2025-IS6611: Applied Research in Business Analytics

IT Artefact V3

Student Dropout Prediction

GROUP - 7

Sajin Siyad - 124104640

Mrithul Madhu Kumar -124104690

Angela George Kurian - 124111002

Diya Binilal - 124107000

Sneha Musale - 124115975

Anamika Chemmancheri - 124115546

Table Of Contents

1. Introduction

Student dropout is a persistent global issue with significant academic and financial consequences. In the U.S., nearly 40% of undergraduates do not complete their degrees (figure: 1.1), with 24% leaving within the first year (figure: 1.2) a critical time for integration and support (NCES, 2022; Hanson, 2024). This early attrition disproportionately affects first-generation, low-income, and minority students who face financial stress, academic gaps, and limited support.

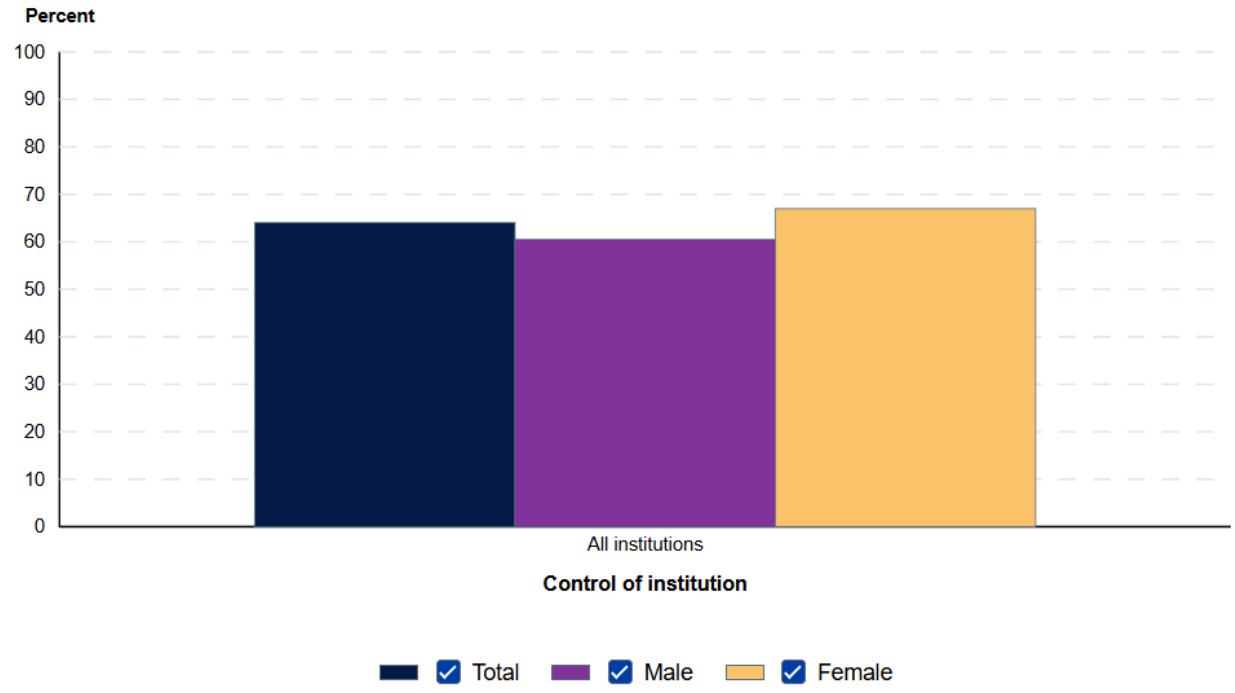


Figure 1.1 Graduation rate of full time degree students (~ 40 % dropout) (NCES, 2022)

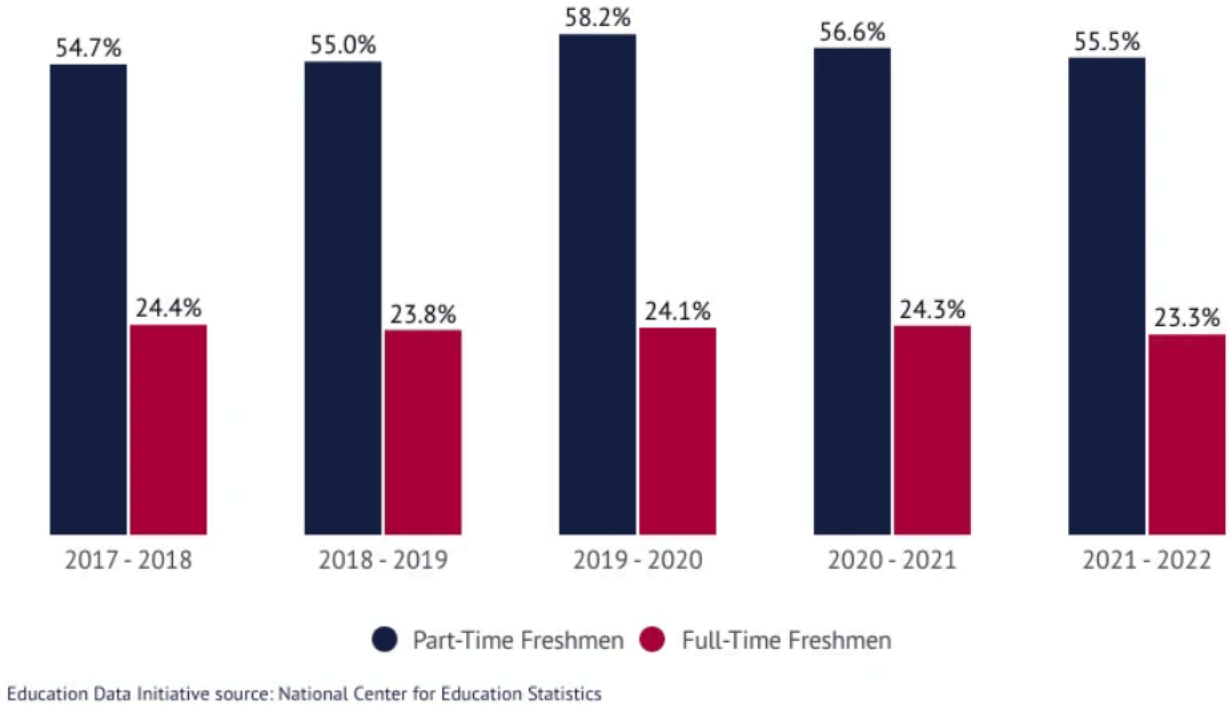


Figure 1.2 : 12 month dropout rate (Hanson, 2024)

Economically, dropouts earn 35% less and are 20% more likely to be unemployed (Hanson, 2024). Institutions also lose over $16 billion annually due to attrition (Raisman, 2013), emphasizing the urgent need for proactive retention strategies.

2. Approach

Figure 2.1: Illustrating approach to our solution

This visual map, Figure 2.1 summarizes our journey from problem understanding to developing an effective solution.

Let’s explore each step of the journey in detail, unpacking how every phase contributed to building our student dropout prediction system.

3. Stakeholder Map, Personas, and User Journey

Behind every successful intervention is a network of people who make it possible. The stakeholder map below captures the ecosystem of advisors, students, data teams, and decision-makers who collectively shape, support, and benefit from student success.

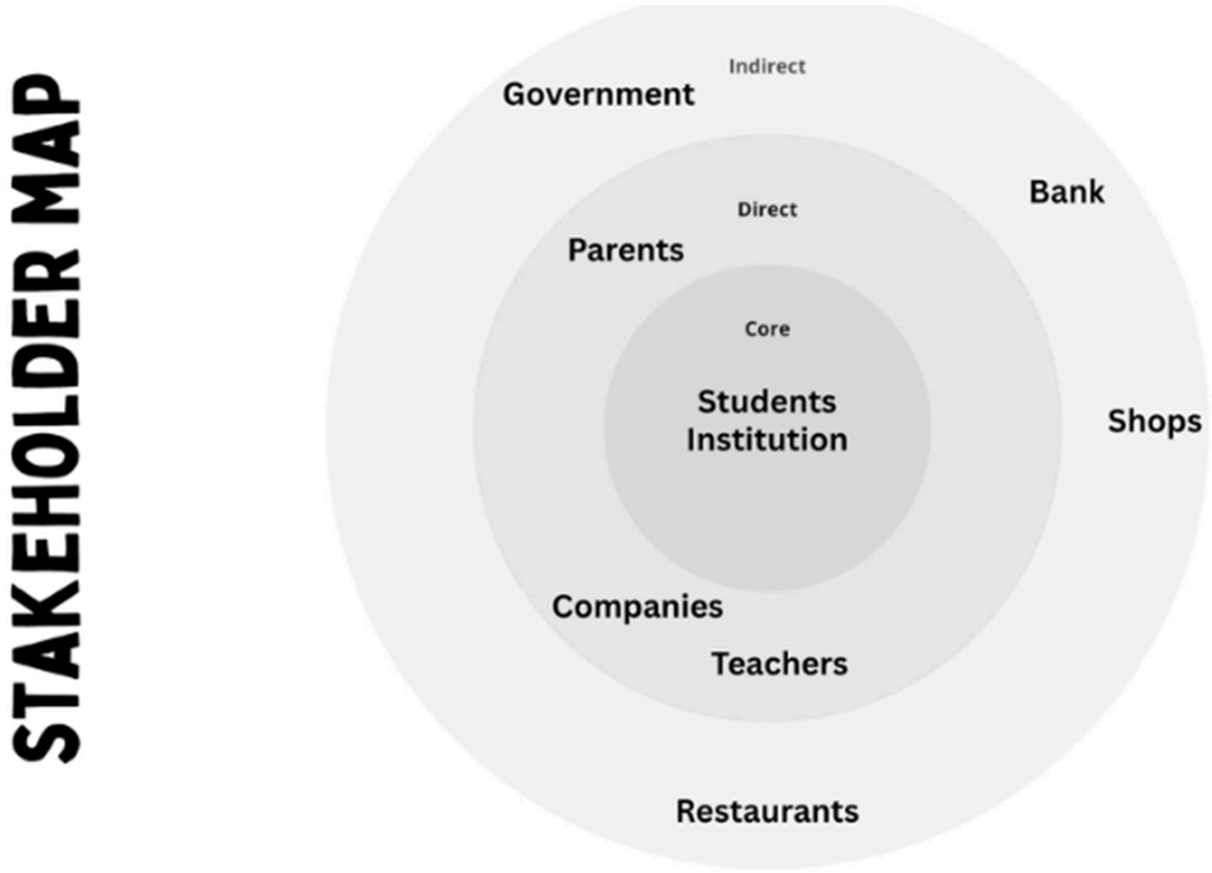


Figure 3.1: Stakeholder Map

As shown in Figure 3.1, We identified two stakeholder groups:

- **Primary stakeholders** include Academic advisors, Students and faculty members who directly shape a student's daily learning environment and academic journey.

- **Secondary stakeholders** such as university administrators, policymakers, parents, financial aid teams, and even potential employers are indirectly affected by the student's academic performance and well-being.

To humanize our design process, we developed two representative personas that reflect real-world challenges:

Figure 3.2: Persona of Emily

- **Emily Johnson** (Figure . 3.2): A high-achieving student quietly struggling with academic stress, often undetected by existing systems.

Figure 3.3: Persona of Peter

- **Peter Jane** (Figure 3.3): A course coordinator burdened by fragmented data sources, making it difficult to monitor student well-being holistically.

These personas helped us understand problems faced by stakeholders. To gain deeper insight into their experiences, we used empathy mapping as in Figure 3.4, identifying what each persona says, does, thinks, and feels. This uncovered hidden pain points, internal motivations, and behavioral patterns that raw data alone could not reveal.

