PROJECT TITLE

AI-Powered EduPredict – Student Performance & Analytics System

THE DOMAIN OF THE PROJECT

Artificial Intelligence & Machine Learning

TEAM MENTORS

Vara Prasad Sir

Tarun Joseph Sir

TEAM MEMBER

1. Mr. G Vivek chary - B.Tech 4th Year pursuing in Computer Science and Engineering

INTERNSHIP

Python & Machine Learning Applications

PERIOD OF THE INTERNSHIP

April 2025 to September 2025

Declaration

The project titled **"AI-POWERED EDUPREDICT-STUDENT PERFORMANCE & ANALYTICS SYSTEM"** has been mentored by **Tarun Joseph Sir, Vara Prasad Sir** organised by SURE Trust, from April 2025 to September 2025, for the benefit of the educated unemployed rural youth for gaining hands-on experience in working on industry relevant projects that would take them closer to the prospective employer. I declare that to the best of my knowledge the members of the team mentioned below, have worked on it successfully and enhanced their practical knowledge in the domain.

Team Members:

1. Mr. G Vivek Chary

Prof. Radhakumari
Executive Director & Founder
SURE Trust

TABLE OF CONTENTS

Abstract

In modern education systems, monitoring and evaluating student performance is crucial for identifying at-risk learners early and providing timely interventions. Traditional evaluation methods are reactive and often fail to offer insights into why a student might succeed or fail. To address this gap, we developed an AI-powered EduPredict-Student Performance & Analytics System that predicts whether a student is likely to Pass or Fail based on subject scores, academic behaviour, and demographic factors.

This project leverages Random Forest and Logistic Regression models trained on academic scores, attendance, study habits, and extracurricular activities. SHAP (SHapley Additive explanations) is integrated to provide explainability for predictions, while visual analysis and textual explanations help teachers understand the reasons behind each decision.

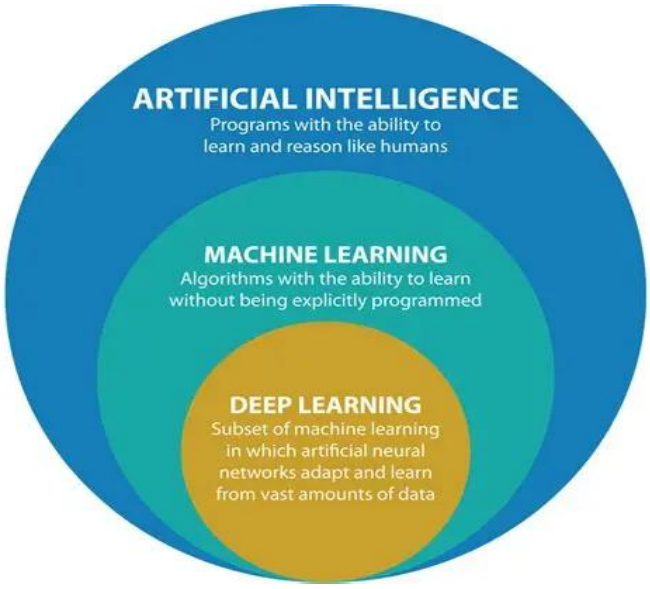
The system also supports authentication, batch prediction, and centralized storage of results, making it a scalable and practical early intervention tool.

🧠 Artificial Intelligence (AI)

Artificial Intelligence refers to the simulation of human intelligence in machines that are programmed to think, learn, and make decisions. AI systems can perform tasks such as reasoning, problem-solving, perception, and language understanding. In education, AI enables personalized learning, predictive analytics, and intelligent tutoring systems.

Key Concepts:

- Mimics human cognitive functions like learning and decision-making.
- Enables automation of complex tasks without explicit programming.
- Powers applications like chatbots, recommendation engines, and predictive systems.



📊 Machine Learning (ML)

Machine Learning is a subset of AI that focuses on building systems that learn from data and improve over time without being explicitly programmed. ML algorithms identify patterns in historical data and use them to make predictions or decisions.

Types of ML:

- **Supervised Learning**: Uses labeled data to train models that predict outcomes, such as classifying students as pass or fail based on input features.
- **Unsupervised Learning:** Works with unlabeled data to discover hidden patterns or groupings, like clustering students based on study habits.
- **Reinforcement Learning:** Learns optimal actions through trial and error by receiving feedback, often used in dynamic environments like game playing or robotics.

🧠 Deep Learning

Deep Learning is a specialized branch of Machine Learning that uses multi-layered neural networks to automatically learn complex patterns from large datasets. It mimics the way the human brain processes information, enabling systems to perform tasks like image recognition, speech processing, and natural language understanding with high accuracy.

🔍 Key Characteristics:

- Uses Artificial Neural Networks (ANNs) with multiple hidden layers (hence “deep”).
- Learns features directly from raw data—no manual feature engineering required.
- Requires large datasets and high computational power (often GPU-based).
- Common architectures include CNNs (for images), RNNs/LSTMs (for sequences), and Transformers (for language tasks).

🛠️ Technologies and Tools Used

Programming Language

- Python: Core language for data preprocessing, model building, and integration.

Machine Learning Libraries

- Scikit-learn: Used for implementing Random Forest and Logistic Regression models.
- SHAP (SHapley Additive exPlanations): For model interpretability and explainability.

Data Handling

- Pandas: For data manipulation and analysis.
- NumPy: For numerical operations and array handling.

Visualization

- Matplotlib / Seaborn: For plotting confusion matrices and other evaluation visuals.
- SHAP plots: To visualize feature importance and model decisions.

Web Interface

- Streamlit: Used to build the user-friendly interface for single and batch predictions.

Authentication & Storage

- SQLite: Lightweight database to store user login credentials and prediction history securely.

Project Objectives

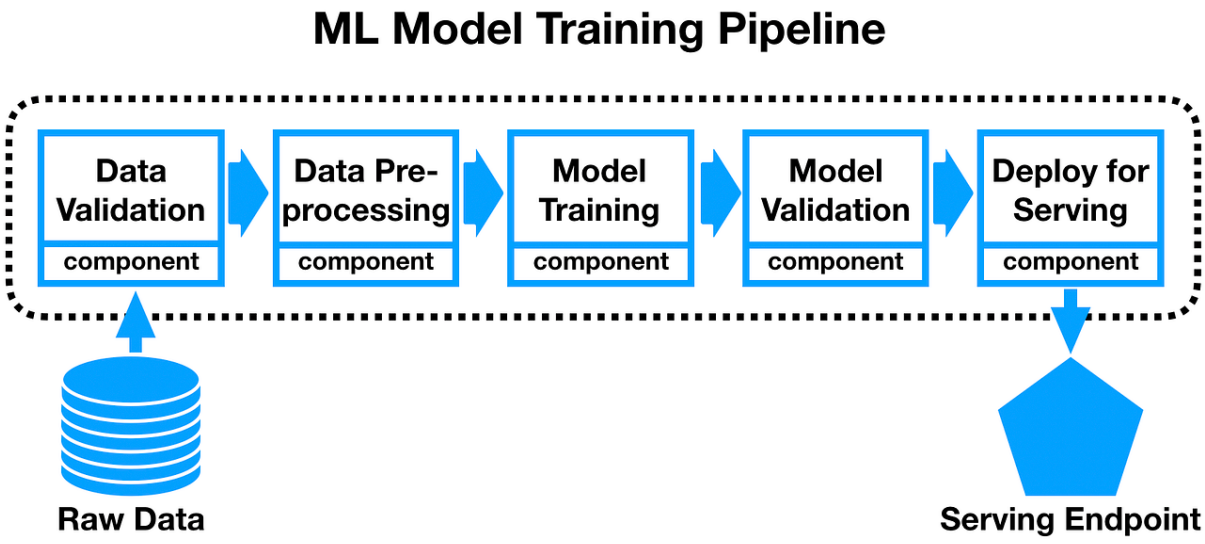
- **Predict Student Outcomes Using ML Models**
  Develop machine learning models (Random Forest, Logistic Regression) to classify students as likely to pass or fail based on academic and behavioural data. This helps in proactive academic planning.
- **Provide Interpretable Explanations Using SHAP**
  Integrate SHAP (SHapley Additive explanations) to visualize and explain the influence of each feature on the prediction. This ensures transparency and builds trust in the system's decisions.
- **Enable Single and Batch Predictions via Interface**
  Build a user-friendly interface using Streamlit that allows teachers to input individual or multiple student records for prediction. This improves usability and saves time.
- **Securely Store Predictions with Authentication**
  Implement a centralized database (e.g., SQLite) with login authentication to store prediction results securely. This protects sensitive student data and supports future analysis.
- **Support Early Intervention for Educators**
  Provide actionable insights to teachers and administrators to identify at-risk students early. This enables timely support, reduces dropout rates, and improves overall academic outcomes.

Methodology

The methodology adopted in this project follows a systematic approach to build an AI-powered Student Performance & Analytics System.

The process is divided into multiple phases as outlined below:

- Data Collection & Preprocessing
- Feature Engineering
- Model Selection & Training
- Explainability with SHAP
- User Interface Development
- Authentication & Database Integration
- Testing & Evaluation



Dataset Overview

The dataset used in this project contains academic, behavioral, and demographic information of students. It has been carefully designed to capture multiple aspects of student performance, including subject-wise exam scores, study habits, attendance, and participation in extracurricular activities.

The target variable pass_fail indicates whether a student is predicted to pass (1) or fail (0), based on academic performance across subjects. The dataset is balanced to ensure fair representation of both pass and fail cases.

