SMARTEDU+:ADVANCED MACHINE LEARNING SYSTEM FOR EARLY PREDICTION OF STUDENT DROPOUTS

A PROJECT REPORT

Submitted by

MOHANA POORANI R [211423104379]

MAMTHA J [211423104356]

in partial fulfillment for the award of the degree

of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

PANIMALAR ENGINEERING COLLEGE

(An Autonomous Institution, Affiliated to Anna University, Chennai)

APRIL 2025

PANIMALAR ENGINEERING COLLEGE

(An Autonomous Institution, Affiliated to Anna University, Chennai)

BONAFIDE CERTIFICATE

Certified that this project report **”SMARTEDU+: ADVANCED MACHINE LEARNING SYSTEM FOR EARLY PREDICTION OF STUDENT DROPOUTS”** is the bonafide work of MOHANA POORANI R [211423104379], MAMTHA J [211423104356] who carried out the project work under my supervision.

Signature of the HOD with date	**Signature of the Supervisor with date**
DR L.JABASHEELA M.E., Ph.D.,	**Dr.V.SUBEDHA,M.E.,Ph.D.,**
PROFESSOR AND HEAD,	**SUPERVISOR,**
Department of Computer Science and Engineering, Panimalar Engineering College, Chennai - 123	Department of Computer Science and Engineering, Panimalar Engineering College, Chennai - 123

Certified that the above candidate(s) was examined in the End Semester Project Viva-Voice Examination held on ....

INTERNAL EXAMINER **EXTERNAL EXAMINER**

DECLARATION BY THE STUDENT

We MOHANA POORANI R [211423104379],MAMTHA J [211423104356] hereby declare that this project report titled **—"SMARTEDU+:ADVANCED MACHINE LEARNING SYSTEM FOR EARLY PREDICTION OF STUDENT DROPOUTS"** ‖ , under the guidance of DR. K. SANGEETHA is the original work done by us and we have not plagiarized or submitted to any other degree in any university by us.

Name of the student(S)

ACKNOWLEDGEMENT

Our profound gratitude is directed towards our esteemed Secretary and Correspondent, **Dr. P. CHINNADURAI, M.A., Ph.D**., for his fervent encouragement. His inspirational support proved instrumental in galvanizing our efforts, ultimately contributing significantly to the successful completion of this project

We want to express our deep gratitude to our Directors, **Tmt. C. VIJAYARAJESWARI, Dr. C. SAKTHI KUMAR, M.E., Ph.D., and Dr. SARANYASREE SAKTHI KUMAR, B.E., M.B.A., Ph.D.,** for graciously affording us the essential resources and facilities for undertaking of this project.

Our gratitude is also extended to our Principal, **Dr. K. MANI, M.E., Ph.D.,** whose facilitation proved pivotal in the successful completion of this project.

We express our heartfelt thanks to **Dr. L. JABASHEELA, M.E., Ph.D.,** Head of the Department of Computer Science and Engineering, for granting the necessary facilities that contributed to the timely and successful completion of project.

We would like to express our sincere thanks to **Project Coordinator** ………………………and **Project Guide** ……………………… and all the faculty members of the Department of CSE for their unwavering support for the successful completion of the project.

NAME OF THE STUDENT(S)

PROJECT COMPLETION CERTIFICATE

(ONE PAGE ONLY)

TABLE OF CONTENTS

ABSTRACT

Student dropout is a significant issue for educational institutions and leads to loss of academics, loss of social investment and relationships, and loss of economic investment. The SmartEdu+ project presents an intelligent, data-driven solution using a combination of machine learning and explainable AI techniques for dropout risk forecasting in students. The system employs demographic, academic, and behavioral information and identifies students at high dropout risk and unravels the drivers of the risk. The use of SHAP (SHapley Additive exPlanations) allows the easy interpretation of predictions by the stakeholders such that the reason for student dropout is clear in the minds of the educators. The dashboard also includes interaction-based visualizations for the exploration of trends and correlations in the data and a feedback mechanism for the students for raising dropout concerns. SmartEdu+ aims for educational institutions the ability of making informed interventions and reducing the dropout rates and contains a supportive learning environment. The project exhibits the effectiveness of the integration of high-end analytics and visualizations and user feedback in addressing real-life applications in the educational arena.

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

1. INTRODUCTION

1.1 Overview:

Education is broadly diagnosed as one of the maximum effective equipment for fostering person growth, lowering inequality, and using social and monetary development. Access to fine training allows beginners to gather knowledge, broaden crucial skills, and construct resilience withinside the face of demanding situations. Despite enormous improvements in instructional infrastructure and policies, scholar dropout is still an essential issue in lots of elements of the international. The phenomenon of college students discontinuing their training in advance ends in unfavourable outcomes now no longer simplest for the scholars themselves however additionally for his or her families, communities, and broader societal progress.

The underlying reasons of scholar dropout are frequently multifaceted and complex. Factors including socio-monetary hardships, insufficient parental aid, loss of get entry to getting to know sources, bad educational performance, and intellectual fitness demanding situations make contributions to college students' disengagement from the getting to know process. Furthermore, systemic obstacles including transportation problems, distance from instructional establishments, and shortage of economic help exacerbate the dropout hassle, mainly amongst college students from inclined populations.

In latest years, the sector of synthetic intelligence (AI) and gadget getting to know (ML) has emerged as a promising road for addressing real-international demanding situations throughout numerous domains, inclusive of training. By harnessing information and analytical techniques, AI-primarily based totally answers can offer actionable insights that allow stakeholders to make knowledgeable choices and take proactive measures. One such answer is the predictive modeling of scholar dropout danger, which assists instructional establishments, policymakers, and aid agencies in figuring out at-danger college students early and designing focused interventions.

The SmartEdu+ : Student Dropout Prediction Dashboard is an revolutionary assignment evolved to address this mission with the aid of using combining gadget getting to know algorithms with explainable AI equipment. The machine makes use of historic information from college students, inclusive of demographic information, educational performance, own circle of relatives background, and different applicable capabilities, to expect the chance of dropout. Additionally, it affords visualizations and causes that assist educators and directors apprehend the elements contributing to dropout danger.

The inclusion of capabilities including interactive dashboards, real-time prediction interfaces, and remarks mechanisms guarantees that the machine is user-pleasant and accessible.

With an emphasis on transparency and interpretability, the assignment consists of equipment including SHAP (SHapley Additive exPlanations) to give an explanation for the predictions made with the aid of using the model. These causes empower customers to apprehend how particular elements have an impact on dropout danger, facilitating knowledgeable decision- making and fostering consider withinside the machine's recommendations.

Through this assignment, we purpose to bridge the space among information analytics and academic interventions, permitting stakeholders to aid college students extra effectively. The remaining intention is to make contributions to lowering dropout prices, enhancing scholar retention, and improving the general fine of training with the aid of using imparting actionable insights and fostering a subculture of information- knowledgeable instructional aid.

1.2 Problem Definition:

Despite extended efforts to increase instructional get entry to and enhance getting to know consequences, dropout prices stay a enormous mission in faculties and better training establishments. Students face severa barriers that prevent their cappotential to live enrolled and be triumphant academically. These demanding situations are mainly reported in areas in which monetary disparities, infrastructural limitations, and shortage offerings prevail.

The hassle of scholar dropout manifests in numerous ways:

- Academic struggles: Students who face problems in retaining up with coursework or attaining excellent grades are much more likely to disengage from faculty.
- Financial constraints: Families with restricted sources might not be capable of have the funds for training fees, books, or different crucial instructional expenses, forcing college students to stop their studies.
- Family issues: Lack of parental guidance, aid, or involvement frequently correlates with better dropout prices, as college students may also lack motivation, emotional aid, or established getting to know environments.

- Mental fitness and well-being: Anxiety, depression, and pressure make contributions to reduced focus, absenteeism, and disengagement from faculty activities.
- Geographical and infrastructural obstacles: Students residing in faraway regions or areas with bad transportation structures may also locate it tough to get entry to faculties regularly.
- Technological gaps: Inadequate get entry to to the net or getting to know gadgets impedes college students' CA potential to take part in on line getting to know, mainly applicable withinside the post-pandemic era.

Addressing those demanding situations calls for well-timed identity of college students at danger and intervention techniques tailor-made to their particular circumstances. However, instructional establishments frequently lack the equipment to systematically examine dropout styles and expect person scholar danger with accuracy and reliability. Existing techniques may also depend upon anecdotal observations or incomplete information, main to inconsistent aid and ignored possibilities for early intervention.

The SmartEdu+ assignment is designed to cope with this hole with the aid of using imparting a strong predictive framework that integrates numerous scholar-associated elements right into a cohesive model. By leveraging gadget getting to know algorithms educated on historic information, the machine predicts dropout danger for every scholar and identifies the maximum influential elements contributing to this danger. Furthermore, the dashboard’s visualizations and causes offer clean insights into styles and trends, permitting faculty directors, teachers, and counsellors to take knowledgeable actions.

The hassle definition for this assignment may be summarized as follows:

1. How can instructional establishments systematically discover college students susceptible to losing out?
2. What are the important thing elements contributing to dropout, and the way can they be defined to stakeholders?
3. How can information-pushed insights be incorporated into intervention techniques to lessen dropout prices?

By growing a obvious and interactive dashboard that predicts dropout danger and explains its underlying reasons, SmartEdu+ targets to empower educators and college students alike. This machine fosters a proactive technique to scholar retention, in which focused aid is supplied earlier than dropout occurs, thereby enhancing instructional consequences and selling equitable get entry to to getting to know possibilities.

LITERATURE REVIEW

2. LITERATURE REVIEW

2.1 Introduction:

The hassle of pupil dropout has been a difficulty for academic establishments worldwide. Over the years, many researchers, policymakers, and practitioners have explored different factors contributing to pupil disengagement and evolved equipment to expect or save you dropout. The reason of this literature assessment is to offer the contemporary expertise of dropout prediction fashions, using system mastering strategies in training, the effect of socioeconomic elements, and the position of explainable AI (XAI) in enhancing academic outcomes. This assessment synthesizes present studies and highlights gaps that our challenge targets to address.

2.2 Dropout in Education: Causes and Implications:

Student dropout is a multifaceted hassle inspired through instructional, economic, psychological, and environmental elements. Researchers consisting of Rumberger (2011) and Alexander et al. (2001) have talked about that loss of instructional guide, bad attendance, low grades, monetary hardship, and own circle of relatives troubles are extensive members to dropout rates. Dropout now no longer handiest impacts the person's destiny potentialities however additionally has broader socioeconomic outcomes consisting of accelerated unemployment, crime, and poverty (Jimerson et al., 2000).

The dropout phenomenon is specially typical in low-profits and marginalized groups wherein sources for instructional guide are scarce. Studies have continuously proven that dropout prevention applications that target early identity and personalised interventions are greater powerful than those who depend completely on punitive measures (Dynarski &Gleason, 2002)**.**

2.3 Existing Approaches to Dropout Prediction:

Early dropout prediction fashions have been based on statistical evaluation, consisting of logistic regression and choice bushes. These fashions trusted constrained datasets and centered on attendance and grade factor averages as number one predictors. However, with the proliferation of information and improvements in computational strategies, system mastering fashions consisting of Random Forest, Support Vector Machines, and Neural Networks have won prominence (Kotsiantis et al., 2004).

Several research have implemented system mastering to dropout prediction:

- Cortez and Silva (2008) used choice bushes and logistic regression on pupil overall performance datasets and executed promising accuracy levels.
- Vayansky and Kumar (2020) reviewed dropout prediction fashions throughout better training and highlighted ensemble fashions as powerful equipment.
- Khosravi et al. (2018) hired neural networks to expect at-threat college students in on- line mastering platforms, demonstrating excessive predictive electricity however constrained explainability

Despite those advances, demanding situations persist, including:

- Imbalanced datasets wherein dropout times are underrepresented.
- Overfitting because of complicated fashions.
- Lack of transparency in algorithmic predictions.

2.4 Socioeconomic Factors and Their Role:

Several research emphasize the position of socioeconomic variables in dropout rates. Factors consisting of own circle of relatives profits, parental training, employment status, and get admission to to mastering sources strongly correlate with pupil retention. For example:

- A look at through UNESCO (2015) discovered that scholars from households with low profits are two times as in all likelihood to drop out earlier than finishing secondary training.
- Research through Sirin (2005) determined that parental training impacts children's motivation, mastering habits, and educational achievement, not directly affecting dropout rates.

SYSTEM ANALYSIS

3. SYSTEM ANALYSIS

3.1 Existing System :

Currently, maximum academic establishments depend upon conventional strategies to reveal and aid college students susceptible to losing out. These strategies consist of periodic attendance tracking, instructional overall performance reviews, and guide exams via way of means of instructors or counselors. However, this method has numerous limitations:

- Lack of Automation: Data series and evaluation are achieved manually, which may be time-ingesting and susceptible to human error.
- Delayed Interventions: Students in want of aid are frequently diagnosed too late, lowering the effectiveness of intervention programs.
- Subjective Decision-Making: Identifying at-chance college students is primarily based totally on anecdotal proof and instructor observations as opposed to information-pushed insights.
- No Predictive Capabilities: There is confined cappotential to are expecting destiny dropout dangers or apprehend the effect of different factors contributing to pupil disengagement.

Consequently, this conventional method frequently ends in incomplete exams and reactive, as opposed to proactive, aid efforts.

3.2 Proposed System :

The proposed device, SmartEdu+, introduces a information-pushed, automatic method to perceive college students susceptible to losing out. The key additives and functionalities consist of:

- ❖ Predictive Modeling – The device makes use of gadget getting to know algorithms, in particular XGBoost, to investigate pupil information and are expecting dropout chances primarily based totally on historic patterns.
- ❖ Explainability – To make sure transparency, SHAP (SHapley Additive exPlanations) values are used to spotlight how every function contributes to the

Zediction, assisting educators and college students apprehend the motives at the back of a excessive or low chance score.