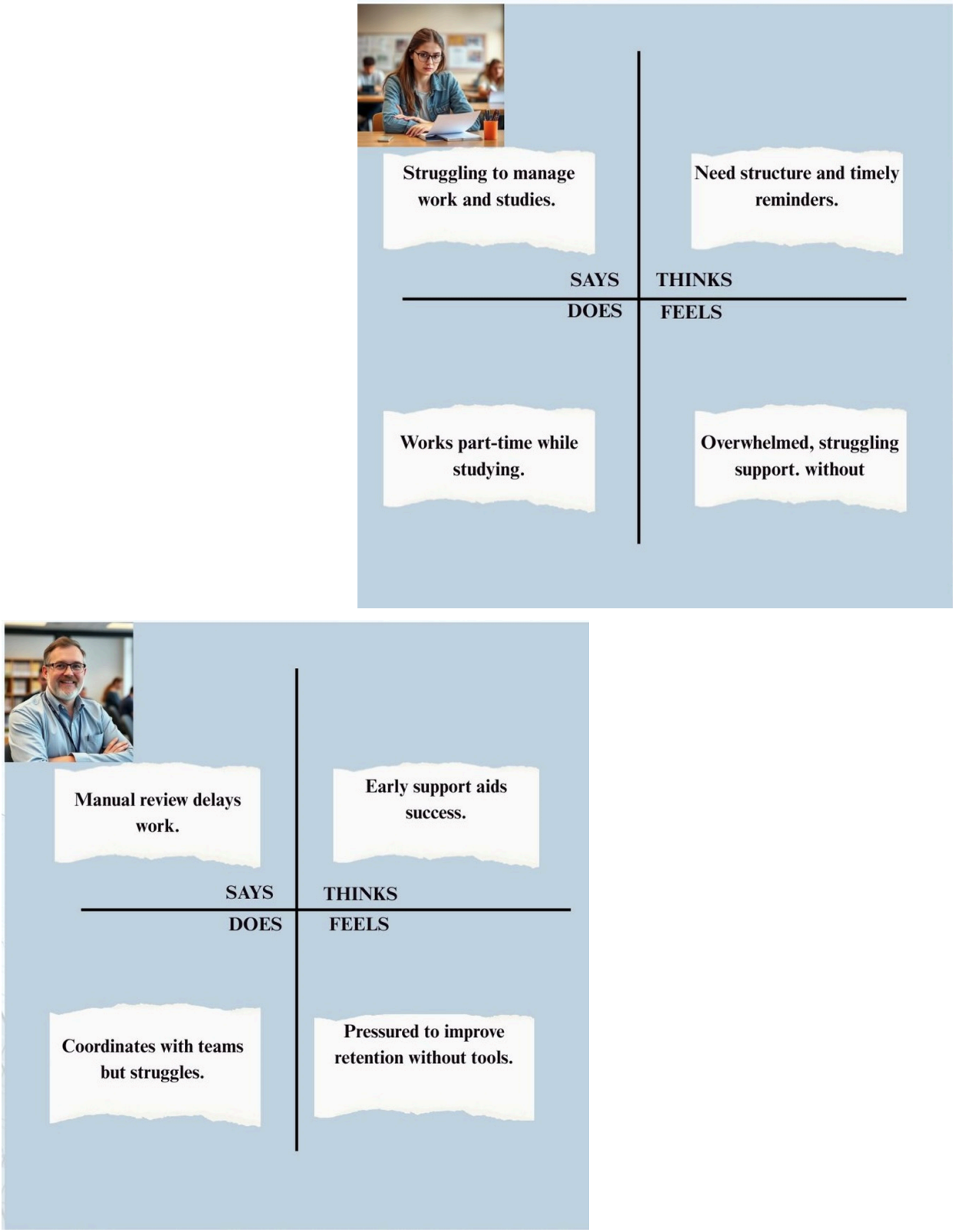


Figure 3.4: Empathy map of personas

The User Journey Map shown in Figure 3.5 highlights how students move from excitement at enrollment to feeling unseen after dropout.

User Journey Map

	ENROLLMENT	ACADEMIC START	MID-TERM PHASE	SUPPORT PHASE	OUTCOME PHASE
STUDENT JOURNEY	Joins Course Orientation	Starts Classes Challenges arise	Disengages Struggles	Hesitates help Unaware of support	Withdraws Feels unseen.
UNIVERSITY RESPONSE	Registers student, Onboards	Tracks manually	Late detection	Offers generic Support	Post-dropout analysis
PAIN POINTS	👨‍🎓:Excited😃& Unsure 🤔 🏛 : Low insight 🔍	👨‍🎓:Struggling😟& Silent😶 🏛:No visibility 🚫👀	👨‍🎓: Isolated 🚶‍♂️ & Overwhelmed 🏛 : Reacts late 🕒	👨‍🎓: Unaware 🧭& Hesitant 👩‍🦰 🏛 Too broad 🎯📉	👨‍🎓:Let down 😔 🏛 : Loses 💸📉
IMPROVEMENT	Early assessments	Engagement tracking	Real-time alerts Predictive analytics	Targeted help	Pre-dropout actions

Figure 3.5: User Journey Mapping of student and university

While institutions collect data across various touchpoints, these remain underutilized. The current support systems are often **reactive, delayed, and generalized**, failing to detect early signs of disengagement such as reduced participation, poor sleep, or emerging mental health concerns.

Our analysis made one thing clear: there is an urgent need for **personalized, early-stage insights** delivered in a timely and actionable format, empowering advisors and stakeholders to intervene **before** students fall through the cracks.

4. As-Is Process: Gaps and Limitations

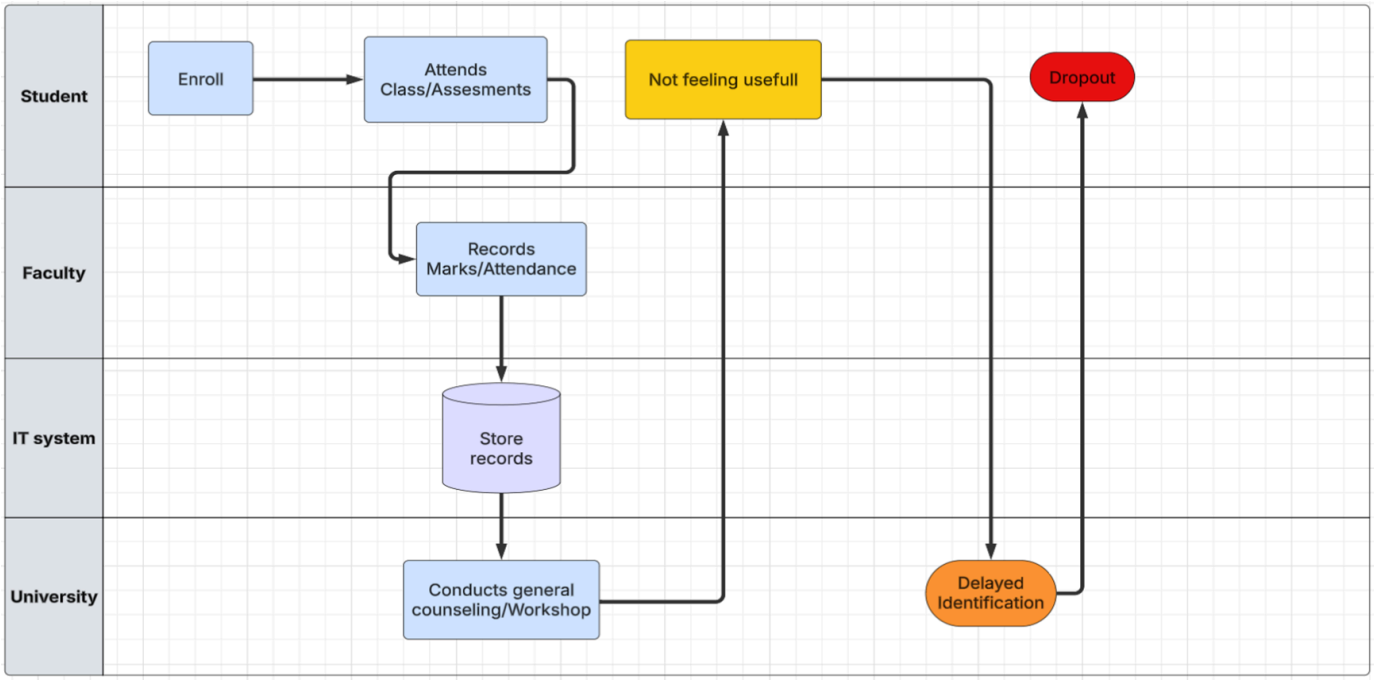


Figure 4.1: Swimlane diagram of current process of student intervention

Our swimlane diagram (Figure 4.1) of the current system illustrates a reactive, disconnected support model:

- Students attend lectures and submit assignments.
- Faculty record attendance and grades, which are stored in siloed institutional databases.
- Advisors step in only after a visible academic decline, often during midterms or finals.
- Interventions (like counseling or remedial workshops) are broad, non-targeted, and too late to be effective.

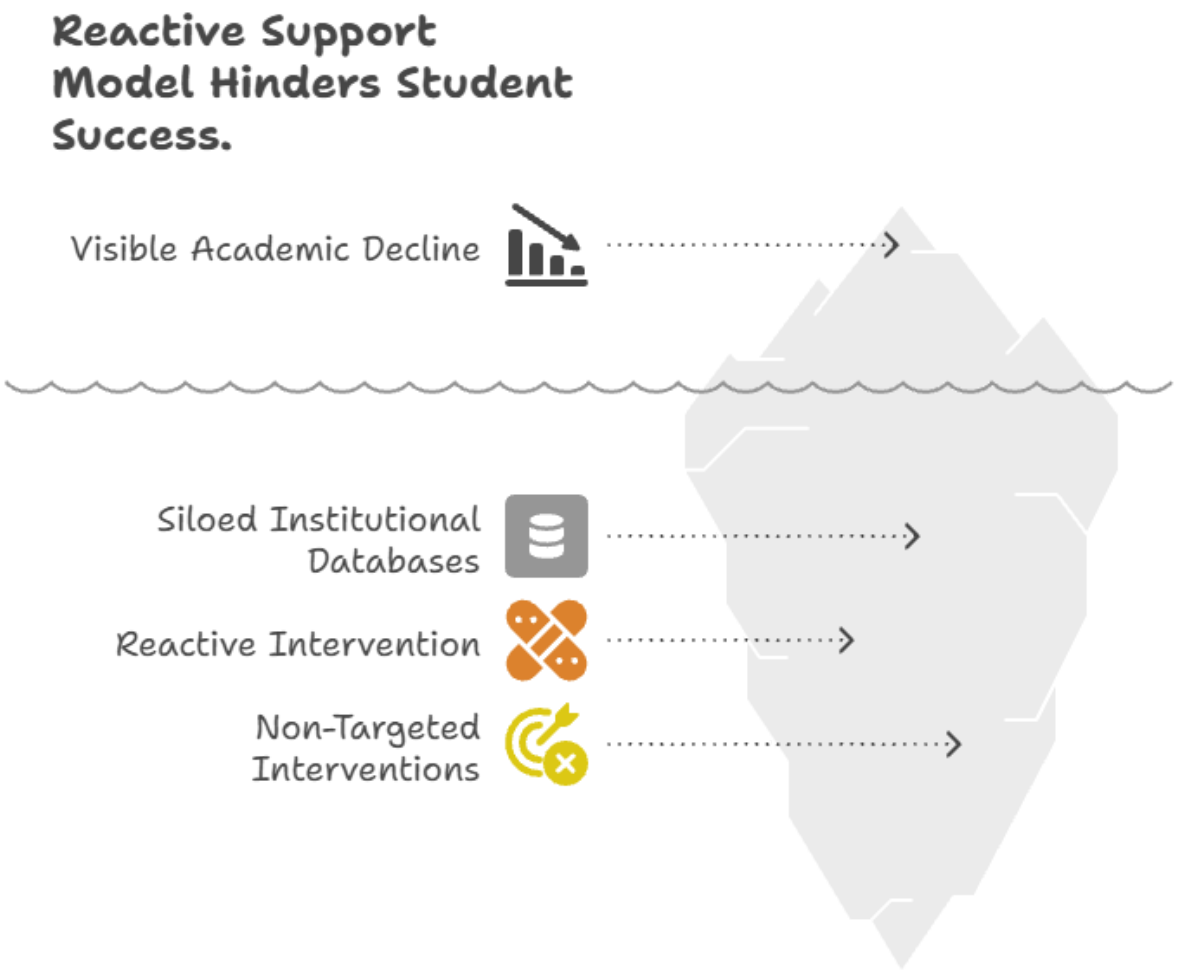


Figure 4.2: Illustration of the issues identified in the current system

Critical early-warning signs such as a gradual decline in attendance, minimal LMS engagement, or socio-emotional struggles go undetected due to poor system integration and a lack of predictive monitoring.

As a result:

- Opportunities for proactive intervention are missed.
- Student morale declines unnoticed.
- Institutional reputation and retention metrics suffer.

5. Problem Statement

"Spotting struggling students early is hard for universities, which leads to late support and more dropouts."

Identifying struggling students early is a significant challenge for universities, often resulting in delayed support and higher dropout rates. Although institutions collect large amounts of student data, they lack real-time analytics and predictive tools to act on it effectively. As a result, support tends to come too late after students are already at risk.

Academic staff are often overwhelmed with administrative work and use disconnected systems, making it difficult to identify and assist at-risk students early on. Existing solutions are either too complex or not well integrated into daily academic workflows. What advisors really need are simple, intuitive tools that offer early, personalized insights, not just raw data.

Without timely and user-friendly support systems, many students continue to slip through the cracks, and institutions face the ongoing costs of preventable dropouts.