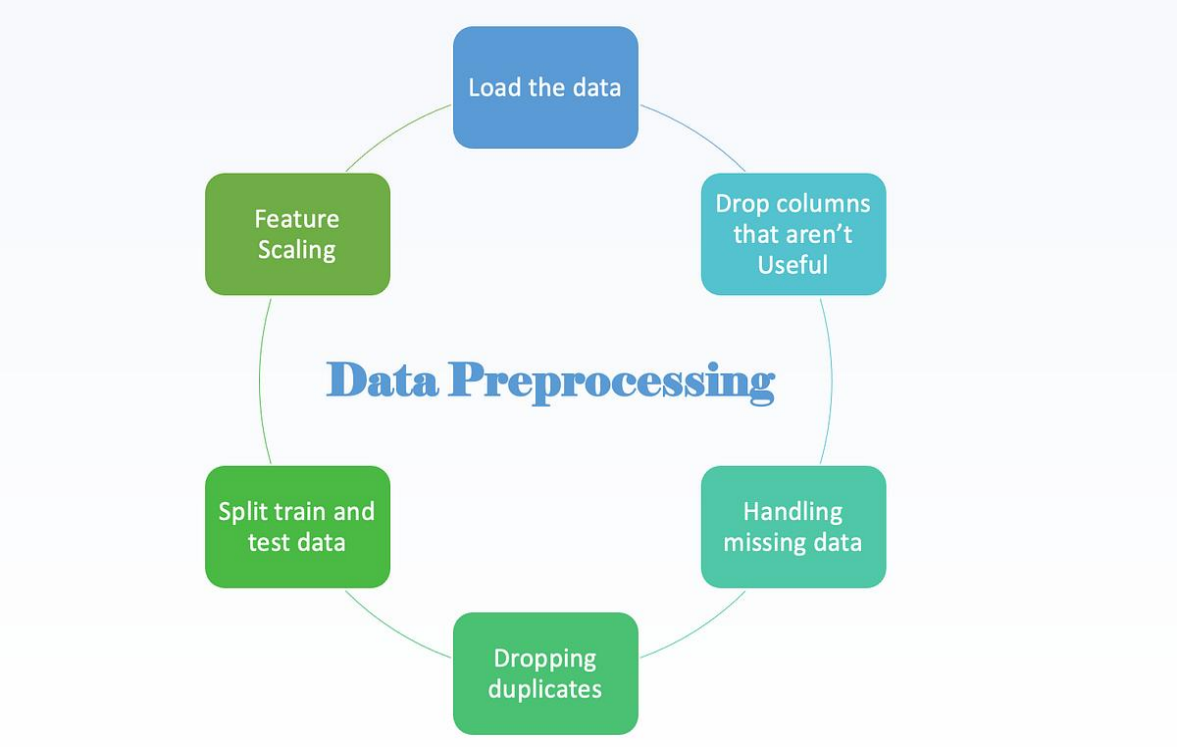
Below is the list of features included in the dataset:

Feature Name	Description	Data Type
gender	Gender of the student	Categorical (encoded)
part_time_job	Whether the student has a part-time job	Categorical (encoded)
extracurricular_activities	Participation in extracurricular activities	Categorical (encoded)
career_aspiration	Student's future career goal	Categorical (encoded)
absence_days	Number of school days missed	Integer
weekly_self_study_hours	Weekly hours spent on self-study	Integer
math_score	Score in Mathematics	Integer (0–100)
history_score	Score in History	Integer (0–100)
physics_score	Score in Physics	Integer (0–100)

Feature Name	Description	Data Type
chemistry_score	Score in Chemistry	Integer (0–100)
biology_score	Score in Biology	Integer (0–100)
english_score	Score in English	Integer (0–100)
geography_score	Score in Geography	Integer (0–100)
pass_fail	Target variable: 1 = Pass, 0 = Fail	Binary

Data Preprocessing

Data preprocessing was an essential step to ensure the dataset was clean, consistent, and ready for model training.



The following steps were performed:

1. **Data Cleaning**
   - Removed irrelevant columns such as id, first_name, last_name, and email as they do not contribute to predicting student performance.

- Ensured there were no duplicate rows in the dataset.
- Handled any missing values by filling them with default values (e.g., zeros or encoded categories) instead of dropping rows, to preserve data.

1. **Feature Encoding**

   - Categorical variables (gender, part_time_job, extracurricular_activities, and career_aspiration) were encoded using Label Encoding.
   - Encoders were saved as .pkl files to ensure consistency between training and prediction phases.

2. **Target Variable Creation**

   - The target column pass_fail was generated based on subject scores.
   - A student was considered Pass (1) if they scored at least 40 marks in 6 out of 7 subjects. Otherwise, they were labeled Fail (0).

3. **Feature Selection**

   - Selected subject-wise scores (math_score, history_score, physics_score, chemistry_score, biology_score, english_score, geography_score) along with academic behavior features (absence_days, weekly_self_study_hours), and demographic factors (gender, part_time_job, extracurricular_activities, career_aspiration).
   - Ensured all selected features were relevant to student performance prediction.

4. **Train-Test Split**

   - The dataset was split into 80% training data and 20% testing data using stratified sampling to maintain the distribution of Pass/Fail labels.

5. **Data Balancing**

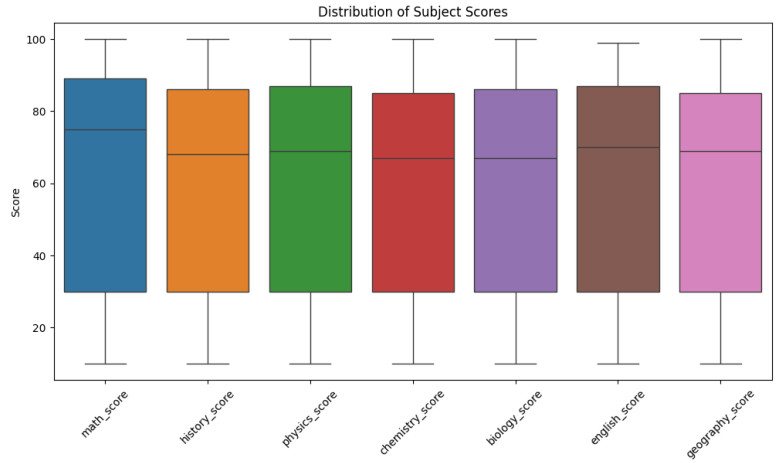
   - To avoid class imbalance, SMOTE (Synthetic Minority Oversampling Technique) was explored, ensuring both Pass and Fail classes were represented fairly in training.

Data Visualization

Data visualization was performed to gain insights into the dataset and better understand patterns in student performance. Various plots and charts were created to analyze subject scores, attendance, and other academic behavior factors.

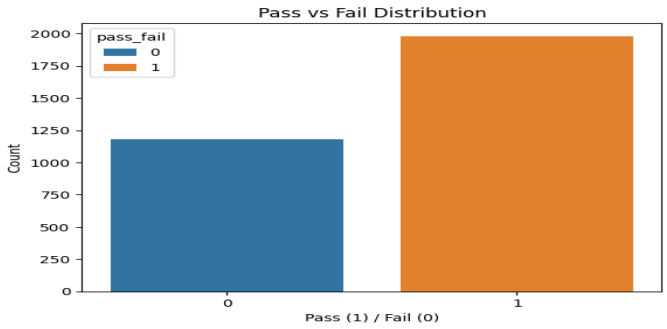
Distribution of Subject Scores

Histograms and boxplots were used to visualize the distribution of scores across all subjects.This helped identify the spread of marks, detect outliers, and observe performance trends in different subjects.



Pass vs. Fail Comparison

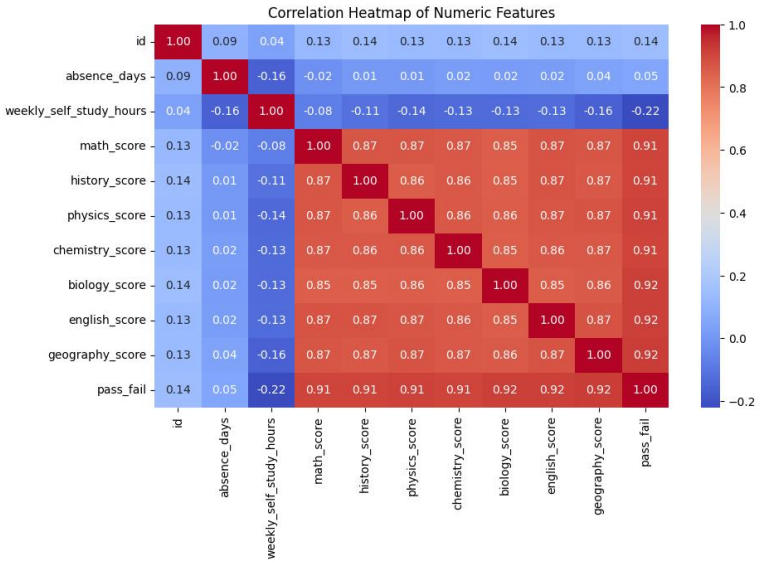
Bar charts were plotted to compare the number of students who passed vs. failed.This provided clarity on dataset balance and highlighted the importance of certain features.



Correlation Heatmap

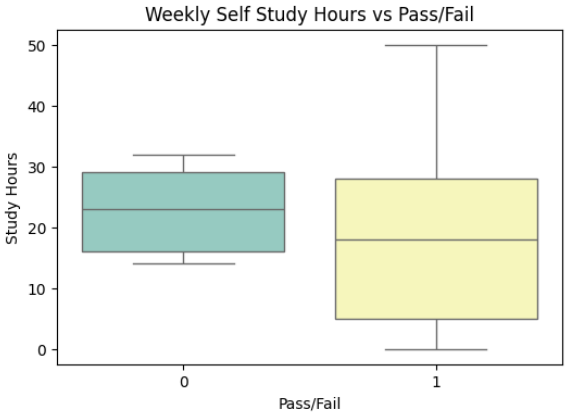
A heatmap was generated to analyze correlations among subject scores and

other numerical features.Strong correlations were observed between science subjects (Physics, Chemistry, Biology), while weak correlations were seen with extracurricular and demographic features.