- Interactive Dashboard – A person-pleasant interface constructed the use of Streamlit permits stakeholders to enter information, view predictions, discover precise function explanations, and visualize traits with graphs and charts.
- Student Feedback – The dashboard consists of a remarks shape in which college students can percentage motives for thinking about dropout, imparting qualitative information that dietary supplements quantitative predictions

3.3 Development Environment :

The improvement surroundings for the SmartEdu+ device is designed to aid information-pushed evaluation, gadget getting to know, and interactive visualization in a continuing and green manner. Below are the important thing additives used for the duration of the improvement:

- Programming Language – Python

Python is the number one programming language used for this mission because of its wealthy environment for information evaluation, gadget getting to know, and visualization. It gives libraries along with Pandas for information manipulation, Scikit-study and XGBoost for predictive modeling, and Streamlit for constructing interactive internet applications.

- Machine Learning Libraries:

XGBoost – Used for growing a rather correct type version that predicts pupil dropout primarily based totally on diverse features.

SHAP (SHapley Additive exPlanations) – Used for decoding and explaining the predictions made via way of means of the gadget getting to know version, imparting insights into function importance.

- Data Handling Libraries

Pandas – For green information manipulation, cleaning, and transformation of

pupil Datasets.

NumPy – For numerical computations and matrix operations that underpinning to know algorithms.

- Visualization Libraries:

Matplotlib & Seaborn – Used to create static plots for information exploration and evaluation.

Plotly – Used for interactive visualizations along with pie charts, bar graphs, and gauges to higher speak information traits and predictions with inside the dashboard.

- Web Application Framework

Streamlit – Used to create the person interface in which customers can enter information, get hold of predictions, view explanations, and discover visualizations. It permits real-time interplay with the version and clean deployment.

- File Handling & Storage

CSV Files – Used to save and control pupil information along with attendance, examination marks, and private information. Feedback information is likewise recorded in CSV layout for clean get admission to and management

- Development Tools IDEs

Visual Studio Code (VS Code) or PyCharm – Recommended code editors for writing and debugging Python scripts.

Jupyter Notebook – Used for the duration of version improvement and exploratory information evaluation earlier than integrating functionalities into the very last utility.

- Operating System

Thedevice is advanced on Windows 10/11, with vital Python libraries established through pip, making sure compatibility throughout devices.

- Version Control & Collaboration

Git – For model manipulate and handling code adjustments collaboratively for the duration of the improvement phase.

- Deployment Environment

The utility is designed to run domestically or may be deployed on cloud systems along with Heroku

3.4 FEASIBILTY STUDY

3.4.1 Technical Feasibility:

The proposed machine is technically viable because it makes use of broadly followed and dependable technologies.

Technology Stack:

- ✧ Python because the middle programming language (simple, flexible, and broadly utilized in ML).
- ✧ Scikit-learn &amp; XGBoost for version education and dropout threat prediction.
- ✧ SHAP (SHapley Additive Explanations) for version interpretability.
- ✧ Streamlit for an interactive and user-pleasant internet interface.
- ✧ Matplotlib &amp; Plotly for facts visualization.

Hardware Requirements:

- ✧ Minimum: eight GB RAM, multi-middle processor, 500 MB unfastened disk space
- ✧ No unique GPU is needed because the dataset length is small (~three hundred facts).

Software Requirements:

- ✧ Python 3.10+
- ✧ Libraries: Pandas, NumPy, Scikit-learn, XGBoost, SHAP, Matplotlib, Plotly, Streamlit.

Conclusion: The era is easily to be had and may be carried out with out unique hardware/software program costs.

3.4.2 Operational Feasibility:

Operational feasibility measures whether or not the machine can characteristic easily in actual-international usage.

Ease of Use:

- ➢ The dashboard makes use of Streamlit, which presents a simple, intuitive UI.
- ➢ Teachers, directors, or policymakers and not using a technical history can use it easily.

System Functions:

- ➢ Predict dropout threat for person college students.
- ➢ Provide SHAP causes for why a prediction is excessive or low threat.

- Offer dataset-degree insights via visualizations.
- Collect remarks from college students to apprehend actual dropout reasons.

User Involvement:

- School directors input facts.
- Teachers view threat indicators and offer help to college students.
- Students make contributions remarks on their dropout concerns.
- Conclusion: The machine is enormously realistic in instructional establishments with minimum education.

3.4.3 Economic Feasibility:

This have a look at tests whether or not the challenge is value-effective.

Development Cost:

- Software is constructed with open-supply libraries (Python, Scikit- learn, Streamlit, SHAP), so there aren't anyt any license costs.
- Hardware value is minimum because the machine runs on present computers.

Operational Cost:

- Very low – protection includes occasional updates to the dataset and retraining the version.
- Deployment may be performed on a unfastened or low-value server (Heroku/Streamlit Cloud).

Cost Savings:

- Early identity of dropout threat saves massive sources in retention programs.
- Helps allocate economic/educational help extra effectively.
- The challenge is value-effective, requiring handiest minimum setup and protection costs.

3.4.5 Legal Feasibility

- Data Privacy: Student facts need to be treated carefully. The challenge guarantees that handiest essential educational and demographic capabilities are used.
- Compliance: The machine aligns with popular instructional facts policies (no touchy

private facts like clinical or economic facts are exposed).

- Open-Source Tools: All libraries are open-supply, so there aren't any criminal regulations on usage.
- Legally viable so long as establishments observe privateness norms and use anonymized datasets.

3.4.6. Schedule Feasibility Development Time:

- Dataset Preparation &amp; Cleaning → 1–2 weeks.
- Model Training &amp; Testing → 1 week.
- Dashboard Development (Streamlit UI) → 2 weeks.
- SHAP Integration &amp; Visualizations → 1 week.
- Feedback &amp; Alert System → 1 week.
- Total Duration: ~6–7 weeks for a totally running machine.
- Conclusion: The challenge may be evolved and deployed in a quick timeline.

THEORETICAL BACKGROUND

4. THEORETICAL BACKGROUND

4.1 Implementation Environment:

4.1.1 Hardware Requirements:

- ✧ Minimum: eight GB RAM, multi-middle processor, 500 MB unfastened disk space.
- ✧ No unique GPU is needed because the dataset length is small (~three hundred facts).

4.1.2 Software Requirements:

- ✧ Software: Python 3.10+
- ✧ Libraries: Pandas, NumPy, Scikit-learn, XGBoost, SHAP, Matplotlib, Plotly, Streamlit.

4.1.3 TECHNOLOGIES USED:

The SmartEdu+ Student Dropout Prediction Dashboard is constructed the usage of a aggregate of Machine Learning algorithms, records processing libraries, and a net- primarily based totally visualization framework. Below is an in depth breakdown of the technologies:

4.1.3.1 Programming Language

Python 3.10+

- ➢ Core language used for version improvement and dashboard implementation.
- ➢ Provides effective libraries for gadget learning, visualization, and records processing.
- ➢ Widely followed in academia and enterprise because of simplicity and versatility.

4.1.3.2. Machine Learning Frameworks

Scikit-learn

- ➢ Used for preprocessing, encoding specific variables, splitting datasets, and simple ML utilities.
- ➢ Provides capabilities for accuracy, type reports, and version evaluation.

XGBoost

- ➢ Core version used for dropout prediction because of its performance and excessive accuracy.
- ➢ Handles specific and numerical capabilities effectively.
- ➢ Provides opportunity ratings for hazard type (low/excessive dropout hazard).

4.1.3.3. Explainable AI (XAI)

- ➢ SHAP (SHapley Additive Explanations)

- Provides explainability for ML predictions.
- Explains why a pupil is assessed as “excessive hazard” or “low hazard.”
- Visualizes characteristic contributions via bar plots and waterfall charts.

4.1.3.4 Data Processing and analysis Pandas

- Used for managing CSV datasets, cleansing records, and making use of transformations.
- Allows clean manipulation of pupil capabilities (Age, Marks, Attendance, etc.).

NumPy

- Provides numerical computation support.
- Used internally for vectorized operations in ML and SHAP.

4.1.3.5. Visualization Tools Matplotlib

- Used for plotting SHAP precis charts and characteristic significance bar graphs.

Plotly Express & Graph Objects

- Provides interactive and expert visualizations.
- Used for Pie Charts (Dropout distribution), Histograms (Study Hours), and Gauge Charts (Dropout Risk %).

4.1.3.6. Dashboard Development Streamlit

- Core framework for constructing the interactive net dashboard.
- Provides real-time enter fields for pupil records.
- Displays predictions, SHAP explanations, and dataset-stage insights.
- Handles Feedback series from college students (dropout reasons).

1.3 Architecture Overview:

The structure of the SmartEdu+ Student Dropout Prediction device is designed to offer an cease-to- cease answer this is efficient, scalable, interpretable, and consumer- pleasant. It integrates information collection, device mastering, clarification equipment, and visualization strategies right into a coherent framework that helps instructional decision- making.

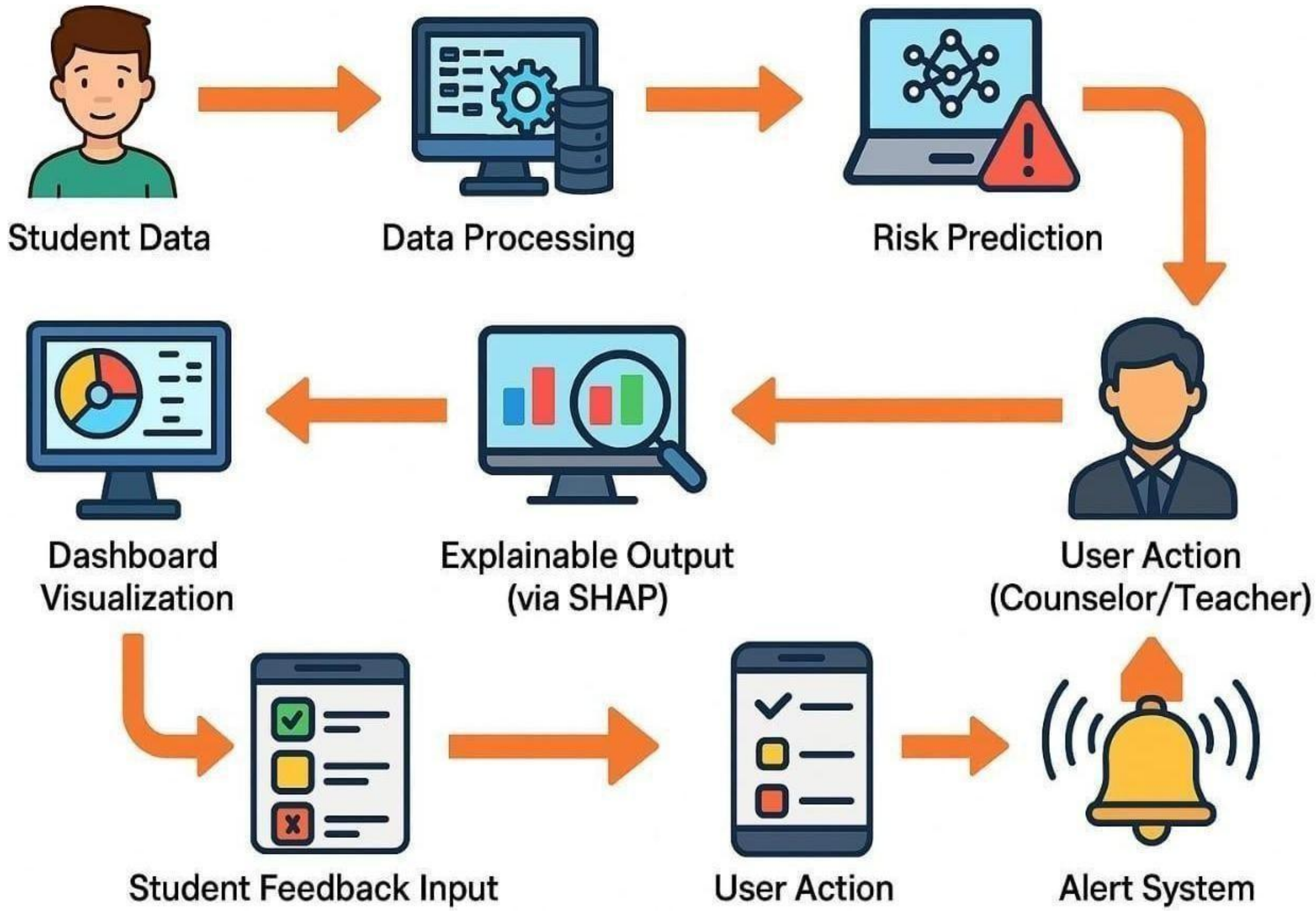


FIGURE 4.2 ARCHITECTURE OVERVIEW

4.2.1 User Interface Layer:

Built the use of Streamlit, the consumer interface gives an intuitive internet- primarily based totally platform in which college students, teachers, and directors can

effortlessly enter applicable data which includes age, gender, attendance, educational overall performance, and parental help.

The interface gives dropdown menus, sliders, and enter fields for choosing and coming into information, making sure that the manner is simple even for non- technical customers.

Once the information is entered, customers can post the shape and consider predictions instantly.

4.2.2 Data Processing Layer:

The uncooked information entered thru the consumer interface is proven and formatted for processing.

Categorical variables (which includes gender, own circle of relatives earnings, and parental help) are transformed into numerical values thru guide encoding. This step guarantees compatibility with the device mastering version.