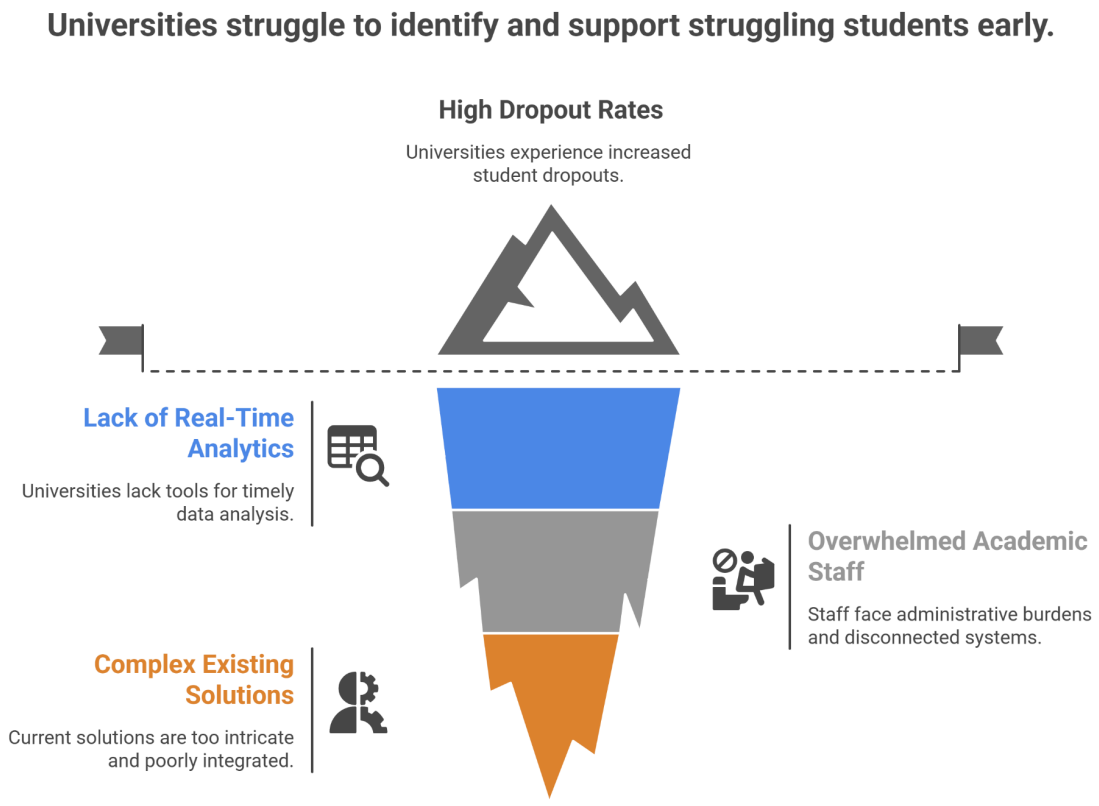


Figure 5.1: Causes of High Student Dropout Rate

6. SWOT Analysis

As we explored the issue of student dropout more deeply, we conducted a SWOT analysis as shown in Figure 6.1 to better understand the education industry and guide the system's development. This helped identify key internal strengths and weaknesses, as well as external opportunities and threats.

The main strengths were the availability of extensive student data and the growing use of analytics in education, which supported the creation of a data-driven solution. However, many institutions lack analytics skills and are often resistant to change, showing the need for a system that balances predictive power with a user-friendly design for non-technical users.

There are strong opportunities as well, such as the increasing demand for student retention tools and rapid progress in machine learning, making such systems timely and relevant. On the other hand, risks like algorithmic bias and cybersecurity threats highlighted the need for ethical design, data transparency, and strong privacy protections.Overall, the SWOT analysis helped shape both the technical and ethical aspects of the project, ensuring the system is useful, responsible, and suitable for the education sector.

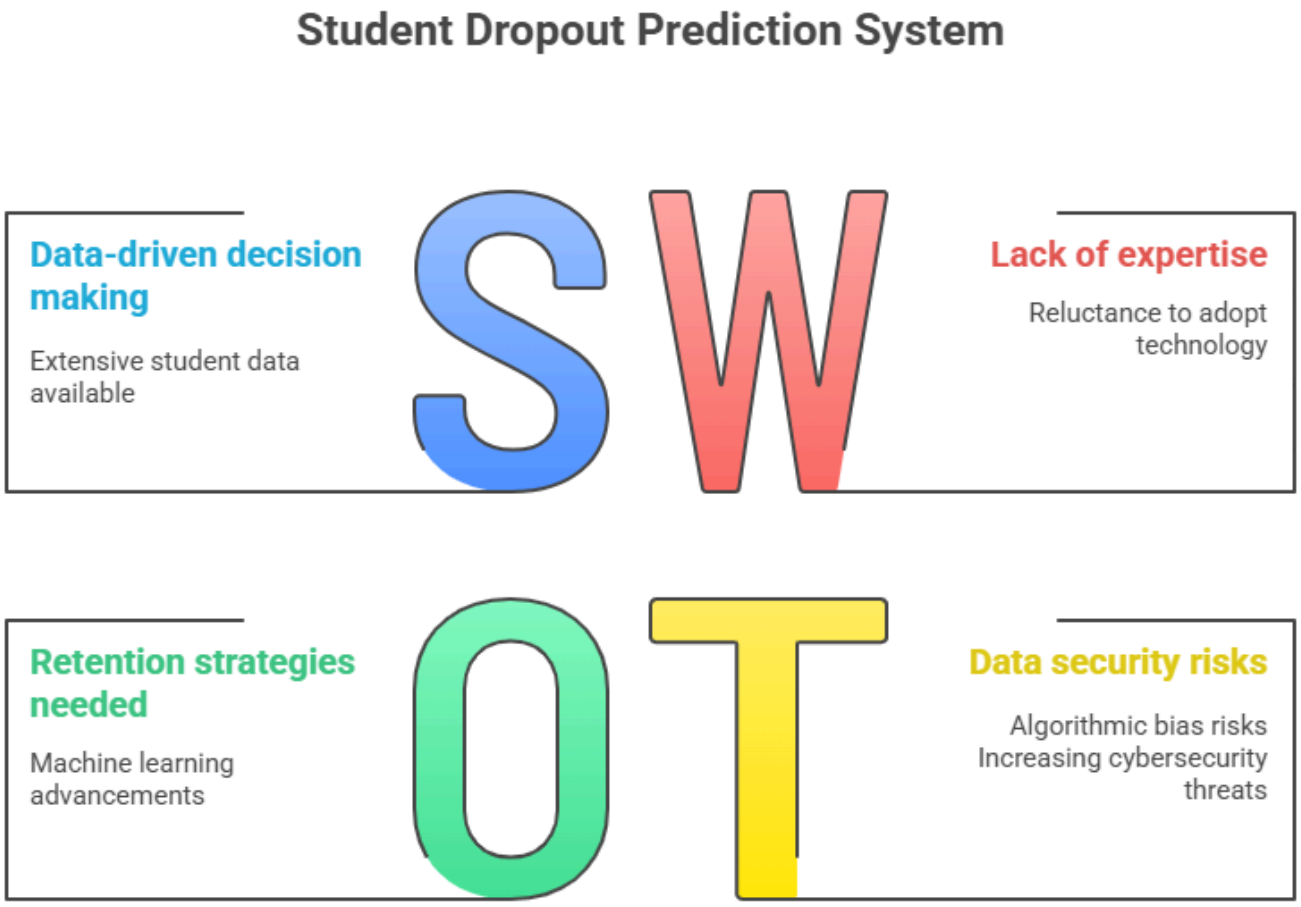


Figure 6.1: SWOT Analysis of the Education Industry

7. How Might We?

Following the analysis of market trends and institutional challenges in student retention, it is evident that both students and universities face growing pressures to address disengagement and dropout rates. Insights from data systems and personalized support mechanisms present an opportunity to transform this landscape.

"How might we enable students and universities to leverage data-driven systems to deliver timely, personalized interventions that reduce dropout risk and improve student retention?"

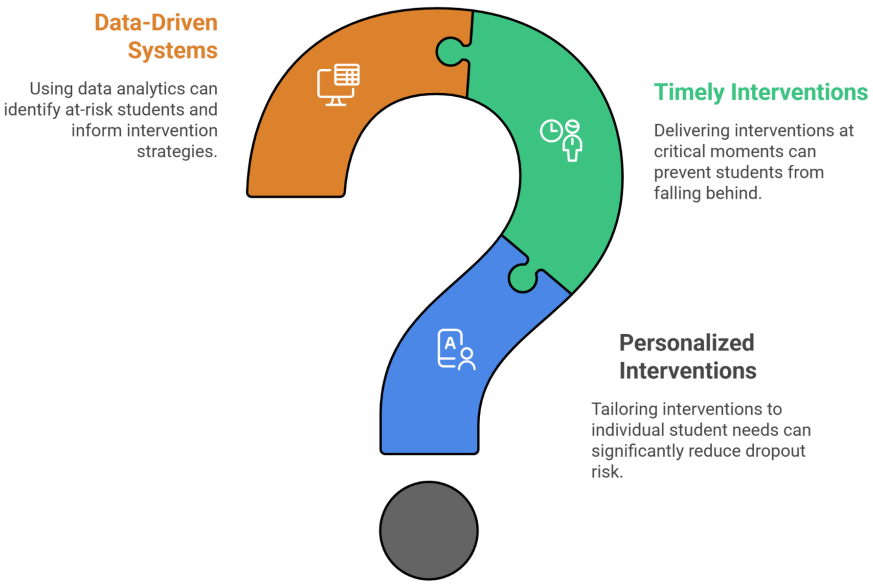


Figure 7.1: How Might We Improve the current situation

8. Proposed Solution

To address the challenge of identifying and supporting at-risk students early, we propose the development of an integrated predictive analytics platform specifically designed for higher education institutions. The **Student Dropout Prediction** leverages machine learning to assess the likelihood of student dropout in real time by analyzing a diverse range of historical and ongoing student data, including academic performance, demographic details, and financial indicators.

At the core of the system is a user-friendly, advisor-facing dashboard that presents risk predictions and key insights clearly and intuitively. This enables academic advisors and student support teams, regardless of technical expertise, to quickly identify students who may be struggling and take personalized, proactive action before disengagement or failure occurs.

Key features of the platform include:

- Real-time risk scoring for individual students based on dynamic data inputs.
- Visual dashboards that highlight trends and flag high-risk cases for immediate attention.

- Integration capabilities with existing student information systems (SIS) for seamless data access and updates.

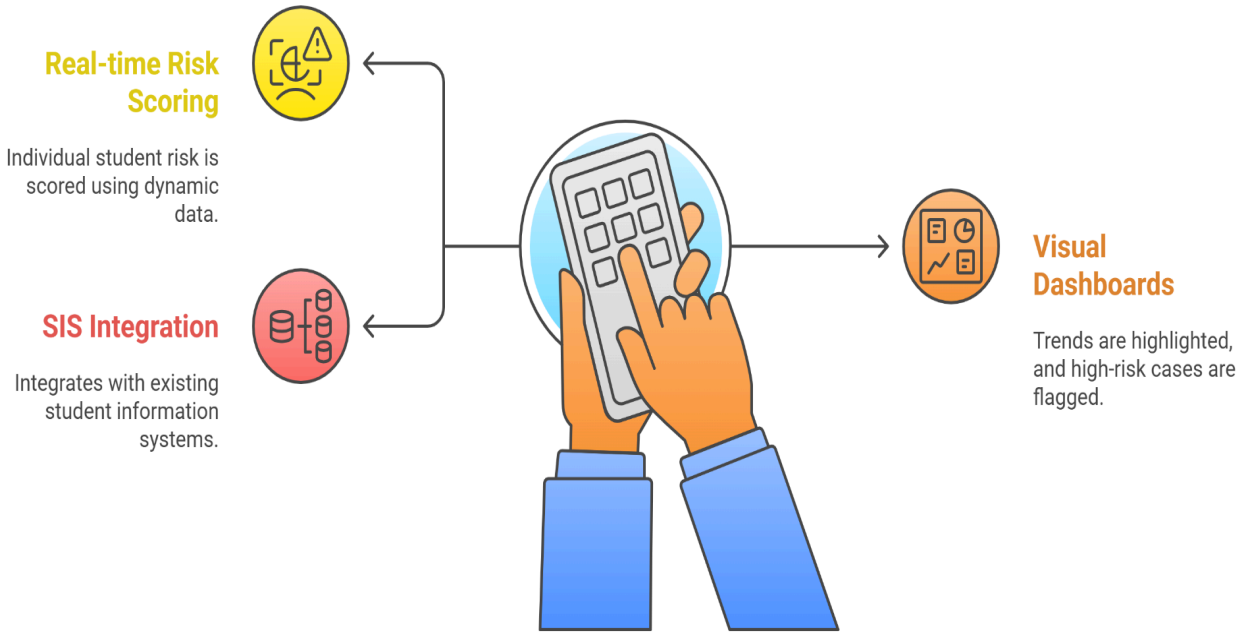


Figure 8.1: Student risk management features

By replacing reactive, manual processes with a data-driven, proactive support model, the platform aims to:

- Enhance student retention and academic performance.
- Reduce institutional dropout rates and associated financial losses.
- Strengthen student-advisor engagement through timely, tailored interventions.
- Improve overall institutional effectiveness and reputation.

