BACHELOR OF SCIENCE IN TECHNICAL AND VOCATIONAL EDUCATION

Smart Academic Forecasting Using Data-Driven Approaches

Most.Nosin Nahar Smita
220032401
Farhana Akter Lipi
220032403
Sultana Razia Shrabonti
220032405
Mishkatul Nayeem
220032406

Department of Technical and Vocational Education
Islamic University of Technology
September, 2025

Smart Academic Forecasting Using Data-Driven Approaches

Most.Nosin Nahar Smita
220032401
Farhana Akter Lipi
220032403
Sultana Razia Shrabonti
220032405
Mishkatul Nayeem
220032406

Department of Technical and Vocational Education
Islamic University of Technology
September, 2025

Declaration of Candidate

This is to certify that the work presented in this thesis is the outcome of the analysis and experiments carried out by **Most.Nosin Nahar Smita**, **Farhana Akter Lipi**, **Sultana Razia Shrabonti**, and **Mishkatul Nayeem** under the supervision of **Shahriar Ivan**, Assistant Professor, Department of Computer Science and Engineering, Islamic University of Technology, Dhaka, Bangladesh. It is also declared that neither this thesis nor any part of it has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others have been acknowledged in the text and a list of references is given.

Shahriar Ivan
Assistant Professor
Department of Computer Science and Engineering
Islamic University of Technology (IUT)
Date: September 30, 2025

Most.Nosin Nahar Smita
Student ID: 220032401
Date: September 30, 2025

Farhana Akter Lipi
Student ID: 220032403
Date: September 30, 2025

Sultana Razia Shrabonti
Student ID: 220032405
Date: September 30, 2025

Mishkatul Nayeem
Student ID: 220032406
Date: September 30, 2025

Dedicated to our supervisor, whose guidance and support were invaluable throughout the completion of this degree.

Contents

List of Figures

List of Tables

List of Abbreviations

DT	Decision Tree
KNN	K-Nearest Neighbors
SVM	Support Vector Machine
PCA	Principal Component Analysis
LDA	Linear Discriminant Analysis
LR	Logistic Regression
RF	Random Forest

Acknowledgement

We would like to express our deepest gratitude to our supervisor, Shahriar Ivan, whose advice, experience and support have been invaluable in this process. His power to give light at the time when the road might have been in doubt and his judicious counsel have influenced this work in a way which we could never have imagined. We are most grateful to his patience and his ability to ask the right question at the right time, when it is important to refine our thoughts and strategy.

We also owe our families a huge debt, which they have been our anchor because of their unwavering support. They have been our encouragements in all our difficulties, giving not only love and care but also that reality check which we sometimes lost in our studies. They have always been a motivating factor through their patience particularly whenever we had a lengthy discussion on complicated issues.

And finally, we want to give our fellow students, our friends, and anyone who has ever been a part of this process, whether by working together, by giving feedback, by just being present to experience our stress and/or celebration. Your support has helped us to continue even when the task appeared to be never-ending.

We could not have done this thesis without each of you and we will always appreciate your assistance.

Abstract

Proper prediction of student academic performance is critical to supporting timely interventions and improving learning results. The article is an evaluation of classical machine learning models, including Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors, and Support Vector Machine, and a transformer-based foundation model based on benchmark academic performance datasets. Transformer-based approach had the best predictive accuracy, whereas Decision Tree provided a good compromise between predictive accuracy and Interpretability. Besides model benchmarking, the study was aimed at exploring the importance of features in addition to analysing Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). These methods pointed out that behavioral attributes like participation, resource use and absence days were the strongest predictors of performance. The consistency of these results was validated on another dataset and the strength of the feature analysis was established. On the whole the study shows that predictive modeling combined with dimensionality reduction gives reliable and explainable information that can be used to build trustworthy academic prediction systems.

Chapter 1

Introduction

1.1 Introductory Information

Education plays a crucial role informing society and developing socio economic growth. Predicting the possible result or outcome of students in their any specific course or overall programs as a whole has become an important domain in educational data mining (EDM). Teachers, administrator, policymakers can take timely action by using accurate forecasting. It helps them to ensure that weaker students get essential support to improve their performance and high achieving students get advance opportunities to grow.

Teacher intuition, manual record analysis or simple statistical models are frequently used in traditional methods to assessing student's academic performance. To manage the massive amount of educational data produced by modern learning management systems and online platforms, these approaches are restricted in scalability and lack the precision necessary. New data-driven approaches have emerged by using the advancement of machine learning and artificial intelligence that can model complex patterns in student behavior and background data, enabling more reliable predictions.

In this thesis, we present a comprehensive study of smart academic prediction using a data-driven approach. We aim to systematically asses the predictive performance of various methods on student academic performance datasets by employing multiple machine learning algorithm – ranging from classical statistical models to modern transformer-based architectures. This allows us to identify the best-performing techniques but also give knowledge about feature importance, dimensionality reduction and the difficulties related to student data.

1.2 Motivations and Scope

The motivation for this study grew from the progressive requirement for personalized education and early interventions in academic environments. Educational institutions require intelligent equipment to forecast student success and failure so that teachers, policymakers and academic counselors can make appropriate decisions. For example, by detecting weaker performing students earlier, institutions can design remedial programs, adjust teaching strategies and provide additional resources in need by detecting weaker performed students earlier. At the same time, early detection of high-achieving students enables institutions to understand their characteristics and provide various nurturing facilities.

The scope of our research includes:

1. Assessing the effectiveness of multiple machine learning algorithms such as logistic Regression, Decision Tree (DT), Random Forest (RF), k-nearest neighbors (KNN), Support Vector Machine (SVM), and TabPFN for predicting academic performance.
2. Comparing the models using key evaluation metrics: Accuracy, RMSE,MAPE, and R^2.
3. Evaluating the consequence of feature selection and dimensionality reduction using Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA).
4. Administering experiments with different subsets of features (top 10, bottom 10, top 5, bottom 5) to explore overlap and redundancy in the dataset.
5. Benchmarking our result against state-of-the-art research, specifically the deep ensemble learning approach proposed by[42].
6. Assessing the strength of our models on an additional dataset to evaluate generalizability.

Our study bridges traditional data driven approaches with modern deep learning strategies by focusing on both classical models and advanced architectures.

1.3 Problem Statement

One of the most difficult aspects in the research of education has always been to predict the academic performance of students. The conventional methods typically use

teacher intuition or manual testing or a simple statistical model, which can barely define the multifactorial nature of demographic, behavioral and contextual variables on learning outcomes. Studies of machine learning done so far have borne potential, yet most are constrained in addressing a limited variety of models, or only measuring accuracy as a measure of evaluation, or ignoring the problem of feature redundancy in learning data. Therefore, the existing forecasting systems are often not generalizable across settings and only give a limited amount of information about what characteristics of students actually lead to performance.

To overcome these weaknesses, the current paper involves a data-heavy, holistic method which combines several classical and state-of-the-art machine learning models, the dimensionality reduction strategies, including Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). They do not only enable us to benchmark predictive accuracy, but feature importance and draw correlations between crucial features and academic outcomes. Moreover, the fact that the results have been validated in two data sets supports the strength of the results. By addressing both the predictive performance and interpretability, this study will help lay a stronger ground in the development of the early-warning systems that will not only be able to aid in the timely intervention but also enhance the success of students.

1.4 Research Challenges

The study faced some challenges that are mentioned below:

1. **Feature Correlation and Redundancy:** The Limited number of features in the dataset makes is difficult for dimensionality reduction methods like PCA and LDA to improve performance effectively. Many features overlap between top and bottom subsets in their forecasting power.
2. **Accuracy versus Interpretability of the Model:** While lightweight models such as Random Forest and Decision Trees are more interpretable, advanced deep learning models like TabPFN and deep ensembles achieve higher accuracy but function as 'black boxes.
3. **Variability of the Dataset:** It is challenging to design universally applicable predicting systems as educational data vary across institutions, regions and contexts.
4. **Underfitting and Overfitting:** Models generalize well without overfitting or underfitting to training data remains a core issue, particularly with smaller

datasets.

5. **Computational Cost:** Advanced deep learning model like deep ensemble and transformer based approaches require higher computational power, which may not be accessible in all educational settings.