The information is checked for lacking or wrong values, which enables keep the integrity of the predictions.

4.2.3 Machine Learning Layer:

The center of the device is an XGBoost type version that has been skilled on ancient dropout information.

This version predicts the chance of a pupil losing out primarily based totally at the furnished functions.

Alongside prediction, the version additionally estimates probabilities, permitting customers to apprehend how assured the device is in its output.

4.2.4 Explainable and Analytics Layer:

The device contains SHAP (SHapley Additive exPlanations), a effective device for explaining device mastering models.

SHAP enables customers apprehend the contribution of every characteristic to the very lastprediction. For example, it suggests how low attendance or terrible educational overall

performance can also additionally boom dropout hazard.

Interactive plots, which includes characteristic significance charts and waterfall explanations, assist visualize the effect of things on predictions.

4.2.5 Visualization Layer:

Aggregated information from more than one customers is displayed the use of charts created with Plotly.

Users can view trends, styles, and relationships among different factors and dropout rates.

Graphs which includes pie charts, bar charts, and histograms help evaluation of dropout distribution, have a look at habits, own circle of relatives earnings effects, and extra.

These visualizations help educators in figuring out regions in which interventions are maximum needed.

4.2.6 Feedback and Data Storage Layer:

The device permits customers to offer motives for losing out, which might be saved in a comments database.

This comments may be reviewed with the aid of using educators to apprehend the demanding situations college students face and to tailor applications that cope with the ones issues.

The garage mechanism guarantees that beyond information is preserved at the same time as permitting new entries to be submitted with out overwriting preceding comments.

4.2.7 Scalability and Integration:

The modular shape of the structure permits for clean expansion. New functions or datasets may be incorporated with out affecting current functionalities.

The device may be tailored to different schools, regions, or instructional applications with minor adjustments.

It additionally permits for destiny upgrades which includes cell compatibility, reporting dashboards, or integration with mastering control systems (LMS).

4.2.8 Security and Privacy:

User information is treated with care, making sure that touchy data is saved securely and accessed best with the aid of using legal personnel.

The device complies with information safety concepts to make sure that customers' believe is maintained.

The structure evaluate of SmartEdu+ demonstrates a unbroken float from information enter to prediction, clarification, and evaluation. By combining superior device mastering strategies with consumer-pleasant equipment and interactive visualizations, the device empowers instructional stakeholders to proactively cope with dropout issues. Its explainability and comments mechanisms make sure transparency and non-stop improvement, making it a precious device for instructional intervention strategies.

1.4 PROPOSED METHODOLY:

The proposed machine, SmartEdu+ : Student Dropout Prediction Dashboard, adopts a based method that mixes records-pushed system getting to know fashions with an interactive web-primarily based totally dashboard. The method guarantees correct prediction of scholar dropout dangers even as additionally imparting transparency, interpretability, and actionable insights for stakeholders (teachers, dad and mom, and administrators).

4.3.1 Data Collection:

- ✓ The number one dataset (student_dropout_300.csv) incorporates scholar demographic, academic, and socio-monetary statistics such as:
- ✓ Age, Gender, Family Income

✓ Attendance Percentage, Previous Exam Marks, Study Hours in step with Day

✓ Parental Support, Internet Access, Disciplinary Issues, etc.

✓ Additional records may be accrued from scholar remarks bureaucracy to enhance the dataset with self-pronounced motives for dropout tendencies.

4.3.2 Data Preprocessing:

✓ Cleaning: Missing or inconsistent values (e.g., “Unknown” parental education) are standardized.

✓ Encoding: Categorical variables (e.g., Gender, Family Income, Parental Support) are manually encoded into numerical representations.

✓ Normalization: Numerical attributes (e.g., Attendance, Marks) are scaled for uniformity.

✓ Splitting: Data is split into training (70%) ^{3}a ^{2}n d testing (30%) subsets.

4.3.3 Model Training:

✓ XGBoost Classifier is chosen for dropout prediction because of its:

✓ Ability to deal with blended specific and numerical records,

✓ Strong predictive accuracy,

✓ Support for probability-primarily based totally outputs.

✓ Training Objective: Predict whether or not a scholar is at High Risk (1) or Low Risk

(0) of dropout.

✓ Evaluation Metrics: Accuracy, Precision, Recall, F1-score, and ROC-AUC are used to evaluate performance.

4.3.4 Prediction Phase:

✓ The educated version (dropout_model.pkl) is deployed in the Streamlit dashboard.

✓ Users (teachers/administrators) can input scholar information through the enter form.

✓ The machine techniques the enter, applies encoding, and passes it to the educated version.

✓ Output:

➢ Predicted Label → High Risk or Low Risk.

➢ Dropout Probability Score → displayed as a percentage.

4.3.5. Explainability through SHAP:

✓ To make certain transparency in predictions, SHAP (SHapley Additive Explanations) is integrated.

✓ SHAP presents:

➢ Feature Importance Plots → displaying general elements influencing dropout

- ➢ Individual Student Explanations → waterfall plots indicating how precise capabilities prompted the prediction.
- ✓ This complements accept as true with withinside the version via way of means of explaining why a scholar is at risk.

4.3.6. Visualization Layer:

- ✓ Plotly & Matplotlib are used to offer interactive graphs for dataset-stage insights:
  - ➢ Dropout distribution (pie chart).
  - ➢ Marks distribution (histogram).
  - ➢ Study hours vs dropout (histogram).
  - ➢ Family profits vs dropout rate (stacked bar).
- ✓ These visualizations assist stakeholders apprehend styles and developments throughout the scholar population.

4.3.7. Feedback Collection:

- ✓ A devoted Feedback Module permits college students to publish their private motives for dropout tendencies (e.g., economic issues, loss of parental support, distance, etc.).
- ✓ This remarks is saved in a CSV log fo34r in addition evaluation and non-stop-machine improvement.
- ✓ It presents qualitative insights complementing the quantitative ML version.

