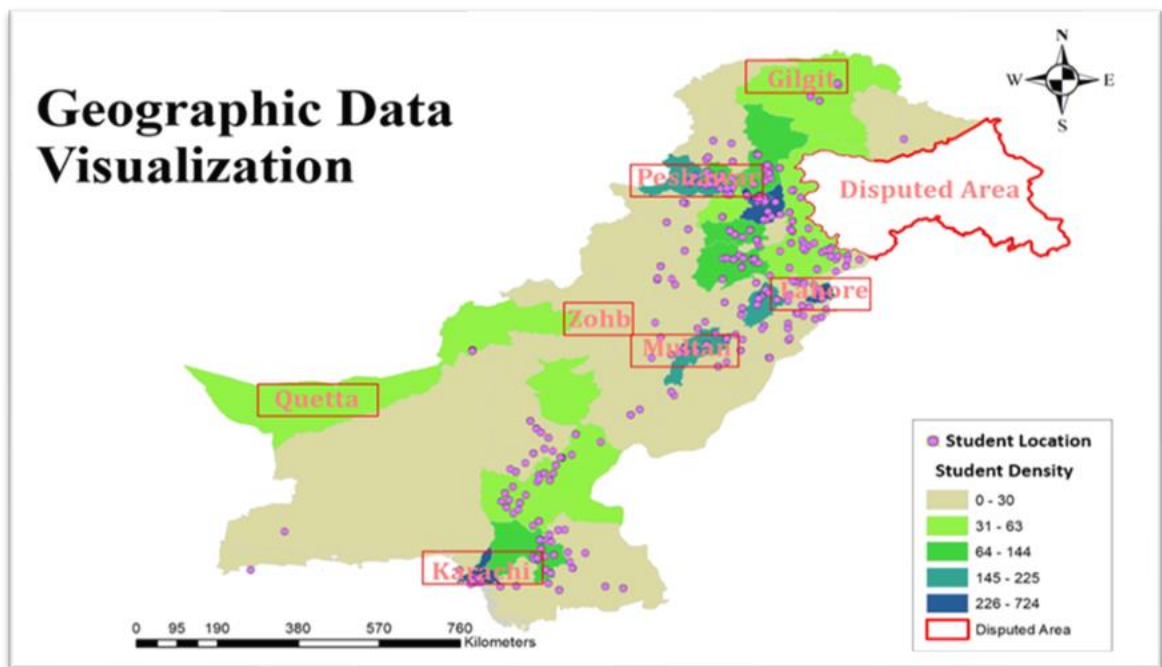


Figure 5: Geographic locations of students residing in highlighted areas.

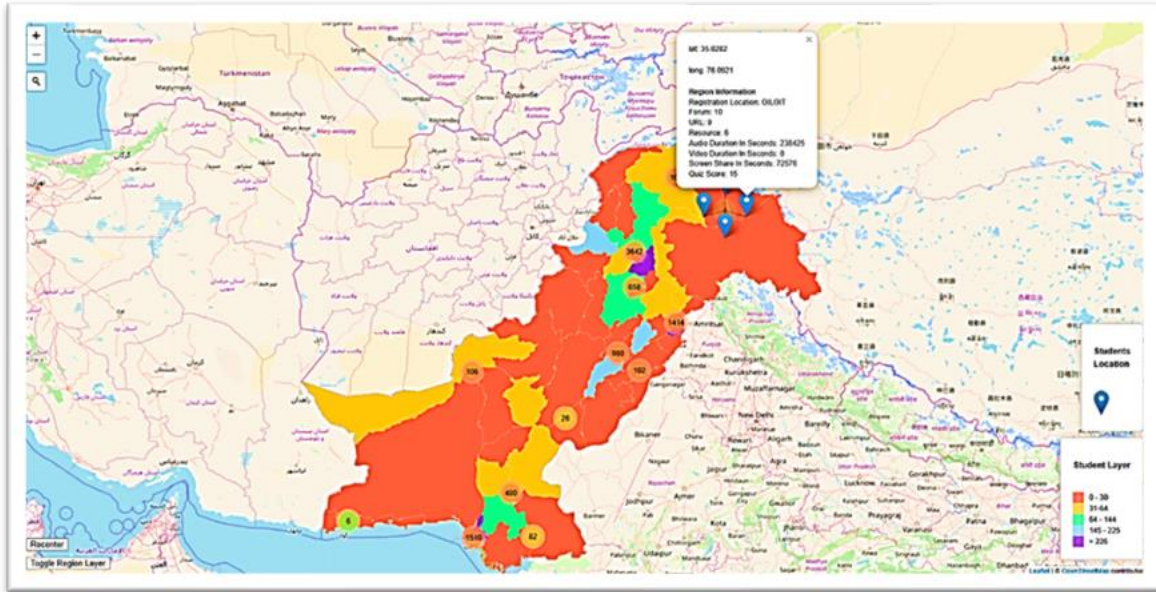


Figure 6: Web map with hierarchical clustering for student information access.

The map in Figure 4 displays different regions, with distinct areas highlighting the distribution and density of students across various regions of the country. Some regions have been labelled for reference purposes.

The map in Figure 5 depicts the locations of students residing in these areas, with individual points indicating the specific residences of each student.

Figure 6 illustrates a web-based map using leaflet that features an interactive hierarchy which includes clicking on a cluster subdivides it into subclusters, and clicking on a subcluster reveals individual pointers. By selecting a pointer, users can access comprehensive information about each student, including their registration location, IP address location, and various data sourced from MS Teams and Moodle.

Interactive maps not only show performance patterns, but also provide institutions with actionable data to personalize interventions, optimize resources, and provide equitable assistance across locations.

2.3.2 Correlation matrix

Correlation matrix heat map shown in Figure 7 use features to predict quiz score. The heatmap displays color-coded numbers, with cool colors indicating negative correlation and warm colors indicating positive correlation. Positive correlations (closer to 1.0) indicate tandem rise or fall, while negative correlations (closer to -1.0) indicate opposite movement. 0.0 correlation indicates no meaningful linear relationship between variables. Correlation matrix indicates weak positive relationships between variables. Weak correlations show that variables are associated, but not strongly enough to indicate a linear relation between them. However, ML algorithms may still detect patterns and use these features to accurately predict the target variable.