1.5 Contribution

The thesis attempts to address these issues:

1. **Comprehensive Benchmarking:** We compare the assumptions of various machine learning models (Logistic Regression, Decision Tree, Random Forest, SVM, KNN and a transformer-based approach) to a set of databases concerning student academic performance and use a uniform set of performance measures.
2. **Feature Selection Analysis:** In this study, we examine influence of reducing the dimensions using PCA and LDA based on varying subsets of features, the difficulties of small features space and redundancy in learning databases.
3. **Transformer-Based Models:** We show that a transformer-based model can be successfully applied to academic performance prediction, and it can be competitive enough with classical machine learning models.
4. **Features Importance insights:** With systematic analysis, we can understand which features (e.g., behavioral (e.g., resource usage, participation, and absence days) are most predictive of student performance, and which contribute minimally, hence give more insights into learning behaviors.
5. **Practical guidance:** We present visual and comparative analysis, which indicate the trade-offs of accuracy, interpretability, and computational efficiency, as well as valuable advice in the application of forecasting systems to real educational situations.

1.6 Organization

The remaining part of this thesis is organized in the following way:

- **Chapter 2** covers Related works on academic prediction detailing the working procedure of machine learning methods and the deep learning methods including ensemble model in [42].

- **Chapter 3** Methodology contains the datasets, the models and the evaluation metrics, the preprocessing and criteria used in the experiments along with feature selection and other steps in the model build for our experiments.
- **Chapter 4** covers the experimental results, the competition between classical and newly proposed models, the effects of model complexity on model performance, and the gold standard for comparison in this study.
- **Chapter 5** provides the fundamentals with the research. It elaborates the impact and the impact of the findings, discusses the findings, highlights the gaps, and formulates the scope for better research.
- **Chapter 6** compiles the referenced works to construct the central argument for the study.

Chapter 2

Related Works

2.1 Overview: Educational Data Mining (EDM) and Learning Analytics

The educational Data Mining and Learning Analytics have become a significant topic of research that is trying to convert raw educational data into something that can be done with. [38] made one of the first extensive reviews of EDM, identifying prediction, clustering, relationship mining and the creation of discovery models as activities. [7] highlighted the ways educational data mining would be applied to create adaptive learning settings and maximize teaching effectiveness. The underlying works attracted the focus on the issues of scalability, sparsity of data, and real-life implementation. The most recent systematic reviews including [33] emphasized the need of empirical evidence and reproducibility. Workflow management research, such as the application of Petri nets to model complex processes [43], offers conceptual parallels to how student learning activities can be modeled and analyzed systematically in EDM. [5] also paid attention to the prediction of academic success in higher education and provided a summary of best practices and popular machine learning algorithms. These surveys prove the maturity of EDM as a discipline as well as identifying gaps including lack of interpretability and generalization across contexts. Reviews and surveys have explained how these methods help to build early-warning systems and feature selection methods [3], [5]. Our work is based on these surveys and makes specific attention to predictive modeling using modern machine learning algorithms and dimensionality reduction methods.

2.2 Predictive Modelling on Education: Classical and Ensemble.

EDM has always been concerned with the prediction of student success. Early techniques used classical supervised learning techniques. Decision trees, including the [35], gave interpretable rule-based models of classification problems, and are still popular because of their explicability and ease of use in educational data. The use of logistic regression has been widely used in binary classification, including pass or fail, with [25] presenting an extensive discussion of the technique. K-Nearest Neighbors K-NN was originally suggested by [16] and this method has also been used to make predictions based on student data using instance-based methods. Cortes [15] and Boser [10] presented Support Vector Machines which became very popular because of its ability to perform well on high dimensional data. In addition to these, those techniques of ensemble learning have greatly enhanced predictive power. Random Forests [11] proved to be very robust with overfitting resistance. This trend was followed by boosting methods, which provide enhanced training, the ability to work with categorical variables and good performance on tabular data XGBoost [13], [28], [34] demonstrate. Other previous EDM-specific studies that use classification to predict academic outcomes include [29] which have positive findings but with small sample sizes, and a study by [40] which employed even smaller sample sizes but with promising results. Deep ensemble learning on prediction tasks on students is also a recent development. Recent studies have also explored classification and prediction of student performance data using multiple machine learning algorithms, showing that even relatively simple approaches can provide useful insights for academic planning [32]. Other previous EDM-specific studies that use classification to predict academic outcomes include [27], [31], all of which reported promising results with varying datasets. [9] also applied data mining approaches for predicting student performance. More recent studies have incorporated e-book reading behavior as predictive features, as in [12], while [46] explored classification of student performance using AI techniques. Designed early warning systems to identify at-risk students using learning analytics. Indicatively, [42] demonstrated how deep ensembles were able to learn intricate relationships between features of students and excel compared to the traditional baselines. Beyond educational data, ensemble deep learning techniques have been comprehensively studied in time series analysis, with surveys highlighting open issues and future directions, which can inform the adaptation of such methods to longitudinal student data [39].

2.3 Tabular Data Transformer and Foundation Models.

Conventional deep learning models could not deal with tabular data because of the problem of overfitting and the absence of inductive biases to handle categorical and mixed-type data. But newer architectures that have been designed explicitly to support tabular learning have been developed.The introduction of the transformer architecture, originally by [44], has had a profound impact on sequence modeling and inspired subsequent tabular learning adaptations. Work on vision transformers, such as [18], demonstrated the versatility of transformer architectures beyond NLP, strengthening the case for applying similar ideas in tabular and educational domains. Recent evaluations of deep learning models for tabular data by [21] emphasized the importance of proper baselines and suggested that simple architectures can sometimes rival transformers in practice. Attentive feature selection was proposed in TabNet [6], making it possible to interpret features and achieve a good predictive accuracy. Most recently, [24] proposed a pretrained transformer that is capable of solving small tabular classification tasks in one forward pass without any expensive hyperparameter tuning. This model has been used in education research where data are sometimes small in scale or of medium size. These developments demonstrate a paradigm shift in which foundation models are able to compete and even outdo conventional ensemble learners. [19] also contributed with work on automated machine learning frameworks which help streamline the process of model selection and optimization for educational data. Transfer learning has also been investigated in domains such as industrial IoT networks, providing empirical evidence of how pretrained models can be adapted across contexts [45], a principle relevant to the adaptation of foundation models for educational datasets. [23] further provided theoretical foundations for kernel methods which remain influential in many classification settings. Extensions such as SAINT [41] leveraged row attention and contrastive pretraining to improve neural networks for tabular data, providing another foundation for exploring educational applications. These developments demonstrate a paradigm shift in which foundation models are able to compete and even outdo conventional ensemble learners.

2.4 Fairness, Interpretability, and Privacy of EDM.

The consequences of educational predictions are high stakes; therefore, interpretability and fairness are of great importance. [37] came up with LIME (Local Interpretable Model-agnostic Explanations), which enables the black-box models to be locally ex-

plained. In 2017, [30] developed SHAP (SHapley Additive exPlanations), which offers a consistent model to explain predictions by features. It has also become popular to be fair in educational prediction, and [36] conducted a review of fairness-conscious machine learning in the educational setting. The works emphasize that predictive models must not merely be true but must also be fair between demographic groups. Moreover, the topic of privacy-preserving data mining has recently been addressed by [4], who addressed the concept of differential privacy, as well as federated learning, to ensure the privacy of sensitive student data.

2.5 Dimensionality Reduction of EDM: PCA vs. LDA.

Educational data also includes correlated variables, such as demographics, behavioral traces and assessment data. Dimensionality reduction is used to reduce redundancy, improve model performance, and make them more interpretable. Principal Component Analysis (PCA), which has as far been reviewed by [26], is a linear projection technique that maximizes variance. It is unsupervised, commonly applied to the noise reduction and visualization purposes, though no label information is utilized. Principal Component Analysis has been reviewed in detail by [1], who provided an accessible introduction and highlighted its applications in reducing dimensionality across diverse datasets. The original form of LDA was developed by [20] and is a supervised method that transforms features into a dimension where the samples are separated by their classes. The concept of discriminative dimensionality reduction was further advanced by [8], who compared eigenfaces and Fisherfaces, illustrating the advantages of class-specific projections similar to LDA. In methods of classification such as student success prediction, LDA can perform better than PCA in that it retains discriminative characteristics associated with labels. These tendencies are proven by empirical research in EDM. PCA has been demonstrated to be efficient when there are numerous redundant variables, whilst LDA is more suitable where the aim is to perform classification. Experiments involving PCA or LDA with classifiers like logistic regression, SVMs or Random Forests demonstrate an increase in predictive power and a reduction in computing costs. [22] in his book on statistical learning also highlighted the value of both supervised and unsupervised dimensionality reduction methods in data mining.