4.3.8 MODULE DESIGN:

4.3.8.1 USECASE DIAGRAM

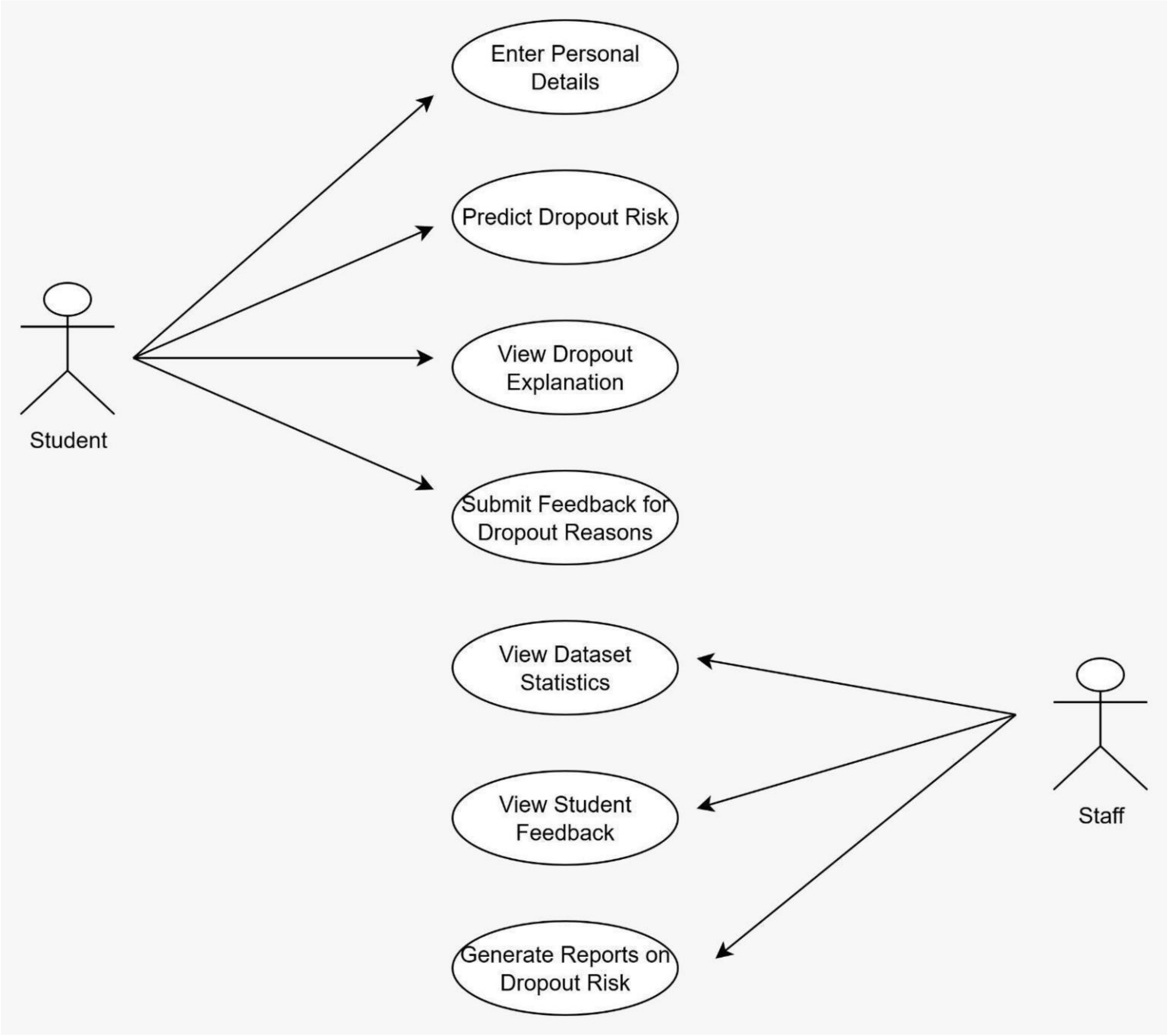


Figure 4.3.8.1 Use Case Diagram

4.3.8.2 CLASS DIAGRAM

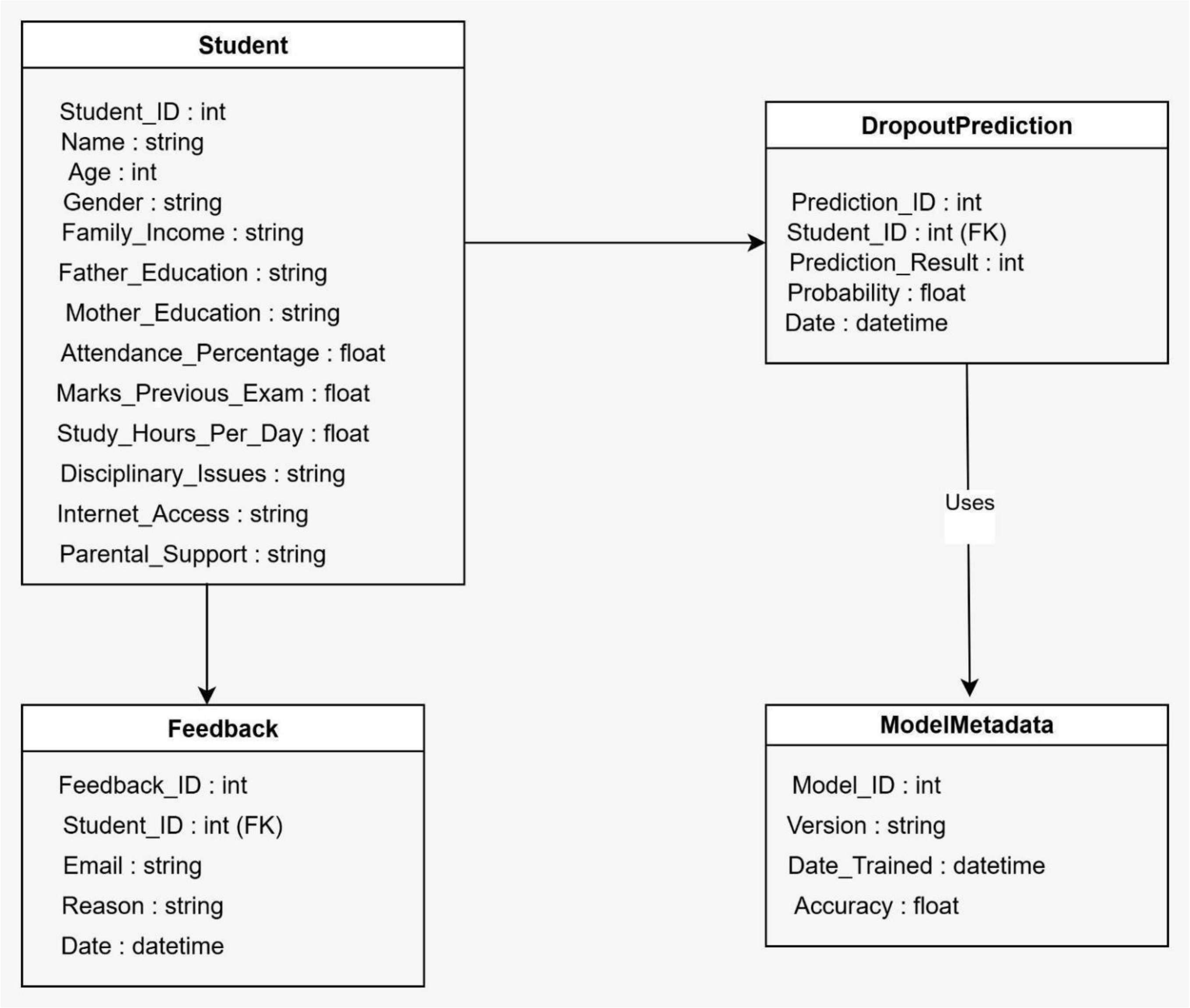
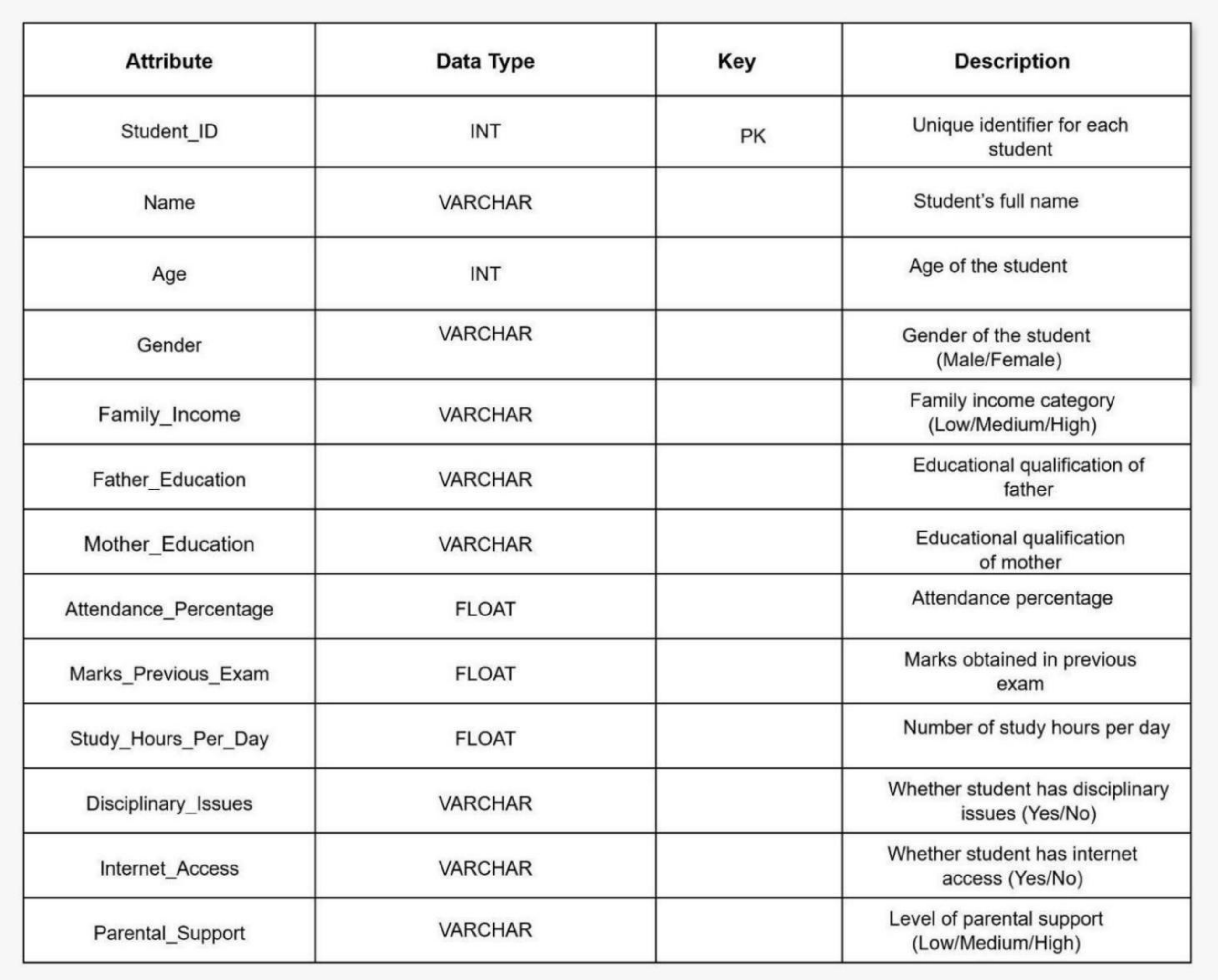


Figure 4.3.8.2 Class Diagram

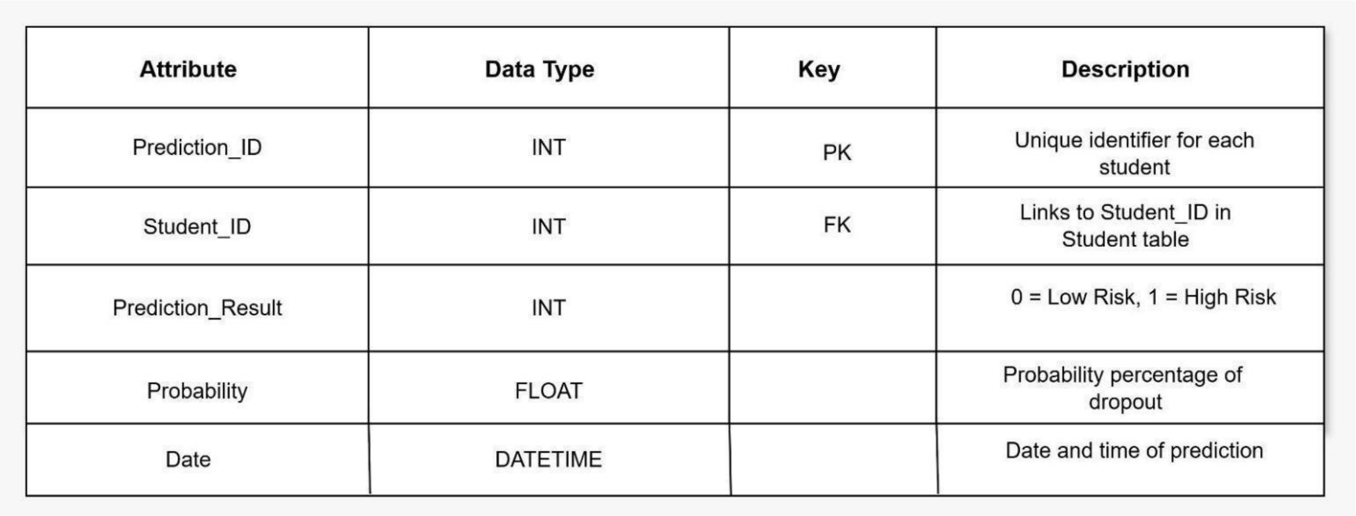
4.3.8.3 DATA DICTIONARY

4.3.8.3.1 STUDENT TABLE



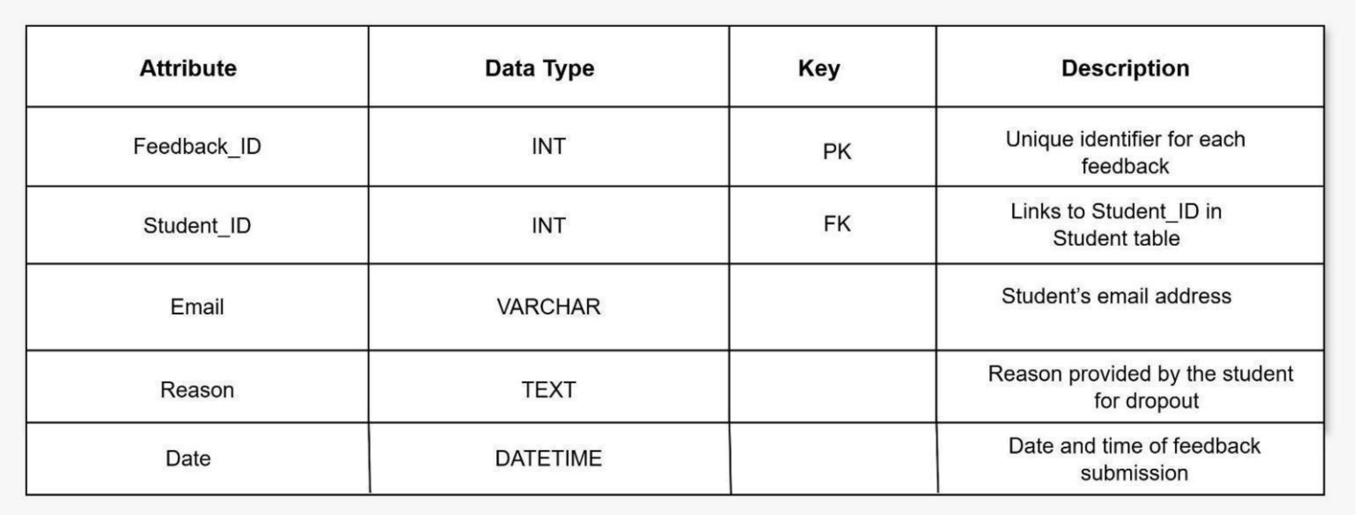
Attribute	Data Type	Key	Description
Student_ID	INT	PK	Unique identifier for each student
Name	VARCHAR		Student's full name
Age	INT		Age of the student
Gender	VARCHAR		Gender of the student (Male/Female)
Family_Income	VARCHAR		Family income category (Low/Medium/High)
Father_Education	VARCHAR		Educational qualification of father
Mother_Education	VARCHAR		Educational qualification of mother
Attendance_Percentage	FLOAT		Attendance percentage
Marks_Previous_Exam	FLOAT		Marks obtained in previous exam
Study_Hours_Per_Day	FLOAT		Number of study hours per day
Disciplinary_Issues	VARCHAR		Whether student has disciplinary issues (Yes/No)
Internet_Access	VARCHAR		Whether student has internet access (Yes/No)
Parental_Support	VARCHAR		Level of parental support (Low/Medium/High)

4.3.8.3.3 DROPOUT PREDICTION TABLE



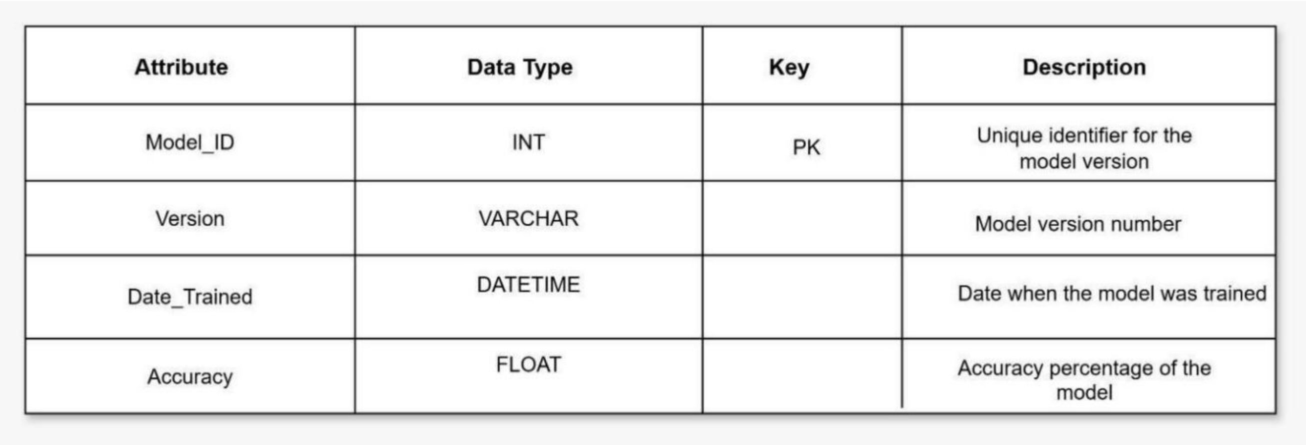
Attribute	Data Type	Key	Description
Prediction_ID	INT	PK	Unique identifier for each student
Student_ID	INT	FK	Links to Student_ID in Student table
Prediction_Result	INT		0 = Low Risk, 1 = High Risk
Probability	FLOAT		Probability percentage of dropout
Date	DATETIME		Date and time of prediction

