



Enhancing Student Performance Prediction Using Real-Time Data and Explainable Artificial Intelligence in Higher Education

P.H.H.C.Dewduni
Reg. No.: MS24010276

A THESIS
SUBMITTED TO
SRI LANKA INSTITUTE OF INFORMATION TECHNOLOGY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE IN INFORMATION TECHNOLOGY

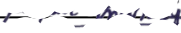
December 2025

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.



Prof. Pradeep Abeygunawardhana

Approved for MSc. Research Project:



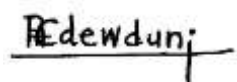
MSc in IT Programme Co-ordinator, SLIIT

Approved for MSc:

Head of Graduate Studies, FoC, SLIIT

DECLARATION

This is to certify that the work is entirely my own and not of any other person, unless explicitly acknowledged (including citation of published and unpublished sources). The work has not previously been submitted in any form to the Sri Lanka Institute of Information Technology or to any other institution for assessment for any other purpose.



Sign:

P.H.H.C.Dewduni

Date:11/12/2025.....

ABSTRACT

Enhancing Student Performance Prediction Using Real-Time Data and Explainable Artificial Intelligence in Higher Education

Dewduni.P.H.H.C

MSc. in Information Technology

Supervisor: Prof. Pradeep Abeygunawardhana

December 2025

Accurate prediction of student performance has become a strategic priority in higher education, driven by increasing learner diversity, rising dropout rates, and limited institutional resources. Traditional statistical models based on historical records are limited in capturing the dynamic and multidimensional nature of student learning. This study proposes a real-time predictive framework that integrates machine learning with explainable artificial intelligence to achieve both predictive accuracy and interpretability. Data were collected from 2175 undergraduate students from Faculty of Humanities & Social Sciences, University of Ruhuna across five course modules, combining institutional academic and learning management system records with a structured survey capturing financial and psychosocial indicators. After a systematic preprocessing pipeline, five machine learning classifiers were developed and evaluated, including Logistic Regression, Decision Tree, K Nearest Neighbours, Support Vector Machine, and Random Forest. Model performance was assessed using accuracy, precision, recall, F1 score, and area under the curve. Random Forest achieved the highest accuracy of 96.92%, while Support Vector Machine achieved the strongest discriminative capability with the highest area under the curve. To address interpretability, SHapley Additive exPlanations provided global feature attributions, and Local Interpretable Model Agnostic Explanations generated case-specific insights. Results consistently identified attendance and assignment performance as dominant predictors, while behavioural engagement and financial and psychological dimensions offered complementary contributions. A decision support dashboard was designed to operationalize these insights by enabling early identification of at-risk learners, explaining contributing factors, and generating individualized reports for timely intervention. The proposed framework demonstrates that robust predictive accuracy and interpretability can be achieved simultaneously, providing a practical tool for enhancing student retention, equitable support, and efficient resource allocation in higher education.

Keywords: *Student performance prediction, machine learning, explainable artificial intelligence, learning analytics, dashboard visualization, higher education*

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to my research supervisor, Prof. Pradeep Abeygunawardhana, for his invaluable guidance, continuous support, and constructive feedback throughout this research journey. His expertise, encouragement, and patience have been instrumental in shaping the direction of this study and in helping me overcome challenges along the way.

I am also sincerely grateful to the Sri Lanka Institute of Information Technology (SLIIT), particularly the Faculty of Graduate Studies and Research, for providing an enabling academic environment and the necessary resources to carry out this research. I extend my appreciation to the faculty members and colleagues whose insightful suggestions and constructive criticism contributed significantly to the refinement of this work.

I would like to extend my sincere appreciation to the University of Ruhuna, Faculty of Humanities and Social Sciences, for kindly granting permission to carry out the data collection for this study. I am profoundly thankful to Dr. Kokila Ramanayaka for his generous support and guidance during this process. My heartfelt thanks also go to all the academic and non-academic members of the faculty for their kind cooperation, and to the students who participated in the survey for their valuable time and thoughtful contributions that made this research possible.

Finally, I wish to express my heartfelt thanks to my family and friends for their unwavering love, motivation, and belief in my abilities. Their patience and constant encouragement have been a source of strength and inspiration throughout this academic journey.