2.6 Other Related Applications and Contexts

While much of the EDM literature focuses on academic performance, there are related applications in different domains that emphasize the versatility of machine learning methods. [14] applied augmented reality in teaching probability and sampling, demonstrating how innovative data-driven approaches support learning. [3] explored fusion visualization techniques using machine learning for three-dimensional imaging, representing the cross-domain application of these techniques. [36] investigated clinical interventions using predictive analytics, highlighting that predictive modeling approaches extend well beyond education. These examples reinforce that the techniques reviewed here have generalizability across fields. Additionally, [17] examined student dropout prediction, which relates closely to early-warning approaches such as [2]. These connections illustrate the breadth of predictive modeling applications in both educational and non-educational contexts.

2.7 Summary

Previous research in Educational Data Mining (EDM) and Learning Analytics has relied on classical machine learning models (LR, DT, RF, KNN, SVM) as a strong foundation for predicting student performance. In line with the fact that behavioral features are often more predictive than population features, the widely used xAPI-Edu-Data benchmark highlights this trend, and recent deep ensemble studies provide state-of-the-art comparative studies. While ensemble and deep methods improve performance on data-rich text, tabular foundation models such as TabPFN provide competitive accuracy with minimal tuning, especially for low-dimensional datasets. In terms of dimensionality reduction, LDA generally outperforms PCA in small-feature systems because of its supervised nature. Our study contributes by systematically comparing classical models and TabPFN, analyzing the PCA vs. LDA trade-offs, and benchmarking against a deep fusion method on the same dataset.

Chapter 3

Methodology

3.1 Overview of the Methodology

In this research, we analyze different machine learning methods to see how well they can predict student performance. The objective is to create a smart and data-driven solution that produces accurate results and makes interpretation easy. To do this, we combined a mix of different methods, along with traditional models, ensemble techniques, and modern transformer-based methods. In our research, we go forward sequentially, acquiring and refining data, reducing the number of features, building and evaluating models, and finally performing evaluation and comparison of results.

First, we used the xAPI-Edu-Data dataset, which is very popular and contains information about students' demographic background, educational history, parental involvement with students, and student behavior in class. Given the domain knowledge, one might argue that designing models on a single dataset could lead to overfitting models. Therefore, we applied the models to different datasets. In this technique, we were able to evaluate the performance of the models in different educational environments and also reduced the risk of overfitting data from a single institution.

This technique was developed as a pairwise approach where each layer of the method adds value to the layer above. For example, the way data is preprocessed affects the performance of models such as KNN and SVM. Similarly, feature reduction decisions can create a trade-off in the amount of information used for training; that is, some useful information is retained, and the rest is discarded. Putting these steps together, the experiments are not only technically robust but also practical and meaningful for real learning environments.

3.2 Positioning Our Work

Our work contributes by:

1. Comparing a new foundation model (TabPFN) with classical models (LR, DT, RF, KNN, SVM) on the xAPI-Edu-Data dataset and Secondary Dataset (NEU).
2. Analyzing PCA vs. LDA on a dataset with only 16 features, showing why supervised methods (LDA) perform better than unsupervised Methods (PCA).
3. Directly comparing our results with an in-depth study that uses the same dataset [42].

3.3 Chronological Breakdown of Components

At first, we started the study with data collection. We used the xAPI-Edu-Data dataset, which consists of 480 student records. Each record consists of 16 features and a target label that indicates the student's performance as high, medium, or low. These features include demographic details (gender, nationality, birthplace), educational details (stage, grade, department, subject, semester), behavioral information (hand raised, resource usage, announcements viewed, discussion participation), parental involvement (surveys and satisfaction), and attendance (days of absence). Together, they provide a complete picture of students' learning behavior.

After collecting the dataset, the next step was data preprocessing. This meant preparing the raw data for machine learning. The dataset was mostly clean, but some values were missing. For numeric features, missing values were replaced with the mean. For categorical features, missing values were replaced with the most common value. This step ensured that the dataset was ready for analysis without introducing bias.

Next, we transformed the features into machine-readable form. Categorical variables such as gender and nationality were converted to numbers using one-hot encoding. Ordinal features such as parental satisfaction and days absent were assigned ordinal values. Since behavioral features such as hand-raised and assets visited were on very different scales, we used min-max scaling to normalize between 0 and 1. This was especially important for distance-based models such as KNN and SVM.

After the completion of the preprocessing phase, it enabled splitting the dataset into three parts: training (80%), validation (10%), and testing (10%). Stratified sampling was applied to make sure that the class proportions (high, medium, low) were preserved in all sets.

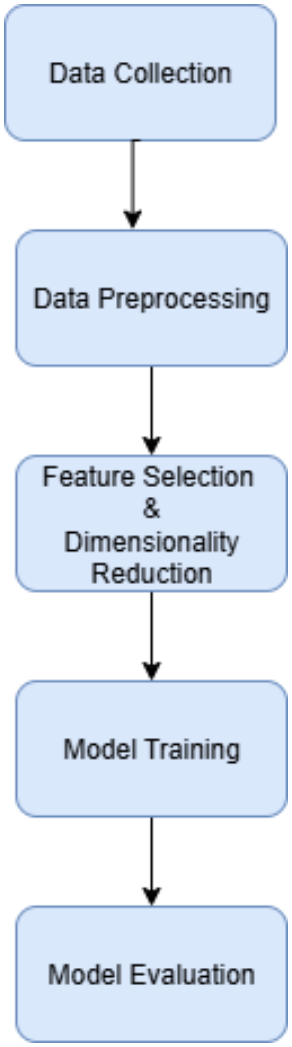


Figure 3.1: Data Preprocessing Pipeline

We tested methods of feature selection and reduction of dimensionality as well. Since the data set contained just 16 features, we were curious to see if having fewer might enhance performance. PCA was used to retain the features with the greatest variance, though it ignores the class labels.

On the other hand, LDA focuses on class separation. The results showed that using the top 10 LDA features generated about 83% accuracy, while using the top 10 PCA features generated only 73%. When we reduced it to 5 features, both methods dropped to about 68%, demonstrating that over-reduction removes useful information.

The next step was model training. We tested several machine learning models like logistic regression (simple and interpretable), decision trees (capture complex patterns), random forests (reduce overfitting and handle interactions), KNN (tested with $k = 5$ and $k = 10$), SVM with RBF kernel (handles nonlinear boundaries), and TabPFN (a modern transformer-based model for tabular data). TabPFN was particularly useful

because it can predict in one step without heavy tuning and performed well on this dataset.

During training, we tuned model parameters using cross-validation. For example, we adjusted the depth and splitting rule in decision trees and random forests, and the regularization strength for SVM. Finally, we evaluated performance using multiple metrics: accuracy (overall accuracy), RMSE (mean root mean square error), MAPE (percentage error), and R^2 (variance explained). Using these two together gave a complete picture of performance, showing not only which models predicted well but also how large their errors were.

3.4 Integration and Interconnections

The methodology is designed as a connected process rather than a set of separate steps. Each stage affects the next one. For example, preprocessing makes sure the data is properly prepared so that the machine learning models can read it correctly. If scaling or encoding is done poorly, models like KNN and SVM may not perform well. In the same way, feature selection using PCA or LDA directly influences model training. Reducing the number of features can sometimes improve results by removing unnecessary noise, but it can also hurt performance if useful information is removed. The choice of models also shows a trade-off between clarity and accuracy. Logistic Regression and Decision Trees are easier to explain and understand, while newer models like TabPFN give better accuracy but are harder to interpret. Finally, the evaluation metrics serve as a bridge between technical results and practical meaning. They help us see not only which model performs best but also why a certain model might be more suitable in real educational settings.

3.5 Clarity and Comprehensiveness in Visual Aids

To make our findings clearer, we included several visual aids alongside the numerical results. A pipeline diagram was used to show the step-by-step process, from collecting the dataset to evaluating the models, helping readers easily follow the workflow. We also created feature importance plots from Random Forest and LDA to highlight which factors, such as the number of raised hands or use of resources, play the biggest role in predicting student performance. In addition, scatter plots from PCA and LDA show how the student groups separate when features are reduced, visually supporting the accuracy differences seen in the numbers. To compare models, we added charts

that summarize accuracy, RMSE, MAPE, and R^2, making it easier to see the strengths and weaknesses of each model. Finally, we used confusion matrices to look at misclassifications, which is important for understanding how well the models perform across the three categories of student achievement.