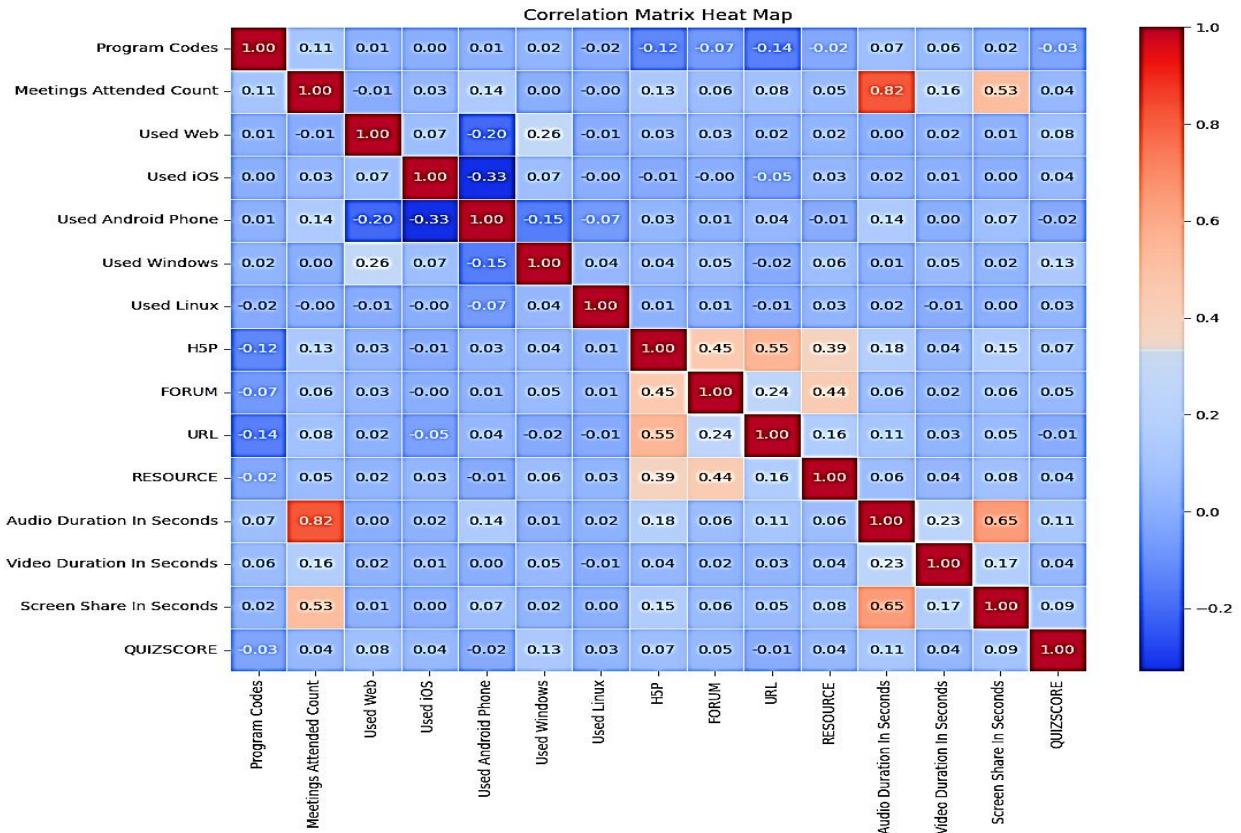


Figure 7: Correlation matrix showing association between variables.

Machine learning models can capture complex patterns, including non-linear correlations, which can enhance predictive power. As a result, even weak correlations might contribute to the model's predictive power. Some features may have non-linear correlations with the target variable that are not reflected in the correlation values. Ensemble approaches, like Random Forest, can handle both linear and non-linear interactions, ensuring enhanced prediction accuracy. Interaction effects of features on the target variable can be detected, enhancing prediction accuracy. Regularized models can handle multicollinearity and prevent overfitting, even if some features are weakly correlated.

2.3.3 Spatial autocorrelation analysis

The Moran's I scatterplot is a spatial statistics method used to visually assess the spatial distribution of a variable and explore spatial patterns. We analyzed spatial autocorrelation using a spatial dataset and Moran's I statistic. Table 1 shows the Moran's I computed value. The Moran's I value of 0.0425 indicates weak positive spatial autocorrelation, with values ranging from -1 to 1.

Negative values indicate spread and dissimilar values are grouped, while positive values indicate clustering together and dispersion. A weak spatial autocorrelation of 0.0425 indicates moderate clustering, suggesting that the distribution of quiz scores among students is spatially dependent. The analysis yielded a p-value of 0.001, indicating statistical significance. Figure 8 shows the scatterplot enlightening the association between the spatially lagged Quiz score and the Quiz score.

Table 1: Spatial Autocorrelation Stats.

Moran's I	0.04251863912834886
Moran's I p-value	0.001

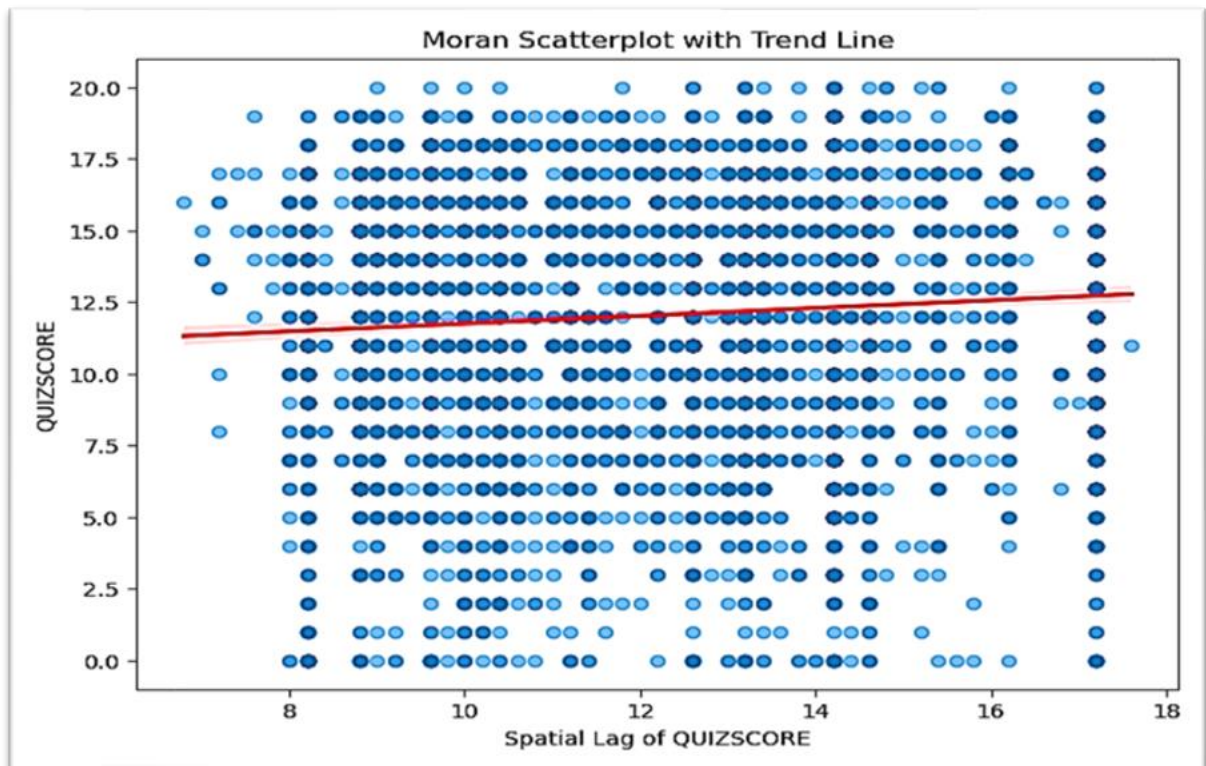


Figure 8: Moran's scatterplot with trend line.

However, the data points are not completely aligned along the trend line, indicating variability in quiz scores for students with similar spatially lagged scores. Clusters closer to the trend line indicate more spatial autocorrelation, while dispersed points indicate lower autocorrelation. Overall, the scatterplot reveals weak positive spatial autocorrelation in students' quiz scores.

2.4 Feature engineering & selection

The technique of selecting or developing features (variables) in a data set to improve machine learning outcomes is known as feature engineering (Domingos, 2012). The selected features, namely program codes, meetings attended count, used web, used iOS, used android phone, used windows, used Linux, h5p, forum, URL, resource, audio duration in seconds, video duration in seconds, and screen share in seconds, will be used to predict the target variable “Quiz score” in the feature engineering step.

Students’ program codes encoded using techniques such as one-hot encoding or label encoding to put them into a numerical format appropriate for modelling. Meetings Attended Count used as is because it shows the number of meetings attended by each student, which can be indicative of their participation level. The elements identifying the usage of multiple devices such as used web, used iOS, used android phone, used windows, used Linux preserved as they are, signifying the mode of access for Microsoft Teams.