TABLE OF CONTENTS

DECLARATION.....	ii
ABSTRACT	iii
ACKNOWLEDGEMENT.....	iv
TABLE OF CONTENTS	v
List of Figures	ix
List of Tables.....	xii
Chapter 1 Introduction.....	1
1.1 Background of Study.....	1
1.2 Problem Background.....	4
1.3 Problem Statement.....	6
1.4 Research Questions	7
1.4.1 Main Research Question.....	7
1.4.2 Sub-Questions.....	7
1.5 Research Objectives	8
1.5.1 Main Objective	8
1.5.2 Specific Objectives.....	9
1.6 Research Contributions	10
1.7 Significance of the Study.....	11
1.8 Chapter Summary	12
Chapter 2 Literature Review	13
2.1 Introduction to Educational Data Mining (EDM)	13
2.2 Traditional Approaches to Student Performance Prediction	14
2.3 Machine Learning in Student Performance Prediction.....	15
2.4 Behavioural, Financial, and Psychosocial Indicators	16
2.4.1 Behavioural Indicators.....	17
2.4.2 Financial Indicators	17
2.4.3 Psychosocial Indicators	17
2.4.4 Towards Multidimensional Predictive Models.....	17
2.5 Benefits and Limitations of XAI	18
2.6 Chapter Summary	19
Chapter 3 Methodology	20
3.1 Introduction	20
3.2 Research Framework.....	20
3.3 Data Collection.....	21
3.3.1 Population and Context	21
3.3.2 Sampling Strategy	22
3.3.3 Data Sources and Feature Categories (Academic, Behavioral, Financial, Psychosocial, Accommodation)	23

3.3.4 Ethical Considerations	25
3.3.5 Summary of Features	26
3.4 Data Preprocessing and Feature Engineering	26
3.4.1 Dataset Import and Initial Exploration	27
3.4.2 Cleaning and Standardization of Columns	28
3.4.3 Column Grouping and Structural Preparation	29
3.4.4 Type Conversion	29
3.4.5 Handling Missing Values	30
3.4.6 Identifier Removal and Categorical Feature Encoding	31
3.4.7 Outlier Detection and Treatment	32
3.4.8 Correlation Analysis	33
3.5 Statistical Analysis	34
3.5.1 Descriptive Statistical Analysis	34
3.5.2 Inferential Statistical Tests	35
1. Chi-Square Test of Independence	35
3.6 Model Development	37
3.6.1 Logistic Regression	38
3.6.2 Decision Tree	45
3.6.3 K-Nearest Neighbours (KNN)	52
3.6.4 Support Vector Machine (SVM)	58
3.6.5 Random Forest	65
3.7 Hyperparameter Tuning	72
3.7.1 Logistic Regression	72
3.7.2 Decision Tree	72
3.7.3 K-Nearest Neighbours (KNN)	73
3.7.4 Support Vector Machine (SVM)	73
3.7.5 Random Forest	73
3.7.6 Comparative Outcome of Tuning	74
3.8 Model Evaluation	74
3.8.1 Accuracy	75
3.8.2 Precision	75
3.8.3 Recall	76
3.8.4 F1-Score	76
3.8.5 Area Under the Curve (AUC)	77
3.9 Explainability and Dashboard Integration	77
3.9.1 Explainability with SHAP	77
3.9.2 Dashboard Design and Integration	78
3.10 Development Environment and Tools	78
3.10.1 Python	78