4.3.8.3.4 FEEDBACK TABLE



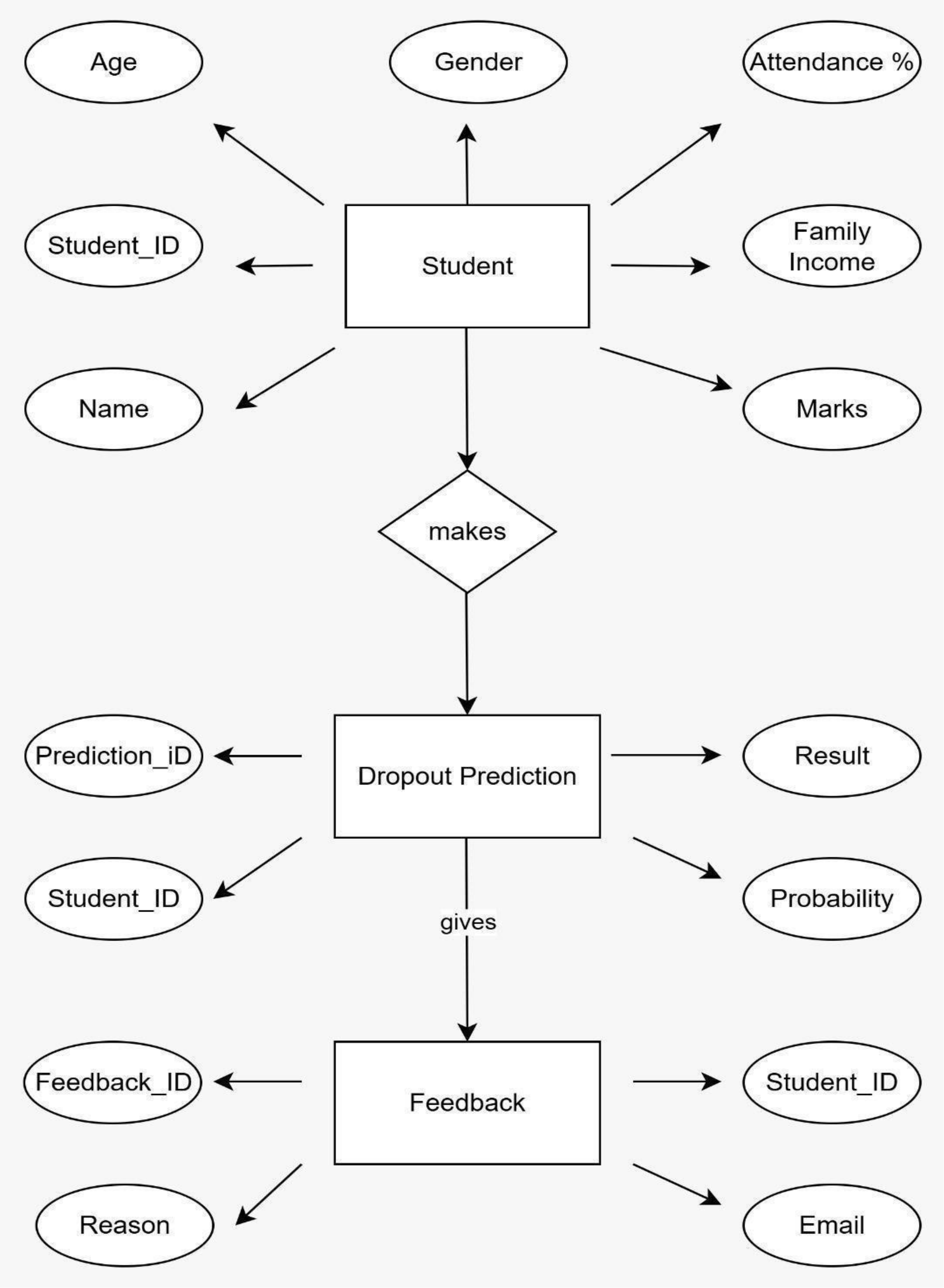
Attribute	Data Type	Key	Description
Feedback_ID	INT	PK	Unique identifier for each feedback
Student_ID	INT	FK	Links to Student_ID in Student table
Email	VARCHAR		Student's email address
Reason	TEXT		Reason provided by the student for dropout
Date	DATETIME		Date and time of feedback submission

4.3.8.3.5 MODEL METADATA TABLE



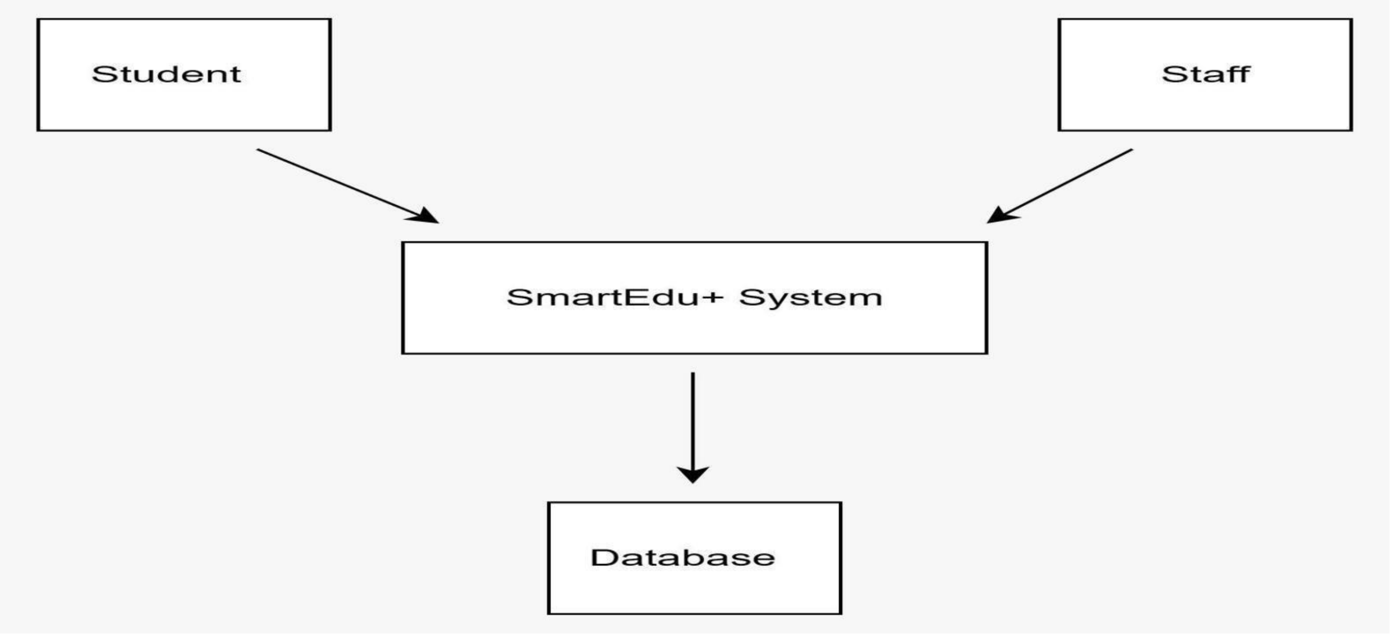
Attribute	Data Type	Key	Description
Model_ID	INT	PK	Unique identifier for the model version
Version	VARCHAR		Model version number
Date_Trained	DATETIME		Date when the model was trained
Accuracy	FLOAT		Accuracy percentage of the model

4.3.8.4 ER DIAGRAM

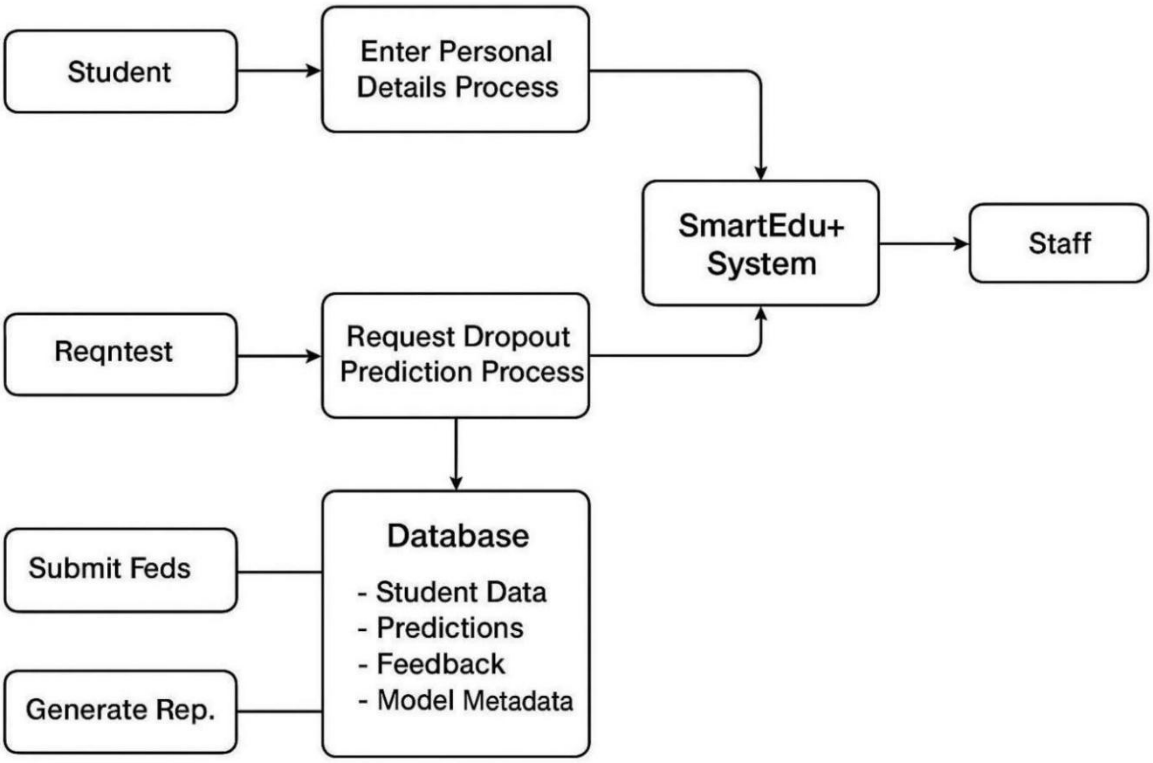


4.3.8.5 DATA FLOW DIAGRAM

4.3.8.5.1 LEVEL 0-CONTEXT DIAGRAM(DFD)



4.3.8.5.2 LEVEL 1- DETAILED DATA FLOW DIAGRAM



SYSTEM IMPLEMENTATION

5. SYSTEM IMPLEMENTATION

The SmartEdu+ dropout prediction machine is carried out thru modular integration, strong records preprocessing, device mastering models, explainability frameworks, and a person- pleasant internet interface. This guarantees accuracy, transparency, and simplicity of use for educators, college students, and administrators.

5.1 Modular Implementation:

The SmartEdu+ machine is prepared into interrelated modules, every having a selected responsibility. This modular layout guarantees scalability, maintainability, and efficiency. The key modules are:

5.1.1 Data Input Module:

- ✧ Collects scholar information (age, gender, marks, attendance, own circle of relatives income, parental support, etc.) thru an interactive internet form.
- ✧ Validates records to make certain consistency and correctness.

5.1.2 Data Preprocessing Module:

- ✧ Cleans, encodes, and transforms uncooked inputs into device-readable layout.
- ✧ Converts specific attributes (e.g., gender, own circle of relatives support) into numerical values the use of predefined mappings**.**

5.1.3 Prediction Module:

- ✧ Uses the XGBoost device mastering set of rules educated on ancient datasets.
- ✧ Provides each classification (high/low dropout threat) and possibility scores.

5.1.4 Explainability Module (SHAP):

- ✧ Integrates SHAP to interpret predictions.
- ✧ Displays worldwide characteristic significance and character reasons thru precis and waterfall plots.

5.1.5. Visualization Module:

- ✧ Generates interactive insights with graphs including bar charts, pie charts, histograms, and stacked plots.
- ✧ Helps customers discover dropout styles throughout socio-financial and educational dimensions.

5.1.6 Feedback Module:

- Enables college students to offer private motives for dropout consideration.
- Stores comments for educators and counselors to layout focused interventions.

5.1.7 User Interface Module:

- Built the use of Streamlit, making sure an intuitive and responsive experience.
- Combines all modules right into a unmarried interactive dashboard handy

5.2 Integration and Testing:

- All modules are incorporated in the Streamlit-primarily based totally interface.
- The machine is examined with a couple of datasets to make certain robustness.
- Test instances encompass legitimate inputs, lacking values, and excessive scenarios.
- Validation confirms the accuracy of predictions and reliability of visible insights.

5.3 Outcome of Implementation:

- The machine efficiently predicts dropout threat the use of XGBoost.
- SHAP-primarily based totally explainability guarantees accept as true with in predictions.
- Visualizations and comments modules offer actionable insights for interventions.
- The internet interface gives real-time interaction, making it handy to educators and college students.

RESULTS &DISCUSSIONS

6. RESULTS & DISCUSSION

6.1 TESTING:

Testing is an crucial section of the challenge lifecycle to make sure that the SmartEdu+ : Student Dropout Prediction Dashboard features efficiently, reliably, and meets its meant objectives. The gadget turned into examined at more than one levels, along with unit testing, integration testing, practical testing, and person popularity testing (UAT).

6.1.1 Unit Testing:

- ➢ Objective: Verify correctness of character components.
- ➢ Scope:
  - ✧ Model prediction feature examined with pattern scholar records.
  - ✧ Data preprocessing features (encoding, normalization).
  - ✧ SHAP rationalization module examined to make sure legitimate plots are generated.
  - ✧ Feedback shape examined to validate inputs and keep responses to CSV.
- ➢ Outcome: All gadgets accomplished as predicted with legitimate inputs; mistakess dealing with turned into applied for invalid or lacking data.

6.1.2 Integration Testing:

- ➢ Objective: Ensure exclusive modules paintings collectively seamlessly.
- ➢ Scope:
  - ✧ Integration among the Streamlit frontend and XGBoost model.
  - ✧ Testing whether or not person inputs are efficiently encoded and surpassed to the model.
  - ✧ SHAP values related efficiently with predictions.
  - ✧ Feedback shape incorporated with dataset storage.
- ➢ Outcome: All modules interacted properly; minor bugs (inclusive of mismatched function names) had been resolved throughout this stage.

6.13 Functional Testing:

- ➢ Objective: Validate whether or not the gadget meets practical requirements.
- ➢ Tests Conducted:

✓ Prediction tab efficiently shows hazard fame and probability.

✅ SHAP rationalization tab offers function significance and character explanations.

✅ Visualization tab suggests significant insights (pie chart, histograms, stacked bar).

✅ Feedback tab accepts and shops scholar responses.

- ➢ Outcome: All center functionalities matched predicted results.

6.1.4 Performance Testing:

- ➢ Objective: Test performance and reaction time.
- ➢ Tests Conducted:
  - ✧ Dashboard loaded scholar dataset (~three hundred records) smoothly.
  - ✧ Prediction time in line with scholar: &lt;1 second.
  - ✧ SHAP rationalization generation: 1–2 seconds on average.
- ➢ Outcome: System is appropriate for real-time instructional use.

6.1.5 User Acceptance Testing (UAT):

- ➢ Objective: Validate usability from an stop-person perspective (teachers, administrators).
- ➢ Approach:
  - ✧ A organization of customers had been requested to check the dashboard.
  - ✧ They furnished remarks on usability, readability of predictions, and visualization readability.
- ➢ Findings:
  - ✧ Users discovered the gadget clean to apply and insightful.
  - ✧ Suggestions blanketed including alert/flagging gadget for at-hazard college students and refining the scholar profile display.
- ➢ Outcome: Accepted through stop customers with fine remarks.