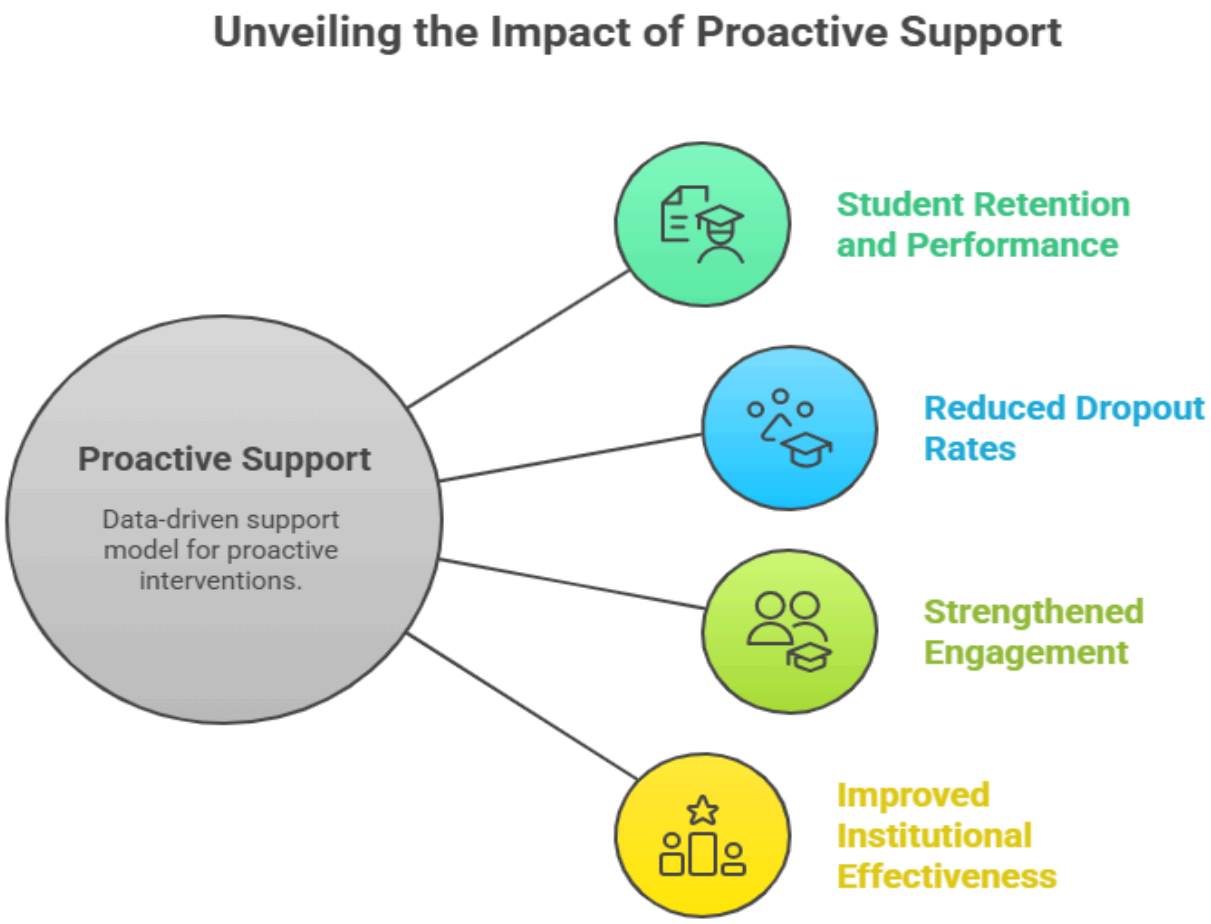


Figure 8.2: Unveiling the impact of proactive support

This solution empowers universities to move beyond generalized assumptions and embrace personalized, evidence-based strategies that prioritize student success at every step of the academic journey.

9. SDG Alignment

Our artefact directly supports the United Nations Sustainable Development Goals:

- **SDG 4: Quality Education** – By enabling equitable, personalized academic support, the platform enhances access to quality education and lifelong learning.

- **SDG 10: Reduced Inequalities** – The predictive system helps institutions identify and support vulnerable students who may otherwise go unnoticed, reducing disparities based on socio-economic or demographic backgrounds.

By democratizing access to data-driven insights and ensuring interventions are timely and targeted, we help create a more inclusive academic environment.

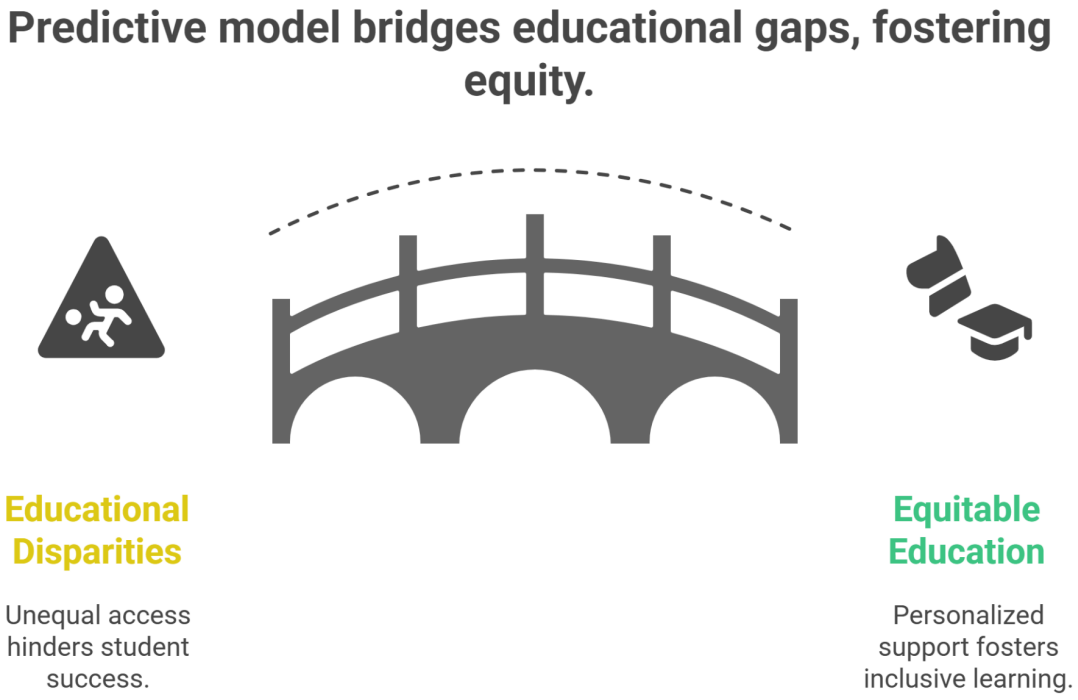


Figure 9.1: Model's role in developing an inclusive academic environment.

10. Data And Pipeline Management

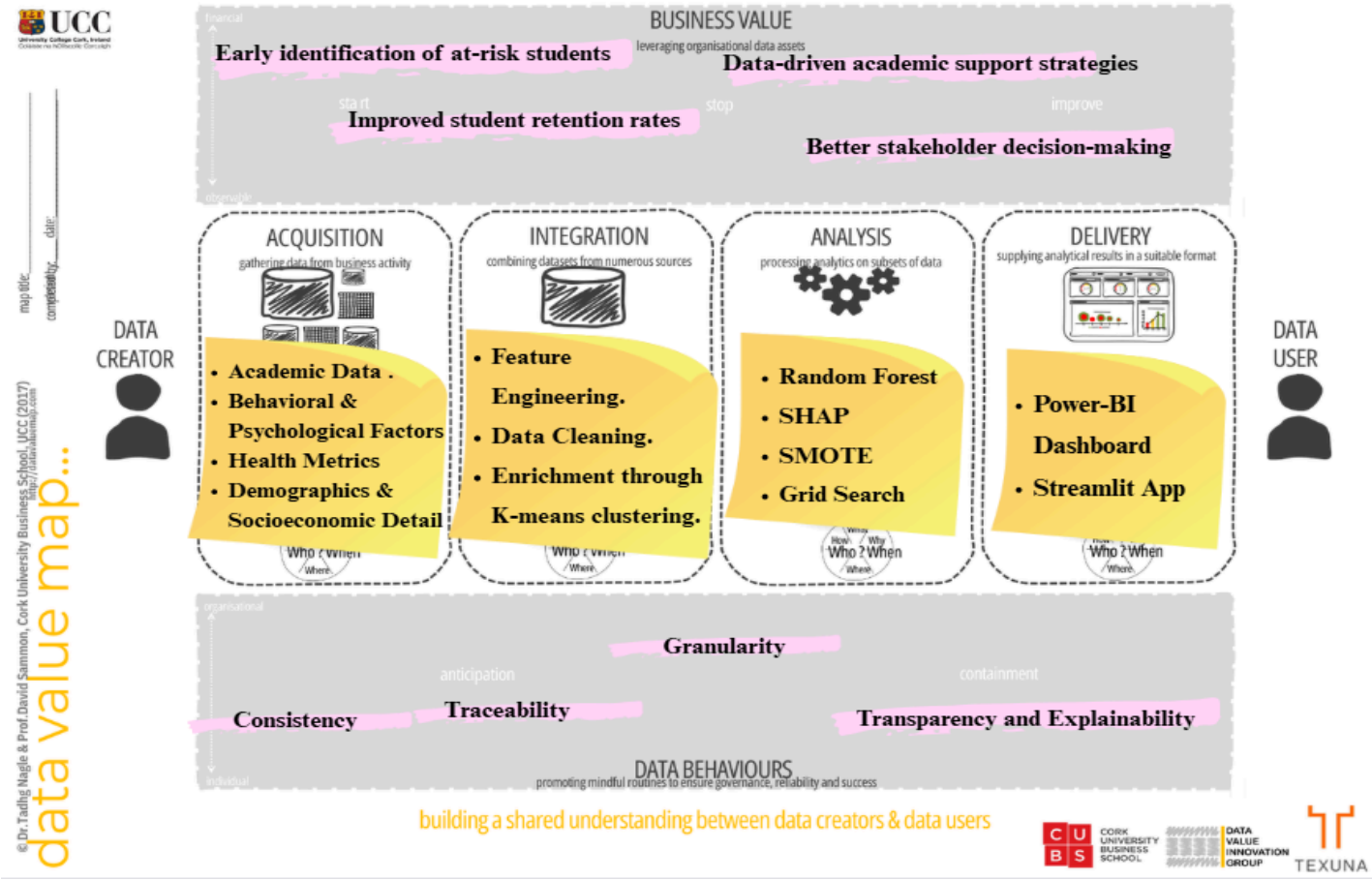


Figure 10.1: DVM.

The different stages of the data lifecycle are portrayed in Figure 10.1.

Acquisition

This project used four datasets capturing diverse aspects of student life: academic records, mental health survey responses, performance evaluations, and simulated health sensor data. Together, they offer a comprehensive view of student behavior, performance, and well-being, forming the basis for dropout prediction modeling.

Dataset Description:

The datasets utilized for our analysis and model development, illustrated in Figure 10.2, include:

1. **edudata**:
   Sourced from a public repository of University of California, Irvine , this dataset contains demographic, academic, and socio-economic information on 4424 undergraduate students. It tracks academic performance over two semesters and labels students as dropped out, enrolled, or graduated. This serves as the core dataset for dropout modeling.

2. **mental_health_data**:
   Collected via an online survey at the International Islamic University Malaysia (IIUM), this dataset includes self-reported mental health conditions such as depression, anxiety, and panic attacks, along with CGPA. It enriches the main dataset with emotional and psychological context.

3. **performance_data**: Comprising 1195 responses, this dataset was from a private Bangladeshi university, combining academic data (CGPA, SGPA, attendance) with behavioral traits like study hours, scholarship status, and extracurriculars. It provides behavioral insights into student performance.

4. **health_data**:
   This simulated dataset includes physiological data from biosensors (e.g., heart rate, blood pressure) and self-reported stress, mood, and academic workload. It also includes a health risk label, used to enrich students' health profiles in the final dataset.