Different sorts of interactions inside the Moodle platform can be represented by H5P, Forum, URL, and Resource. The features that represent the duration of audio, video, and screen sharing during team meetings are audio duration in seconds, video duration in seconds, and screen share in seconds. They can be directly used in the model to evaluate their impact on quiz scores. To summarize, the chosen characteristics will be subjected to appropriate encoding, scaling, or other preprocessing processes to ensure compliance with the chosen prediction model. The feature engineered dataset will then be fed into the machine learning model, where different techniques can be used to train the model and predict students quiz scores based on their feature values.

Some learning algorithms use all available attributes to make predictions, whether they are relevant or not, whilst others use variable selection to exclude uninformative features from the model. All the selected features, such as program codes, meetings attended count, used web, used iOS, used android phone, used windows, used Linux, h5P, forum, URL, resource, audio duration in seconds, video duration in seconds, and screen share in seconds, will be used in our study to develop predictive models using various machine learning (ML) algorithms. The ML algorithms will include, among other things, regression, classification that are appropriate for the prediction task. These algorithms will be trained on a dataset that contains all the attributes to evaluate their performance in predicting the target variable.

Following that, the significance of each feature in the prediction models will be assessed using approaches such as the feature importance graph. The final predictive model will be chosen based on its performance metrics and ability to estimate quiz scores properly. We aim to create an interpretable predictive model that will provide significant insights into the elements impacting student quiz performance.

2.5 Predictive model development

2.5.1 Machine learning basic

Machine learning is defined as a computer's ability to learn from experience (Mitchell, 1997). In most cases, experience is provided in the form of input data. The machine can uncover dependencies in the data that are too complicated for a human to establish by looking at this data. Basically, the core idea of machine learning is to enable the computer to learn from data patterns and relationships to make predictions. Machine learning can be used to discover a latent class structure in unstructured data or to find connections in structured data to create predictions. The latter is the thesis's focus.

2.5.2 Predictive analytics

Predictive analytics is the act of forecasting future occurrences and behaviours in previously unseen data using a model developed from similar prior data (Nyce & Cpcu, 2007; Shmueli & Koppius, 2011). It has numerous applications in domains such as finance, education, healthcare, and law as explained by SAS (2017). All these fields use the same application method. A machine learning system identifies relationships between data attributes using previously acquired data. Based on attributes, the resulting model can predict one of the properties of future data by Eckerson (2007).

The goal is to predict the quiz score of students based on the values of other variables in the dataset. In this context, the quiz score is referred to as the dependent variable, while all other features are the independent variables. The dependent variable, "quiz score," is represented as a numerical value, and the machine learning algorithm aims to create a prediction model that takes the independent variables as input and outputs the predicted quiz score for a given student.

The act of creating a prediction model from previously known data is called training, and such data is called training data or a training set. After the model is created, it must be applied to another data set to test its effectiveness. Data used for such a purpose is called test data or test set. The reason for using two different sets is to ensure that the model is flexible enough to be used on data sets other than the one it was built with. Otherwise, overfitting may occur, which occurs when a model is accurate with its initial data set but performs badly with subsequent data sets due to being overly sophisticated (Srivastava et al., 2014). A common method to avoid overfitting is to divide the input data set into training and test sets.

To evaluate the model with test data, the model is used to predict the dependent variable in the test set. Then, the predicted values and actual values of the dependent variable are compared. Evaluation is more complicated than looking at the number of correct predictions.

2.6 Selected methods

There are numerous algorithms to create a prediction model. This thesis uses different algorithms: Random Forest, Decision trees, Support Vector Machine, Light GBM, K-Nearest Neighbors, Logistic Regression and Neural network. While they all essentially have the same task, which is predicting a dependent variable based on independent variables, they are based on different mathematical methods. Table 2 provides description of selected models for predicting students' performance. These combined methods form a comprehensive toolkit, enabling customized choices to meet research objectives, whether focusing on accuracy, efficiency, interpretability, or adaptability.

2.6.1 Predictive model evaluation

2.6.1.1 Training and testing data split

The process of training and evaluating our predictive model involves binary classification method. The goal of classification task is to predict whether a student passes or fails based on a threshold of 10. Students with scores above 10 will be classified as "pass," while those with scores equal to or below 10 will be classified as "fail."

Table 2: Brief description of ML models.

Random Forest	Random Forest is a machine learning ensemble approach that combines decision trees to improve prediction accuracy while reducing overfitting.
Decision Tree	A decision tree is a supervised machine learning technique that makes choices or predictions based on input features using a tree-like structure.
Support Vector Machine	A Support Vector Machine (SVM) is a supervised machine learning technique that finds an ideal hyperplane that best splits data points into different classes.
K- Nearest Neighbors	K-Nearest Neighbors (KNN) is a supervised machine learning technique that classifies data points in a feature space based on the majority class among their nearest neighbors.
Light GBM	Light GBM is a gradient boosting framework that uses tree-based learning methods to manage huge datasets efficiently and quickly.
Neural Network	A neural network is a computational model inspired by the human brain that is made up of interconnected nodes (neurons) organized in layers, where information is processed and transformed via weighted connections to make predictions or perform tasks such as image recognition or natural language processing.
Logistic Regression	Logistic regression is a binary classification statistical model that assesses the likelihood of an event occurring based on input data and a logistic function.

2.6.1.2 Data split

We divided the dataset into two distinct subsets to ensure an unbiased evaluation of the predictive model's performance: a training set and an evaluation (or testing) set. The training set contains 70% of the original data, whereas the evaluation set contains 30%. This data splitting allows us to train the model on a sufficiently substantial piece of the data while also measuring its generalization abilities on a different set. We reduce the risk of overfitting by using this data split strategy, which ensures that the model learns meaningful patterns that can be applied to unseen data rather than simply memorizing the training samples.

2.6.1.3 Binary classification method and threshold

We used a variety of machine learning techniques for the binary classification problem, including K-Nearest Neighbours (KNN), Random Forest, Decision Trees, Light GBM, Support Vector Machine (SVM), and others. The threshold of ten was established as the point of choice for classifying students as passing or failing. Predictions with probabilities more than or equal to 10 are given to the "pass" category, while predictions with probabilities equal to or less than 10 are assigned to the "failure" category. In the subsequent section, we will present the results of our predictive models' performance on the evaluation set and discuss the evaluation metrics used to assess their effectiveness. Table 3 shows example data about students who passed or failed, along with other information about students.

2.6.1.4 Evaluation metrics

The accuracy, precision, recall, and F1-score metrics are used to evaluate the performance and thus application of various classification methods, such as (K-Nearest Neighbors (KNN), Random Forest, Decision Trees, Light GBM, Support Vector Machine (SVM), and others.

Table 3: Example data about students who passed or failed in quiz

Registration no	Quiz Scores	Passed
Sample A	2	0
Sample B	9	0
Sample C	11	1

Accuracy: Accuracy is a key metric that quantifies the fraction of correctly identified instances in the evaluation set. It is calculated as:

$$\text{Accuracy} = (\text{Number of Correctly Classified Instances}) / (\text{Total Number of Instances})$$

Precision: Precision assesses the model's ability to correctly identify positive instances among those labelled as positive (including true and false positives). It is calculated as:

$$\text{Precision} = (\text{True Positives}) / (\text{True Positives} + \text{False Positives})$$

Recall: Recall assesses the model's ability to accurately identify positive cases among all genuinely positive instances (including true positives and false negatives). It is calculated as:

$$\text{Recall} = (\text{True Positives}) / (\text{True Positives} + \text{False Negatives})$$

F1-score: The harmonic mean of precision and recall is used to get the F1 score. It considers both false positives and false negatives. It is calculated as:

$$\text{F1 Score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

RESULTS AND DISCUSSIONS

3.1 Predictive model comparison

3.1.1 Performance analysis of different models

It involves evaluating the performance of various predictive models when they are applied to the binary classification job of predicting student pass/fail outcomes. Decision Tree, K-Nearest Neighbors (KNN), Random Forest, Light GBM, Support Vector Machine (SVM), Logistic Regression, and Neural Network were the machine learning techniques employed.