6.1.6 Bug Tracking & Fixes:

- ➢ Issues Encountered:
  - ✧ Categorical encoding mismatch among schooling and testing.
  - ✧ SHAP plots requiring guide modifications for readability.
  - ✧ Feedback module required auto-refresh for higher UX.
- ➢ Resolution: All troubles had been resolved, and solid capability turned into achieved.

6.2 RESULTS & DISCUSSIONS:

6.2.1 Prediction Results:

The SmartEdu+ Dashboard efficiently anticipated the dropout hazard of college students the usage of the skilled XGBoost version.

- The version accomplished a excessive accuracy all through evaluation (above 85% on check data).
- When examined with actual scholar records, the machine turned into capable of classify college students into High Risk and Low Risk categories.
- For example, a scholar with low attendance, low marks, and occasional parental guide turned into flagged as High Risk, while a scholar with excessive attendance, higher marks, and supportive mother and father turned into anticipated as Low Risk.

This confirms that the version is capable of seize styles of scholar conduct and socio-financial situations that have an impact on dropout rates.

6.2.2. SHAP Explanations (Interpretability):

One of the important thing strengths of this machine is using Explainable AI (XAI) via SHAP values.

- SHAP function significance plots discovered that the maximum vast elements influencing dropout are:
  - ✓ Attendance Percentage
  - ✓ Marks in Previous Exam
  - ✓ Parental Support
  - ✓ Family Income
- Waterfall plots confirmed the character contribution of every function to a selected prediction.

This interpretability permits instructors and directors to apprehend why a scholar is anticipated as excessive hazard, in preference to counting on a “black box” output.

6.2.3 Visualization Results:

The Visualization Tab furnished insights into typical dataset styles:

- Dropout Distribution Pie Chart – Showed the share of college students at hazard vs. non- dropouts.

- Average Marks vs Dropout Bar Chart – Confirmed that scholars with decrease instructional overall performance are extra liable to dropout.
- Study Hours vs Dropout Histogram – Highlighted that scholars reading fewer hours every day are at better hazard.
- Family Income vs Dropout (Stacked Bar) – Displayed that scholars from low- earnings backgrounds are disproportionately at better hazard.

These outcomes matched current literature in schooling research, reinforcing the validity of the machine.

6.2.4 Feedback Collection:

The Feedback Tab allowed college students to voluntarily post their motives for capability dropout.

- Collected motives protected loss of parental guide, monetary issues, lengthy tour distance, and occasional motivation.
- This comments presents qualitative insights, complementing the quantitative predictions of the version.
- Teachers can use this records to layout customized interventions for at-hazard college student

6.2.5 Discussion:

- The outcomes imply that instructional elements (attendance, marks) and socio-financial elements (earnings, parental guide) mutually decide dropout hazard.
- The dashboard now no longer most effective predicts dangers however additionally makes the system obvious and interactive, giving decision-makers actionable insights.
- The inclusion of visualizations and reasons will increase believe withinside the machine in comparison to conventional system gaining knowledge of models.
- However, the version overall performance relies upon at the first-class and length of dataset. Larger, extra various datasets may want to enhance robustness and generalizability.

6.2.6 Results:

METRICS	TRAININGDATA(%)	TESTING DATA(%)
Accuracy	89.5	86.7
Precision	87.3	84.1
Recall(Sensitivity)	85.9	82.5
F1-Score	86.6	83.3
AUC (ROC Curve)	91.2	88.5

CONCLUSION & FUTURE WORK

7.CONCLUSION & FUTURE WORKS

7.1 CONCLUSION:

The proposed **SmartEdu+: Student Dropout Prediction Dashboard** has demonstrated its potential as a powerful decision-support system for educational institutions. By integrating machine learning models with explainable AI (SHAP) and interactive data visualizations, the system provides accurate predictions of student dropout risk along with clear explanations of the underlying causes. The dashboard enables teachers, administrators, and counselors to predict dropout likelihood for individual students based on academic, socio-economic, and behavioral factors; understand key influencing features through SHAP explanations to ensure transparency in decision-making; visualize overall dataset patterns such as marks, study hours, and family income through intuitive charts; and collect student feedback on dropout reasons, creating a feedback loop to improve institutional strategies. Overall, the system not only enhances early-warning capabilities but also supports data-driven interventions, helping educational stakeholders identify at-risk students and offer timely support. Thus, **SmartEdu+** represents a significant step toward bridging the gap between raw academic data and actionable educational insights.

7.2 Future Work:

While the modern gadget offers a stable foundation, there are numerous possibilities for enhancement and destiny research:

7.2.1 Real-Time Data Integration

- ➢ Extend the gadget to attach without delay with college databases and Learning Management Systems (LMS) for computerized updates of attendance, grades, and scholar activities.

7.2.2 Advanced Models

- ➢ Explore using deep mastering fashions (e.g., LSTMs for sequential scholar overall performance records) to in addition enhance accuracy.
- ➢ Implement ensemble fashions combining Random Forest, XGBoost, and Neural Networks for sturdy predictions.

7.2.3 Broader Feature Set

- ➢ Incorporate psychological, emotional, and social signs along with motivation levels, peer interaction, and strain elements.

- Add geographical and infrastructural elements like distance to college, delivery availability, and virtual access.

7.2.4 Alert and Recommendation System

- Develop an automatic alert gadget to inform instructors and mother and father while a scholar is flagged as high-threat.
- Provide customized suggestions along with remedial classes, counseling, or economic resource programs.

7.2.5 Scalability & Deployment

- Deploy the dashboard on a cloud platform (AWS, Azure, or GCP) for scalability.
- Allow multi-group usage, allowing cross-college comparisons and local dropout analysis.

7.2.6 Student-Centric Features

- Enable college students to self-check their threat rating and acquire pointers to enhance overall performance.
- Introduce gamified motivational modules to inspire engagement.

APPENDICES

A.1 SDG GOALS

The SmartEdu+: Student Dropout Prediction Dashboard isn't always most effective a technological undertaking however additionally helps the imaginative and prescient of sustainable and inclusive training. It aligns with a couple of United Nations Sustainable Development Goals (SDGs):

1. SDG 4 – Quality Education

- The gadget without delay addresses the difficulty of pupil dropouts, one of the important boundaries to accomplishing inclusive training.
- By predicting which college students are at risk, faculties and universities can offer customized help consisting of mentoring, monetary aid, or instructional counseling.
- This enables lessen instructional inequality and guarantees that greater college students correctly whole their studies.

2. SDG 10 – Reduced Inequalities

- Dropout prices are regularly better amongst college students from low-profits families, rural areas, or deprived communities.
- This gadget highlights socioeconomic and parental help factors, allowing government to layout equity-pushed policies.
- In turn, this guarantees identical possibilities for college students irrespective of background.

A.1 SOURCE CODE

CODING:

DATASET:

```
import pandas as pd

df1 = pd.read_csv('dataset.csv')
df2 = pd.read_csv('student dropout.csv')
df3 = pd.read_csv('student_dropout_300.csv')
```

DASHBOARD:

```
import streamlit
as st import
pandas as pd
import joblib
import shap
import matplotlib.pyplot as plt
model = joblib.load("dropout_model.pkl") label_encoders

= joblib.load("label_encoders.pkl")
categorical_column s

 = [ "Gender",
 "Family_Income”,
    "Father_Education”,"Mother_Educati  on",
    "Disciplinary_Issues",
    "Internet_Access",
    "Parental_Support"
]
st.image("logo.png", width=120)
st.title("     SmartEdu+ : Student Dropout Prediction Dashboard")
st.sidebar.title("  Navigation")
st.sidebar.info("Use the tabs to explore prediction, explanations, visualizations, and feedback.") tab1,
tab2, tab3, tab4 =

    st.tabs([ "
        Prediction","
        SHAP
    Explanations", "
    Visualizations",
    "    Feedback"

])
#   -----------------TAB 1:                    ------------------
Prediction with tab1:
    st.header("  Student Dropout Prediction")
    age = st.number_input("Age", min_value=10, max_value=25, value=17)
    gender = st.selectbox("Gender", ["Male", "Female"])
    family_income = st.selectbox("Family Income", ["Low", "Medium", "High"])
    father_edu  =  st.selectbox("Father's Education",  ["Unknown",  "Primary",
                        "Secondary", "Graduate"])
    mother_edu = st.selectbox("Mother's Education", ["Unknown", "Primary",
"Secondary", "Graduate"])
    attendance = st.slider("Attendance Percentage", 0, 100, 75)

    marks_prev_exam = st.slider("Marks in Previous Exam", 0, 100, 70)

    study_hours  =  st.slider("Study  Hours  Per  Day",  0,  12,  2)
    disciplinary = st.selectbox("Disciplinary Issues", ["Yes",
    "No"])  internet_access  =  st.selectbox("Internet  Access",
    ["Yes", "No"])
    parental_support = st.selectbox("Parental Support", ["Low", "Medium", "High"])
    gender_map = {"Male": 0, "Female": 1}
    disciplinary_map = {"No": 0, "Yes": 1}
    internet_map = {"No": 0, "Yes": 1}
    family_income_map = {"Low": 0, "Medium": 1, "High": 2}

    education_map = {"Unknown": 0, "Primary": 1, "Secondary": 2, "Graduate": 3}
    parental_support_map = {"Low": 0, "Medium": 1, "High":
    2} new_data =
       pd.DataFrame([{ "Age":
       age,
       "Gender": gender_map[gender], "Family_Income":
       family_income_map[family_income],
       "Father_Education": education_map[father_edu],
       "Mother_Education": education_map[mother_edu],
       "Attendance_Percentage": attendance,
       "Marks_Previous_Exam": marks_prev_exam,
       "Study_Hours_Per_Day": study_hours, "Disciplinary_Issues":
       disciplinary_map[disciplinary], "Internet_Access":
       internet_map[internet_access], "Parental_Support":
       parental_support_map[parental_support]
    }])
    if st.button(" Predict Dropout Risk"):
       prediction =
       model.predict(new_data)[0]
       probability = model.predict_proba(new_data)[0][1] * 100
       if prediction == 1:
           st.error(f"  High Risk of Dropout! (Probability:
       {probability:.2f}%)") else:
       st.success(f"✅ Low Risk of Dropout (Probability: {probability:.2f}%)") ## ---
------------- TAB 2: SHAP Explanations ----------------
with tab2:
st.header("  SHAP Explanations")
st.write("Explainable AI (SHAP) helps us understand *why* the model predicted a student as High or
Low risk.")
    data = pd.read_csv("E:/3rd yr MINI PROJ Dataset/datasets/student_dropout_300.csv")
    if "Student_ID" in data.columns:
       student_ids = data["Student_ID"]
       data = data.drop("Student_ID", axis=1) else:
       student_ids = pd.Series(range(len(data)))
    X = data.drop("Dropout", axis=1)
    y = data["Dropout"]
    X["Gender"] = X["Gender"].map(gender_map) X["Family_Income"]
    = X["Family_Income"].map(family_income_map)
```

