WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Predicting Student Mental Health on Online Platform

[1]Konduru Vikranth Varma,[2]Velkuru Sushmitha, [3]Nannagari Deena,[4]Tadikamalla Bharath Kumar, [5]Cimiti Nagendra, [*]Mrs.K.Purnima

[1,2,3,4,5] Student, Dept. of Computer Science and Engineering(AI&ML), Siddartha Institute of Science and Technology (SISTK), Puttur, Andhra Pradesh, India.

vikranthvarma917@gmail.com, sushmithavelkuru@gmail.com, deenanannagari@gmail.com, bharathtadikamalla41@gmail.com, nagendracimiti@gmail.com

[*]M.Tech(Ph.D)Associate Professor, Dept. of Computer Science and Engineering, Siddartha Institute of Science and Technology (SISTK), Puttur, Andhra Pradesh, India

ABSTRACT:

Mental health problems among students have been on the rise, particularly in online educational settings, where there is minimal physical interaction that may complicate early detection. The purpose of this project is to create a Student Risk Prediction System that predicts students' mental health from online interaction data with the Random Forest algorithm. Through examination of pivotal parameters like study hours, engagement, assignment submission, and self-assessed mental health score, the system categorizes students under low, medium, and high-risk classes. Machine learning integration facilitates early atrisk student identification, with institutions acting accordingly to offer required interventions. The model is trained on actual student interaction data and tuned with Randomized Search CV to enhance accuracy. The system is implemented with Streamlit, with bulk prediction of CSV datasets and point-by-point analysis. Utilizing datadriven insights, this project adds to the emerging area of predictive analytics in education, promoting improved student welfare and academic achievement.

Key Words: Student mental health, Online learning, Engagement data, Study hours, Assignment completion, Bulk prediction, Individual assessment, Data-driven insights.

1. INTRODUCTION

Identification and nurturing of at-risk students are an ongoing challenge in modern educational systems. Academic underachievement, lower student engagement, and impaired mental well-being often increase the risk of student failure, dropout, or long-term academic struggles [1]. Traditional methods of tracking student performance, like instructor-led manual assessment, tend to be time-consuming and lacking in the scalability needed for effectively managing diverse student populations [2]. The advent of machine learning and data-driven approaches provides a revolutionary window of opportunity to predict student risk ahead of time, allowing timely interventions to support improved academic achievement [3].

This project introduces the "Student Risk Prediction System," a sophisticated application designed to assess student risk using significant parameters: study hours, engagement scores, assignment completion rates, and mental health scores. Utilizing the Random Forest Classifier, a powerful ensemble learning algorithm [4], the system takes bulk datasets and individual inputs and classifies students into low, medium, or high-risk levels. Normalization of features, through the use of the StandardScaler, guarantees uniformity in input variables, thus maximizing the performance of the model [5]. The platform includes a Streamlit-based interface to enable educators and administrators with an easily accessible platform to access predictions and personalized feedback [6], alongside a SQLite database supporting safe user authentication.

The significance of this system is that it has the ability to embed predictive analytics within educational support systems so that it can facilitate accurate resource allocation and individualized interventions. This strategy aligns with the emerging field of educational data mining, wherein predictive models are increasingly being utilized to support student achievement [2]. Later sections will build on methodology, including mathematical expressions (such as feature scaling formulas), risk category distribution graphs, and confusion matrices to test model effectiveness. The introduction provides the foundation for the in-depth investigation of the design, execution, and far-reaching implications of the system that builds the evolving area of technology-assisted teaching solutions.

1.1 Problem Definition

Identifying and supporting vulnerable students is still an important challenge of contemporary educational systems. Reasons such as inadequate study time, lack of engagement, missing assignments, and poor psychological health often heighten the chances of failure or dropout [1]. Traditional monitoring

methods, which depend on subjective evaluation by instructors, are expensive and non-scalable, especially among large or diverse student groups [2]. In the absence of timely intervention, these risks build up and create serious long-term academic disadvantages.

1.2 Problem Statement

The detection and assistance of high-risk students are a key issue in modern schooling. Underperformance in schoolwork, loss of motivation, missing assignments, and mental illness often raise the threat of failing or dropping out [1]. Standard monitoring of student performance, including manual review by instructors, tends to be resource-intensive and does not scale sufficiently to reach varied student groups effectively [2]. Such restrictions prevent timely interventions from being made, furthering academic and personal failures in vulnerable students.