3.1.1.1 Decision tree

We present a study of the Decision Tree model's performance on the dataset. The analysis's purpose was to predict quiz scores based on a variety of independent factors. The model was trained, tested, and its performance was evaluated using a variety of indicators. The dataset was loaded into the Python environment using the panda's package during the data preprocessing step. The model's relevant independent variables were identified and chosen. These features included meeting attendance count, program codes, web usage, iOS use, Android phone use, Windows use, Linux use, H5P, forum, URL, resource, audio use, and so on. Audio duration is measured in seconds, video duration is measured in seconds, and screen share duration is measured in seconds. The model's dependent variable was "Quiz score."

The data was then divided into a training set (70% of the data) and an evaluation set (30% of the data) using scikit-learn's `train_test_split` function. To ensure that the model operates optimally, we used a data standardization technique to standardize the independent variables

using scikit-learn's StandardScaler. This step was necessary to bring all features to the same scale and minimize biases towards bigger magnitude features. The scikit-learn DecisionTreeClassifier was chosen as the prediction model for this challenge. The model was built, and the training time was calculated. After that, it was trained in a standardized training set. On the evaluation set, the trained Decision Tree model was utilized to predict quiz scores. The predicted time was measured and output is shown in Table 4.

3.1.1.2 K-nearest neighbors (knn)

In this section, we give a complete performance study of the dataset's K-Nearest Neighbours (KNN) model. The KNN algorithm is a non-parametric, instance-based classification approach for predicting categorical outcomes based on feature vector similarity. The panda's library was used to load the dataset into the Python environment. During the preprocessing step, any rows with missing values were eliminated to assure data quality. The target variable "Quiz score" was translated into custom binary labels for the binary classification challenge. Quiz scores of 10 or less were labelled as "Low" (labelled as 0), while scores of 10 or more were labelled as "High" (labelled as 1). We chose a set of independent variables (features) that we thought were important for the KNN model.

The features chosen were the same as those used in the decision tree model. The dataset was divided into training and evaluation sets, with 70% used for training and 30% for assessment. For this, we used the train_test_split function from the scikit-learn library. Before training the KNN model, the features were standardized using scikit-learn's StandardScaler to ensure uniform scales and increase model performance. The number of neighbours (k) was set to 5, suggesting that the model produces predictions based on a particular sample's five nearest

neighbours. The trained KNN model was tested on the evaluation set. Table 5 shows the performance of model using the metrics.

3.1.1.3 Light GBM

In this study, we analyzed a dataset using machine learning techniques, specifically LightGBM as the classifier. The dataset was loaded, and any missing values were eliminated. For model training and evaluation, we concentrated on using the same features. The data was divided into two sets: training and evaluation, with a 70%-30% split. We standardized the features so that they were all on the same scale. The training data was then used to train a LightGBM classifier. The trained LightGBM model was used to make predictions on the evaluation set, and various evaluation metrics were calculated to evaluate the model's performance. Accuracy, precision, recall, and F1 score were among the criteria used. Table 6 shows calculated evaluation metrics.

3.1.1.4 Logistic regression

We investigated the dataset in this research using the Logistic Regression model, a prominent linear classification approach. The dataset was preprocessed, and important features for model training and evaluation were chosen. The dataset was loaded, and the features of interest for analysis were extracted. Using a 70%-30% split, the data was divided into training and evaluation sets. The characteristics in the training set were standardized using the StandardScaler to achieve uniform scaling. The training set was then used to train a Logistic Regression model. To predict quiz scores on the assessment set, the trained Logistic Regression model was used. Various assessment measures, such as accuracy, precision,

recall, and F1 score, were generated to analyse the model's performance. Table 7 shows the assessment measures to analyse the model's performance.

3.1.1.5 Neural network

We employed a Neural Network (NN) model to predict quiz scores based on a variety of independent factors. Using pandas, we load the dataset from a CSV file. The dataset's independent variables (features) are chosen, and the dependent variable (QUIZSCORE) is extracted. Using the `train_test_split` function from `sklearn.model_selection`, the dataset is split into a training set (`X_train, y_train`) and an evaluation set (`X_eval, y_eval`). Using `StandardScaler` from `sklearn.preprocessing`, the data is standardised to have a zero mean and unit variance. This phase guarantees that the features are all on the same scale, which can help with training. `Sequential` from `tensorflow.keras` is used to define the Neural Network model.

It is made up of three layers which include two dense levels with ReLU activation functions and a single output layer with a linear activation function. The "adam" optimizer and the "mean_squared_error" loss function are used to compile the model. The fit function is used to train the model on the training set (`X_train, y_train`). The training procedure lasts 50 epochs and has a batch size of 32. After that, the trained model is used to forecast quiz scores for the evaluation set (`X_eval`). `y_pred` stores the projected quiz scores. Based on a threshold value of 10, the anticipated quiz results are translated into binary numbers (0 or 1). Table 9 presents the evaluation metrics.

3.1.1.6 Random forest

The Random Forest Classifier was used to predict "Quiz score" based on several independent characteristics. The dataset is read from a csv file and loaded into a panda Data Frame, which includes features and the goal variable (Quiz score). The Data Frame is used to extract the independent variables (features) and the target variable (Quiz score). Additional preprocessing was performed to remove missing values, outliers, and inconsistencies. Using a test size of 30% and a train size of 70%, the data is divided into training and evaluation sets. This facilitates model training and evaluation. The StandardScaler is used to apply standardization to the features to ensure consistent scaling across different features.

For reproducibility, a Random Forest Classifier model is built with default hyperparameters and a random seed. On the training set, the model is trained, and the training time is recorded. The trained model is then used to predict "Quiz score" on the evaluation set, and the time it takes to forecast is recorded. Based on a threshold of 10, a binary classification is done on the projected "Quiz score" and the actual "Quiz score," classifying them as pass or fail. Model performance is presented in Table 10.

3.1.1.7 Support vector machine

SVM is a strong supervised machine learning method that may be used to perform classification and regression problems. We use pandas to import the dataset, which comprises several features (independent variables) and the target variable "Quiz score." Missing values and outliers were removed by data preprocessing. Using `train_test_split` from `sklearn.model_selection`, the data is subsequently split into training and evaluation sets.

StandardScaler is used to standardize the independent variables. The fit () method is then used to train the model on the training data. The time spent on model training is tracked.

Table 4: Decision Tree evaluation metrics.

Model Training time	0.0156seconds
Model Predicting time	0.0156seconds
Accuracy	0.6192
Precision	0.7495
Recall	0.7310
F1 Score	0.7401

Table 5: KNN evaluation metrics.

Model Training Time	0.0156 seconds
Model Prediction Time	0.0807 seconds
Accuracy	0.6542
Precision	0.7172
Recall	0.8120
F1-score	0.7617

Table 6: Light GBM evaluation metrics.

Model Training Time	1.6066 seconds
Model Predicting Time	0.0313 seconds
Accuracy	0.6981
Precision	0.7555
Recall	0.8768
F1 Score	0.8117

Table 7: Logistic Regression evaluation metrics.

Model Training Time	0.2150 seconds
Model Predicting Time	0.1096 seconds
Accuracy	0.7266
Precision	0.7467
Recall	0.9557
F1 Score	0.8384

Table 8: Neural Network evaluation metrics.