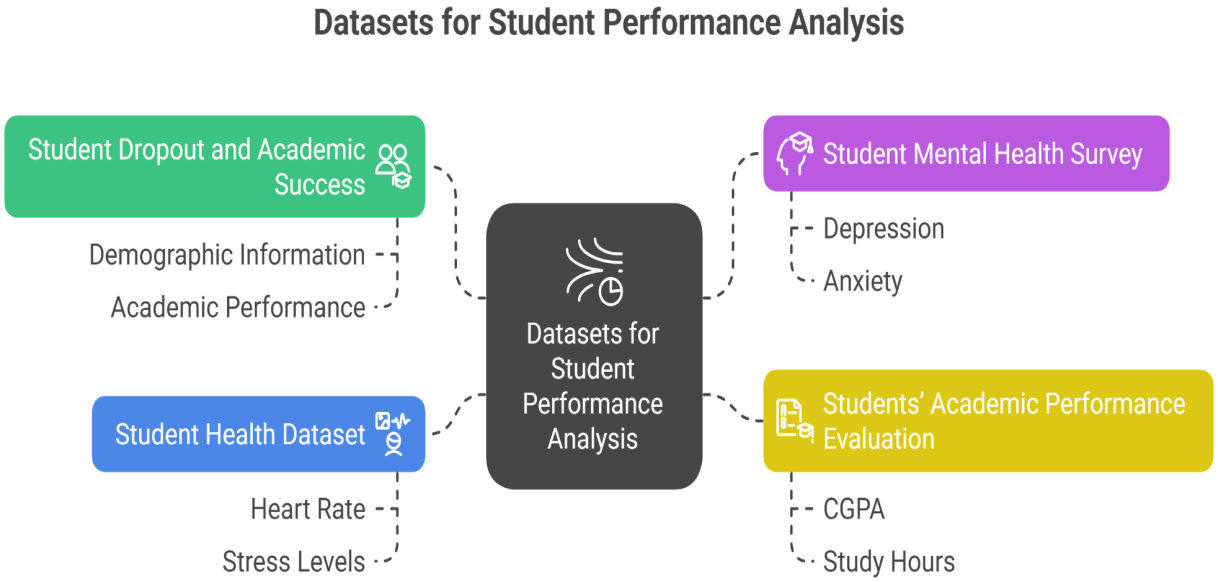


Figure 10.2 : Datasets Acquired

Integration

The integration phase focused on combining data from four distinct sources into a single enriched dataset (**edudata_final**). Since the datasets lacked a shared student identifier, traditional key-based joins were not feasible.

To address this, we applied a clustering-based integration strategy using **K-Means clustering** as shown in figure 10.3. Each dataset (**edudata**, **mental_health_data**, **performance_data**, and **health_data**) was clustered independently, grouping students with similar characteristics within each domain. These cluster IDs served as semantic links, allowing us to enrich **edudata** with relevant features from matching clusters across the other datasets.

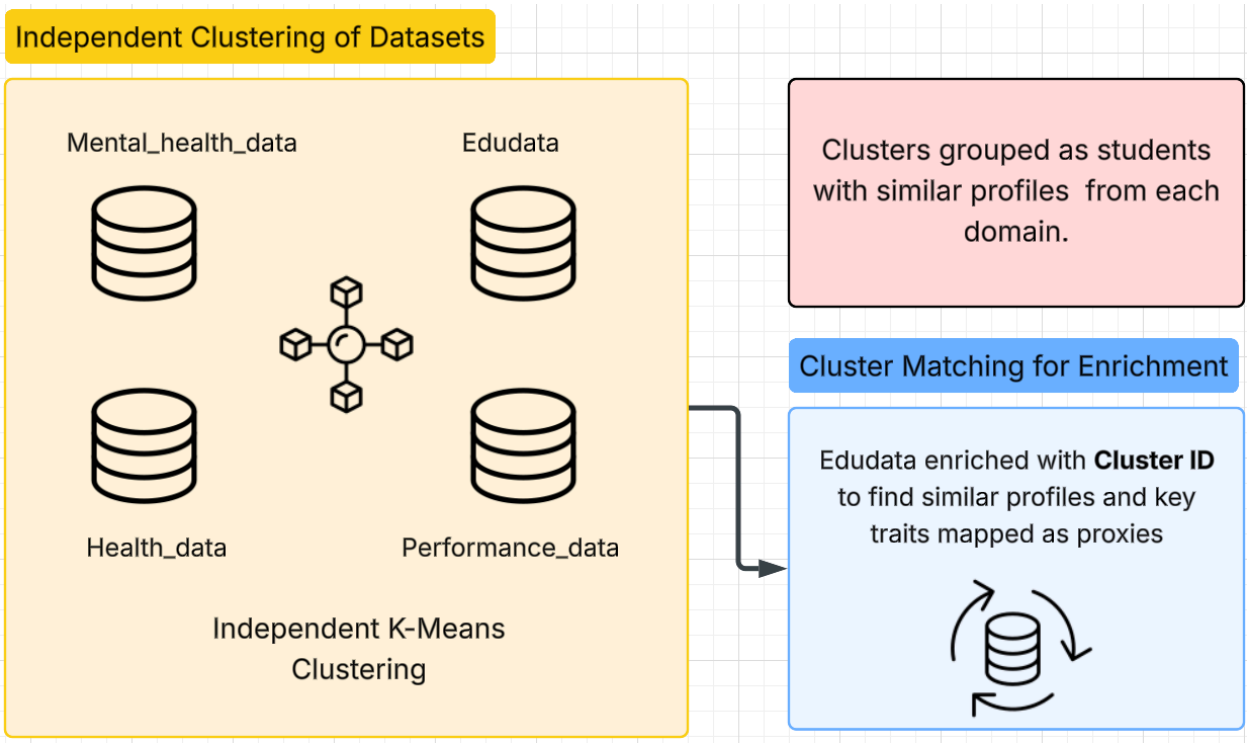


Figure 10.3: Clustering based enrichment

For each student in **edudata**, values such as stress_levels, study_habits, and mental_health_indicators were generated by sampling from corresponding clusters of other datasets. Enriched features were then inverse-transformed to restore original value scales for interpretability.

Before clustering, extensive **preprocessing** was carried out to ensure consistency and model-readiness across datasets. This included standardizing column names and formats, handling missing values and outliers, and converting categorical values to numeric. Manual **feature engineering** was performed, such as constructing **Performance_Ratio_avg** by aggregating semester data, and scaling GPA to a 0–4 format. **Redundant or highly correlated columns** were removed to reduce dimensionality, avoid multicollinearity, and improve clustering quality.

Numerical features were **normalized** using MinMax scaling for clustering. After enrichment, the dataset was further enhanced with synthetic identifiers (**Student_ID**, **Name**, **Email**) using **Synthetic Data Vault (SDV)**, and derived demographic fields like **Living_With_Family** and **StudentWorker**, based on cluster profiles.

This approach effectively simulates institutional data silos while enabling cohesive student-level integration. The final dataset supports reliable dropout prediction and multidimensional analysis (figure 10.4).

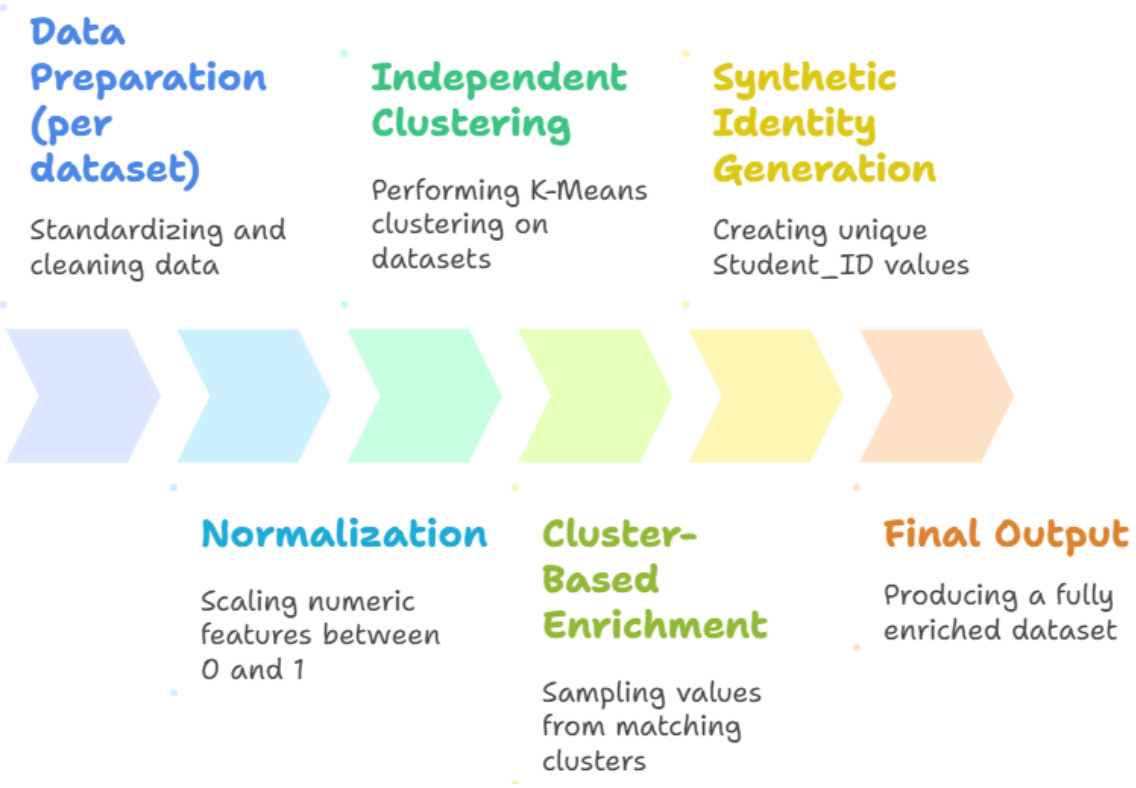


Figure 10.4: Integration workflow

Data Behaviors

To ensure transparency and data integrity, all sensitive identifiers were anonymized or synthesized. Adopting a **data-first mindset**, we applied consistent preprocessing and feature engineering across all sources. In addition to static attributes like GPA and age, we modeled behavioral patterns such as academic engagement, study intensity, and health risk. These dynamic signals added contextual depth, enhancing both robustness and the ethical quality of our predictions.

Analysis

The analysis phase involved building the final predictive model using the fully preprocessed dataset. Only records labeled as *Graduate* or *Dropout* were used for training, while data labeled *Enrolled* was reserved for future prediction. To address class imbalance, the **Synthetic Minority**

Oversampling Technique (SMOTE) was applied, ensuring a more balanced and robust training process.

Feature Selection:

To improve model interpretability and efficiency, **Recursive Feature Elimination (RFE)** was conducted using **SHAP (SHapley Additive exPlanations)** values in combination with **five-fold cross-validation**(figure 10.5).