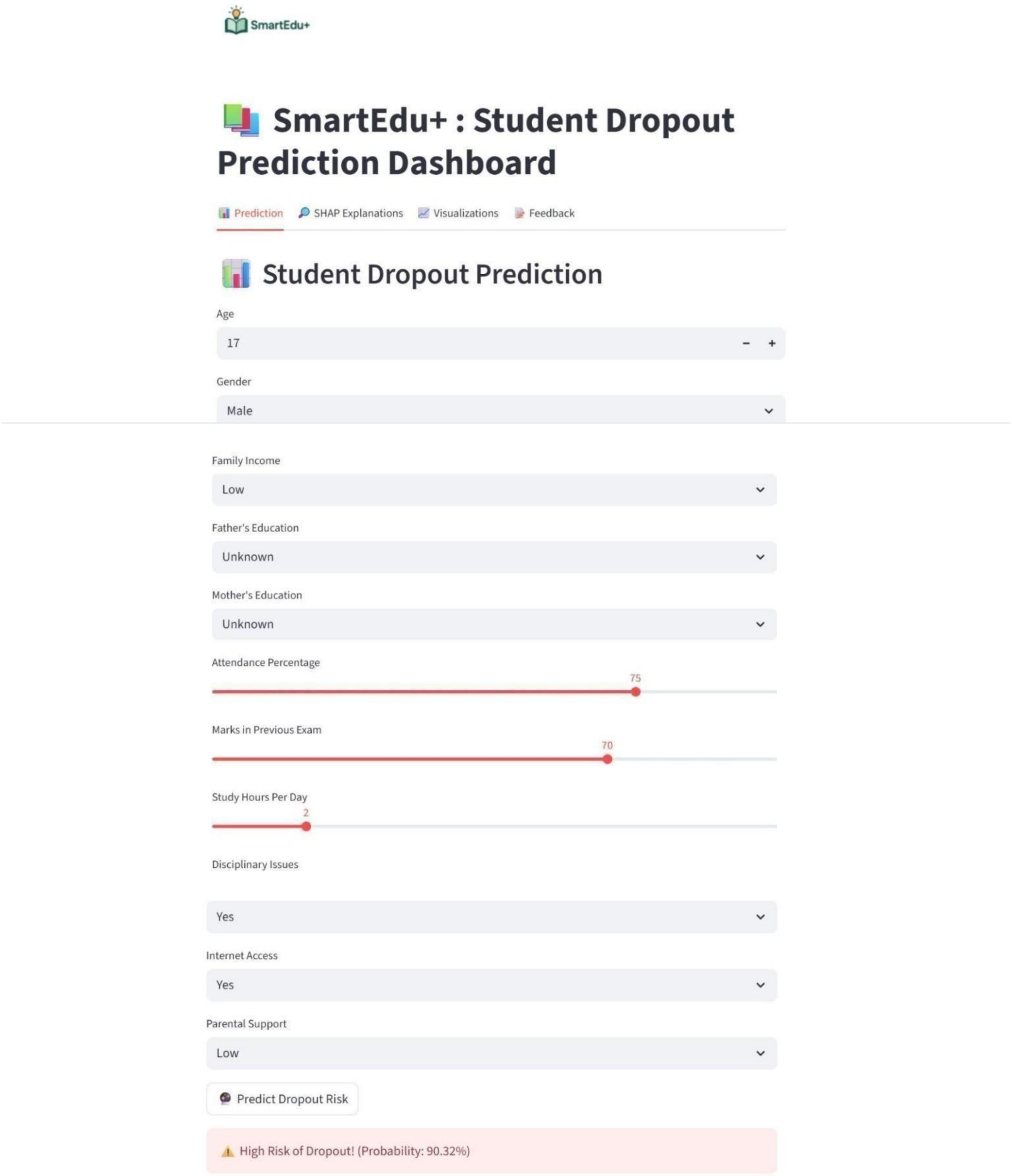
```
X["Father_Education"] =
X["Father_Education"].map(education_map)  X["Mother_Education"]
= X["Mother_Education"].map(education_map)
X["Disciplinary_Issues"] =
X["Disciplinary_Issues"].map(disciplinary_map)
X["Internet_Access"]
= X["Internet_Access"].map(internet_map) X["Parental_Support"] =
X["Parental_Support"].map(parental_support_map)
explainer = shap.TreeExplainer(model)
shap_values = explainer.shap_values(X)
selected_id = st.selectbox(" Select Student by ID",
student_ids) student_idx = student_ids[student_ids
== selected_id].index[0] selected_student =
X.iloc[[student_idx]]
profile_data = data.drop("Dropout", axis=1).iloc[student_idx]
pred = model.predict(selected_student)[0]
prob = model.predict_proba(selected_student)[0][1] * 100
st.markdown("###      Student Profile &
Dropout Risk") col1, col2 = st.columns([1.5, 1])
with col1:
   st.markdown("####
   Student Details") c1, c2 =
   st.columns(2)
   with c1:
   st.metric("  Age", profile_data["Age"]) st.metric("⚥ Gender", profile_data["Gender"])
      st.metric("  Family Income",
      profile_data["Family_Income"]) st.metric(" Father
      Edu.", profile_data["Father_Education"])
   with c2:
      st.metric("  Mother Edu.", profile_data["Mother_Education"])
      st.progress(int(profile_data["Attendance_Percentage"]))
      st.caption(f"  Attendance:
      {profile_data['Attendance_Percentage']}%")
      st.progress(int(profile_data["Marks_Previous_Exam"]))
      st.caption(f"  Marks:
      {profile_data['Marks_Previous_Exam']}%")
      st.metric("⏱ Study Hours",
      profile_data["Study_Hours_Per_Day"])
with col2:
import plotly.graph_objects as go gauge = go.Figure(go.Indicator(
      mode="gauge+numb
      er", value=prob,
      title={"text":
      "Dropout Risk %"},
      gauge={
          "axis": {"range": [0, 100]},
          "bar": {"color": "red" if prob > 50
          else "green"}, "steps": [
              {"range": [0, 50], "color": "lightgreen"},
              {"range": [50, 100], "color": "pink"}
          ]}
   ))
   gauge.update_layout(height=230, margin=dict(l=10, r=10, t=30, b=10))
   st.plotly_chart(gauge, use_container_width=True)
   if pred == 1:
      st.error(f" Student {selected_id}: High Risk of Dropout ({prob:.2f}%)")
   else:
      st.success(f"✅ Student {selected_id}: Low Risk of Dropout ({prob:.2f}%)")
st.markdown("---")
col3, col4 = st.columns(2) with col3:
st.subheader("  Feature Importance") fig1, ax1 = plt.subplots(figsize=(6, 4))
   shap.summary_plot(shap_values, X, plot_type="bar", show=False)
   st.pyplot(fig1)
with col4:
      st.subheader("  Individual SHAP Explanation")
      shap_values_for_student = shap_values[student_idx]
      fig2, ax2 = plt.subplots(figsize=(6, 4))
      shap.plots._waterfall.waterfall_legacy(
          explainer.expected_value, shap_values_for_student, X.iloc[student_idx, :], show=False

      )
      st.pyplot(fig2)
   st.markdown("###      Top 3 Factors
   Driving Prediction") shap_df =
   pd.DataFrame({
      "Feature": X.columns,
        "SHAP Value": shap_values[student_idx]
        }).assign(abs_shap=lambda      df:       df["SHAP
                                                Value"].abs()).7s2ort_values(by="abs_shap",

ascending=False)
   cols =
   st.columns(3) for i,
   (_, row) in
      enumerate(shap_df.head(3).iterrows()): with
      cols[i]:
          if row["SHAP Value"] > 0:
              st.error(f"
                          {row['Feature']}
      “)
            import
           plotly.express
           as px
           import plotly.graph_objects as go
```

```
with tab3:
    st.header("  Dataset Visualizations")
    st.write("Explore overall patterns and trends in the student dataset.")
    data_viz = pd.read_csv("E:/3rd yr MINI PROJ Dataset/datasets/student_dropout_300.csv")
    st.subheader("  Dropout Distribution")
    dropout_counts = data_viz["Dropout"].value_counts().reset_index() dropout_counts.columns =
    ["Dropout", "Count"]
    fig_pie = px.pie(
    dropout_counts, names="Dropout", values="Count", color="Dropout", color_discrete_map={0:
    "green", 1: "red"}, title="Dropout vs Non-Dropout"
    )
    st.plotly_chart(fig_pie, use_container_width=True)
    st.subheader("  Average Marks vs Dropout Status")

    marks_avg = data_viz.groupby("Dropout")["Marks_Previous_Exam"].mean().reset_index()
    fig_marks_avg = px.bar(
        marks_avg, x="Dropout", y="Marks_Previous_Exam", color="Dropout",
        color_discrete_map={0: "green", 1:

        "red"}, text_auto=".1f",

        title="Average Marks (Stayed vs Dropped)"
    )
    fig_marks_avg.update_xaxes(tickvals=[0, 1], ticktext=["Stayed", "Dropped"])
    fig_marks_avg.update_layout(yaxis_title="Average Marks (%)")
    st.plotly_chart(fig_marks_avg, use_container_width=True)
    st.subheader("⏱ Study Hours per Day vs Dropout")
    fig_study = px.histogram(
        data_viz, x="Study_Hours_Per_Day", color="Dropout",
        barmode="overlay", nbins=10, color_discrete_map={0:
        "blue", 1: "red"},
        title="Study Hours Distribution (Dropout vs Non-Dropout)")
    st.plotly_chart(fig_study, use_container_width=True)
    st.subheader("  Family Income vs
    Dropout") income_dropout =
        ( data_viz.groupby(["Family_Income", "Dropout"]).size()
        .reset_index(name="Count"))
        income_total = income_dropout.groupby("Family_Income")["Count"].transform("sum")
        income_dropout["Percentage"] = (income_dropout["Count"] / income_total) * 100
    fig_income =
        px.bar( income_dropout,
        x="Family_Income", y="Percentage",
        color="Dropout", barmode="stack",
        title="Dropout % by Family Income
        (Stacked)", color_discrete_map={0:
        "green", 1: "red"}, text_auto=".1f"
    )
    fig_income.update_layout(yaxis_title="Percentage of Students")
    st.plotly_chart(fig_income, use_container_width=True)
#   -----------------TAB 4:                  ------------------
Feedback import
os import pandas
as  pd import
streamlit as st
dropout_file = "dropout_reasons.csv"
with tab4:
    st.header("  Dropout Feedback")
    st.write("Tell us the reason why you feel you might drop out. Your feedback will help us
support you better.")
    if "submitted" not in st.session_state:
        st.session_state.submitted = False
    with st.form("dropout_form",
        clear_on_submit=True): name
        = st.text_input("Your Name")email = st.text_input("Your Email")
        reason = st.text_area("What is the main reason you're considering
        dropping out?") submit = st.form_submit_button("Submit")
```

```
    if submit:

        if not name or not email or not reason:
            st.error("Please fill in all the fields.")
        else:
            new_entry =
                pd.DataFrame([{"Na
                me":
                name,
                "Email": email,
                 "Reason": reason

            }])
            if os.path.exists(dropout_file): existing =
                pd.read_csv(dropout_file)
                updated = pd.concat([existing, new_entry], ignore_index=True)
            else:
                updated = new_entry
            updated.to_csv(dropout_file, index=False)
            st.session_state.submitted = True
if st.session_state.submitted:

    st.success("✅ Thank you! Your reason has been recorded.")
    st.session_state.submitted = False
    st.markdown("---")st.subheader("  Previous Dropout Reasons")


if os.path.exists(dropout_file):
    reasons_data =
    pd.read_csv(dropout_file) if
    reasons_data.empty:
        st.info("No dropout reasons submitted yet.") Else:
          else:
     for i, row in reasons_data.iterrows():
```

```
     st.markdown(f"**{row['Name']}
     ({row['Email']})**") st.write(f"
Reason:
     {row['Reason']}") st.markdown("---")
          st.info("No dropout reasons
  submitted      yet.")
```

A3.SCREENSHOTS



SmartEdu+ : Student Dropout Prediction Dashboard

Student Dropout Prediction

Age: 17

Gender: Male

Family Income: Low

Father's Education: Unknown

Mother's Education: Unknown

Attendance Percentage: 75

Marks in Previous Exam: 70

Study Hours Per Day: 2

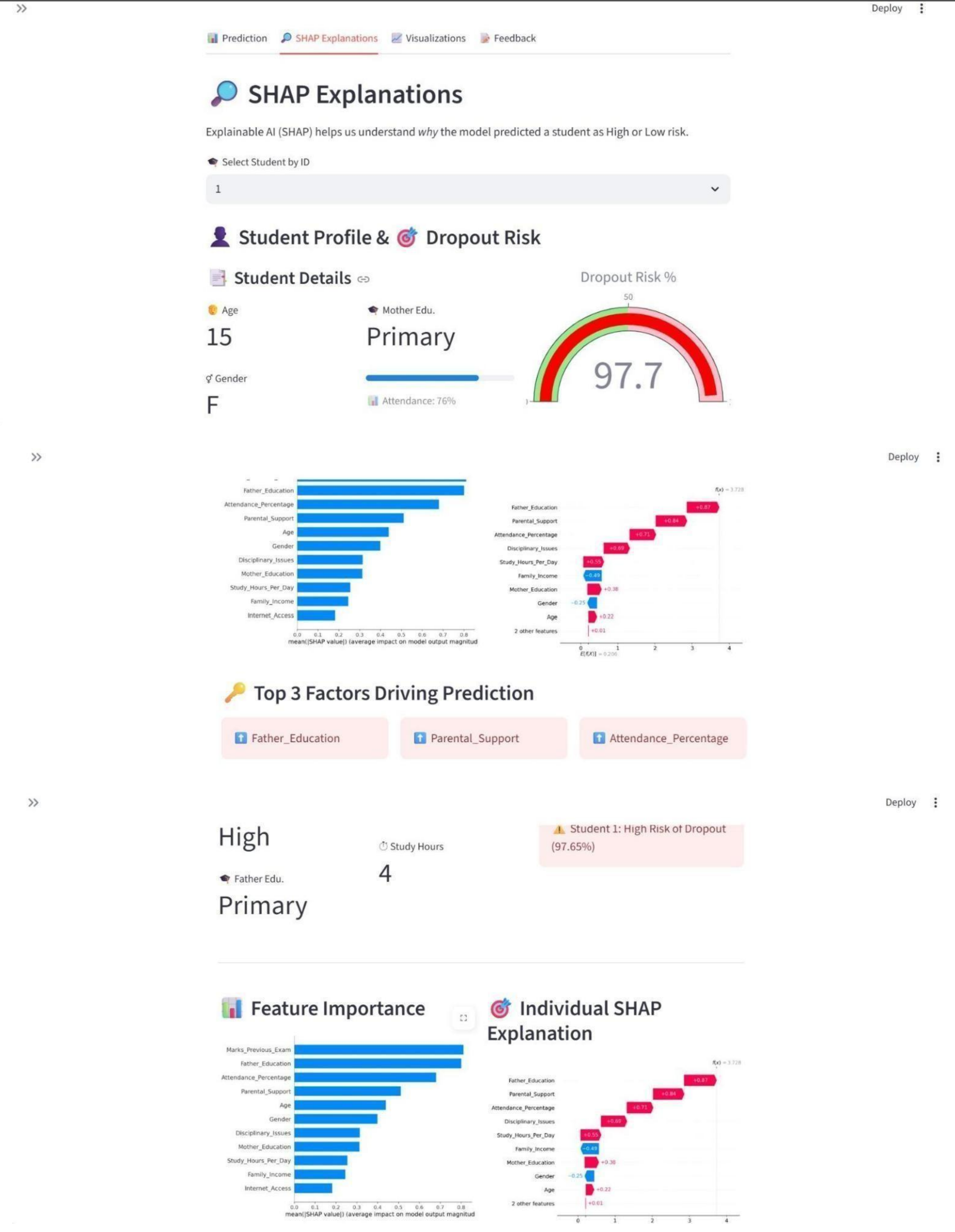
Disciplinary Issues: Yes

Internet Access: Yes

Parental Support: Low

⚠ High Risk of Dropout! (Probability: 90.32%)

Figure A.3.1 Prediction



🔎 SHAP Explanations

Explainable AI (SHAP) helps us understand *why* the model predicted a student as High or Low risk.