3.6 Summary of the Methodology

In summary, this research followed a well-structured and clear process from both the preprocessing (data cleaning and feature reduction) to the model testing (training various types of models and performance evaluation using different approaches). We have taken the xAPI-Edu-Data dataset and another dataset in order to verify the reliability of the outcome and for better consistency.

By practicing PCA and LDA features, we examine how reducing features affects performance. We also compared a diversity of models, from simple ones like logistic regression to advanced ones like TabPFN, to cover both traditional and modern approaches.

We used multiple evaluation procedures and used visual representations to make the results both accurate and easy to understand for research and practical use. This solid methodology gives us a complete set of findings, which are explained in detail in the next chapter.

Chapter 4

Results and Discussion

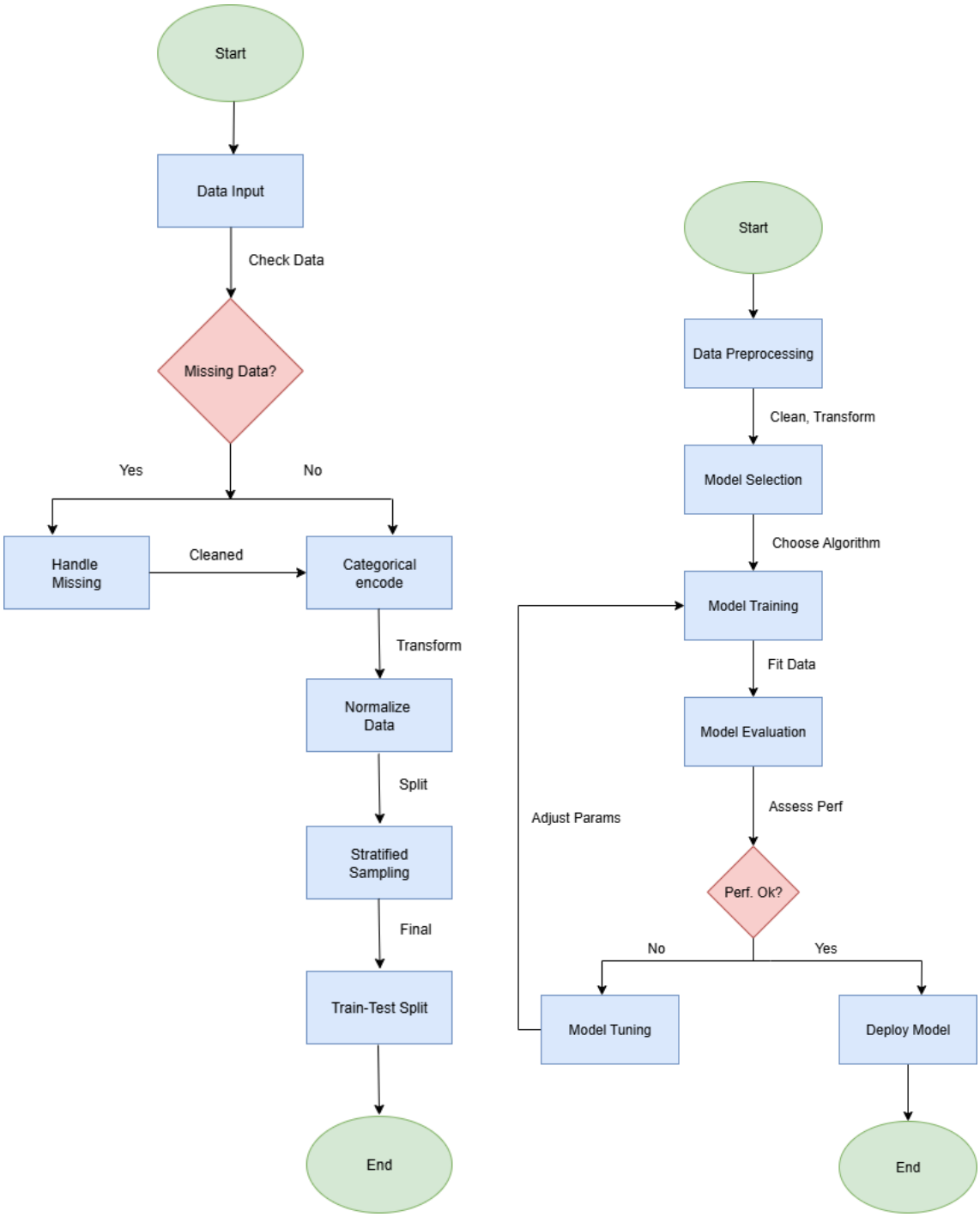
4.1 Datasets and Experimental Setup

This study was carried out by the examination on two different datasets that represent distinct academic contexts. The xAPI-Edu-Data dataset was the first dataset of this experiment that was also adopted by [42] for their deep ensemble learning approach to determine student academic performance. This dataset is highly significant because it includes both behavioral and demographic features such as gender, nationality, place of birth, stage ID, grade ID, section ID, topic, semester, relation, raised hands, visited resources, announcements view, discussion participation, parental survey responses, parental satisfaction, student absence days, and final class labels. These variable shows the details profiles of 480 students on their academic performance.

The second dataset was taken from the work of [46] at Near East University. A questionnaire was used to collect the data across 101 students in three different courses. It divides the data into three categories, which are personal, family and demographic data and finally shows the final performance class level. The dataset includes 33 features, and it was used to measure classical artificial intelligence methods, including Decision Trees, RBF Neural Networks, and Logistic Regression, where 70-88% accuracy was gained by RBFNN based on the courses. The dataset was repurposed for this study to compare TabPFN with findings published in the literature and assess the generalizability of our models.

A complete pipeline of data preprocessing was used on both datasets to ensure reliability and validity across all the models. Using an appropriate encoding scheme, the categorical features were encoded to a numerical representation that helps the algorithms to efficiently and correctly interpret non-numeric features. For preventing the

variables with larger ranges, controlling distance or weight weight-biased model, continuous variables were normalized to a common scale. Additionally, through imputation approaches, missing data were carefully handled to prevent bias and information loss.



(a) Preprocessing **(b)** Model Evaluation

Figure 4.1: Comparison of preprocessing and model evaluation steps.

After preprocessing of the data, the whole dataset was split into subsets where 80% data was used for training, 10% for validation and 10% for testing. Stratified sampling was used to ensure that the proportional proportion of classes was maintained across all subsets, hence mitigating possible issues of class imbalance.

To capture a diverse learning framework, a different set of models was used for the experiment. These included the following: K-Nearest Neighbors with k = 5 and k = 10 (instance-based learning with different neighborhood sizes), Random Forest (an ensemble of decision trees leveraging bagging for improved generalization), Logistic Regression (a linear baseline model), Decision Tree (a straightforward, interpretable tree-based learner), Support Vector Machine with an RBF kernel (a margin-based model capable of handling non-linear decision boundaries), and TabPFN (a cutting-edge transformer-like architecture pre-trained for tabular data tasks). A variety of learning approaches, including linear, tree-based, instance-based, margin-based, and advanced deep learning, were included in this purposefully wide model selection.

Table 4.1: Dataset description and experimental setup summary

Dataset	**Source**	**Feature Size**	**Samples**	**Split (Train/-Val/Test)**	**Preprocessing**
xAPI-Edu-Data	Tang et.al. [42]	16	480	80% / 10% / 10%	Categorical encoding, normalization, missing data handling, stratified sampling
Secondary Academic Dataset	Yilmaz et.al. [46]	33	350	80% / 10% / 10%	Categorical encoding, normalization, missing data handling, stratified sampling

Table 4.2: Comparison of Features between Dataset 1 and Dataset 2

Dataset 1 Feature	**Dataset 2 Feature**
Gender	Sex
Nationality	Country
Place of Birth	Birthplace
StageID	EducationLevel
GradeID	YearGrade
SectionID	ClassSection
Topic	Subject
Semester	Term
Relation	GuardianRelation
RaiseHands	Participation
VisitedResources	ResourceUsage
AnnouncementsView	NoticeBoardViews
Discussion	ForumPosts
ParentAnsweringSurvey	ParentSurveyResponse
ParentSchoolSatisfaction	ParentFeedback
StudentAbsenceDays	Absenteeism
Class	FinalResult

4.2 Presenting the Results

4.2.1 Accuracy and Error Results

The outcomes of the experiment revealed a strong variation of the model performance of the selected classifiers to the xAPI-Edu-Data dataset as shown in Figure 4.3 and Table 4.3. The linear model LR provided an accuracy of 77.08 % and its RMSE of 0.7971 and a MAPE of 23.78 %. The model managed to achieve a very small variance that was expressed via the R^2 of 0.0311 which implies that a simple linear relationship could not effectively demonstrate the trend in student performance.