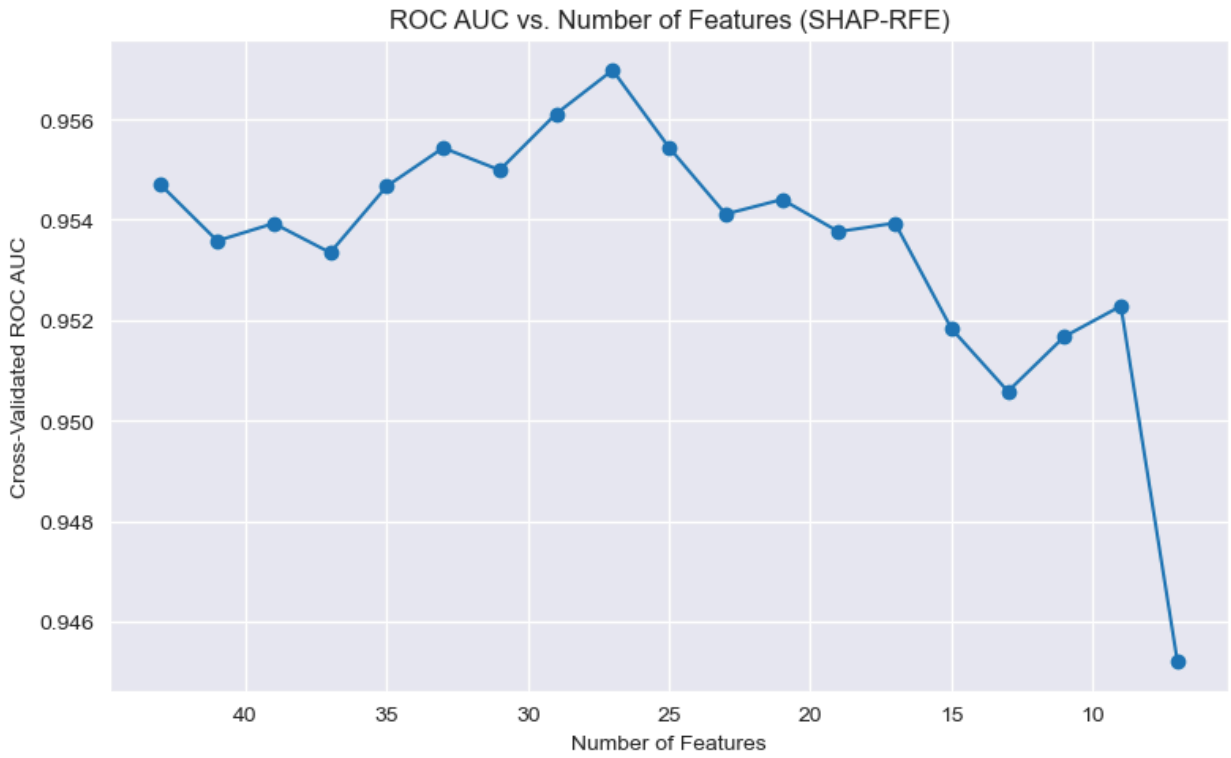


Figure 10.5 : RFE with SHAP and CV along with model performance

SHAP values, derived from a Random Forest model trained on SMOTE-resampled data, provided a robust ranking of feature importance. The least important features were iteratively removed, with model performance evaluated at each step using ROC AUC. Performance remained stable as features were reduced, with the optimal AUC achieved using the top 30 features, which were retained for the final model (Figure 10.6).

```
Final Features to Eliminate:
 ['Daily_Study_Hours', 'Curricular units 1st sem (credited)', 'Displaced', 'Physical_Activity', 'Curricular units 2nd sem (credited)',
   'Previous qualification', 'Marital status', 'Daytime/evening attendance', 'Health_Issues_Reported', 'Nationality', 'International',
   'Educational special needs']
Final Features to Keep (for modeling):
 ['Application mode', 'Application order', 'Course', 'Previous qualification (grade)', "Mother's qualification", "Father's
 qualification", "Mother's occupation", "Father's occupation", 'Admission grade', 'Debtor', 'Tuition fees up to date', 'Gender',
 'Scholarship holder', 'Age at enrollment', 'Unemployment rate', 'Inflation rate', 'GDP', 'Performance_Ratio_avg', 'GPA_avg_scaled',
 'Health_Risk_Proxy', 'Stress_Level_Proxy', 'Study_Hours_Proxy', 'Sleep_Quality_Proxy', 'Mental_Depression', 'Mental_Anxiety',
 'Social_Media_Hours', 'Transport_User', 'Avg_Attendance', 'Teacher_Consultancy', 'StudentWorker', 'Living_With_Family']
```

Figure 10.6 : Final features

Model Building

Multiple classification algorithms were initially evaluated, including Logistic Regression, SVM, Random Forest, and K-Nearest Neighbors. Random Forest was selected due to its superior accuracy, recall, and F1-score, particularly for predicting dropouts (Figure 10.7).

Model	Accuracy	precision	recall	f1-score	support
Random Forest	0.87	0.84	0.75	0.79	284
		0.89	0.93	0.91	601
SVM	0.86	0.89	0.65	0.75	284
		0.86	0.6	0.91	601
Logistic Regression	0.82	0.70	0.82	0.75	284
		0.91	0.83	0.87	601
KNN	0.82	0.88	0.57	0.69	316
		0.80	0.96	0.87	569

Figure 10.7 : Consolidated Classification report of multiple models

To further optimize performance, **hyperparameter-tuning** was conducted using both GridSearchCV and RandomizedSearchCV. This improved the model's recall and F1-score for the dropout class to 92%, aligning with the goal of minimizing false negatives (Figure 10.8).

```
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.95      0.95       427
           1       0.93      0.92      0.92       284

    accuracy                           0.94       711
   macro avg       0.94      0.94      0.94       711
weighted avg       0.94      0.94      0.94       711

Confusion Matrix:
[[406  21]
 [ 22 262]]
```

Figure 10.8 : Final Model Classification Report

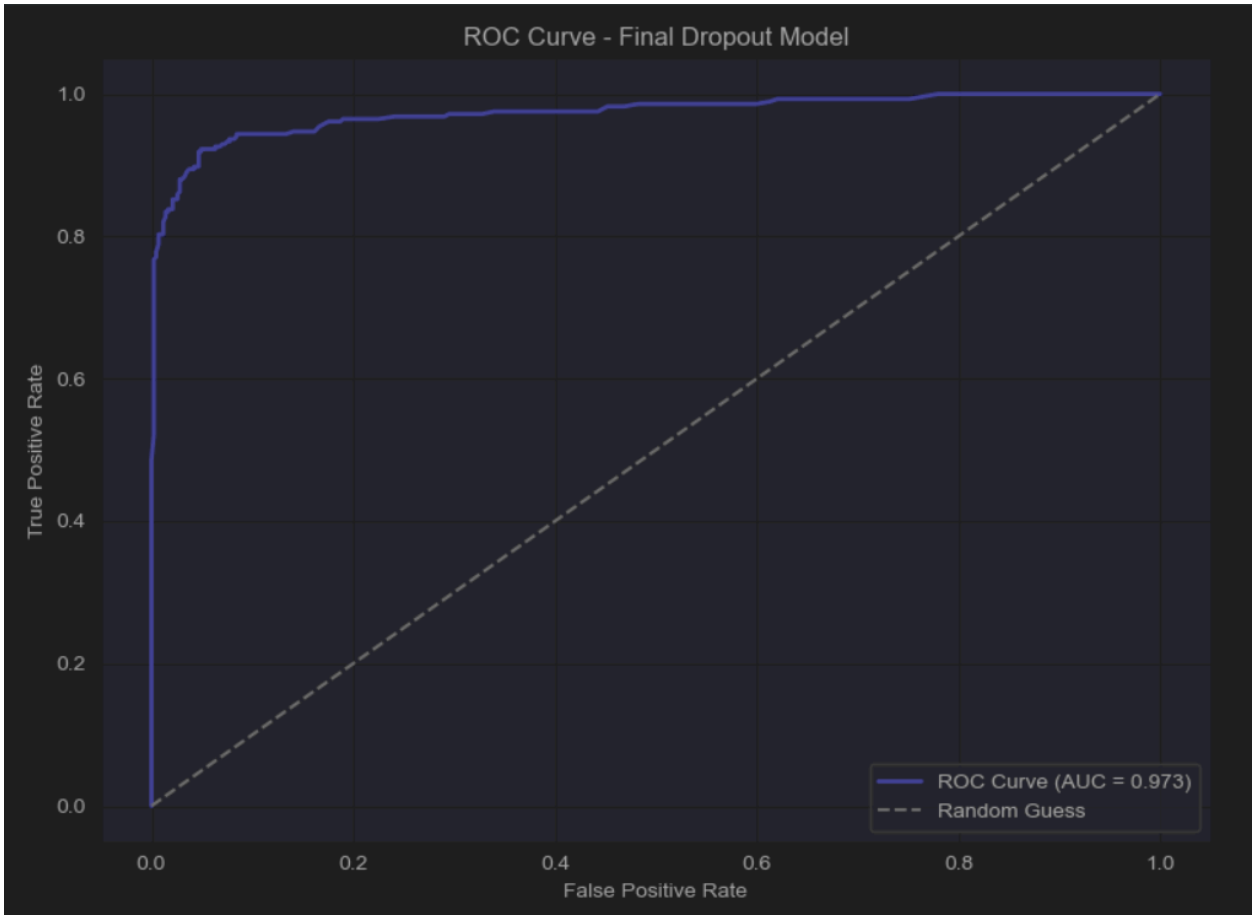


Figure 10.9 : ROC Curve

The final model achieved a ROC AUC score of 0.973 (Figure 10.9), indicating excellent class separation. To ensure the model was not overfitted, all evaluation metrics were assessed on a held-out test set and validated through cross-validation during feature selection and tuning. The model showed consistent performance across folds, confirming its generalizability. Outliers were addressed during preprocessing to reduce bias and enhance robustness. Once validated, the model was exported as a **model.joblib** file for deployment (figure 10.10).

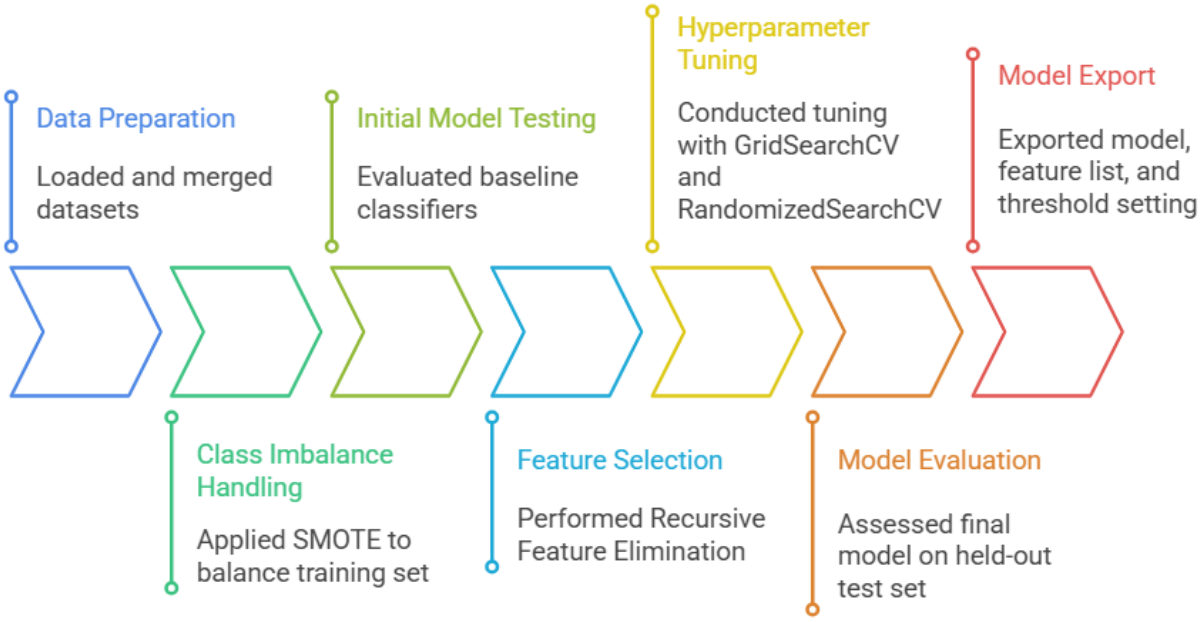


Figure 10.10: Illustration of Analysis Workflow

11. Delivery

Our goal was to equip academic advisors, counselors, and administrators with actionable, data-driven insights. We built the dashboard using Streamlit and Power BI, combining enterprise scalability with user-friendly design to ensure accessibility for non-technical users. This stage transformed raw student data and model outputs into clear insights that support early intervention. By emphasizing transparency and explainability, we encouraged informed action. The artefact delivers business value by reducing student attrition, improving outcomes, and protecting institutional revenue through timely, data-driven support.

11.1 Streamlit Interface

The **Student Dropout Prediction** web application, built with Streamlit and hosted on Streamlit Cloud, serves as a user-friendly and intelligent interface to identify students at risk of dropping out. Powered by a machine learning model and integrated with a MySQL database and SHAP explainability, the system supports both individual analysis and large-scale batch processing and provides critical insights for academic stakeholders. The application acts as a medium for the users to upload student dataset to the predictive model.

Key Features and Working of the Dropout Prediction Streamlit App

The **Streamlit web application** offers two main modes for student dropout prediction:

1. Individual and Batch Prediction

- **Individual Prediction (via Student ID):**
  Users can input a Student ID to fetch data from a connected **MySQL database**. A pre-trained **Random Forest model**, loaded via joblib, predicts whether the student is at risk of dropping out. Results include:

  - A clear prediction with dropout probability.

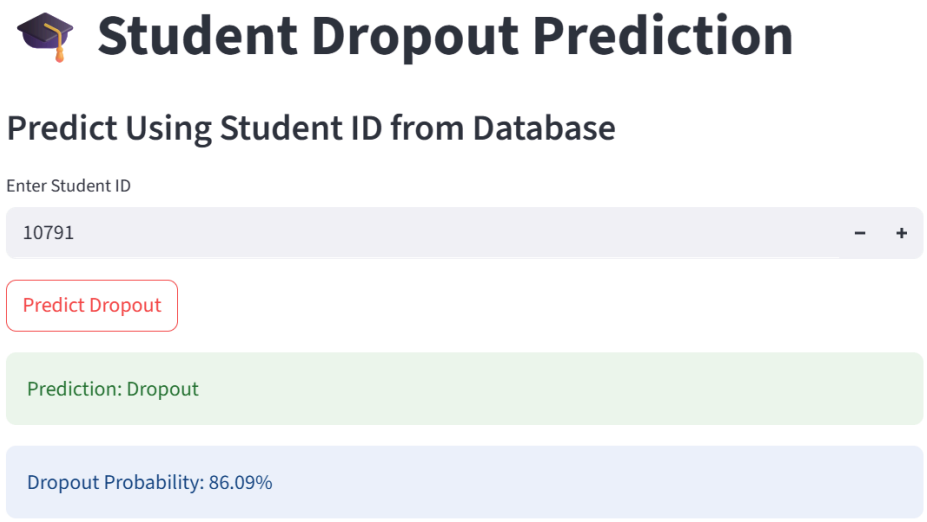


Figure 11.1.1: Student ID based dropout prediction

  - A **SHAP chart** displaying the top 10 features contributing to the prediction (e.g., GPA, Mental Health, Transport Usage).