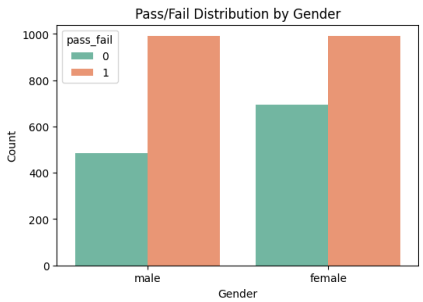


Weekly Self-Study Hours

Visual analysis indicated that students with higher self-study hours performed better, confirming its importance as a feature in the model.

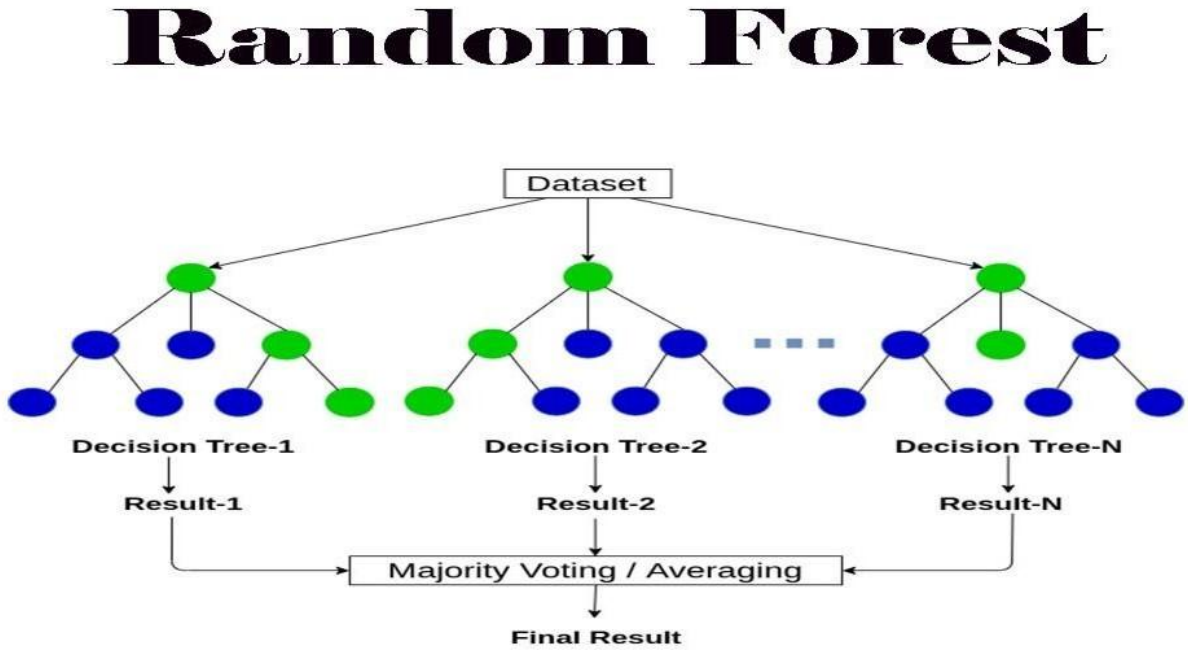


Pass / Fail Distribution by Gender



Model Description

We implemented a Random Forest Classifier, an ensemble learning technique that builds multiple decision trees and merges their outputs to improve prediction accuracy and reduce overfitting. This model was selected for its versatility and reliability in educational data analysis.



Why Random Forest?

- **Handles Mixed Data Types**: Efficiently processes both categorical (e.g., gender, activity participation) and numerical (e.g., scores, attendance) features.
- **Robust Performance:** Delivers consistent results even on moderately sized and imbalanced datasets.
- **Captures Complex Patterns:** Effectively models non-linear relationships and feature interactions, which are common in student behaviour and performance data.

Explainability with SHAP (SHapley Additive Explanations)

To ensure transparency in model predictions, the project integrates SHAP (SHapley Additive Explanations). SHAP provides a way to understand how individual features contribute to the model's output for both single and batch predictions.

- It assigns each feature a contribution value (positive or negative) showing whether it pushed the prediction towards Pass or Fail.
- Teachers can visually interpret results using SHAP decision plots, which illustrate how subject scores, attendance, study hours, and other features influence predictions.
- For example, a low math score or high absence days may strongly contribute towards a "Fail" prediction, while regular self-study and good English performance may push the result towards "Pass".
- This improves trust and accountability, as teachers can understand not just the prediction but also the reason behind it.

User Interface Development

The User Interface (UI) of the AI-Powered EduPredict system was designed to ensure simplicity, accessibility, and interactivity for both teachers and administrators. The UI was implemented using Streamlit, a lightweight Python framework for building interactive web applications.

1. Design Goals

- Provide a clean and user-friendly interface for educators with minimal technical background.
- Enable quick data input, either through manual entry of student scores or bulk uploads (CSV/Excel).
- Display visual analytics for easy interpretation of student performance.

2. Key Features

- **Login & Authentication**: Secure access for teachers and administrators using role-based login connected to a SQLite database.
- **Single Student Prediction**: Teachers can enter a student's demographic and academic details (e.g., subject scores, study hours, absence days) and instantly get a Pass/Fail prediction along with visual and textual explanations.
- **Batch Prediction**: Teachers can upload a CSV file with multiple student records. The system predicts outcomes for all students at once, generates explanations, and allows downloading the results as a CSV.
- **Visual Feedback**: Subject-wise bar charts are displayed, highlighting weak subjects in red, and a SHAP decision plot is provided to explain feature contributions.
- **Prediction Logging**: Each prediction (input, result, probability, explanation) is securely stored in the database for monitoring and review.

- **Automatic Explanations:** The system generates text-based summaries alongside graphs, making it easier to identify at-risk students.
- **Admin Panel**: Admin users can view, filter, and export all predictions made by different teachers, providing oversight and performance tracking

3. Visualization & Feedback

- **Bar Charts:** Indicate weak subjects (score < 40).
- **SHAP Plots:** Show how features influence predictions.
- **Status Indicators:** Color-coded feedback (e.g., green for Pass, red for Fail).

Authentication & Database Integration

To ensure secure access and role-based functionality, the EduPredict system integrates an authentication mechanism and a backend database for storing credentials and predictions. This design provides accountability, security, and centralized data management.

Design Goals

- Provide secure login for teachers and administrators.
- Restrict access to certain features (e.g., Admin Panel) based on user roles.
- Maintain a history of predictions for monitoring and auditing purposes.

Key Features

- **Role-Based Authentication**: Users must log in with valid credentials. Admins can access the Admin Panel, while teachers are restricted to student predictions.
- **SQLite Database Integration**: A lightweight yet powerful database stores user credentials, prediction records, and explanations.
- **Password Security**: Passwords are stored using hashing techniques, ensuring user privacy and security.
- **Prediction Storage**: Each prediction (input data, result, probability, explanation, and timestamp) is saved in the database for future reference.

Database Structure

Users Table

```
CREATE TABLE IF NOT EXISTS users (
   id INTEGER PRIMARY KEY AUTOINCREMENT,
   username TEXT UNIQUE,
   password_hash TEXT,
   role TEXT,
   subject TEXT   );
```

- **username** → Unique teacher/admin login name
- **password_hash** → Encrypted password storage
- **role** → User role (teacher/admin)
- **subject** → Teacher's subject specialization (only for teachers)

Predictions Table

```
CREATE TABLE IF NOT EXISTS predictions (
  id INTEGER PRIMARY KEY AUTOINCREMENT,
  username TEXT,
  student_name TEXT,
  input_json TEXT,
  result TEXT,
  probability REAL,
  explanation TEXT,
  timestamp DATETIME DEFAULT CURRENT_TIMESTAMP
);
```

- **username** → User who made the prediction
- **student_name** → Name of the student (optional input by teacher)
- **input_json** → Stored features (marks, attendance, study hours, etc.)
- **result** → Pass/Fail prediction outcome
- **probability** → Model confidence score
- **explanation** → Auto-generated explanation (using SHAP + rules)
- **timestamp** → Date and time of prediction

Benefits

- Prevents unauthorized access to sensitive educational data.
- Provides administrators with a complete view of all predictions and system usage.
- Ensures transparency by storing both model outputs and explanations.

Testing & Evaluation

The developed **AI-Powered EduPredict system** was thoroughly tested and evaluated to ensure reliability, accuracy, and usability. The following approaches were adopted:

Unit Testing

- Individual modules such as data preprocessing, model training, SHAP explanation, and visualization were tested independently.
- Edge cases such as missing values, invalid input, and low sample size were handled to ensure robustness.

Integration Testing

- The integration of the Random Forest model with SHAP explanations and the Streamlit-based user interface was tested.
- Proper interaction between model predictions, explanation visualizations, and database storage was verified.

System Testing

- End-to-end functionality was validated, including input of student data, model prediction, generation of SHAP-based explanations, and role-based authentication.
- Both individual prediction and batch prediction modes were tested.

Performance Evaluation

- The model was evaluated using standard machine learning metrics: **Accuracy, Precision, Recall, F1-Score, and ROC-AUC**.
- The Random Forest model achieved **high accuracy** while maintaining interpretability through SHAP explanations.

User Acceptance Testing (UAT)

- Faculty members and students were invited to test the system.
- Feedback highlighted the ease of understanding SHAP-based explanations and the usefulness of visual insights.

Results & Screenshots

Model Performance

The Random Forest model was trained and evaluated on the student performance dataset. The model achieved a strong performance with **97% accuracy**, demonstrating its ability to predict student outcomes (Pass/Fail) reliably.