Model Training Time	3.4478 seconds
Model Predicting Time	0.1440 seconds
Accuracy	0.7244
Precision	0.7496
Recall	0.9438
F1 Score	0.8356

Table 9: Random Forest evaluation metrics.

Model Training time	0.6959 seconds
Model Predicting time	0.0493 seconds
Accuracy	0.7010
Precision	0.7663
Recall	0.8591
F1 Score	0.8100

Table 10: SVM Evaluation Metrics.

Model Training Time	0.5415 seconds
Model Predicting Time	0.1406 seconds
Accuracy	0.7420
Precision	0.7420
Recall	1.0000
F1 Score	0.8519

3.2 Selection of the best performing model

The first stage in analyzing the results is to evaluate the prediction performance of the machine learning approaches. Table 12 shows comparison and performance of each predictive model. SVM and Logistic Regression have the highest precision at 74.20% and 74.67%, respectively, indicating they have fewer false positives. Neural Network is not far behind at 74.96%, followed by Random Forest at 76.63%. These models are well-suited for applications where minimizing false alarms is critical.

SVM stands out with a perfect recall of 100%, capturing all positive cases. Neural Network follows closely with 94.38%, indicating its proficiency in identifying actual positive instances. Logistic Regression also exhibits high recall at 95.57%. KNN and Decision Tree lag in recall. SVM has the highest F1 score of 85.19%, indicating a good balance between precision and recall. Neural Network also strikes a good balance with an F1 score of 83.56%. Logistic Regression follows with an F1 score of 83.84%. Random Forest performs well with an F1 score of 81.00%.

The fastest training times are observed for KNN (0.0156 seconds) and Decision Tree (0.0156 seconds). Logistic Regression (0.2150 seconds) and SVM (0.5415 seconds) are reasonably quick to train. The slowest training time is for the Neural Network (3.4478 seconds).

The fastest prediction times are for Decision Tree (0.0156 seconds) and Random Forest (0.0493 seconds). Logistic Regression (0.1096 seconds) and Neural Network (0.1440 seconds) have moderate prediction times. KNN has the slowest prediction time (0.0807 seconds).

Table 11: Evaluation metrics comparison of predictive models.

Model	Accuracy	Precision	Recall	F1 Score	Training Time(s)	Prediction Time(s)
Neural Network	0.7244	0.7496	0.9438	0.8356	3.4478sec	0.1440sec
SVM	0.7420	0.7420	1.0000	0.8519	0.5415 sec	0.1406 sec
Random Forest	0.7010	0.7663	0.8591	0.8100	0.6959 sec	0.0493 sec
Logistic Regression	0.7266	0.7467	0.9557	0.8384	0.2150 sec	0.1096 sec
KNN	0.6542	0.7172	0.8120	0.7617	0.0156sec	0.0807sec
Light GBM	0.6981	0.7555	0.8768	0.8117	1.6066 sec	0.0313 sec
Decision Tree	0.6192	0.7495	0.7310	0.7401	0.0156 sec	0.0156 sec

In short, SVM excels in recall and F1 score, indicating it captures all positive cases with a balanced precision. Logistic Regression and Neural Network also show strong performance in terms of precision and recall. Random Forest strikes a good balance between precision and recall while offering relatively faster prediction times.

3.3 Features importance graphs

We calculated feature importance using Models to learn which features were most influential in predicting quiz scores. The graph analysis gives intriguing insights on the impact of various characteristics in predicting quiz scores. Understanding the relative relevance of these characteristics might provide significant insights for improving educational practices and optimizing learning outcomes.

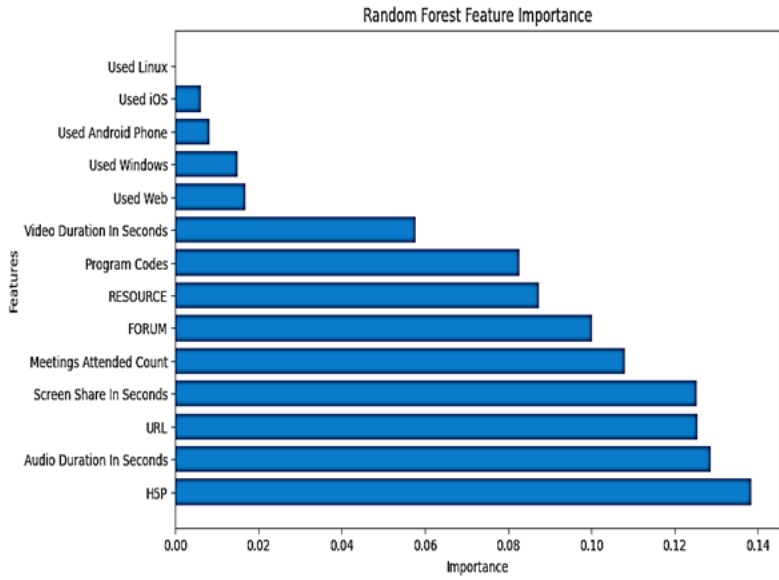


Figure 9: Random Forest features importance graph.

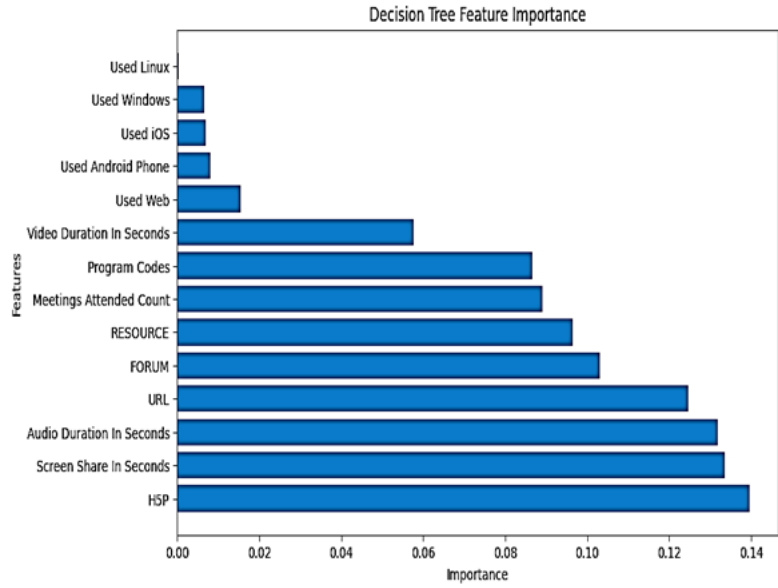


Figure 10: Decision Tree features importance graph.

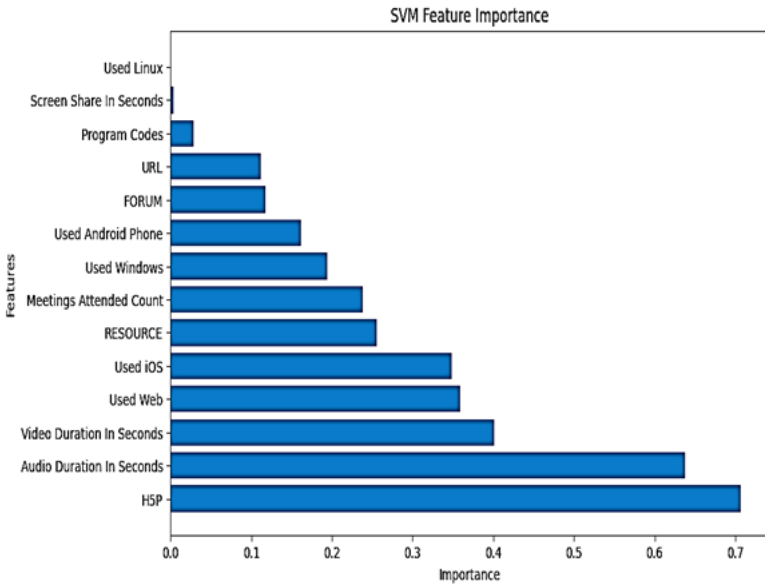


Figure 11: SVM features importance graph.