In the case with RMSE of 0.6455, the Decision Tree model achieved 83.33 % of accuracy. MAPE of 13.37 %. The R^2 value is 0.3647, which is a significant gain 17 compared to LR since trees are capable of establishing the non-linear interaction between features. This shows that the behavioral variables of importance such as visited resources, raised hands and absence days are complexly interconnected and can be captured by DT.

Random Forest classifier with MAPE of 18.06 % and an RMSE of 0.7706 was an 81.25 % accurate classifier. The R^2 value was lower than the DT, at 0.0946, and this value does not capture on deeper variance inside of this dataset, although the ensemble method reduced over-fitting. Random Forest was also predictable, but there is a possibility that

it has averaged several trees and removed any important signal in the very predictive features.

Table 4.3: Summary of model results across Accuracy, RMSE, MAPE, and R^2

Model	**Accuracy**	**RMSE**	**MAPE (%)**	**R^2**
Logistic Regression	0.7708	0.7971	23.78	0.0311
Random Forest	0.8125	0.7706	18.06	0.0946
Decision Tree	0.8333	0.6455	13.37	0.3647
KNN-5	0.6458	1.0206	32.29	-0.5884
KNN-10	0.6562	0.9682	27.95	-0.4295
SVM (RBF)	0.8229	0.6997	19.27	0.2535
TabPFN	0.8438	0.6614	15.45	0.3329

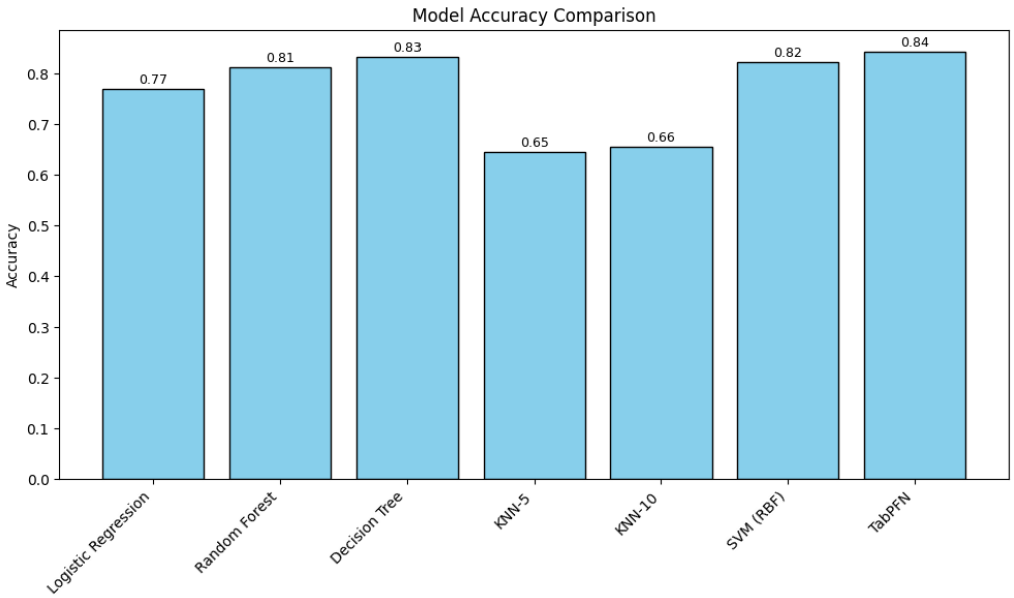
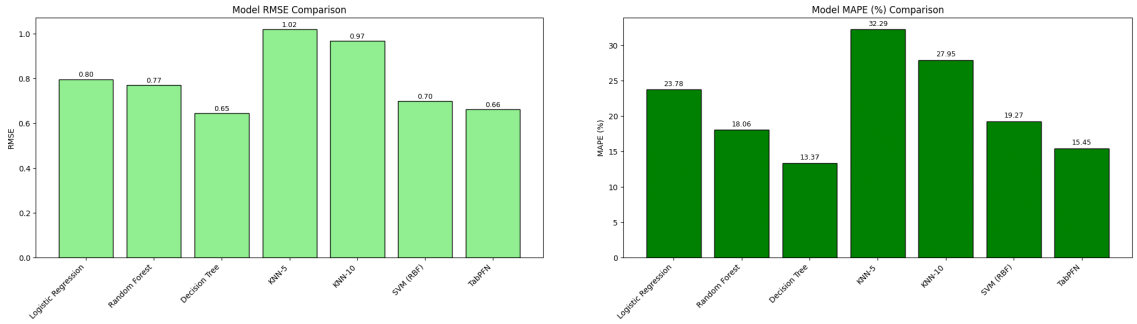


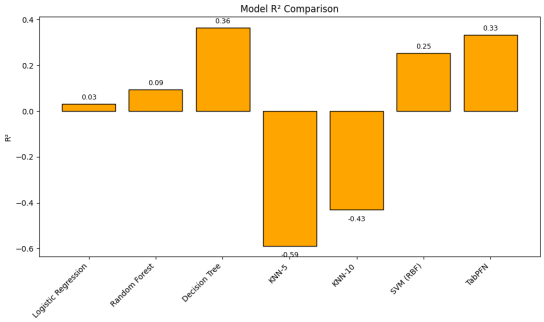
Figure 4.2: Comparison of Model Accuracies

K-Nearest Neighbors was the worst performing model. KNN with k=5 gave an accuracy of 64.58 % with RMSE of 1.0206, MAPE of 32.29 %, and a R^2 of -0.5884. Even though increasing the number of neighbors to k=10 makes the result somewhat better, the accuracy is valued at 65.62 %, RMSE = 0.9682, MAPE = 27.95 %, and R^2 = 0.4295. This type of educational dataset does not fit well with instance-based learning, where the distributions of the features are highly overlapping, which is evidenced by negative R^2 values, which indicates that the model is not doing as well as a simple mean predictor.



(a) RMSE

(b) MAPE



(c) R^2

Figure 4.3: Comparison of RMSE, MAPE, and R^2 across models. (a) RMSE, (b) MAPE, (c) R^2.

The Support Vector Machine using RBF kernel gave good results, with an accuracy of 82.29 %, RMSE of 0.6997, MAPE of 19.27 %, and R^2 of 0.2535. This shows that feature based classification was able to successfully divide academic performance groups using boundaries. Even though it failed to significantly out-perform tree based methods, SVM margin maximization principle allowed it to generalize better than the KNN and LR.

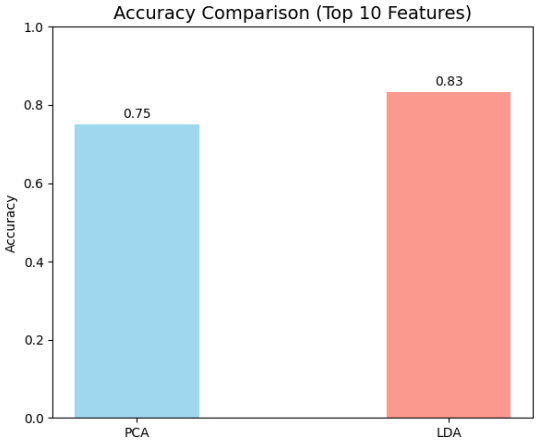
More Recent Transformer based model TabPFN gives the best results on the xAPI dataset. It has an accuracy of 84.38 % resulting in RMSE of 0.6614, MAPE of 15.45 %, and R^2 of 0.3329. The findings suggest that the transformer based frameworks that are structured data oriented can be able to identify complex patterns of student performance in addition to generalization. TabPFN performed marginally better than DT and it was more stable. TabPFN provided better accuracy and less predictable error dispersion compared to the Decision Tree.