Confusion Matrix:

Actual \ Predicted	Fail (0)	Pass (1)
Fail (0)	210	6
Pass (1)	3	413

Classification Report:

Class	Precision	Recall	F1-Score	Support
0 (Fail)	0.97	0.97	0.97	216
1 (Pass)	0.99	0.99	0.99	416
Accuracy			0.97	632
Macro Avg	0.98	0.98	0.98	632
Weighted Avg	0.98	0.98	0.98	632

User Interface (Screenshots)

The system was deployed with an interactive Streamlit interface consisting of:

➢ Registration Page

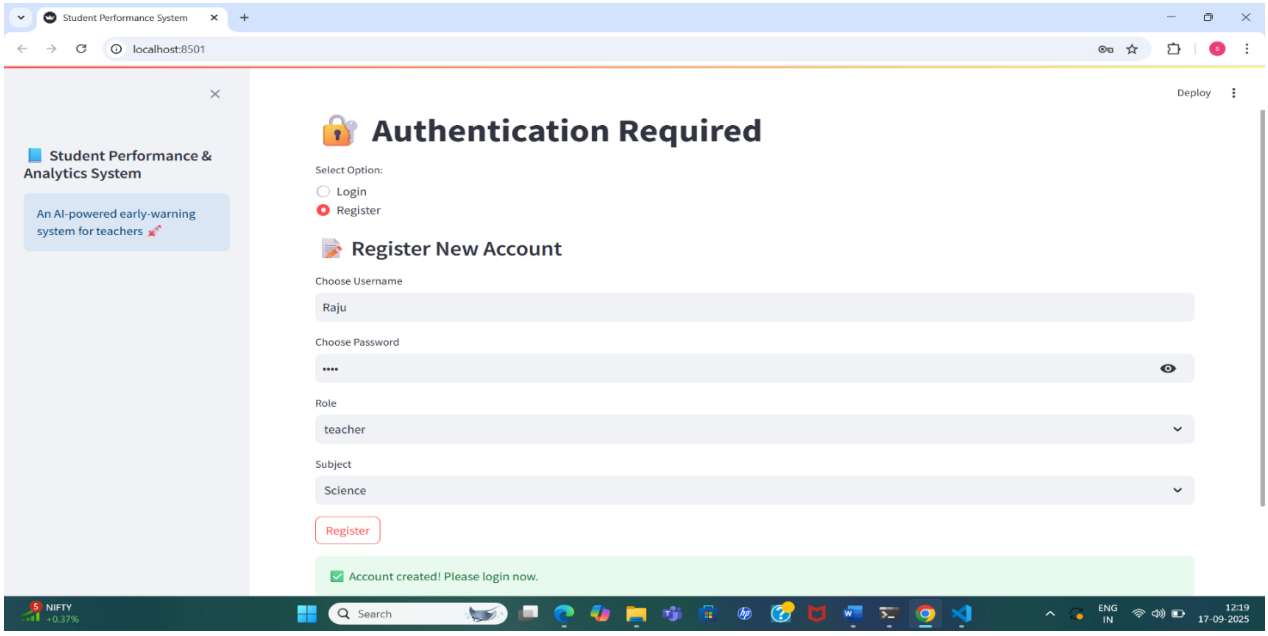


Fig 1: Teacher's Registration

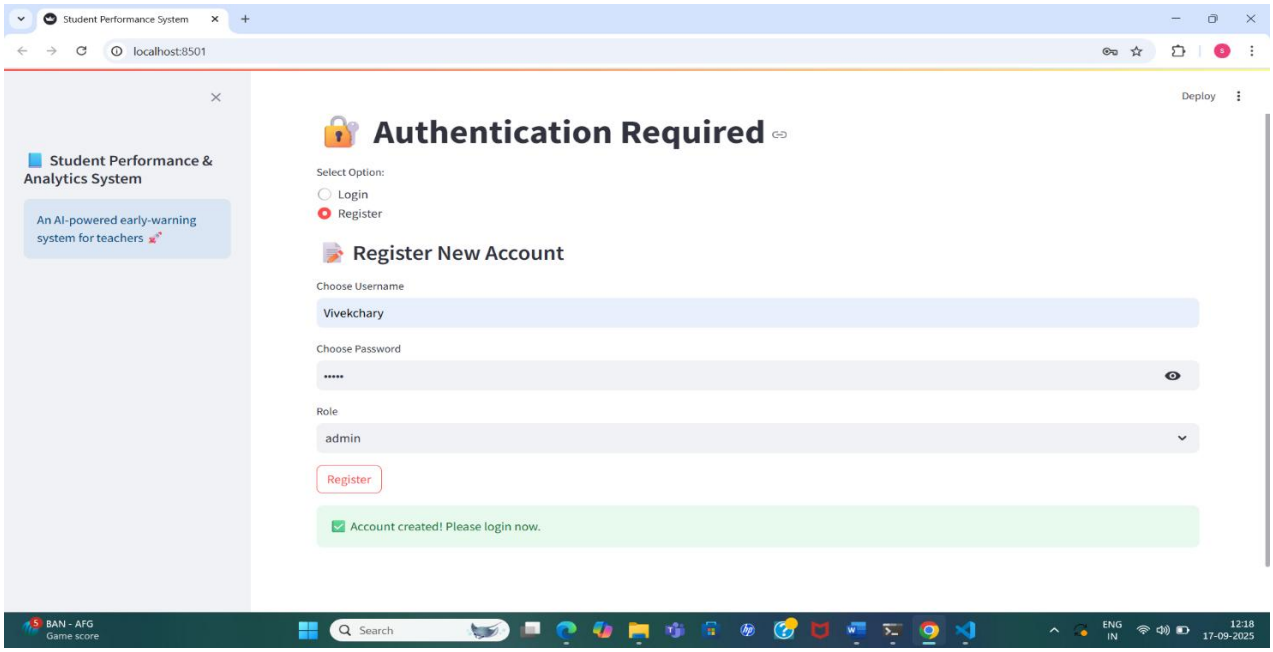


Fig 2: Adminstrator's Registration

➢ Login Page

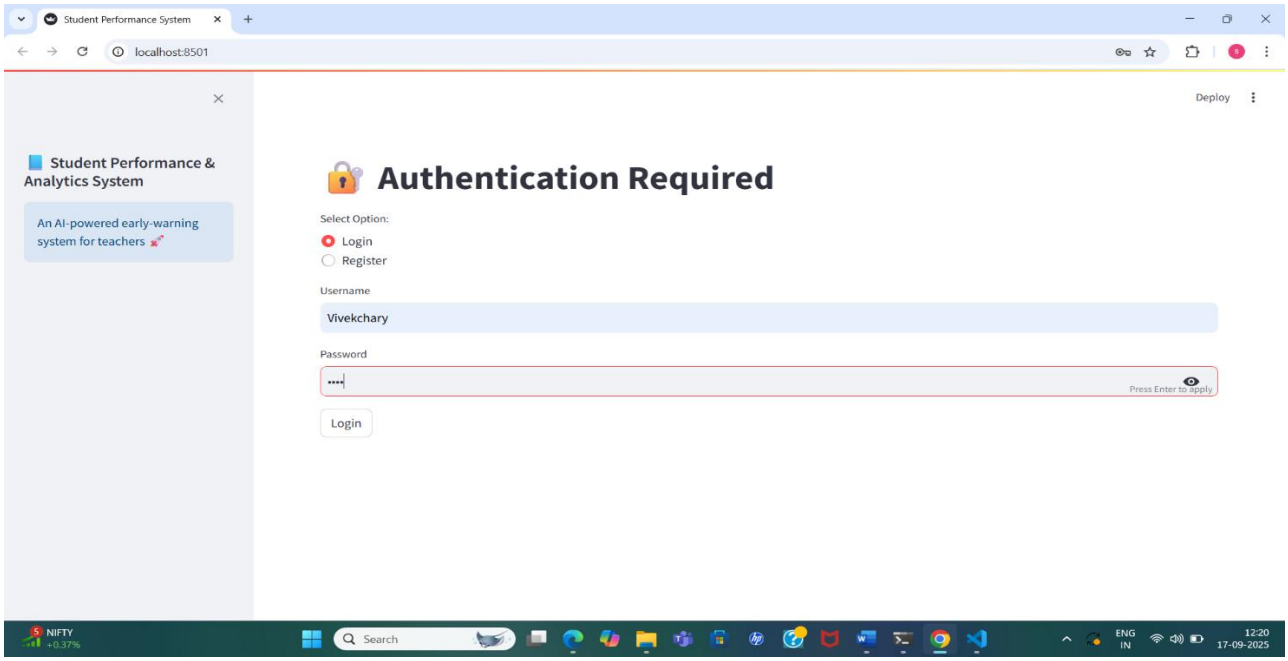
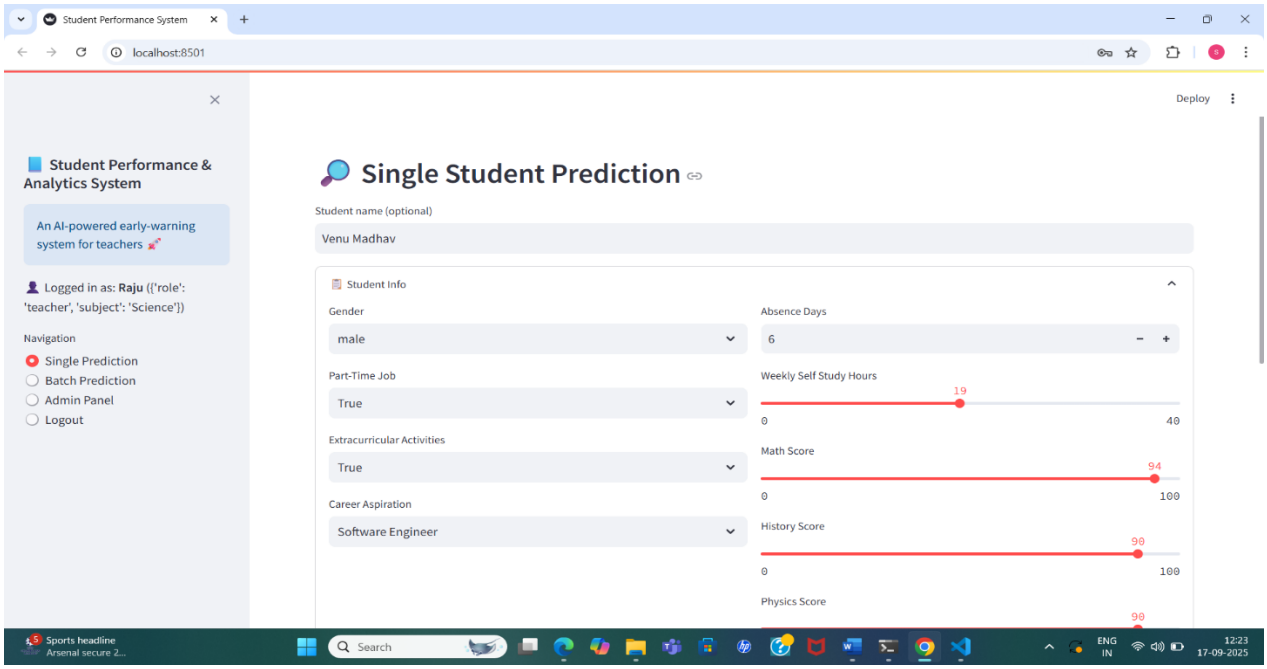


Fig 3: Role-based authentication system with teacher and admin login.