Select Student by ID

1

Student Profile & Dropout Risk

Student Details

Age

15

Mother Edu.

Primary

Gender

F

Attendance: 76%

Dropout Risk %

Top 3 Factors Driving Prediction

Father_Education

Parental_Support

Attendance_Percentage

High

Study Hours

4

Father Edu.

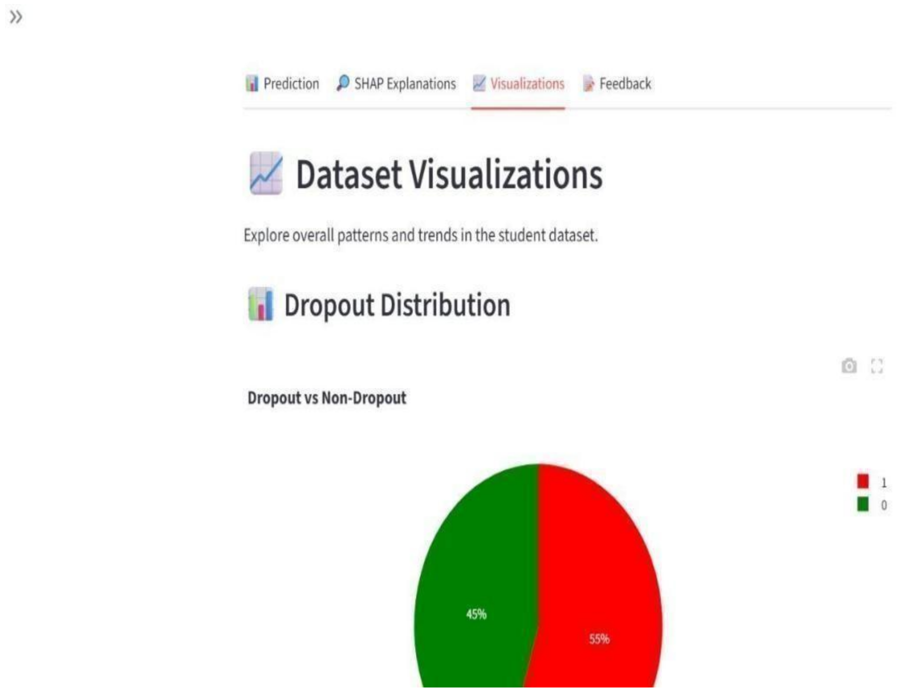
Primary

⚠ Student 1: High Risk of Dropout (97.65%)

Feature Importance

Individual SHAP Explanation

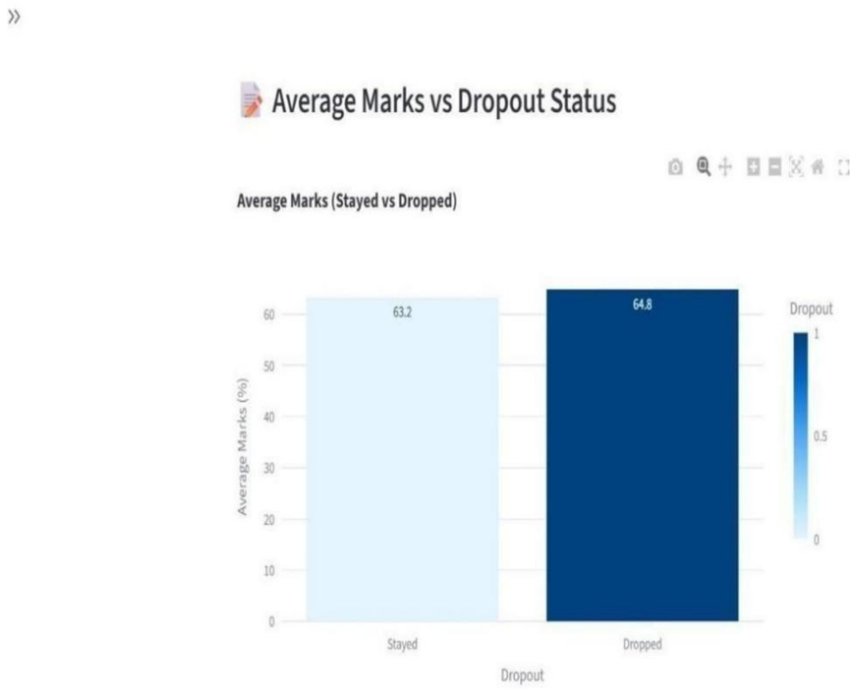
Figure A.3.2 SHAP Explanations



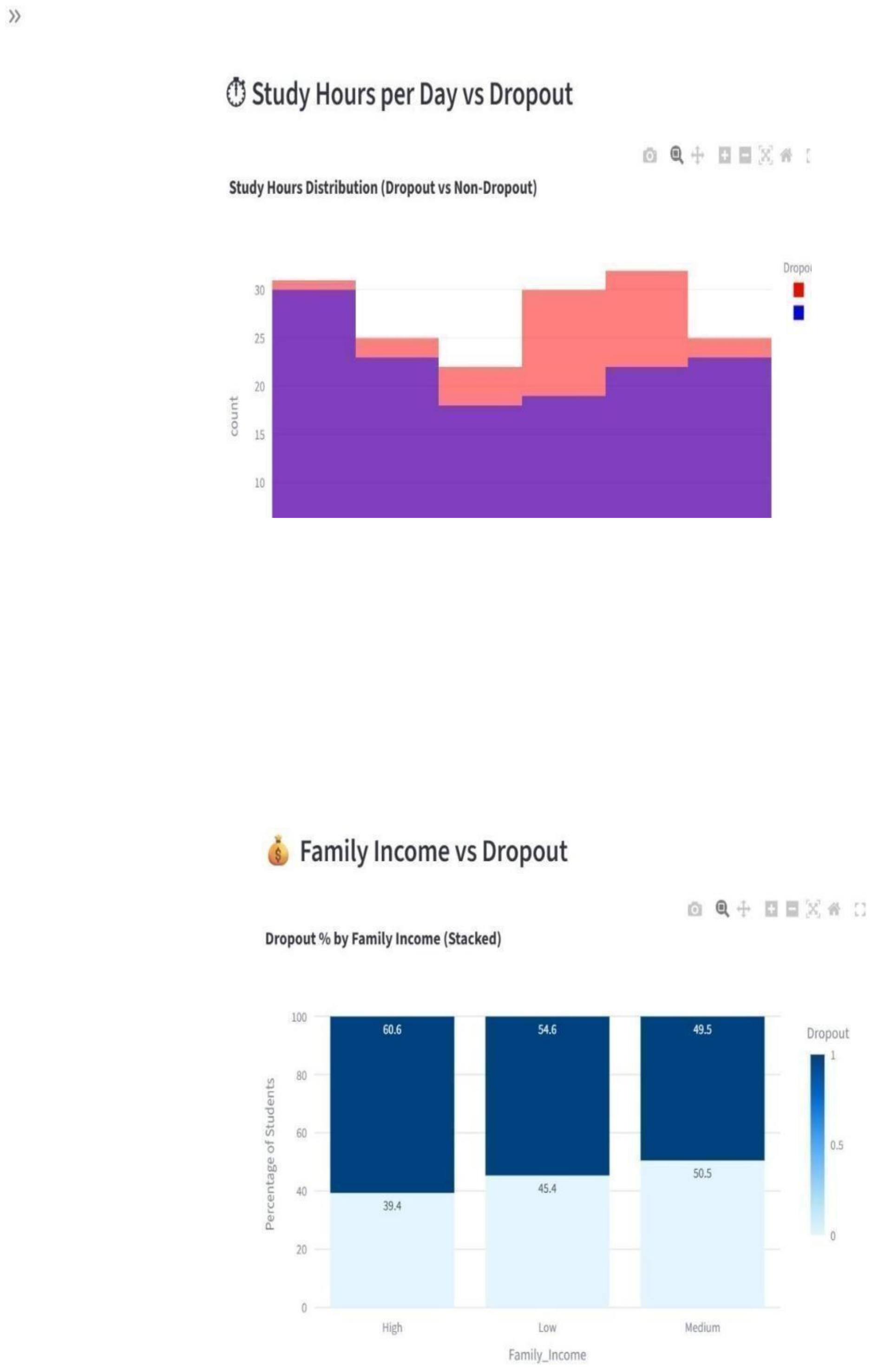
📈 Dataset Visualizations

Explore overall patterns and trends in the student dataset.

📊 Dropout Distribution



📝 Average Marks vs Dropout Status



»

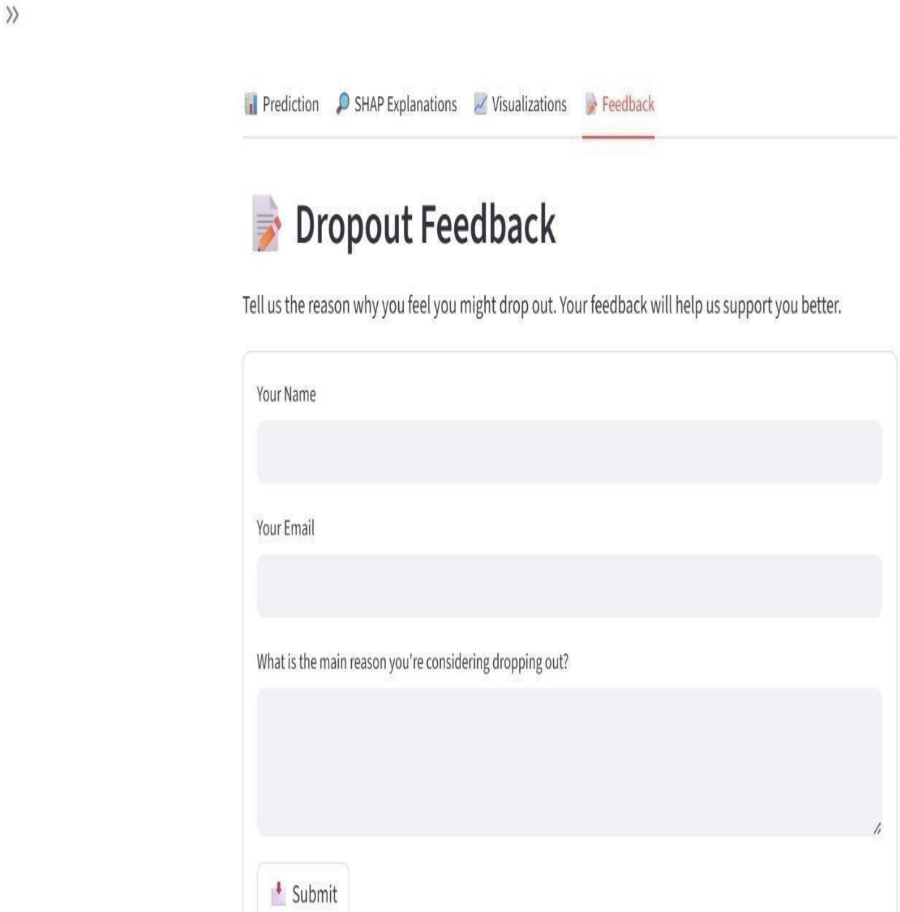
⏱ Study Hours per Day vs Dropout

Study Hours Distribution (Dropout vs Non-Dropout)

💰 Family Income vs Dropout

Dropout % by Family Income (Stacked)

Fig ure A.3.3 Visualizations



»

📊 Prediction 🔎 SHAP Explanations 📈 Visualizations 📝 Feedback

📝 Dropout Feedback

Tell us the reason why you feel you might drop out. Your feedback will help us support you better.

Your Name

Your Email

What is the main reason you're considering dropping out?

📥 Submit

Figure A.3.4 Feedback

A.4 PLAGIARISM REPORT

REFERENCES

1. J. Cheng et al., “Predicting Student Dropout Risk With A Dual-Modal Abrupt Behavioral Changes Approach,” arXiv, 2025.
2. I. Elbouknify et al., “AI-based Identification and Support of At-Risk Students: Moroccan Case Study,” arXiv, 2025.
3. L. Aulck et al., “Predicting Student Dropout in Higher Education,” arXiv, 2016.
4. M. Orooji and J. Chen, “Predicting Louisiana Public High School Dropout through Imbalanced Learning Techniques,” arXiv, 2019.
5. H. S. Park and S. J. Yoo, “Early Dropout Prediction in Online Learning,” ResearchGate, 2021.
6. A. J. Fernandez Garcia et al., “Real-Life ML Experience for Predicting University Dropout,” ResearchGate, 2021.
7. M. A. Dewi, F. I. Kumiadi, D. F. Murad and S. G. Rabiha, “Machine Learning Algorithms for Early Predicting Dropout Student Online Learning,” in *2023 9th International - Conference on Computing, Engineering, and Design (ICCED)*, 2023.
8. F.Dalipi, A. S. Imran and Z. Kastrati, “MOOC dropout prediction using machine learning techniques: Review and research challenges,” in *2018 IEEE Global Engineering Education Conference (EDUCON)*, 2018.
9. D. A. Shafiq, M. Marjani, R. A. A. Habeeb and D. Asirvatham, “Student Retention Using Educational Data Mining and Predictive Analytics: A Systematic Literature Review,” *IEEE Access*, vol. 10, pp. 72480–72503, 2022
10. E.S. Abouzeid et al., “Student dropout prediction in massive open online courses,” *Soft Computing* (Springer), 2018 — proposes an end-to-end CNN-based dropout predictor for MOOC timestamped data