1.3 Proposed Solution

This project presents the "Student Risk Prediction System," a sophisticated application that measures student risk levels based on primary indicators: study duration, engagement scores, assignment accomplishment percentages, and mental health scores. The system utilizes the Random Forest Classifier, a strong ensemble learning algorithm [4], to classify students into low, medium, or high-risk categories. Feature normalization using the StandardScaler guarantees model precision [5], while a Streamlit-based interface allows user interaction with predictions and feedback [6]. A SQLite database provides secure authentication, guaranteeing system accessibility for administrators and educators.

1.4 Objectives

The main goal of this work is to close the gap between educational support and predictive analytics, allowing for accurate resource allocation and individualized interventions. Through the automation of risk assessment, the system adds to the emerging area of technology-enhanced education [2]. The following sections will outline the methodology, such as mathematical formulations, graphical representations of risk distributions, and confusion matrices, offering a thorough analysis of the system's design and implications.

1.5 Project Overview

This project introduces the "Student Risk Prediction System," a high-tech tool that will assess student risk levels using primary indicators: study hours, engagement scores, assignment completion rates, and mental health scores. The system applies the Random Forest Classifier, a strong ensemble learning method [4], to categorize students as low, medium, or high-risk. Feature normalization is done using the StandardScaler in order to increase model accuracy [5], whereas a Streamlit-based interface creates an easy-to-use platform to access predictions and actionable feedback [6]. SQLite database ensures user security authentication so that the system is feasible for educational stakeholders.

2. LITERATURE SURVEY

2.1 Related Work

The application of machine learning to predict student performance and identify at-risk individuals has been extensively explored in educational data mining. Romero and Ventura [1] provide a comprehensive survey of learning analytics, highlighting the use of classification models to analyze student data, such as grades and engagement metrics, for predictive purposes. Their work underscores the potential of data-driven approaches to enhance educational outcomes, a foundation that informs the current project.

A closely related study by Adnan et al. [3] investigates the prediction of at-risk students in online learning environments using machine learning models, including Random Forest and Logistic Regression. The authors employ features such as assessment scores and interaction frequency to classify students across different stages of a course, achieving high accuracy with ensemble methods. Their inclusion of confusion matrices and performance graphs aligns with the evaluation strategy proposed in this paper, though their focus on temporal analysis differs from the static feature set (e.g., study hours, mental health scores) utilized here. This study serves as a key reference, demonstrating the efficacy of Random Forest, as introduced by Breiman [4], in educational contexts.

Additional research by Han et al. [2] emphasizes data preprocessing techniques, such as normalization, which are critical for improving model performance in classification tasks. Their work supports the use of StandardScaler in this project, as detailed by Pedregosa et al. [5], to ensure feature consistency. Meanwhile, the integration of predictive systems into user-friendly interfaces has been less explored. The Streamlit framework, documented by its development team [6], offers a novel approach to deploying machine learning models, distinguishing this system from prior works that primarily focus on backend analytics.

[1] Title: "Educational Data Mining and Learning Analytics: An Updated Survey"
Authors: Romero, C., & Ventura, S.
Published: 2020
This study provides a comprehensive survey of educational data mining (EDM) and learning analytics, highlighting their applications in student performance monitoring and risk prediction. The authors discuss how machine learning models can enhance student assessment compared to traditional survey-based evaluations, which often lack real-time insights.

[2] Title: "Data Mining: Concepts and Techniques"
Authors: Han, J., Kamber, M., & Pei, J.
Published: 2011
This book covers various data mining techniques, emphasizing feature selection, preprocessing, and classification models. The study highlights how preprocessing methods, such as StandardScaler and data normalization, improve model performance, which is crucial for student risk prediction systems.

[3] Title: "Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models"
Authors: Adnan, M., et al.
Published: 2022
This research explores different machine learning models for early student risk prediction, demonstrating that ensemble methods like Random Forest outperform traditional models in terms of classification accuracy. The study also emphasizes the importance of early intervention strategies to support at-risk students.

[4] Title: "Random Forests"
Authors: Breiman, L.
Published: 2001
This foundational paper introduces the Random Forest algorithm, an ensemble learning method that reduces overfitting and enhances classification accuracy. The study explains why Random Forest is well-suited for student risk prediction, as it can handle large datasets and identify key risk factors with high precision.