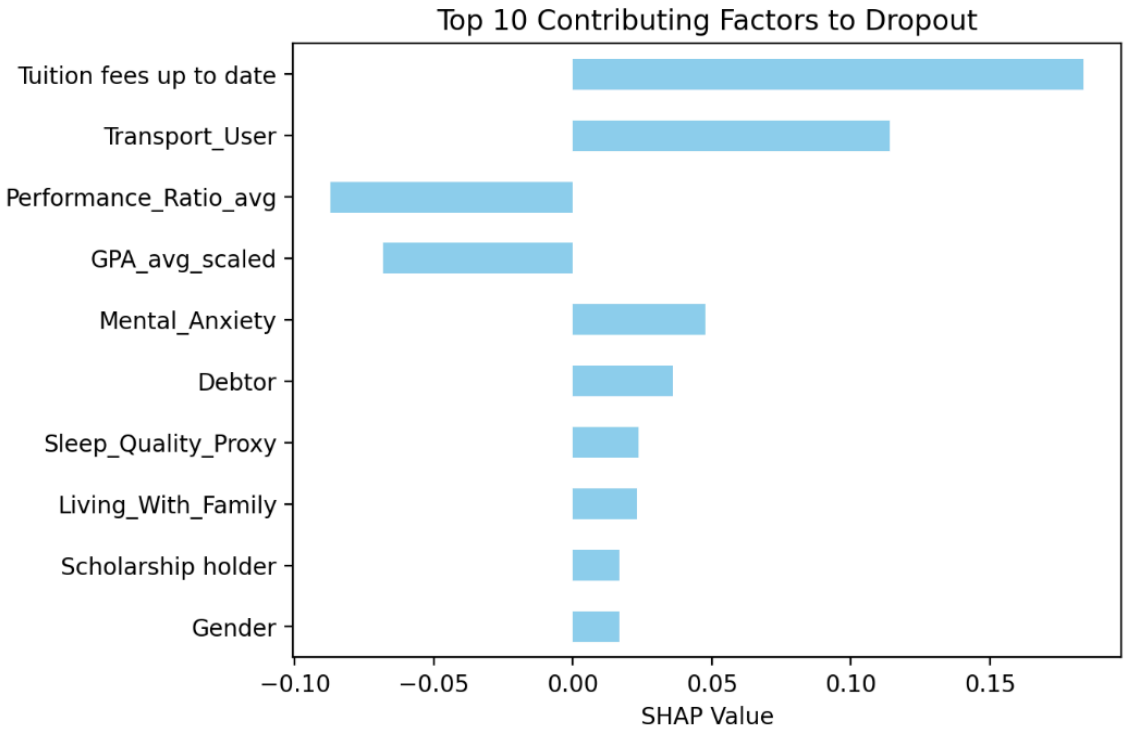
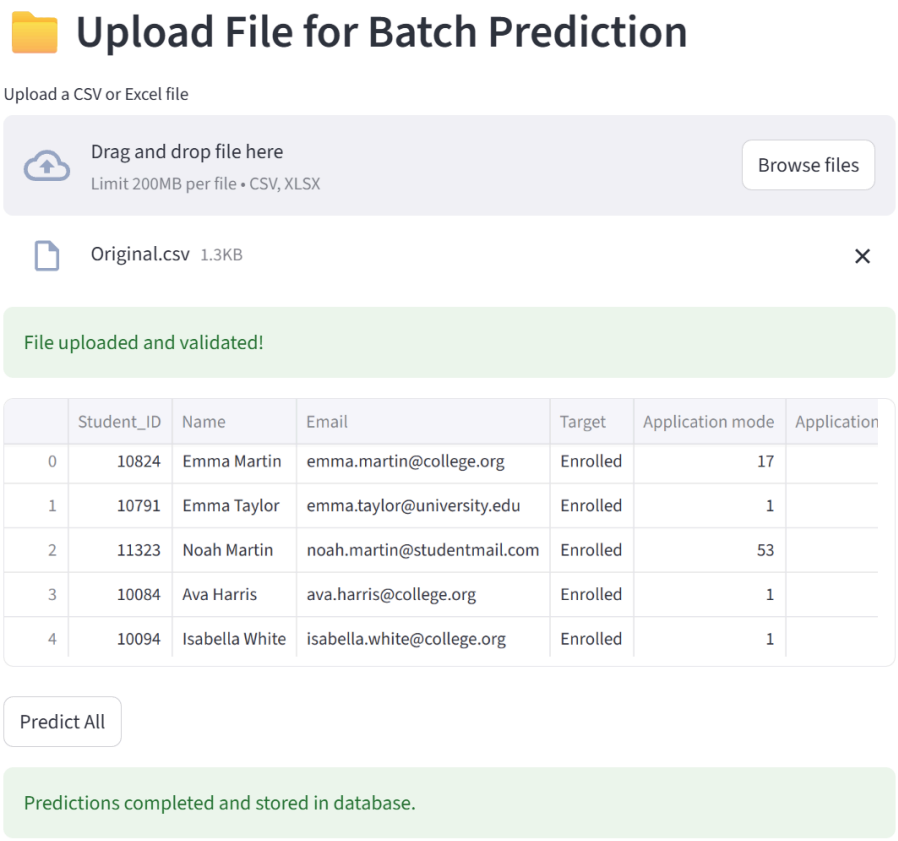


Figure 11.1.2: SHAP chart of top 10 risk factors of individual student

- **Batch Prediction (Bulk Upload):**
  Users can upload CSV or Excel files containing student data. The file is validated, records are stored or updated in the database, and predictions are generated for each student. This allows institutions to assess dropout risks for multiple students at once.



	Student_ID	Name	Email	Target	Application mode	Application
0	10824	Emma Martin	emma.martin@college.org	Enrolled	17	
1	10791	Emma Taylor	emma.taylor@university.edu	Enrolled	1	
2	11323	Noah Martin	noah.martin@studentmail.com	Enrolled	53	
3	10084	Ava Harris	ava.harris@college.org	Enrolled	1	
4	10094	Isabella White	isabella.white@college.org	Enrolled	1	

Figure 11.1.3: Batch data upload section

The output includes:

- A table of predictions with probabilities.

Prediction Results

	Student_ID	Prediction_Result	Dropout_Probability
1	10791	Dropout	86.09
2	11323	Dropout	73.31
0	10824	Dropout	67.29
3	10084	Not Dropout	14.29
4	10094	Not Dropout	7.14

Figure 11.1.4: Prediction table showing dropout probabilities of batch data of students

- Dropout vs. Non-Dropout count chart.

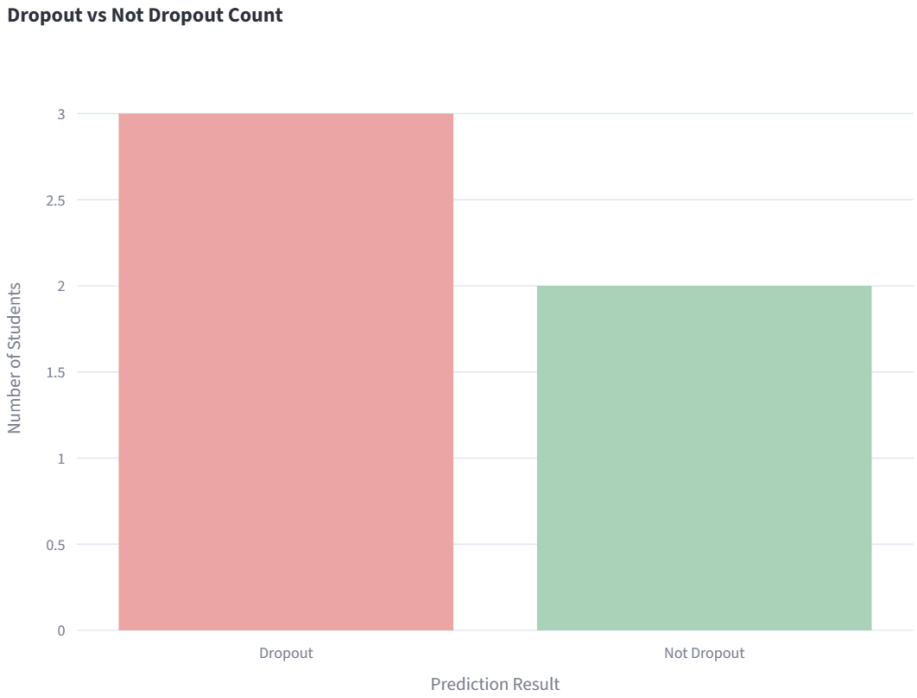


Figure 11.1.5: Dropout and Non Dropout students count comparison chart

- SHAP summary of top predictive features across the dataset.

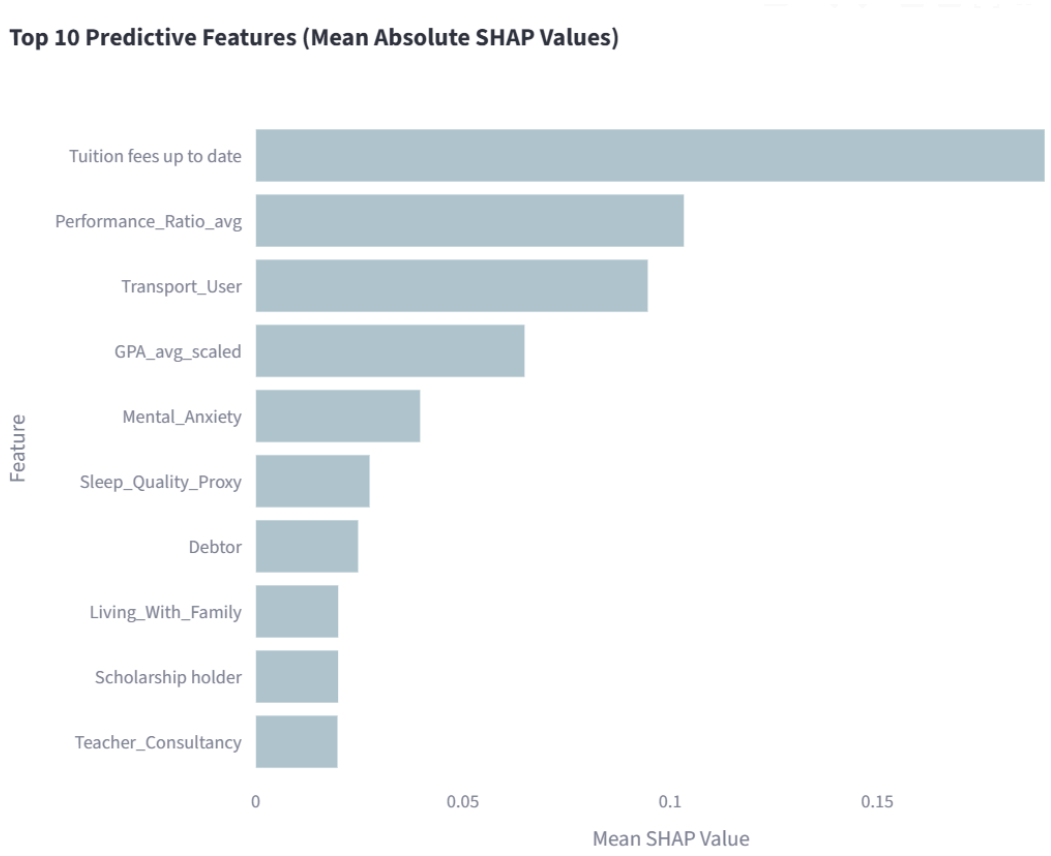


Figure 11.1.6: Summary chart of top dropout risk factors of the uploaded batch dataset dropout students

2. Database and Explainability Integration

The app is fully connected to a **cloud-hosted MySQL database**, where:

- Input data is fetched from the student_data table.
- Predictions and SHAP explanations are stored in the predictions table.

3. Integration with Power BI

The uploaded datasets also feed into a **Power BI dashboard**, enabling deeper, interactive insights. Streamlit thus acts as both a prediction tool and a gateway for **bulk data ingestion** into a broader analytics ecosystem.

Designed with non-technical users in mind, this app combines real-time predictions, intuitive visualizations, SHAP-based transparency, and Power BI integration empowering institutions to make **data-driven decisions** to reduce student dropout.