TabPFN achieved 77 % accuracy on the second dataset of. Near East University [22]. This accuracy was higher than the 70 % overall accuracy of RBF Neural Networks, and falls within the 70 to 88 % accuracy of RBF Neural Networks in the original study. This illustrates that TabPFN still has predictive power even when it is working with smaller

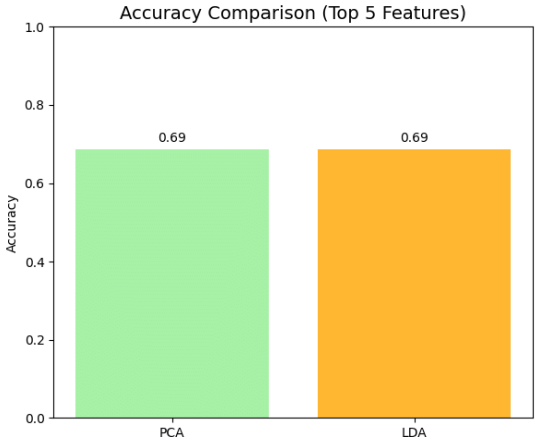
datasets (also questionnaire based). Its competitive performance without requiring extensive parameter tuning makes it a viable substitute for neural-based approaches in educational forecasting.

4.2.2 Feature Selection and Dimensionality Reduction Results

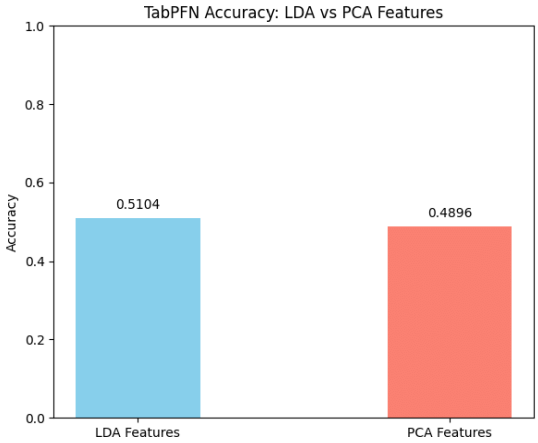
Results of Feature Selection and Dimensionality Reduction is shown in 4.4. The feature selection process of PCA and LDA also gave another perspective of this experiment. Although the best 10 features picked using LDA, the model achieved an accuracy of 83%. In comparison, the 10 features with the highest rank chosen by PCA only achieved 73% of accuracy.



(a) Top 10 Features



(b) Top 5 Features



(c) Bottom 5 Features

Figure 4.4: Comparison of PCA and LDA model performances across Top 10, Top 5, and Bottom 5 selected features.

PCA and LDA as feature selection analysis further revealed some information. Using the LDA to select the top 10 features, the models obtained 83% accuracy, whereas the PCA-based top 10 features obtained only 73% accuracy. The reason is that LDA is controlled, it maximizes differences across 19 classes that are significant in classification tasks such as predicting academic performance. This is the reason why LDA is better. Alternatively, PCA only preserves total variance, which may not be similar to predictive data.

PCA and LDA both reached a 68% accuracy when the number of features was again trimmed down to the top five. This significant drop implies that a broader scope of the interconnected factors determine academic achievement. Absence days, parental satisfaction, and visited resources are important predictors that have a significant impact on academic outcomes, but they were removed as a result of assertive reduction of the dataset.

Table 4.4: Accuracy of Different Feature Reduction Strategies

Reduction Strategy	**Accuracy (%)**
Full Features (16)	Baseline
Top-10 PCA	73
Top-10 LDA	83
Top-5 PCA	68
Top-5 LDA	68

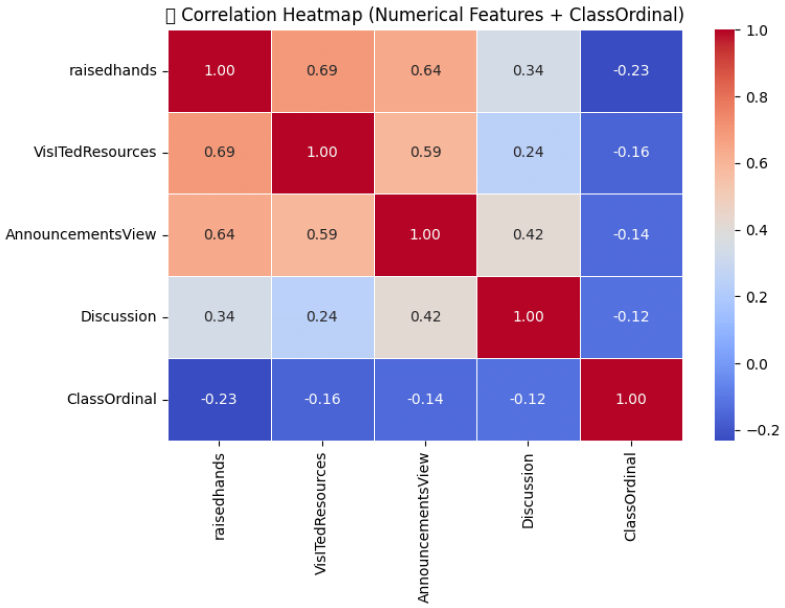
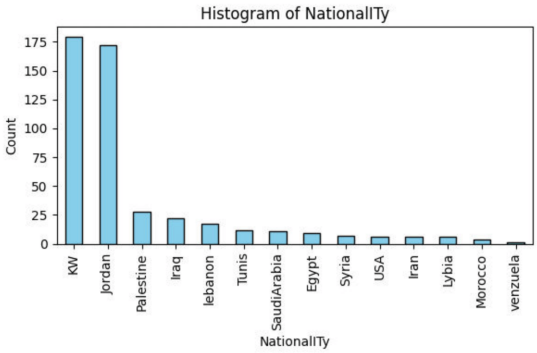
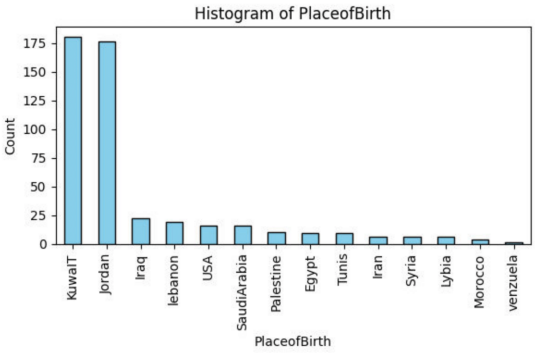


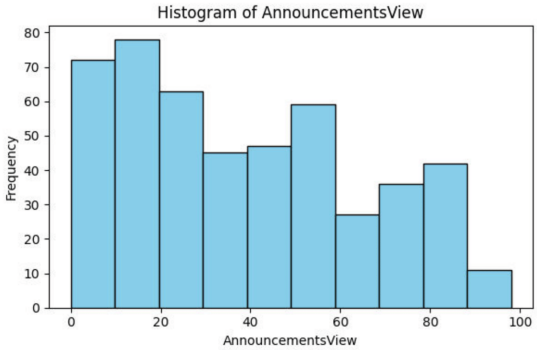
Figure 4.5: Feature correlations with student performance.