➢ Single Student Prediction Form



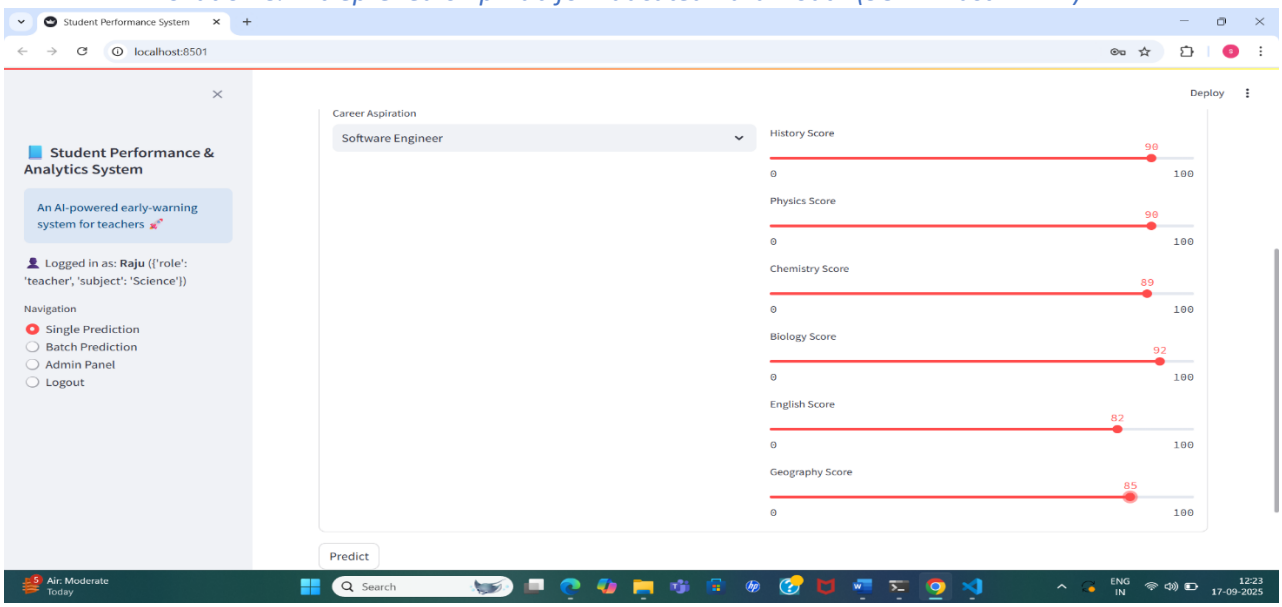


Fig 4 & 5: Teacher enters subject-wise marks and details for performance prediction.

➢ Single Prediction Result

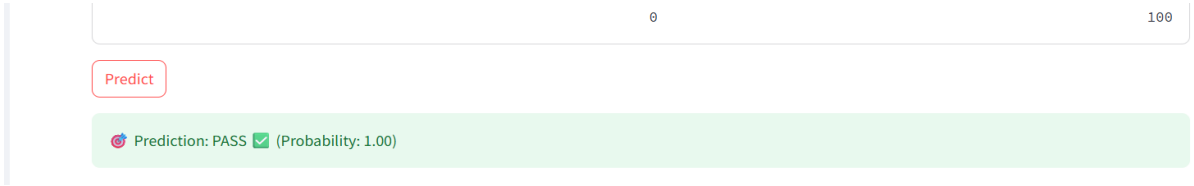


Fig 6: Model output showing Pass/Fail prediction, probability.

➢ Subject-wise Bar Chart

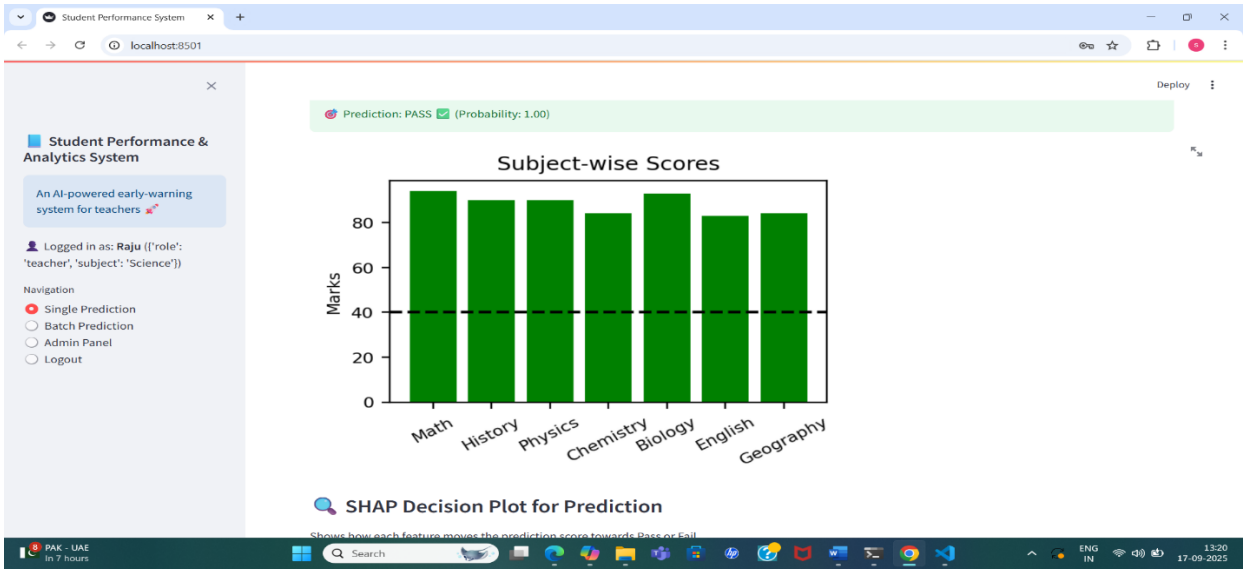
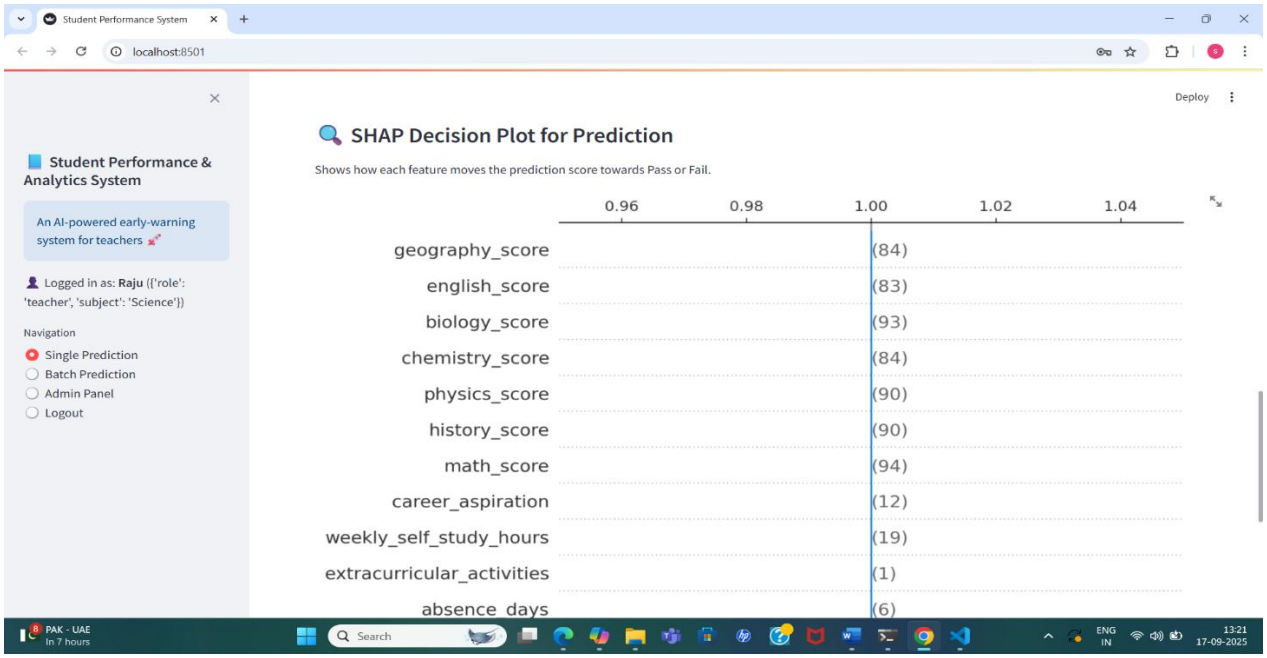


Fig 7: Visualization of student scores with pass mark indicator.