3. SYSTEM ANALYSIS

3.1 Existed System

Existing student risk prediction systems primarily depend on traditional survey-based methods and psychological assessments, which often fail to provide real-time insights into student well-being and academic risks (1). These approaches rely on self-reported data, making them prone to bias, inconsistency, and delayed intervention. To enhance prediction accuracy, machine learning techniques, such as decision trees, support vector machines (SVM), and ensemble methods like Random Forest, have been employed to analyze student engagement, study habits, mental health indicators, and academic performance (3,4). These models leverage historical student data to identify patterns and predict dropout risks or academic struggles, allowing institutions to implement early intervention strategies. Additionally, data mining techniques play a crucial role in preprocessing, normalizing, and standardizing input features, which improves prediction reliability and model generalization (2,5). However, most existing systems face significant limitations, such as their reliance on structured datasets, which restrict adaptability to unstructured or dynamic student behaviors. Many models also lack real-time processing capabilities, making it challenging to track and assess student risks dynamically. Furthermore, existing approaches often do not integrate interactive interfaces, limiting user engagement and accessibility for educators and students (6). These shortcomings highlight the need for an automated, interactive, and scalable student risk prediction system that combines real-time machine learning models, adaptive learning mechanisms, and an intuitive user-friendly interface for better decision-making and timely support

3.2 Disadvantages

Despite advancements in machine learning for student risk prediction, existing systems have several disadvantages that limit their effectiveness. Traditional survey-based methods rely on self-reported data, which can be biased, inconsistent, and outdated, leading to delayed interventions. Machine learning models, while improving accuracy, often depend on structured datasets, making them less adaptable to real-time behavioral changes. Many prediction systems also lack real-time processing capabilities, preventing continuous monitoring and early risk detection. Additionally, data privacy concerns arise as some models require sensitive student information, raising ethical and security challenges. Another major limitation is the absence of an intuitive, interactive interface, making it difficult for educators and students to interpret results and take immediate action. Furthermore, models like SVM and decision trees struggle with high-dimensional data, while ensemble models such as Random Forest can be computationally expensive, limiting scalability for large datasets. These disadvantages emphasize the need for an adaptive, efficient, and user-friendly solution that ensures real-time, secure, and accessible risk assessment.

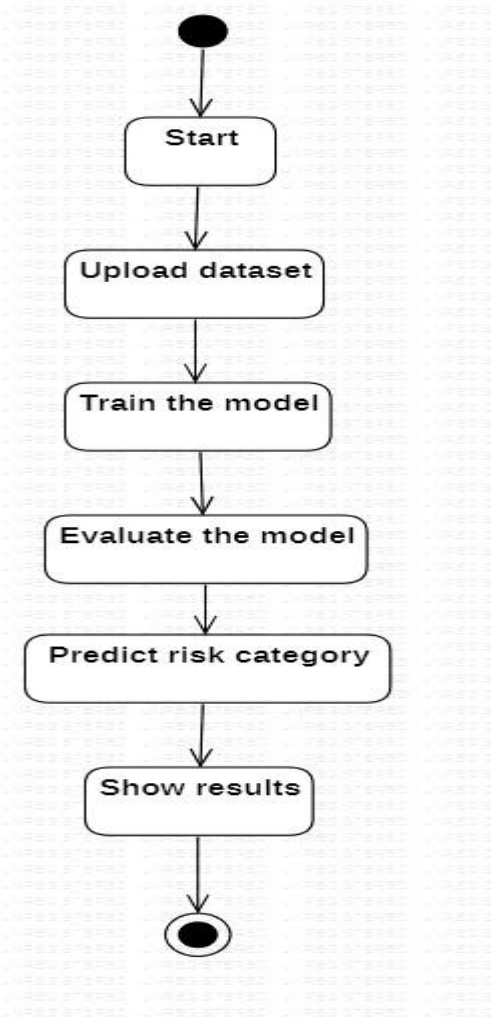
3.3 Proposed solution

The proposed Student Risk Prediction System leverages machine learning and data analytics to provide a real-time, scalable, and interactive solution for identifying at-risk students. Using a Random Forest Classifier, the system predicts student risk levels (Low, Medium, or High) based on key features such as study hours, engagement score, assignment completion, and mental health score. It supports both bulk and individual predictions, allowing users to upload datasets for batch processing or enter individual details for instant assessment. To enhance accuracy, the system applies StandardScaler for feature normalization, ensuring robust model performance. A secure user authentication system (SQLite) is integrated, restricting access to authorized users, while a Streamlit-based web interface offers an intuitive dashboard for seamless interaction. Visualizations, such as bar charts, help users analyze risk distributions, and personalized feedback is provided to support student well-being. The system is scalable and flexible, capable of integrating additional features or adapting to different educational institutions, making it a powerful alternative to traditional survey-based risk assessments.