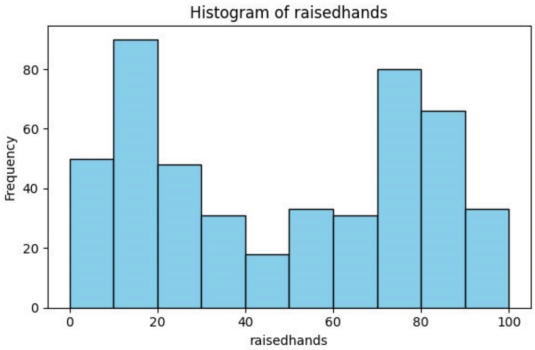
(a) Nationality



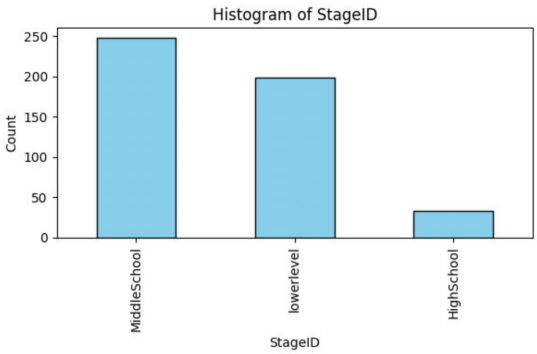
(b) Placeofbirth



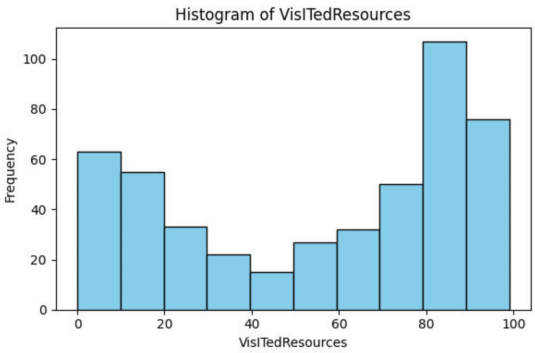
(a) Announcementview



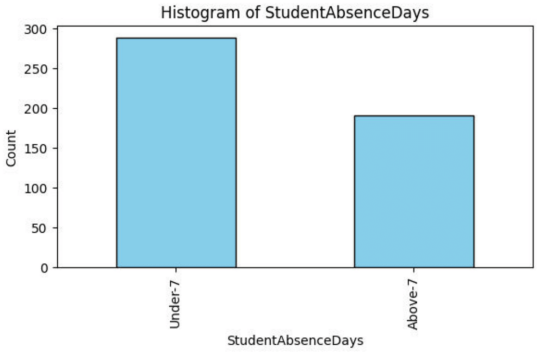
(b) Raisehands



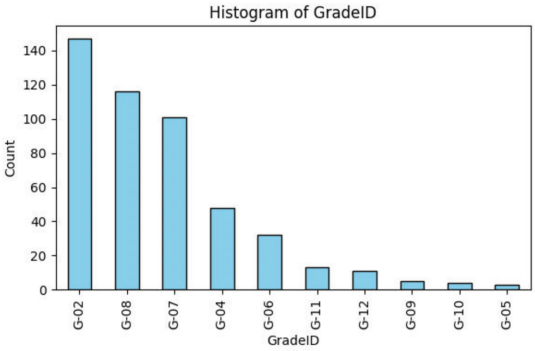
(c) Stageid



(d) Visitedresource



(e) Studentabsenseday



(f) Gradeid

Figure 4.7: Histograms of important features.

Interestingly, due to the number of features, only 16 in the dataset, the effort to compare best and worst features gave identical sets. This overlap points to the limitations of the feature-based experiments in small datasets. The fact that PCA and LDA differ obviously in the top 10 cases, however, indicates that supervised feature reduction is more appropriate in this type of educational problem.

4.3 Interpreting the Results

The findings indicate that simple linear models including LR are not sufficient to make the academic performance complexity prediction. Interacting non-linearly such as behavioral and demographic variables are most effectively represented using tree-based and modern models, such as TabPFN. TabPFN was superior in generalization over the dispersion of errors but the most powerful candidates were DT and TabPFN.

The performance difference is also noticeable compared to that of [42] as shown in 4.5. Their deep ensemble model gave an R^2 of 0.7430, MAPE of 9.75 %, and RMSE of 1.66. Their R^2 was greater than twice that of our best performing TabPFN model and their MAPE was much lower although we had a larger RMSE. This means that their method reduced the percentage-based prediction error and enhanced the explanatory power. However, their model required computationally expensive pools of various deep belief networks and complex optimization methods such as the particle swarm optimization. By contrast, we demonstrate that lighter machine learning models can still achieve a competitive accuracy at much lower cost of computation which is more applicable to resource limited educational institutions.

The PCA and LDA experiments provide a considerable rationale behind the difference in the model performance. Critical predictors are eliminated by reducing the quantity of features beneath a specified threshold. An example is the absence days and parental satisfaction which are extrinsic factors that affect learning, but characteristics such as raised hands and resources visited which assess student engagement. Since the field of prediction of academic performance is multifactorial in nature, dimensionality reduction must be applied carefully. The best results of LDA demonstrate that the supervised methods are more suitable to maintain discriminative power.

According to testing on the second dataset, decision tree and TabPFN remained reliable, whereas, KNN remained poor. This demonstrates the cross context strength of our results. Our models continue to be under explaining as much variance as the results of [42]. though, which demonstrates the potential of advanced deep learning ensembles when power consumption is not a problem.

The second dataset was tested with an accuracy of 77, which confirmed that TabPFN continues working rationally well [46]. This makes it higher than their average combination of multi-course results of 70% and within the band of RBFNN performance (70 to 88)% as mentioned in the original study as shown in 4.8. Importantly, TabPFN proved its adaptability to different academic settings due to its ability to show competitive results without the need to optimize hyperparameters, but RBFNN was dataset-specifically optimized (learning rate, iterations, and clusters).

Table 4.5: Side-by-side comparison of the results

Model / Dataset	**xAPI-Edu-Data**	**Near East Univ. Dataset**	**Notes**
TabPFN (Proposed)	Accuracy = 84.38% RMSE = 0.6614 MAPE = 15.45% R^2 = 0.3329	Accuracy = 77% (within 70–88% band of RBFNN)	Competitive accuracy, stable, minimal tuning
Tang et al. [42] (Deep Ensemble)	RMSE = 1.6562 MAPE = 9.75% R^2 = 0.7430 PLCC = 0.8619 SROCC = 0.8582 CCC = 0.8571	Not tested	Very strong, but computationally heavy (deep belief networks + PSO)
Yılmaz & Şekeroğlu [46] (RBFNN)	–	Accuracy = 70–88% depending on course	High but dataset-specific, required tuning

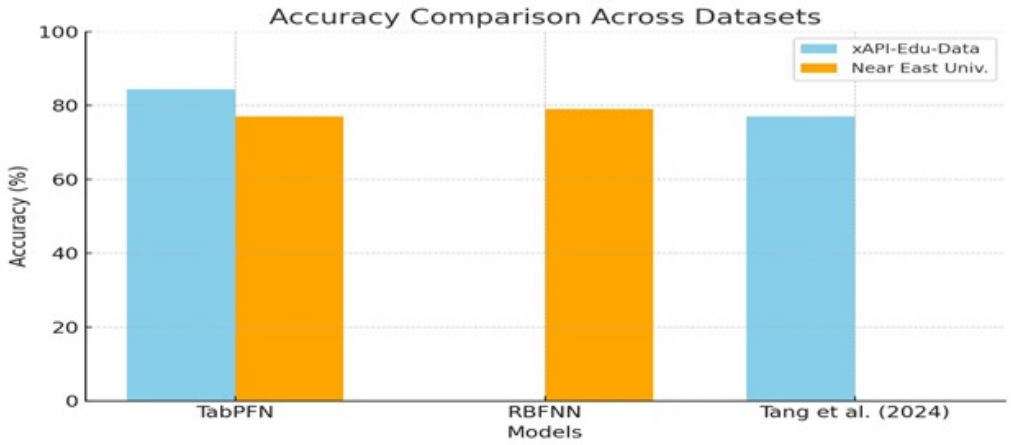


Figure 4.8: Visualization of accuracy comparison across datasets

This demonstrates that RBFNN [46] has a high dataset-specific performance, and deep ensemble models such as [42] might be more able to explain the variance ((R^2 = 0.7430), but TabPFN can offer an advantageous compromise: effective competence in both datasets with minimum tuning and reduced computational expense.

4.4 Challenges and Limitations

The metrics show the effectiveness of our model but also reveal some limitations. The R^2 coefficients (even those of the strongest-performing models) show that a large component of variance in student performance is still inexplicable. This is an indication that aspects not directly represented in the model's features but clearly also influencing learning behavior, like intrinsic motivation or socio-economic status or real-life situations, are crucial.

The practical interpretation of the RMSE is that it is a measure of the error in prediction. In cases where the model predicts near a critical at-risk cutoff for a student, this inherent error should be taken into account. As such, predictions are to be viewed as probabilistic warning signs rather than definitive facts, and interventions should target at least those students around the risk threshold rather than merely a subset of students scoring above a cut-off.

Another was the trade-off between model accuracy and interpretability. Although TabPFN produced better accuracy, it is more complex for its internal functioning to be interpreted than that of the decision tree. This "black-box" behavior is a significant drawback in an educational setting, where explanation of the reasons for a prediction is essential in order to design the intervention. In addition, the scale and type of data limited this study. The generalizability of the results is restricted by the small size of the second data set (n=101), and capturing bias may be introduced in questionnaires as a result of self-report.

4.5 Implications and Significance of Findings

This study has a few crucial educational and research implications. We have demonstrated the feasibility of an early warning system based on data-driven interventions. Correctly identifying those students most at risk early in the school year allows for targeted resources in the form of tutoring, counselling or other support to move away from a reactive approach to student care to one that is preventative.

This study demonstrates that sophisticated models, such as TabPFN, are feasible for analysing students' tabulated data. TabPFN performs better with a few hyper-params without heavy tuning; as such, it is both a practical and powerful way to train against traditional ML models or complex neural networks. The achievement is a new high for the domain and tends to promote efficient yet elaborate procedures in educational data mining.