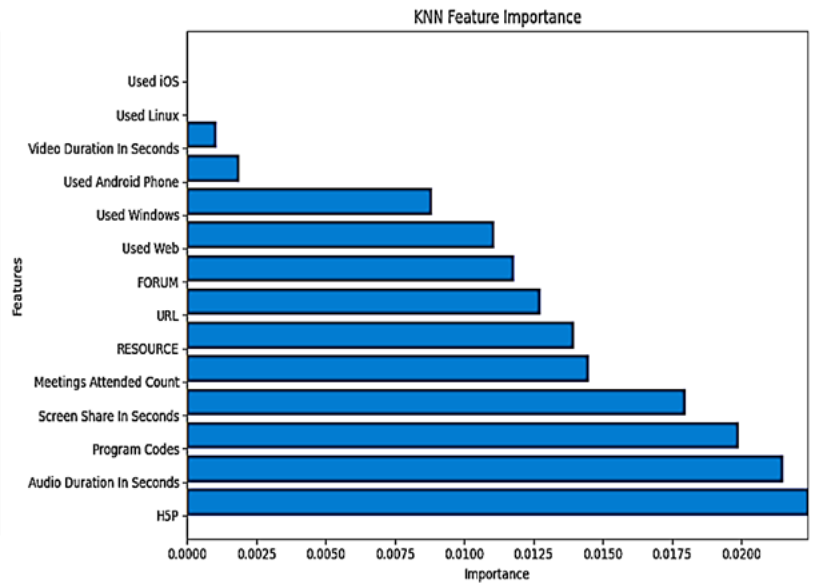


Figure 12: KNN features importance graph.

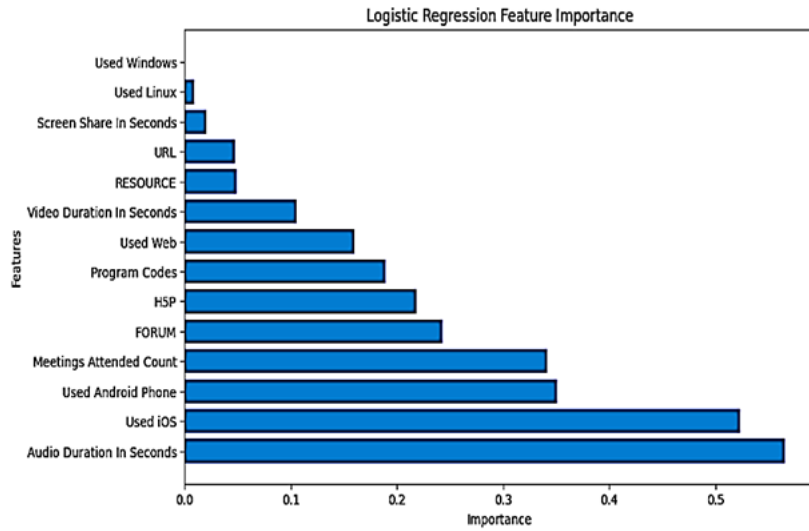


Figure 13: Logistic Regression features importance graph.

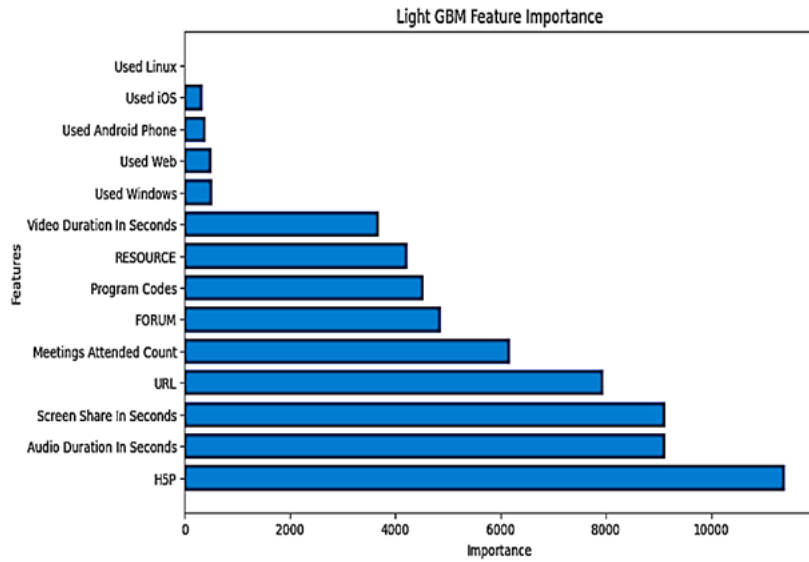


Figure 14: Light GBM features importance graph.

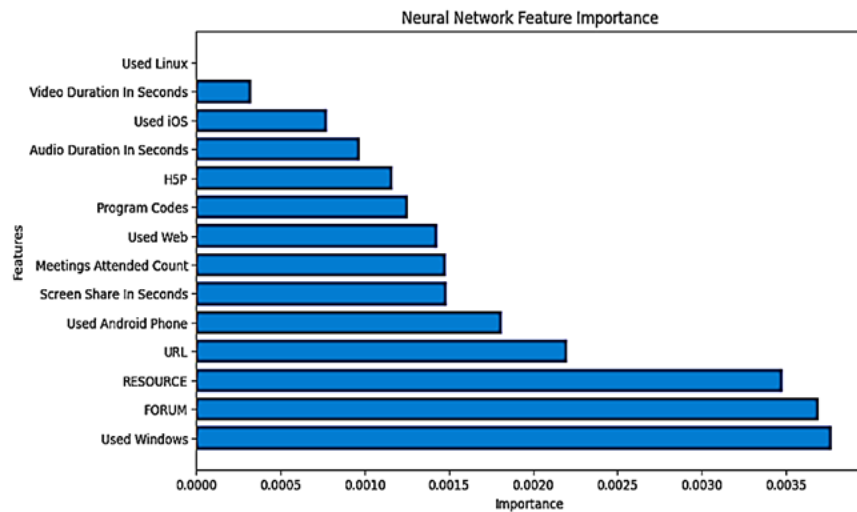


Figure 15: Neural Network features importance graph.

CONCLUSIONS

4.1 Conclusions

This study highlights the potential of machine learning and spatial learning analytics in revolutionizing online and distance education. By merging data from platforms like Moodle and MS Teams, predictive models were developed to predict student performance based on their engagement with course materials. Interactive maps and exploratory data analysis revealed valuable insights into regional trends and student distribution, enabling targeted interventions and resource allocation. Furthermore, the correlation matrix and spatial autocorrelation analysis provided a deeper understanding of the relationships between variables and the spatial dependence of quiz scores. In model comparisons, the Neural Network achieved an accuracy of 0.7244 and a recall of 0.9438, indicating its strong ability to differentiate between learning outcomes, despite its longer training time. The SVM model stood out as the top performer with an accuracy of 0.7420, a perfect recall of 1.0000, and an F1 score of 0.8519, excelling in correctly categorizing positive cases and demonstrating efficiency with shorter training and prediction times. Other models, including Random Forest, Logistic Regression, and Light GBM, also performed well. Logistic Regression showed a strong precision-recall trade-off and quick processing times. KNN and Decision Tree had accuracies of 0.6542 and 0.6192, respectively, indicating potential for accurate predictions. Most other research (Jayaprakash et al., 2014; Kabakchieva, 2013; Cortez & Silva, 2008) does not include recall metrics, focusing solely on accuracy. For example, Kabackchieva (2013) reported 63.1% accuracy with decision trees, while other studies achieved higher accuracy rates with Neural Networks, using factors like SAT scores, past GPAs, and admission test grades, which are reliable predictors of student achievement. SVM emerges as the best model, providing an excellent balance between predictive performance and practical applicability. This study contributes to learning analytics by demonstrating the potential of spatial data in predicting and improving student outcomes, aligning with United Nations Sustainable Development Goal 4 for accessible and equitable education.