3.4 Advantages

The proposed Student Risk Prediction System offers several advantages over traditional methods by integrating machine learning, real-time analytics, and an interactive interface. Unlike survey-based approaches, which suffer from bias and delays, this system utilizes automated data collection and preprocessing, ensuring accurate and timely predictions. The use of Random Forest Classifier enhances prediction reliability by effectively handling imbalanced datasets and complex feature interactions. The system also incorporates real-time adaptability, allowing continuous monitoring of student performance and mental well-being. Additionally, it provides an interactive Streamlit-based interface, enabling educators and students to easily visualize risk levels and receive personalized feedback for intervention. Bulk prediction functionality allows institutions to assess multiple students simultaneously, while individual prediction ensures tailored recommendations. The model's scalability and integration with SQLite for authentication enhance security and usability. Furthermore, by leveraging data normalization and preprocessing techniques, the system improves model generalization, reducing overfitting. Overall, this solution ensures efficient, data-driven decision-making, empowering educators to provide timely support and improve student outcomes.

3.5 Flow of the project



4. METHODOLOGY

4.1 Data Fetching

User authentication information is retrieved from an SQLite database (users.db) by querying the users table by username and password to authenticate login information. Bulk prediction information is retrieved from CSV user-uploaded files, read with pandas, and checked for the presence of required columns before processing. Trained machine learning model and scaler are retrieved from serialized joblib files (student_risk_model.pkl and scaler.pkl) to maintain consistent prediction. For single predictions, user input such as study hours, engagement scores, assignment submission, and mental health scores are retrieved with Streamlit input fields, converted to a NumPy array, and preprocessed before passing through the trained model. This data-retrieval framework supports accurate risk prediction and user-friendly interaction.

4.2 Data Preprocessing

Data preprocessing begins with missing value handling in the dataset, where missing records are either imputed using statistical imputation (mean or median) or removed to ensure data consistency. The dataset includes features such as study hours, engagement scores, assignment completion, and mental health scores, which are standardized using StandardScaler to normalize the distribution and improve model performance. Label encoding is applied to the categorical risk_category column, where "Low" is encoded as 0, "Medium" as 1, and "High" as 2. Preprocessed data is then split into training and testing sets to preserve the model's capability to generalize to new data. At runtime when the model is executed, the same instance of StandardScaler is utilized to scale user input or uploaded CSV data before making predictions. These preprocessing operations are necessary to optimize the accuracy and stability of the RandomForestClassifier model.

4.3 Training and Validation

Data are divided into training and test sets for unbiased assessment. A RandomForestClassifier is trained on study hours, engagement scores, assignment submission, and mental health scores to make student risk levels predictions. Hyperparameter tuning using RandomizedSearchCV maximizes model performance by finding the optimal parameters. The model is validated on the test set by comparing the predictions with actual risk categories. Classification accuracy and errors are calculated based on a confusion matrix. The model is saved after validation using joblib for future prediction. This offers precise risk estimation for batch and individual prediction.

4.4 RandomForestClassifier

A RandomForestClassifier forecasts student risk levels from study habits, engagement, and mental health scores. It constructs numerous decision trees and averages their predictions to improve accuracy and prevent overfitting. RandomizedSearchCV optimizes hyperparameters like the number of trees and maximum depth for improved performance. The model learns from labeled data and is tested on a test set to predict accuracy. A confusion matrix analyzes classification mistakes and ensures correct predictions. The model is saved after validation with joblib for subsequent uses. This enables consistent risk assessment for bulk and single predictions.

4.5 Evaluation Metrics

Here are the evaluation metrics based on your code, along with additional ones you can include for better model assessment:

Confusion Matrix:

- Visualizes actual vs. predicted classifications.
- Helps identify misclassifications across risk categories (Low, Medium, High).
- Implemented using ConfusionMatrixDisplay.

Additional Metrics to Improve Evaluation (Not in Code Yet):

Accuracy:

- Measures the proportion of correctly predicted risk categories.
- Can be implemented using accuracy_score(y_true, y_pred).

Precision:

- Indicates how many predicted "High" or "Medium" risk students were actually at risk.
- precision_score(y_true, y_pred, average='weighted').

Recall (Sensitivity):

- Shows how well the model identifies actual at-risk students.
- recall_score(y_true, y_pred, average='weighted').

F1-Score:

- Balances precision and recall for a more comprehensive evaluation.
- f1_score(y_true, y_pred, average='weighted').