➢ SHAP Decision Plot



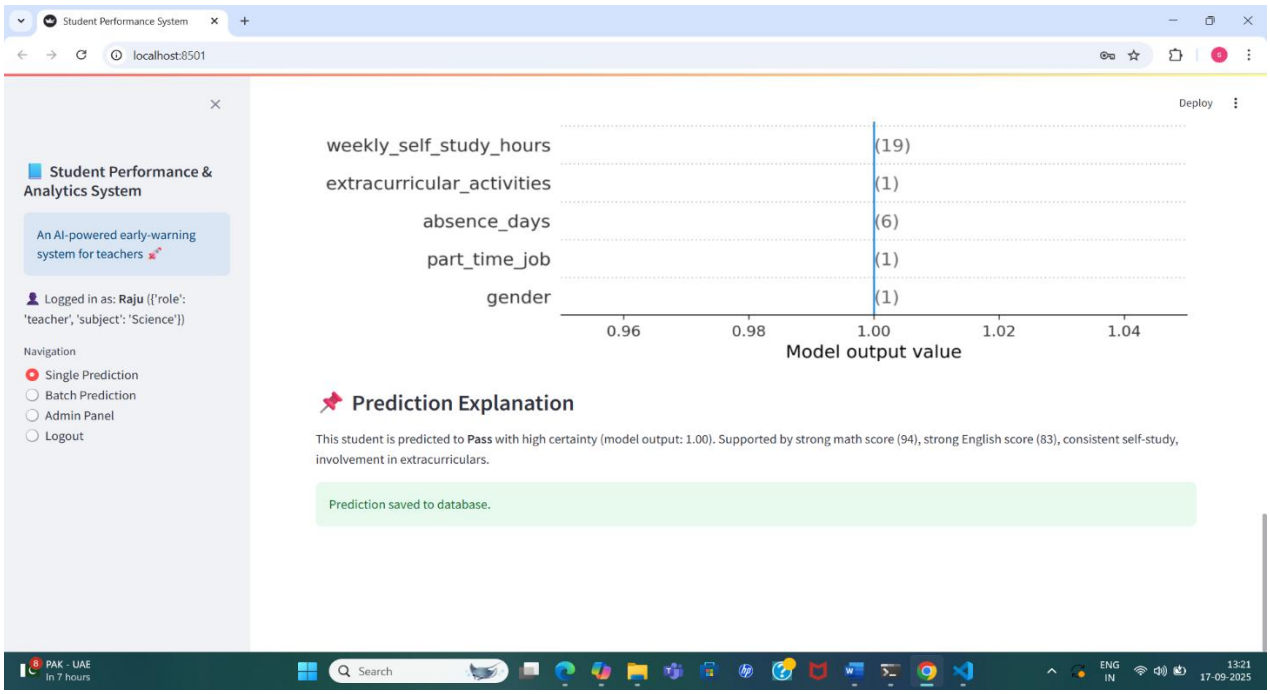
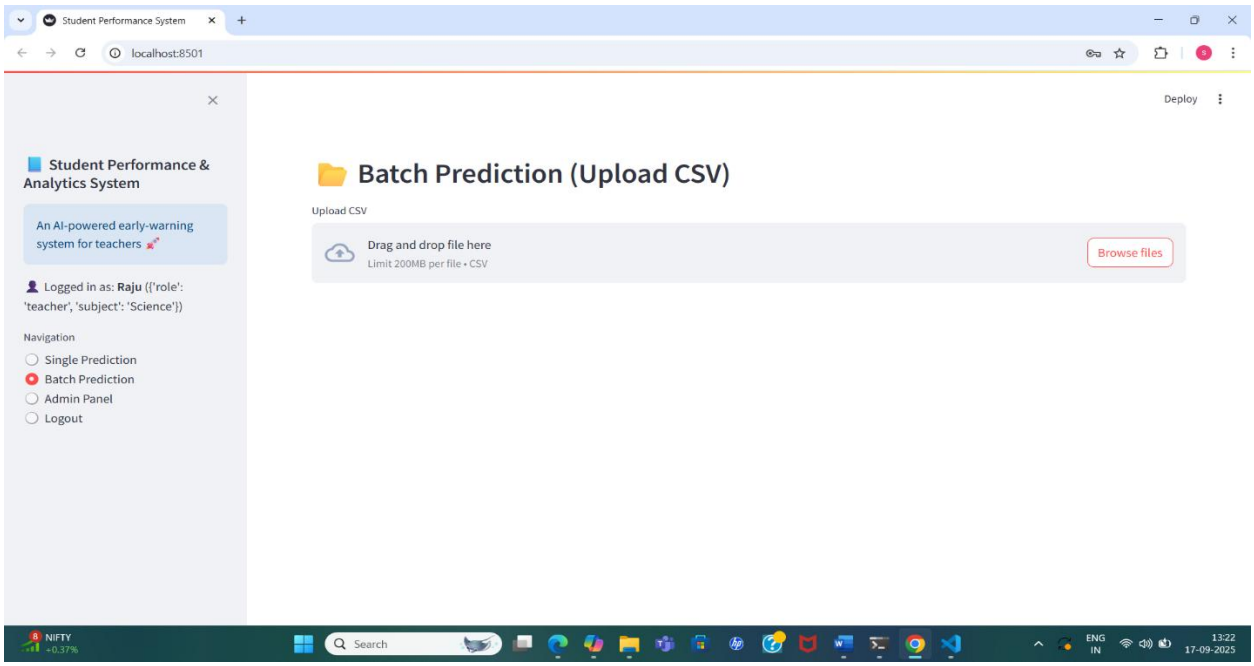


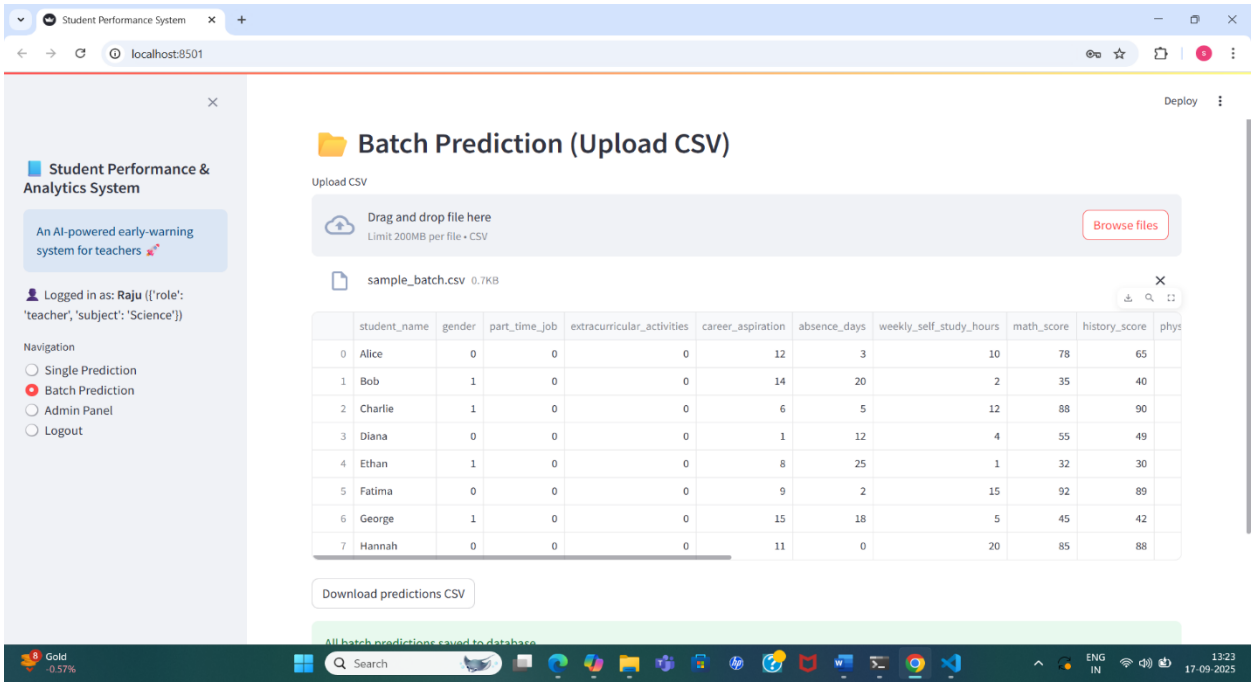
Fig 8 & 9: Explainable AI visualization showing feature contributions to prediction.

➢ Batch Prediction Upload

Fig 10: CSV upload interface for bulk student predictions.



➢ Batch Prediction Results Table



	student_name	gender	part_time_job	extracurricular_activities	career_aspiration	absence_days	weekly_self_study_hours	math_score	history_score
0	Alice	0	0	0	12	3	10	78	65
1	Bob	1	0	0	14	20	2	35	40
2	Charlie	1	0	0	6	5	12	88	90
3	Diana	0	0	0	1	12	4	55	49
4	Ethan	1	0	0	8	25	1	32	30
5	Fatima	0	0	0	9	2	15	92	89
6	George	1	0	0	15	18	5	45	42
7	Hannah	0	0	0	11	0	20	85	88