LIMITATIONS AND FUTURE RESEARCH DIRECTIONS

5.1 Limitations

One of the study's noteworthy shortcomings is the very limited dataset employed for predictive modelling. The lack of association between the characteristics utilized in the predictive models may have hampered their performance. Strong feature correlations frequently contribute to more accurate predictions in predictive modelling. Models may struggle to recognize meaningful patterns when feature correlations are weak, resulting in lower predictive power. The models' overall accuracy and efficacy in identifying key predictors of students' learning outcomes may have suffered because of the lack of strong correlations. Although the selected qualities in this study did not show strong connections, it is feasible that other unknown features or external influences could have a significant impact on students' learning outcomes.

The models' capacity to capture crucial variables influencing learning outcomes may have been hampered by the lack of critical features, resulting in less useful predictions. The information was gathered from an institute that provides online distance education, which opens the door to academic dishonesty and exam cheating to get higher grades. The lack of supervision and ease of access to external resources in online learning settings may have altered the association between certain course aspects and actual quiz scores.

5.2 Future research directions

Future study should concentrate on broadening the dataset by integrating information from several semesters, courses, or educational institutions. Adding extra information on student demographics, historical academic performance, and learning behaviours to the dataset can provide more thorough insights into the factors influencing learning outcomes. Researchers should investigate new methods of engineering or combining features to better capture students' learning behaviours. Developing composite features based on interactions across multiple course resources, as well as combining time-based data, may result in more powerful predictive models.

It is necessary to create methods for identifying and assessing academic dishonesty or cheating in online quizzes and exams. Integrating cheating behaviour factors into prediction models can help them become more resilient and accurate. We can create more accurate and trustworthy predictive models for students' learning outcomes in online learning environments by addressing these limitations and exploring future research initiatives. These developments have the potential to transform educational support systems and assist educators and policymakers in making informed decisions to improve students' learning experiences.

REFERENCES

1. Agrawal, R., & Pandya, M. (2015). "Data Mining with Neural Networks to Predict Students' Academic Achievements." *International Journal of Computer Science and Technology (ICJST)*, 7(2), 86-90. Retrieved from <http://www.ijcst.com/vol72/1/19-richa-shambhulal-agrawal.pdf>.
2. Agrawal, S., Vishwakarma, S. K., & Sharma, A. K. (2017). Using data mining classifier for predicting student's performance in UG level. *International Journal of Computer Applications*, 172(8), 39-44. Retrieved from <https://www.ijcaonline.org/archives/volume172/number8/agrawal-2017-ijca-915201.pdf>
3. Al-Shabandar, R., Hussain, A., Laws, A., Keight, R., Lunn, J., & Radi, N. (2017). Machine learning approaches to predict learning outcomes in Massive open online courses. *2017 International Joint Conference on Neural Networks (IJCNN)*, 713-720. <https://doi.org/10.1109/ijcnn.2017.7965922>
4. Anderson, T., & Anderson, R. (2017). Applications of machine learning to student grade prediction in quantitative business courses. *Global Journal of Business Pedagogy*, 1(3), 13-22. Retrieved from https://www.igbr.org/wp-content/uploads/articles/GJBP_Vol_1_No_3_2017-pgs-13-22.pdf
5. Becker, B. (2013). Learning Analytics: Insights into the Natural Learning Behaviour of Our Students. *Behavioural & Social Sciences Librarian*, 32(1), 63-67. <https://doi.org/10.1080/01639269.2013.751804>
6. Belachew, E. B., & Gobena, F. A. (2017). Student Performance Prediction Model using Machine Learning Approach: The Case of Wolkite University. *International Journal of Advanced Research in Computer Science and Software Engineering*, 7(2), 46-50. <https://doi.org/10.23956/ijarcsse/v7i2/01219>
7. Brooks, C., & Thompson, C. (2022). *Predictive Modelling in Teaching and Learning*. In C. Lang, G. Siemens, A.F. Wise, D. Gašević, & A. Merceron (Eds.), *The Handbook of Learning Analytics*, 2nd ed., pp. 29-37. Vancouver, Canada: SoLAR. ISBN: 978-0-9952408-3-4. Retrieved from <https://www.solaresearch.org/publications/hla-22/>
8. Buenaño-Fernández, D., Gil, D., & Luján-Mora, S. (2019). Application of Machine Learning in Predicting Performance for Computer Engineering Students: A Case Study. *Sustainability*, 11(10), 2833-2851. <https://doi.org/10.3390/su11102833>
9. Bujang, S. D. A., Selamat, A., Ibrahim, R., Krejcar, O., Herrera-Viedma, E., Fujita, H., & Ghani, N. A. M. (2021). Multiclass Prediction Model for Student Grade Prediction Using Machine Learning. *IEEE Access*, 9, 95608-95621. <https://doi.org/10.1109/access.2021.3093563>

10. Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. EUROSIS. Retrieved from https://www.researchgate.net/publication/228780408_Using_data_mining_to_predict_secondary_school_student_performance
11. Dollinger, S. J., Matyja, A. M., & Huber, J. L. (2008). Which factors best account for academic success: Those which college students can control or those they cannot? *Journal of Research in Personality*, 42(4), 872–885. Retrieved from <http://s3.amazonaws.com/academia.edu.documents/32955335/Dollinger>
12. Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78-87. <https://doi.org/10.1145/2347736.2347755>
13. Dondorf, T. (2022). *Learning Analytics for Moodle: Facilitating the Adoption of Data Privacy-Friendly Learning Analytics in Higher Education* [Dissertation, Rheinisch-Westfälische Technische Hochschule Aachen]. RWTH Aachen University. Retrieved from <https://d-nb.info/1257981234/34>
14. Eckerson, W.W. (2007). Predictive Analytics: Extending the Value of Your Data Warehousing Investment. TDWI Best Practices Report, 1, 1-36. Retrieved from <https://tdwi.org/research/2007/01/bpr-1q-predictive-analytics.aspx>
15. Greller, W., & Drachsler, H. (2012). Translating Learning into Numbers: A Generic Framework for Learning Analytics. *Educational Technology & Society*, 15, 42-57. Retrieved from https://www.researchgate.net/publication/234057371_Translating_Learning_into_Numbers_A_Generic_Framework_for_Learning_Analytics
16. Hämmäläinen, W., & Vinni, M. (2006). Comparison of Machine Learning Methods for Intelligent Tutoring Systems. In M. Ikeda, K.D. Ashley, & TW. Chan (Eds.), *Intelligent Tutoring Systems*. ITS 2006. Lecture Notes in Computer Science, Vol. 4053. Springer. https://doi.org/10.1007/11774303_52.
17. Ibrahim, Z., & Rusli, D. (2007). Predicting students' academic performance: Comparing artificial neural network, decision tree and linear regression. In Annual SAS Malaysia Forum, Kuala Lumpur 1-6. Retrieved from https://www.researchgate.net/publication/228894873_Predicting_Students'_Academic_Performance_Comparing_Artificial_Neural_Network_Decision_Tree_and_Linear_Regression
18. Jayaprakash, S. M., Moody, E. W., Laur'ia, E. J., Regan, J. R., & Baron, J. D. (2014). Early Alert of Academically At-Risk Students: An Open-Source Analytics Initiative. *Journal of Learning Analytics*, 1(1), 6–47. Retrieved from <http://www.learning-analytics.info/journals/index.php/JLA/article/viewFile/3249/4011>