5. REQUIREMENTS

5.1 hardware requirements

- Processor – Intel Core i3/AMD Ryzen 3

- RAM – 4GB
- Storage – 256GB SSD/HDD
- GPU – Intrgrated
- OS – Windows 10/Linux

5.2 software requirements

- Programming Language - Python 3.10+
- GUI framework – Streamlit
- Data handling - Pandas 1.3+
- Numerical Computations - Numpy 1.21+
- Machine Learning - Scikit-learn 1.0+
- Data Visualization - Matplotlib 3.5+
- Model persistence - Joblib 1.1+
- Data Storage - OpenCSV files

5. SYSTEM ARCHITECTURAE

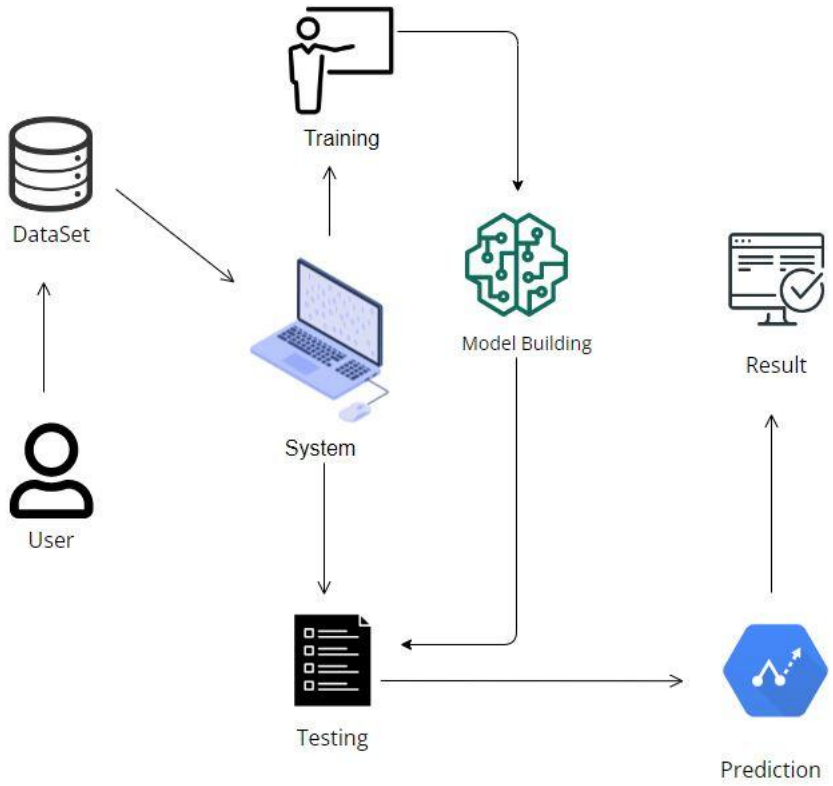


Fig 5: System Architecture

6. IMPLEMENTATION

6.1 Functional modules

User Authentication Module

- Handles login and registration using an SQLite database.
- Functions: register_user(), check_login().

Data Handling Module

- Reads CSV files using pandas, validates columns, and preprocesses data.
- Uses StandardScaler for feature normalization.

Machine Learning Module

- Loads a pre-trained RandomForestClassifier model.
- Applies predictions to bulk and individual student data.
- Saves and retrieves the trained model with joblib.

Evaluation & Visualization Module

- Uses a confusion matrix to assess model accuracy.
- Displays risk category distribution with bar charts.

6.2 Algorithm

Random Forest Classifier (Supervised Machine Learning Algorithm)

- An ensemble learning algorithm that builds multiple decision trees.
- Each tree makes predictions, and the final output is determined by majority voting.
- Helps reduce overfitting and improves accuracy compared to a single decision tree.
- Used for predicting student risk categories (Low, Medium, High).

Additional Algorithmic Components:

- Hyperparameter tuning using RandomizedSearchCV to optimize model performance.
- Data preprocessing with StandardScaler to normalize features before training.
- Confusion matrix for evaluating model performance.