3.10.2 Google Collaboratory	78
3.10.3 NumPy.....	79
3.10.4 Pandas.....	79
3.10.5 Scikit-learn	79
3.10.6 Seaborn and Matplotlib	79
3.10.7 SHAP.....	79
3.10.8 Flash and Dashboard Integration	80
3.11 Chapter Summary	80
Chapter 4 Results and Analysis.....	81
4.1 Statistical Analysis Results.....	81
4.1.1 Descriptive Statistical Analysis	81
4.1.2 Inferential Statistical Tests	82
4.1.3 Visual Exploration of Non-Linear Patterns	83
4.2 Descriptive and Correlational Analysis.....	84
4.2.1 Academic Indicators	84
4.2.2 Distribution of Pre-processed Features.....	85
4.2.3 Correlation Analysis	85
4.3 Comparative Evaluation of Models	86
4.3.1 Accuracy.....	86
4.3.2 Precision	87
4.3.3 Recall.....	88
4.3.4 F1-Score	89
4.3.5 Area Under the Curve (AUC).....	89
4.4 Confusion Matrix Analysis.....	90
4.4.1 Logistic Regression	90
4.4.2 Decision Tree.....	91
4.4.3 K-Nearest Neighbours (KNN).....	92
4.4.4 Support Vector Machine (SVM)	93
4.4.5 Random Forest.....	94
4.5 ROC Curve and AUC Analysis.....	95
4.5.1 Logistic Regression	95
4.5.2 Decision Tree.....	96
4.5.3 K-Nearest Neighbours (KNN).....	96
4.5.4 Support Vector Machine (SVM)	97
4.5.5 Random Forest.....	98
4.6 Dashboard Demonstration	99
4.6.1 Dashboard Layout and Components.....	99
4.6.2 Usability and Accessibility	102
4.6.3 Institutional Value	102

4.7 Chapter Summary	102
Chapter 5 Limitations and Future Work.....	104
5.1 Limitations of the Study	104
5.2 Future Work	105
5.3 Chapter Summary	106
Chapter 6 Conclusion	107
References	108
Appendix	114
Appendix 1: Ethical Approval Form	114
Appendix 2: Dashboard Run Guidelines	115
Appendix 3: Dashboard Code	115

List of Figures

Figure 3.1 Proposed methodology steps for student performance prediction.	21
Figure 3.2 Features Collected and Their Characteristics.....	26
Figure 3.3 preprocessing workflow.....	27
Figure 3.4 Table Descriptive Statistics of the Dataset.....	27
Figure 3.5 Code Implementation for Column Cleaning and Duplicate Removal	28
Figure 3.6 Conversion of LMS Frequency Values from Strings to Numeric Format	28
Figure 3.7 Code for Column Grouping and Structural Preparation.....	29
Figure 3.8 Code Implementation for Type Conversion of Features	30
Figure 3.9 Mode Imputation for Categorical State Variables.....	30
Figure 3.10 KNN Imputation for Numeric Variables.....	30
Figure 3.11 Manual Mapping of Categorical Variables and Removal of Identifiers	32
Figure 3.12 One-Hot Encoding of the Gender Variable.....	32
Figure 3.13 StandardScaler Implementation for Numeric Features	33
Figure 3.14 Distribution of Features After Scaling	33
Figure 3.15 Spearman Correlation Heatmap of Features and Academic Outcomes	34
Figure 3.16 Descriptive statistics Analysis.....	35
Figure 3.17 1. Chi-Square Test of Independence	36
Figure 3.18 T Test comparing attendance between male and female students.....	36
Figure 3.19 T Test comparing financial support between at risk and not at-risk students.	36
Figure 3.20 scatter plots to explore nonlinear relationships between academic factors and student risk classification.....	37
Figure 3.21 Installing and importing required libraries.....	39
Figure 3.22 : Loading the dataset from the mounted drive.....	39
Figure 3.23 Preparing features and target after removing identifiers.	40
Figure 3.24 split preserving class ratios.	41
Figure 3.25 Leakage-safe pipeline (SMOTE → StandardScaler → Logistic Regression).	41
Figure 3.26 Hyperparameter search space for Logistic Regression.	41
Figure 3.27 Grid search with 5-fold cross-validation.....	42
Figure 3.28 Model evaluation.....	43
Figure 3.29 Regularization path analysis	43
Figure 3.30 ROC–AUC and threshold evaluation.....	44
Figure 3.31 Installing and importing required libraries.....	46
Figure 3.32 Loading and preparing the dataset from the mounted drive.....	46
Figure 3.33 Preparing features and target after removing identifiers and state fields.	47
Figure 3.34 Train–test protocol with stratification	47
Figure 3.35 Leakage-safe pipeline (SMOTE → Decision Tree).....	48
Figure 3.36 Hyperparameter search space for Decision Tree.....	48
Figure 3.37 Grid search with 5-fold cross-validation.....	49
Figure 3.38 Model evaluation and performance visualization.	50
Figure 3.39 Training vs testing accuracy by tree depth.....	50
Figure 3.40 ROC curve and AUC for the tuned Decision Tree model.....	51
Figure 3.41 Installing and importing required libraries.....	53
Figure 3.42 Loading and preparing the dataset from the mounted drive.....	54
Figure 3.43 Feature definition and leakage control	54
Figure 3.44 Applying KNN imputation to handle missing values.	55
Figure 3.45 Train–test protocol with stratification	55
Figure 3.46 Standardizing features for kNN distance metrics.....	55
Figure 3.47 SMOTE oversampling applied to training data.....	56