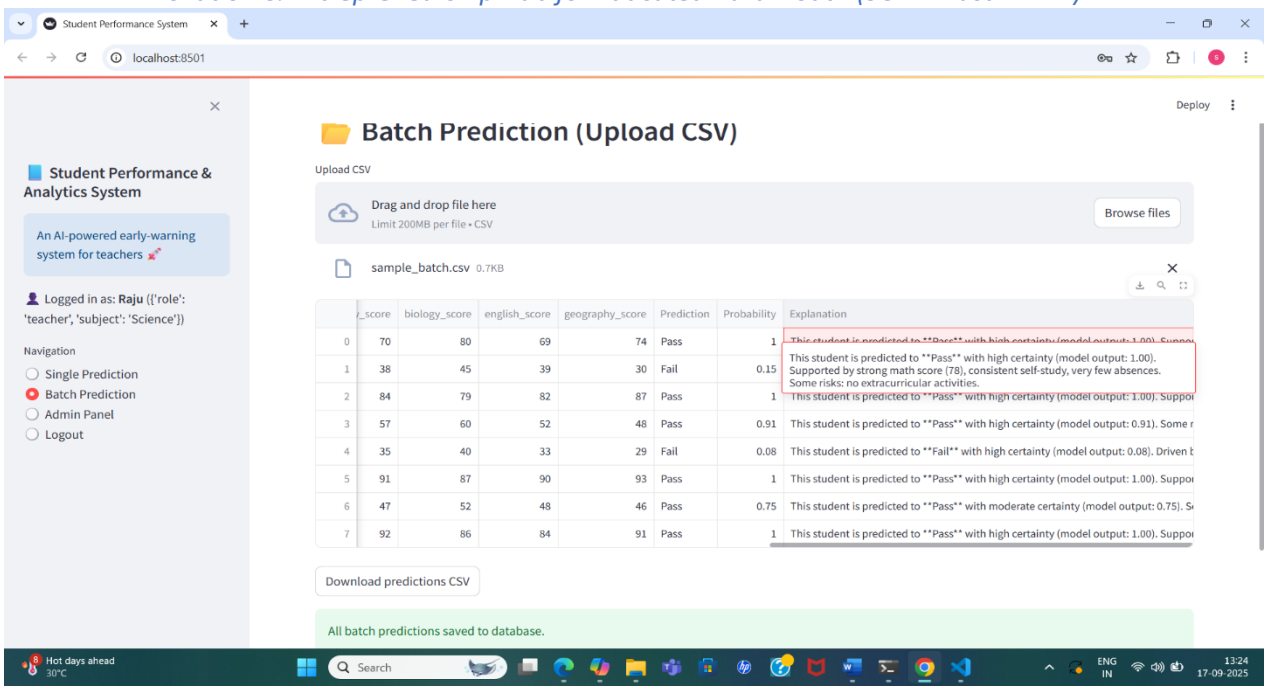


Fig 11 & 12: Predictions for multiple students with explanations and export option.

- **Admin Login**

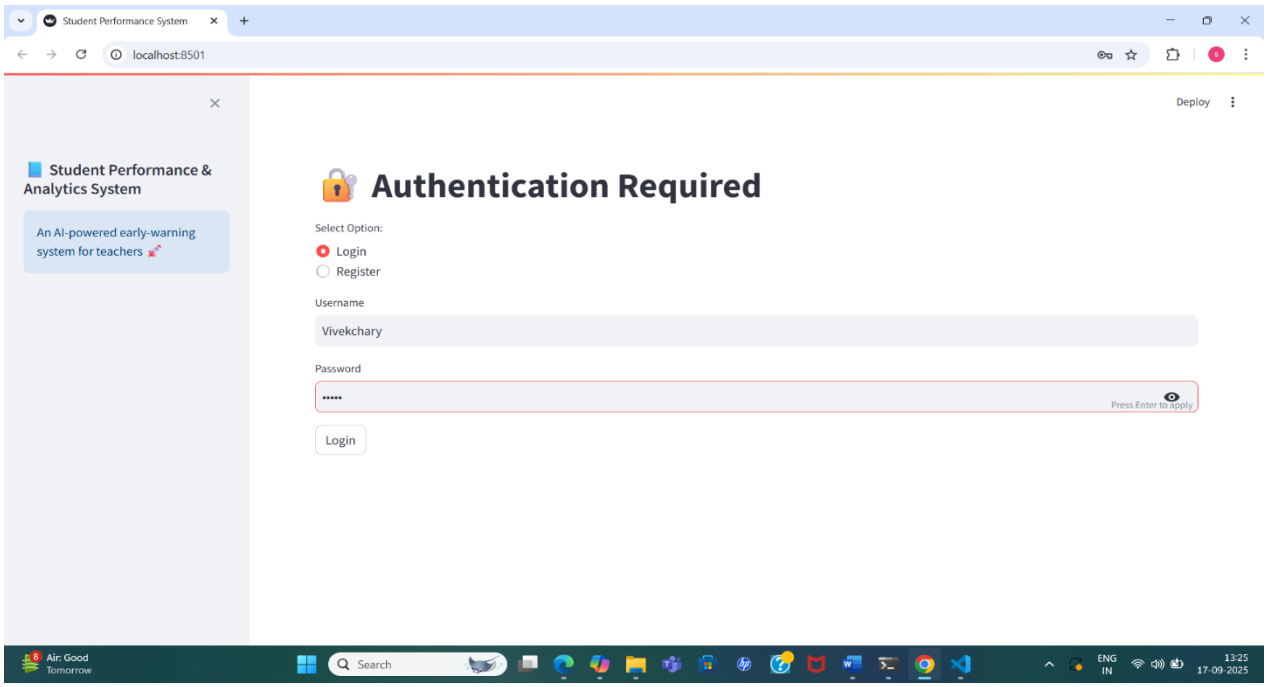


Fig 13: Admin login to view the prediction history

➢ Admin panel Page

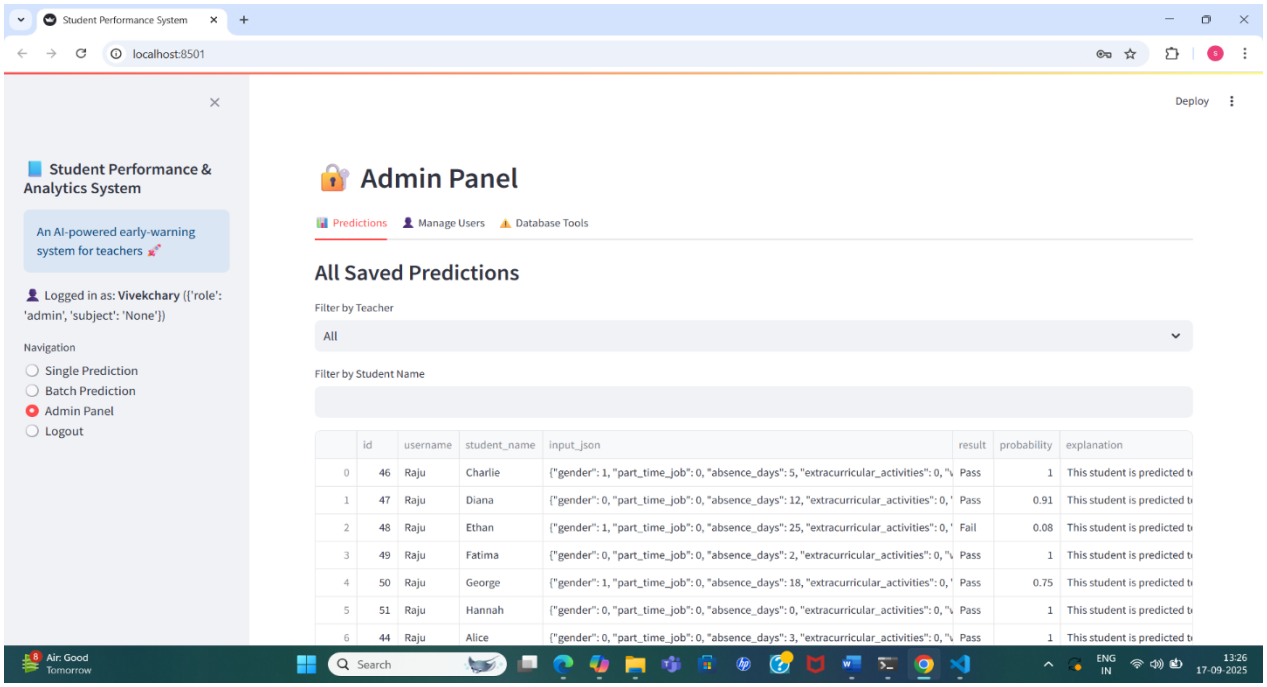


Fig 14: Admin page Centralized database view of all predictions with download option.