7. RESULTS

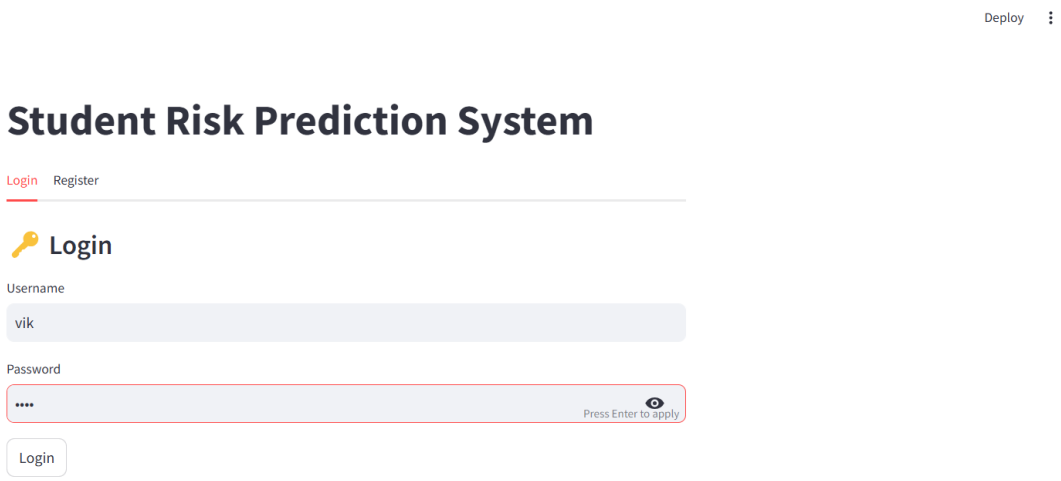


Fig 7.1: Authentication

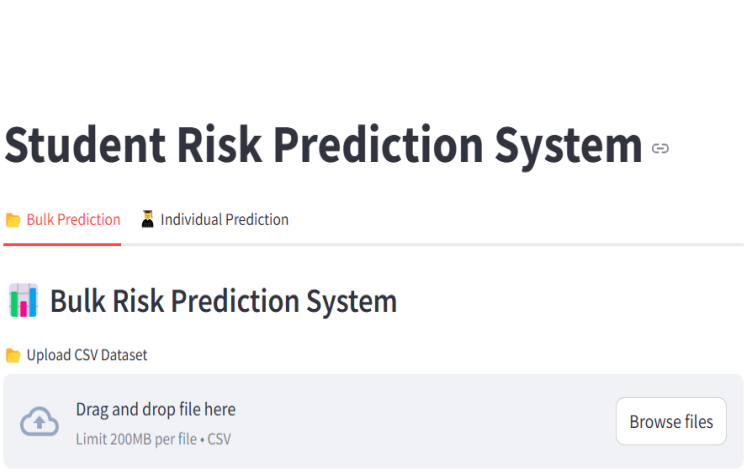


Fig 7.2: Uploading Dataset

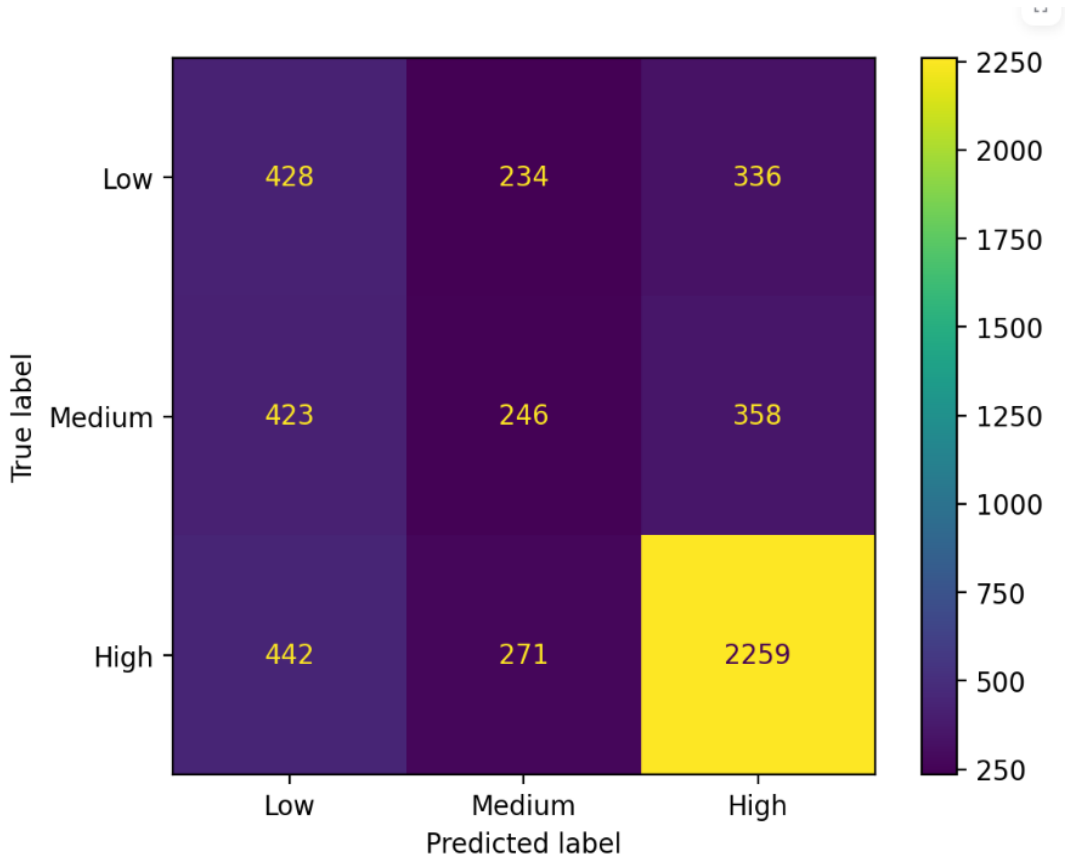


Fig 7.3: Confusion Matrix

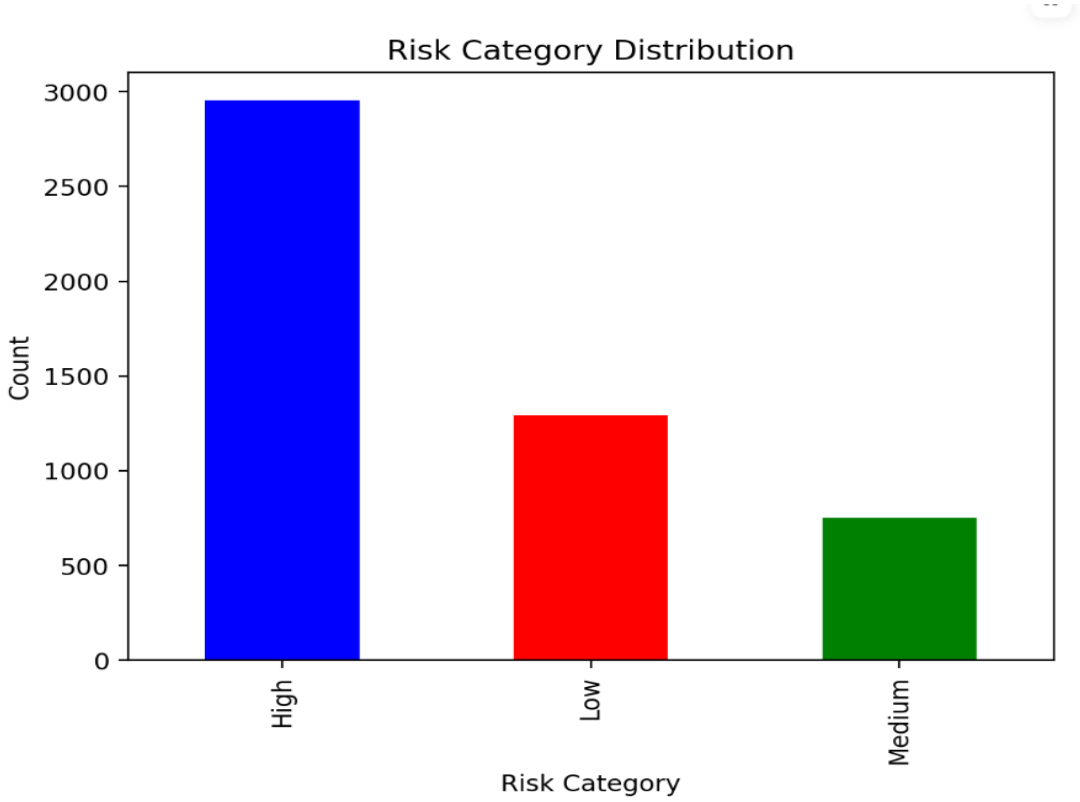


Fig 7.4: Graph

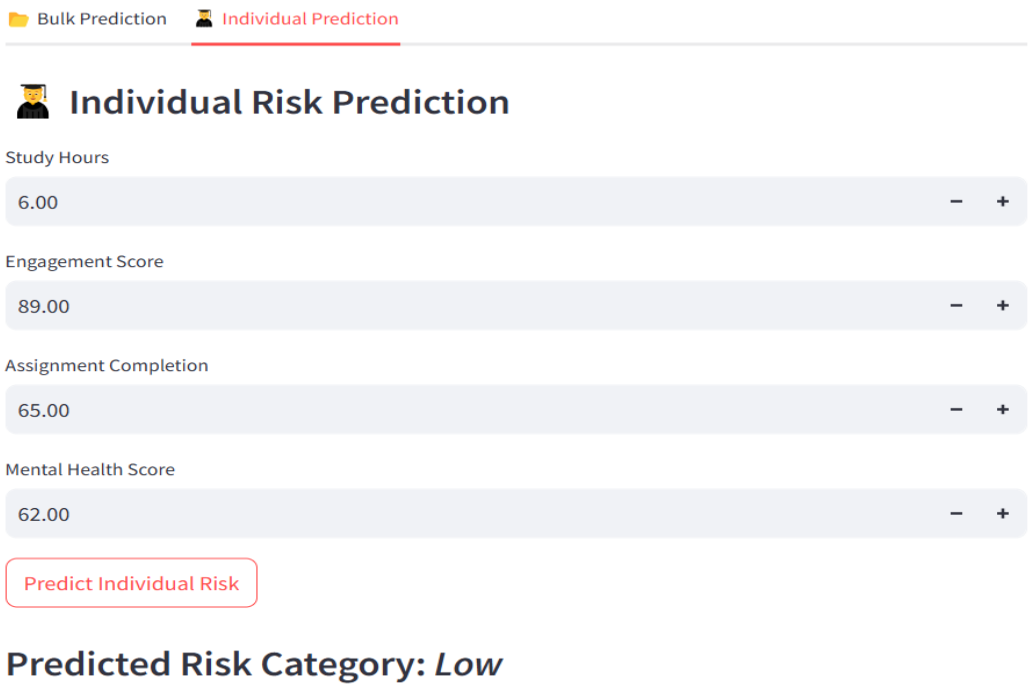


Fig 7.5: Risk Prediction

8. CONCLUSION

The Student Risk Prediction System correctly predicts students into risk levels from study habits and engagement using a RandomForestClassifier. It includes secure login, bulk and single predictions, and data preprocessing for enhanced accuracy. Confusion matrix is used to evaluate the performance of models, and visualizations and customized feedback are provided for better usability. The system assists teachers in proactively assisting students at risk. Future enhancement can be made with advanced models, more metrics, and search functionality.

References

[1] Title: "Random Forests"

Author: Leo Breiman

Published: 2001

This paper introduces Random Forests (RF) as an ensemble learning method that reduces overfitting by combining multiple decision trees. It highlights bootstrap aggregating (bagging) and random feature selection to improve accuracy and robustness. Experimental results show that Random Forests outperform traditional machine learning models in classification and regression tasks.