19. Johnson, D., & Samora, D. (2016). The potential transformation of higher education through computer-based adaptive learning systems. *Global Education Journal*, 2016(1). Retrieved from <http://web.a.ebscohost.com/abstract/112407351>.
20. Kabakchieva, D. (2013). Predicting Student Performance by Using Data Mining Methods for Classification. *Cybernetics and Information Technologies*, 13(1), 61–72. DOI: [10.2478/cait-2013-0006](https://doi.org/10.2478/cait-2013-0006)
21. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-Scale Video Classification with Convolutional Neural Networks. In 2014 *IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1725-1732). Columbus, OH, USA. doi:10.1109/CVPR.2014.223.
22. Kumar, S., Surjeet, & Bharadwaj, B. (2012). Mining Education Data to Predict Student's Retention: A Comparative Study. *International Journal of Computer Science and Information Security*, 10, 113-117. Retrieved from https://www.researchgate.net/publication/221700448_Mining_Education_Data_to_Predict_Student's_Retention_A_comparative_Study
23. Long, P. D., & Siemens, G. (2011). Penetrating the fog: Analytics in learning and education. EDUCAUSE Review Online. Retrieved from <http://www.educause.edu/ero/article/penetrating-fog-analytics-learning-and-education>
24. Minaei, B., Kashy, D.A., Kortemeyer, G., & Punch, W. (2003). Predicting student performance: An application of data mining methods with an educational Web-based system. Proceedings - Frontiers in Education Conference, 1, T2A-13. <https://doi.org/10.1109/FIE.2003.1263284>.
25. Mitchell, T.M. (1997). Machine Learning. McGraw-Hill. Retrieved from <https://www.cin.ufpe.br/~cavmj/Machine%20-%20Learning%20-%20Tom%20Mitchell.pdf>
26. Moucary, C., Khair, M., & Zakhem, W. (2011). Improving Students Performance using Data Clustering and Neural Networks in Foreign-Language based Higher Education. *The Research Bulletin of Jordan ACM*, 2(3), 27–34. Retrieved from <http://ijj.acm.org/volumes/volume2/no3/ijjvol2no3p1.pdf>
27. Nyce, C., & Cpcu, A. (2007). Predictive Analytics White Paper. American Institute for CPCU, Insurance Institute of America. Retrieved from <https://www.the-digital-insurer.com/wp-content/uploads/2013/12/78-Predictive-Modeling-White-Paper.pdf>
28. Olmos, M., & Corrin, L. (2012). Learning Analytics: A Case Study of the Process of Design of Visualizations. *Online Learning*, 16(3), 39-49. <https://doi.org/10.24059/olj.v16i3.273>

29. Papamitsiou, Z., & Economides, A. (2014). Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence. *Educational Technology & Society*, 17, 49-64. Retrieved from https://www.researchgate.net/publication/267510046_Learning_Analytics_and_Educational_Data_Mining_in_Practice_A_Systematic_Literature_Review_of_Empirical_Evidence
30. Pardo, A., & Siemens, G. (2014). Ethical and privacy principles for learning analytics. *British Journal of Educational Technology*, 45(3), 438-450. <https://doi.org/10.1111/bjet.12152>
31. Pardos, Z., Heffernan, N., Anderson, B., & Heffernan, C. (2007). The Effect of Model Granularity on Student Performance Prediction Using Bayesian Networks. In International Conference on User Modeling, Corfu, Greece 435-439. Retrieved from DOI: [10.1007/978-3-540-73078-1_60](https://doi.org/10.1007/978-3-540-73078-1_60)
32. Peña-Ayala, A. (2014). Review: Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications: An International Journal*, 41, 1432-1462. <https://doi.org/10.1016/j.eswa.2013.08.042>.
33. Pojon, M. (2017). Using machine learning to predict student performance (master's thesis). University of Tampere, Finland. Retrieved from <https://core.ac.uk/download/pdf/250148623.pdf>
34. Powers, D.M.W. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, 2(1), 37-63. Retrieved from <https://arxiv.org/abs/2010.16061>
35. Quinn, R. J., & Gray, G. (2019). Prediction of student academic performance using Moodle data from a Further Education setting. *Irish Journal of Technology Enhanced Learning*, 5(1), 1-10. <https://doi.org/10.22554/ijtel.v5i1.57>
36. Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin*, 138(2), 353. Retrieved from <http://emilkirkegaard.dk/en/wp-content/Psychological-correlates>
37. Robinson, A. C., Anderson, C. L., & Quinn, S. D. (2020). Evaluating geovisualization for spatial learning analytics. *International Journal of Cartography*, 6(3), 331-349. <https://doi.org/10.1080/23729333.2020.1735034>
38. Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33(1), 135-146. <https://doi.org/10.1016/j.eswa.2006.04.005>
39. Romero, C., & Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6), 601-618. <https://doi.org/10.1109/tsmcc.2010.2053532>

40. Romero, C., Ventura, S., Hervás, C., & Gonzales, P. (2008). Data mining algorithms to classify students. In *International Conference on Educational Data Mining*, Montreal, Canada ,8-17.Retrieved from https://www.researchgate.net/publication/221570435_Data_Mining_Algorithms_to_Classify_Students
41. SAS. (2017). Predictive Analytics: What it is and Why it Matters. Retrieved from https://www.sas.com/en_us/insights/analytics/predictive-analytics.html
42. Shmueli, G., & Koppius, O.R. (2011). Predictive Analytics in Information Systems Research. *MIS Quarterly*, 35(3), 553-572. <https://doi.org/10.2307/23042796>
43. Siemens, G. (2010). What are learning analytics. *ELEARNSPACE: Learning, networks, knowledge, technology, community*. Retrieved from <http://www.elearnspace.org/blog/2010/08/25/what-are-learning-analytics>
44. Smith, V. C., Lange, A., & Huston, D. R. (2012). Predictive Modeling to Forecast Student Outcomes and Drive Effective Interventions in Online Community College Courses. *Online Learning*, 16(3), 51-61. <https://doi.org/10.24059/olj.v16i3.275>
45. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *The Journal of Machine Learning Research*, 15(1), 1929-1958.Retrieved from <https://jmlr.org/papers/v15/srivastava14a.html>
46. Stapel, M., Zheng, Z., & Pinkwart, N. (2016). An ensemble method to predict student performance in an online math learning environment. In Proceedings of the 9th International Conference on Educational Data Mining, *International Educational Data Mining Society* (pp. 231–238). Retrieved from <https://www.semanticscholar.org/paper/An-Ensemble-Method-to-Predict-Student-Performance-Stapel-Zheng/e1ca981f55484ade30a8132d3a9492994f989319>
47. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann. Retrieved from <https://www.elsevier.com/books/data-mining/witten/978-0-12-804291-5>.
48. Xu, J., Moon, K. H., & Van der Schaar, M. (2017). A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs. *IEEE Journal of Selected Topics in Signal Processing*, 11(5), 742-753. <https://doi.org/10.1109/jstsp.2017.2692560>
49. Yadav, S. K., Bharadwaj, B., & Pal, S. (2012). Mining education data to predict student's retention: A comparative study. *International Journal of Computer Science and Information Security*, 10(2), 113-117.Retrieved from https://www.researchgate.net/publication/221700448_Mining_Education_Data_to_Predict_Student's_Retention_A_comparative_Study