Feature importance analysis (supported by the performance of models like decision trees and feature reduction analysis using LDA) identifies drivers of student success, such as raising hands, visiting resources and attendance. This offers educators and policymakers empirically based guidance to inform strategic decisions, for example, encouraging certain learning behaviours or enhancing student engagement platforms.

Also, the knowledge gained in the course of this study can be used to create adaptive learning systems that track the progress of students and offer recommendations on the basis of these results. Further research can investigate incorporating these predictive systems in current learning management systems, allowing instructors to make interventions based on timely and data-driven decision-making. This work will eventually lead to the larger idea of applying artificial intelligence to build more inclusive, responsive, and effective learning spaces.

Chapter 5

Conclusion

5.1 Rephrasing Research Objectives

This study aimed to discuss the data-driven methods of predicting student academic performance. In particular, the study focused on assessing the performance of classical machine learning models and a transformer-based strategy, as well as on how dimensionality reduction techniques (PCA and LDA) can be used to help locate important features. The study was designed to endorse the approaches of early intervention in education by consideration of both predictive accuracy and feature interpretability.

5.2 Summary of Key Findings

Results indicated that Decision Tree was the best model by accuracy with interpretability as opposed to transformer-based model which had the best overall performance among classical models. The PCA and LDA results indicated that the supervised dimensionality reduction (LDA) is able to retain more discriminative information, compared to PCA, and behavioral characteristics, including participation, resource utilization, and absence days appear to be the most significant predictors. The strength of these feature insights was verified by the validation with the use of a secondary dataset and it was shown that the findings are not data-dependent. The results underscore the importance of integrating predictive models and feature analysis in order to come up with sound academic prediction models.

5.3 Future Research Advice

The feature space can be further expanded in future research adding any academic, behavioral, and socioeconomic factors. High-level ensemble methods, e.g. XGBoost or hybrid deep learning models, might also enhance performance, and longitudinal datasets could be used to understand performance over time. Explainable AI represents another significant way that can be used to increase the transparency of the forecasting systems and be more adopted in the educational institutions.

5.4 Concluding Remarks

This paper proves that machine learning provides a viable and efficient method of predicting academic performance. Although more sophisticated models are more accurate, interpretable models such as Decision Trees are still useful in resource-constrained learning environments. Combining the analysis of the feature makes the predictions more reliable, and the at-risk students can also be intervened beforehand. Finally, the results indicate that an intelligent application of predictive modeling has the potential to facilitate more ethical, data-driven educational activities and establish the foundation of potential future developments in effective and interpretable forecasting systems.

References

[1] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[2] G. Akçapınar, A. Altun, and P. Aşkar, "Using learning analytics to develop early-warning system for at-risk students," *International Journal of Educational Technology in Higher Education*, vol. 16, no. 1, pp. 1–20, 2019.

[3] K. Ali et al., "Fusion visualization technique to improve a three-dimensional isotope-selective ct image based on nuclear resonance fluorescence with a gamma-ct image," *Applied Sciences*, vol. 11, no. 24, p. 11 866, 2021.

[4] A. Allahyari et al., "Comparing efficacy and safety of p013, a proposed pertuzumab biosimilar, with the reference product in her2-positive breast cancer patients: A randomized, phase iii, equivalency clinical trial," *BMC cancer*, vol. 22, no. 1, p. 960, 2022.

[5] E. Alyahyan and D. Düştegör, "Predicting academic success in higher education: Literature review and best practices," *International journal of educational technology in higher education*, vol. 17, no. 1, p. 3, 2020.

[6] S. Ö. Arik and T. Pfister, "Tabnet: Attentive interpretable tabular learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, 2021, pp. 6679–6687.

[7] R. Baker et al., "Data mining for education," *International encyclopedia of education*, vol. 7, no. 3, pp. 112–118, 2010.

[8] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[9] A. Bhusal, "Predicting student's performance through data mining," *arXiv preprint arXiv:2112.01247*, 2021.

[10] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 144–152.

[11] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[12] C.-H. Chen, S. J. Yang, J.-X. Weng, H. Ogata, and C.-Y. Su, "Predicting at-risk university students based on their e-book reading behaviours by using machine learning classifiers," *Australasian Journal of Educational Technology*, vol. 37, no. 4, pp. 130–144, 2021.

[13] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[14] Q. Conley, R. K. Atkinson, F. Nguyen, and B. C. Nelson, "Mantarayar: Leveraging augmented reality to teach probability and sampling," *Computers & Education*, vol. 153, p. 103 895, 2020.

[15] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[16] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE transactions on information theory*, vol. 13, no. 1, pp. 21–27, 1967.

[17] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers, "Predicting students drop out: A case study," in *Proceedings of the 2nd International Conference on Educational Data Mining, EDM 2009, July 1-3, 2009. Cordoba, Spain*, 2009, pp. 41–50.

[18] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[19] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter, "Efficient and robust automated machine learning," *Advances in neural information processing systems*, vol. 28, 2015.

[20] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[21] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," *Advances in neural information processing systems*, vol. 34, pp. 18 932–18 943, 2021.

[22] T. Hastie, *The elements of statistical learning: Data mining, inference, and prediction*, 2009.

[23] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," 2008.

[24] N. Hollmann, S. Müller, K. Eggensperger, and F. Hutter, "Tabpfn: A transformer that solves small tabular classification problems in a second," *arXiv preprint arXiv:2207.01848*, 2022.

[25] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013.

[26] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20 150 202, 2016.

[27] D. Kabakchieva, "Predicting student performance by using data mining methods for classification," *Cybernetics and information technologies*, vol. 13, no. 1, pp. 61–72, 2013.

[28] G. Ke et al., "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.

[29] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Predicting students'performance in distance learning using machine learning techniques," *Applied Artificial Intelligence*, vol. 18, no. 5, pp. 411–426, 2004.

[30] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[31] F. Marbouti, H. A. Diefes-Dux, and K. Madhavan, "Models for early prediction of at-risk students in a course using standards-based grading," *Computers & Education*, vol. 103, pp. 1–15, 2016.

[32] H. Pallathadka, A. Wenda, E. Ramirez-Asís, M. Asís-López, J. Flores-Albornoz, and K. Phasinam, "Classification and prediction of student performance data using various machine learning algorithms," *Materials today: proceedings*, vol. 80, pp. 3782–3785, 2023.

[33] Z. Papamitsiou and A. A. Economides, "Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence," *Journal of educational technology & society*, vol. 17, no. 4, pp. 49–64, 2014.

[34] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: Unbiased boosting with categorical features," *Advances in neural information processing systems*, vol. 31, 2018.

[35] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.

[36] L. V. Ravn-Nielsen et al., “Effect of an in-hospital multifaceted clinical pharmacist intervention on the risk of readmission: A randomized clinical trial,” *JAMA internal medicine*, vol. 178, no. 3, pp. 375–382, 2018.

[37] M. T. Ribeiro, S. Singh, and C. Guestrin, “" why should i trust you?" explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[38] C. Romero and S. Ventura, “Educational data mining: A review of the state of the art,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (applications and reviews)*, vol. 40, no. 6, pp. 601–618, 2010.

[39] M. Sakib, S. Mustajab, and M. Alam, “Ensemble deep learning techniques for time series analysis: A comprehensive review, applications, open issues, challenges, and future directions,” *Cluster Computing*, vol. 28, no. 1, p. 73, 2025.

[40] A. Silva, “Using data mining to predict secondary school student performance,” 2008.

[41] G. Somepalli, M. Goldblum, A. Schwarzschild, C. B. Bruss, and T. Goldstein, “Saint: Improved neural networks for tabular data via row attention and contrastive pre-training,” *arXiv preprint arXiv:2106.01342*, 2021.

[42] B. Tang, S. Li, and C. Zhao, “Predicting the performance of students using deep ensemble learning,” *Journal of Intelligence*, vol. 12, no. 12, p. 124, 2024.

[43] W. M. Van der Aalst, “The application of petri nets to workflow management,” *Journal of circuits, systems, and computers*, vol. 8, no. 01, pp. 21–66, 1998.

[44] A. Vaswani et al., “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

[45] P. Yadav et al., “Investigation and empirical analysis of transfer learning for industrial iot networks,” *IEEE Access*, 2024.

[46] N. Yılmaz and B. Sekeroglu, “Student performance classification using artificial intelligence techniques,” in *International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions*, Springer, 2019, pp. 596–603.