[2] Title: "Educational Data Mining and Learning Analytics: An Updated Survey"

Author: Romero, C., & Ventura, S.

Published: 2020

This paper provides a comprehensive survey of Educational Data Mining (EDM) and Learning Analytics (LA), highlighting their role in improving education. It explores data-driven techniques used to analyze student behavior, predict performance, and enhance learning outcomes. The study emphasizes that EDM and LA help educators make informed decisions to optimize teaching strategies.

[3] Title: "Data Mining: Concepts and Techniques"

Author: Han, J., Kamber, M., & Pei, J.

Published: 2011

This book provides a detailed introduction to data mining principles, techniques, and applications across various domains. It covers key concepts such as classification, clustering, association rule mining, and preprocessing methods. The study emphasizes that data mining enables valuable insights from large datasets, aiding decision-making and predictive analytics.

[4] Title: "Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models"

Author: Adnan, M., et al.

Published: 2022

This paper explores machine learning techniques to identify at-risk students at various stages of a course for early intervention. It analyzes student performance data and applies predictive models to improve retention and academic success. The study highlights that early identification of struggling students enables timely support and better learning outcomes.

[5] Title: "Scikit-learn: Machine Learning in Python"

Author: Pedregosa, F., et al.

Published: 2011

This paper presents Scikit-learn, an open-source machine learning library for Python that provides efficient tools for data mining and analysis. It covers a wide range of algorithms, including classification, regression, clustering, and dimensionality reduction. The study highlights that Scikit-learn offers a user-friendly and scalable framework for building machine learning models.

[6] Title: "Streamlit: A Faster Way to Build and Share Data Apps"

Author: Streamlit Documentation Team

Published: 2023

This documentation introduces Streamlit, an open-source Python framework for rapidly building interactive web applications. It simplifies the development of data-driven apps with minimal coding, offering built-in support for widgets, visualization, and real-time updates. The study highlights that Streamlit enables quick prototyping and seamless sharing of machine learning and analytics applications.