➢ Admin filters predictions

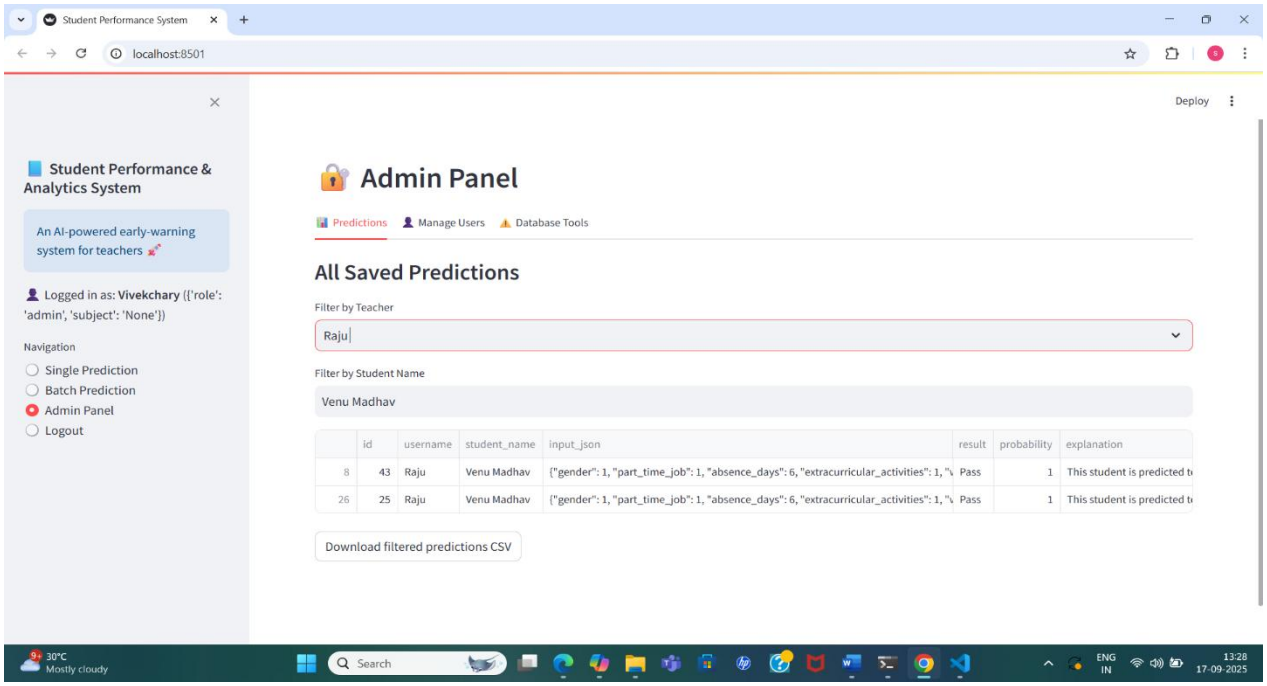


Fig 15: Admin filtering the predictions based in the teachers and student

➢ Admin User Management

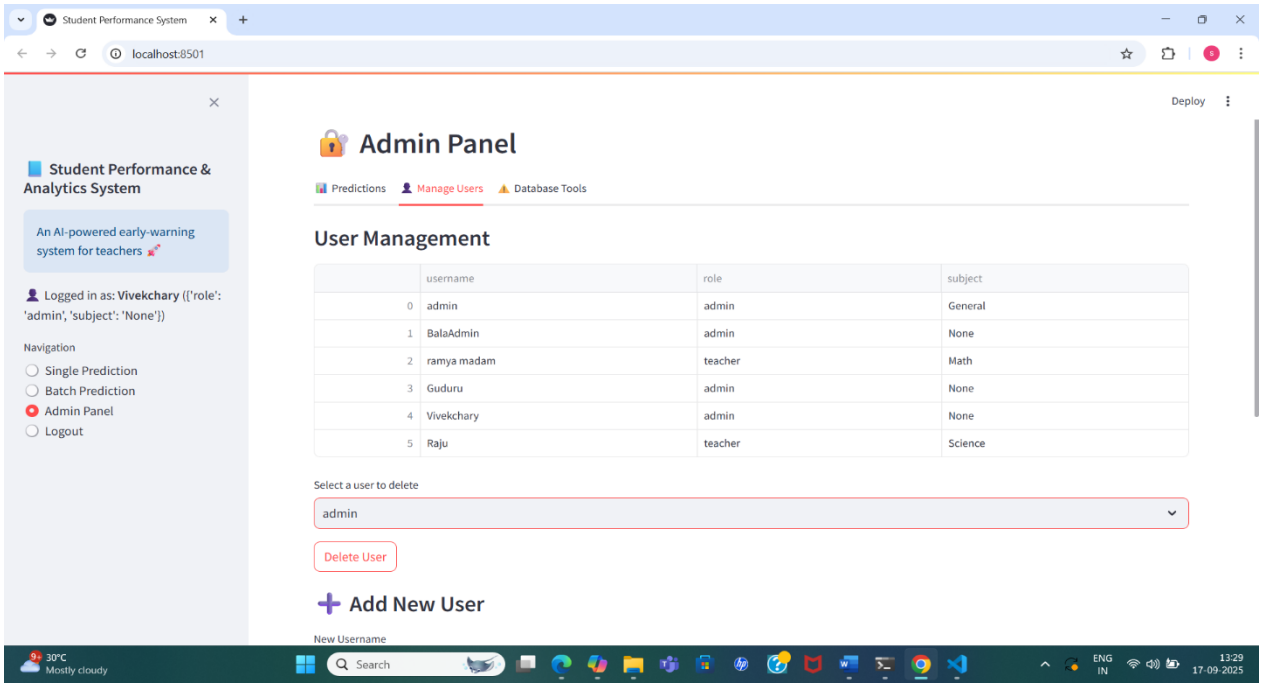


Fig 16: Admin deleting a user

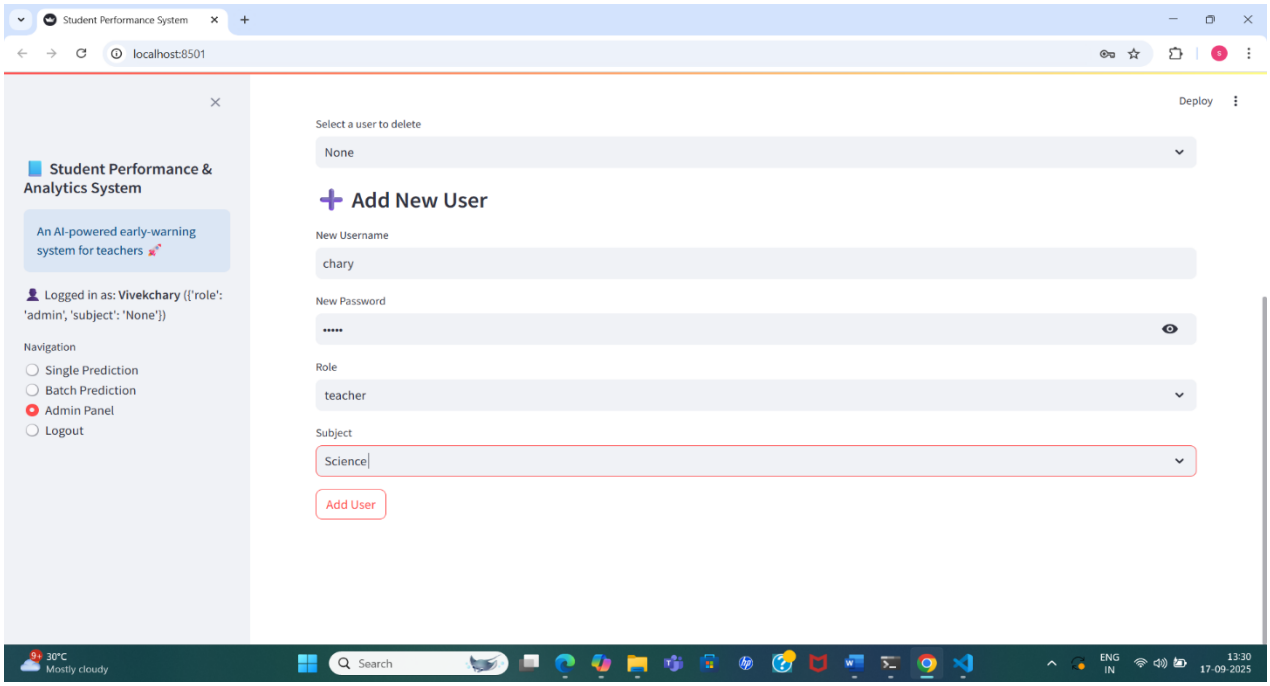


Fig 17: Admin adding a new user

➢ Admin Database Tools

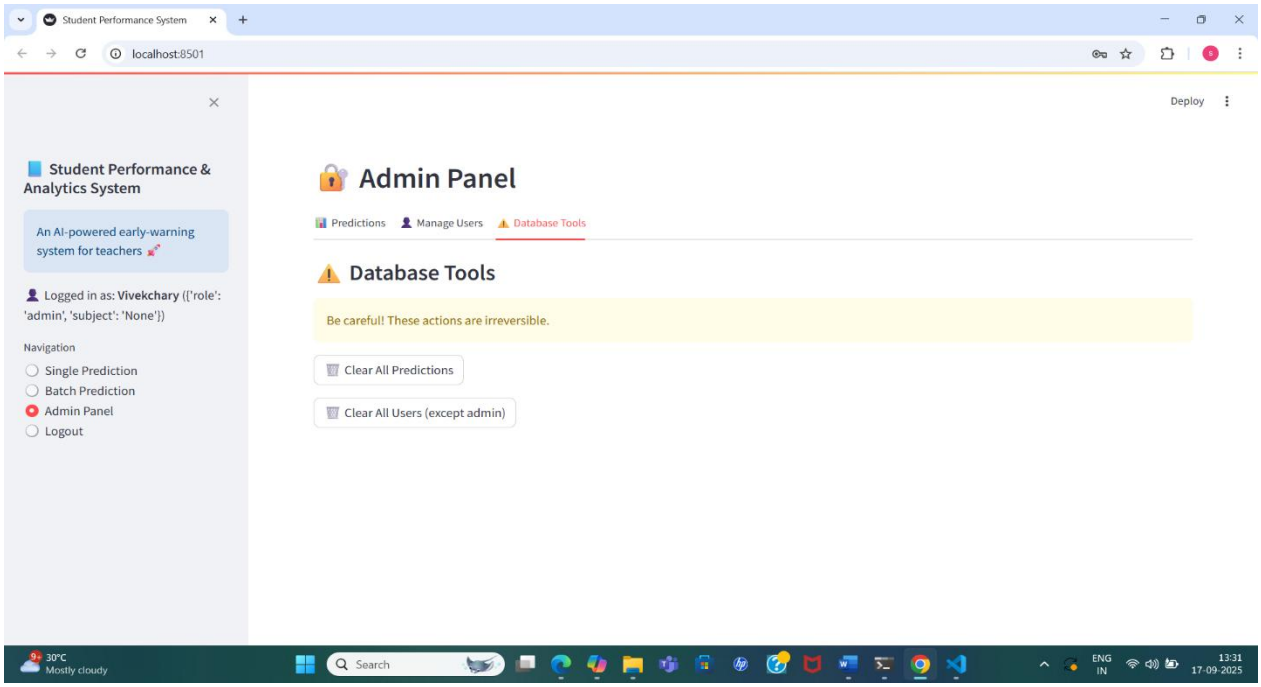


Fig 18: Admin can clear all the predictions and all the users from the database

GitHub Repo URL for Project Files

https://github.com/Vivekchary2607/AI-Powered-EduPredict-Student-Performance-Analytics-System.git

✅ Conclusion

This project demonstrates the practical application of Artificial Intelligence and Machine Learning in predicting student outcomes with transparency and usability. By integrating supervised learning models like Random Forest and Logistic Regression, and enhancing interpretability through SHAP, the system provides actionable insights for educators. The Streamlit-based interface and secure data handling via SQLite ensure accessibility and reliability for end users.

Through this documentation, we've outlined the technical foundation, implementation details, and social impact of the project. Future enhancements—such as cloud deployment, mobile integration, and expanded datasets—can further elevate its effectiveness and reach. This work not only showcases technical proficiency but also reflects a commitment to solving real-world educational challenges using intelligent systems.