Figure 3.48 Grid search with 5-fold cross-validation for kNN optimization.	56
Figure 3.49 Model evaluation and confusion matrix visualization.	57
Figure 3.50 Training vs testing accuracy for varying neighborhood sizes.	57
Figure 3.51 ROC curve and AUC for the tuned kNN model.	58
Figure 3.52 Installing and importing required libraries.	59
Figure 3.53 Loading and preparing the dataset from the mounted drive.	60
Figure 3.54 Feature definition and leakage control	61
Figure 3.55 Applying KNN imputation for missing data.	61
Figure 3.56 Feature standardization using StandardScaler.	61
Figure 3.57 Applying SMOTE for balanced class distribution.	62
Figure 3.58 Train–test split with reproducibility	62
Figure 3.59 Hyperparameter tuning with GridSearchCV	63
Figure 3.60 Model evaluation and confusion matrix visualization.	63
Figure 3.61 Training vs testing accuracy for varying C values.	64
Figure 3.62 ROC curve and AUC for the tuned SVM model.	65
Figure 3.63 Environment and library setup	66
Figure 3.64 Data access and loading	67
Figure 3.65 Feature definition and leakage control	67
Figure 3.66 KNN-based missing value imputation	68
Figure 3.67 Encoding and feature scaling	68
Figure 3.68 Train–test split and class rebalancing	69
Figure 3.69 Hyperparameter tuning with GridSearchCV	69
Figure 3.70 Model evaluation and confusion matrix visualization.	70
Figure 3.71 Comparison of training and testing accuracy.	71
Figure 3.72 Model persistence and deployment preparation.	71
Figure 4.1 Descriptive statistics of academic and behavioural variables	81
Figure 4.2 Chi-Square Test of Independence between financial assistance and risk classification	82
Figure 4.3 T-Test comparing attendance between male and female students	83
Figure 4.4 T-Test comparing financial support between at-risk and not-at-risk students	83
Figure 4.5 Scatter plot of C2_Attendance vs Risk Classification	83
Figure 4.6 Scatter plot of C2_Attendance vs Risk Classification	
Figure 4.7 Scatter plot of Financial Support vs Risk Classification	84
Figure 4.8 Scatter plot of Academic Stress Level vs Risk Classification	
Figure 4.9 Module-level averages for attendance and assessments	84
Figure 4.10 Scaled distributions of academic and behavioral features.	85
Figure 4.11 Correlation heatmap of all features and student performance outcome	86
Figure 4.12 Comparative accuracy of tuned models	87
Figure 4.13 Comparative precision of tuned models.	88
Figure 4.14 Comparative recall of tuned models	88
Figure 4.15 Comparative F1-scores of tuned models.	89
Figure 4.16 Comparative AUC values of tuned models.	90
Figure 4.17 Confusion matrix for Logistic Regression	91
Figure 4.18 Confusion matrix for Decision Tree	92
Figure 4.19 Confusion matrix for KNN	92
Figure 4.20 Confusion matrix for SVM	93
Figure 4.21 Confusion matrix for Random Forest	94
Figure 4.22 ROC curve for Logistic Regression.	95
Figure 4.23 ROC curve for Decision Tree.	96
Figure 4.24 ROC curve for KNN.	97
Figure 4.25 ROC curve for SVM.	97
Figure 4.26 ROC curve for Random Forest.	98
Figure 4.27 Dashboard Home Panel.	99
Figure 4.28 Semester Evaluation – Dataset Upload Interface.	100

Figure 4.29 Semester Evaluation – Risk Distribution and Downloadable Prediction Sheet	100
Figure 4.30 Single Course Prediction – Risk Distribution Overview.	101
Figure 4.31 Single Student Prediction – Risk State, Probability, and Explanation of Contributing Factors. .	102

List of Tables

Table 3.1 Abbreviations of the Confusion Matrix	75
---	----