11.2 Power BI Dashboard

The interactive dashboards created for this project serve as a critical interface between the predictive analytics engine and its real-world application by institutional stakeholders. It comprises two integrated dashboards: **a University Overview Dashboard (Figure 11.2.1)** and an **Individual Student Overview Dashboard (Figure 11.2.2)**. While the backend model and web app automate dropout risk predictions, the Power BI dashboards transform these outputs into visually intuitive, actionable insights tailored for decision-makers. The dashboards are primarily designed for **academic advisors**, **student support staff**, and **university administrators**, enabling them to proactively monitor dropout risk patterns and intervene with the right support measures.

This dual-layered system allows stakeholders to not only identify who is at risk but also understand **why ensuring interventions are data-informed and targeted rather than generic or reactive**. By linking the machine learning predictions to contextual insights, the dashboard empowers human-led academic support, making the artefact both intelligent and empathetic.

Figure 11.2.1 University Overview Dashboard

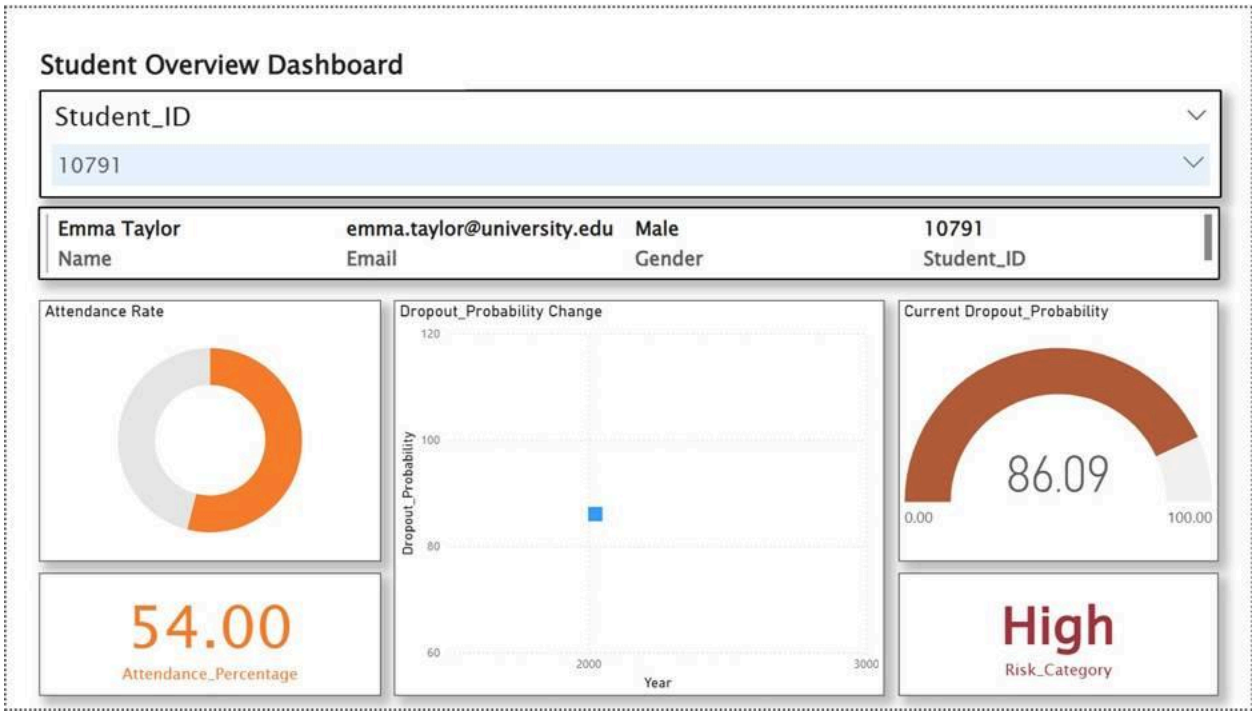


Figure 11.2.2 Student Overview Dashboard

11.2.1 University Overview Dashboard

1. Dynamic KPI Cards

These offer a snapshot of critical institutional metrics to aid quick decision-making:

- Displays the count of the total enrolled students.
- Shows the percentage of students categorized as **"Low Risk" (green)** and **"High Risk" (red)**, calculated using DAX formulas based on dropout probability scores.
- Cards are conditionally formatted to visually signal risk levels.

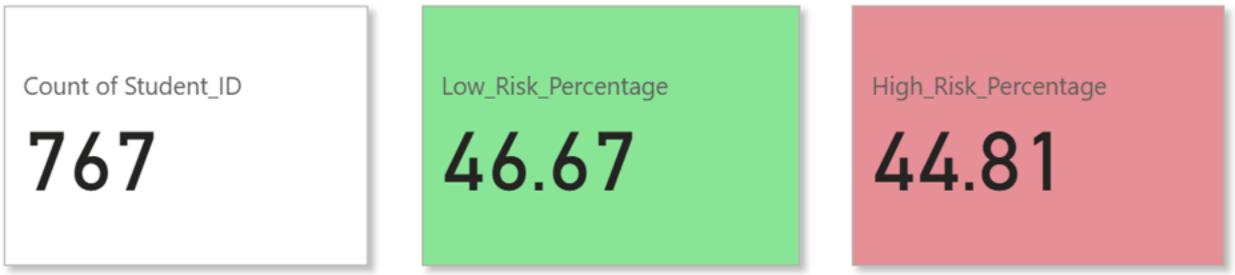


Figure 11.2.1 Dynamic KPI Cards

2. Profile Summary Table

Designed for administrators to quickly identify and focus on at-risk students:

- Lists **Student ID**, **average academic metrics**, and **Risk Category**, with color-coding for emphasis.
- Supports filters to isolate and analyze students by risk classification.

Profile Summary

Student_ID	Average of Performance_Ratio _avg	Risk_Category
10000	0.23	High
10016	0.44	High
10024	0.40	High
10031	0.25	High
10044	0.38	High
10060	0.28	High
10061	0.25	High
10066	0.23	High

Figure 11.2.2 Profile Summary

3. Top Risk Factors Chart (SHAP Values)

It explains the predictive drivers behind student risk:

- Uses **SHAP (SHapley Additive exPlanations)** values to determine feature importance.

- A **bar chart** displays average SHAP values, highlighting impactful factors like **attendance**, **grades**, and **mental health indicators**.

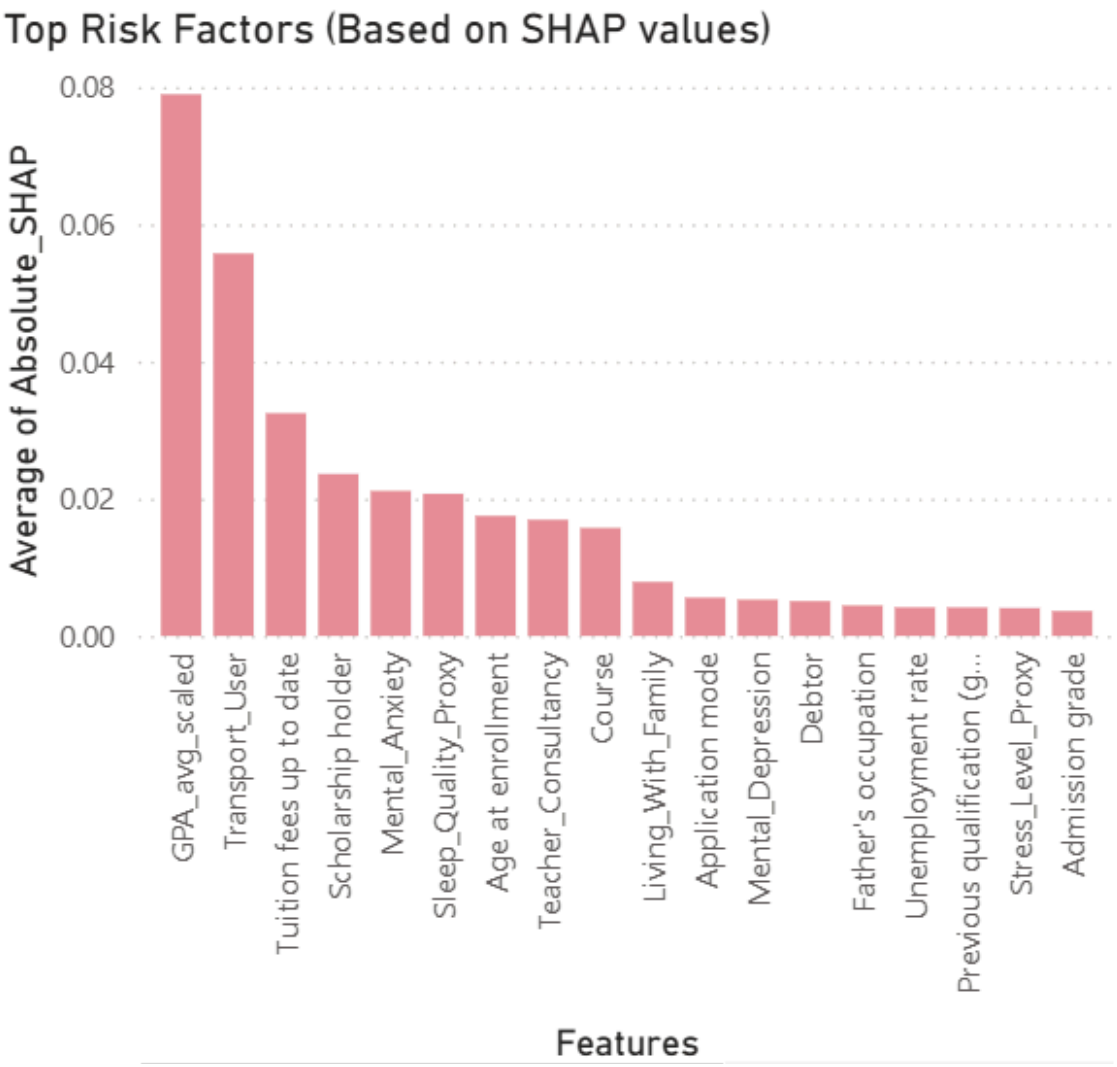


Figure 11.3.1 Top Risk Factors

4. Risk Analysis

This section highlights disparities across demographic lines:

- A **clustered column chart** compares risk categories across gender groups.
- A **donut chart** visualizes the overall breakdown of Low, Moderate, and High-risk students.

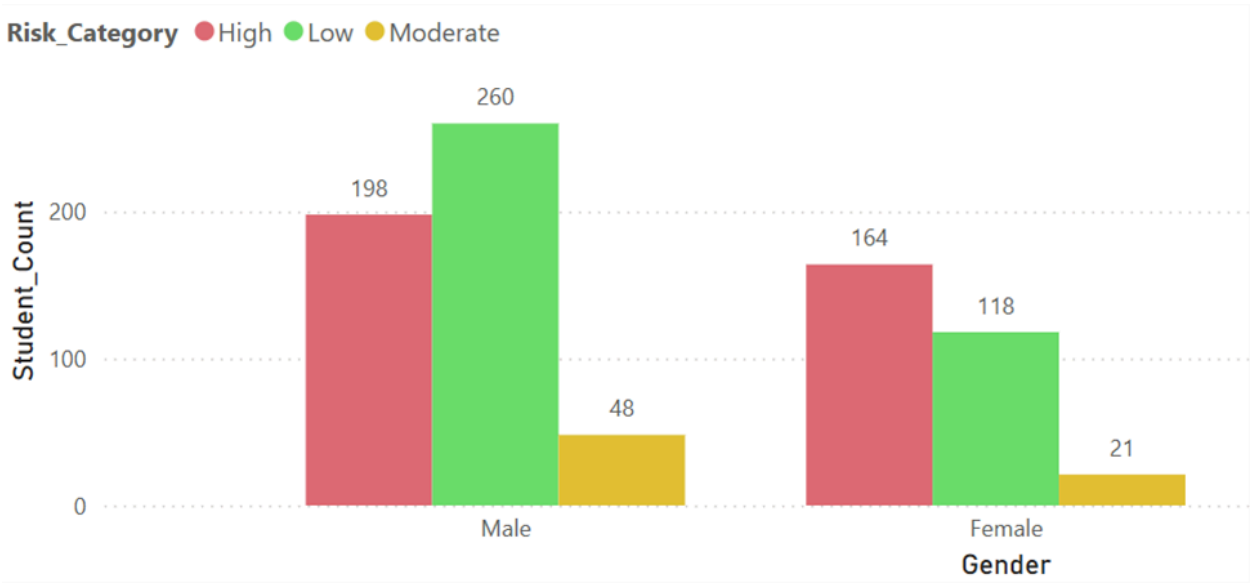


Figure . 11.4.1 Gender Based Risk Analysis

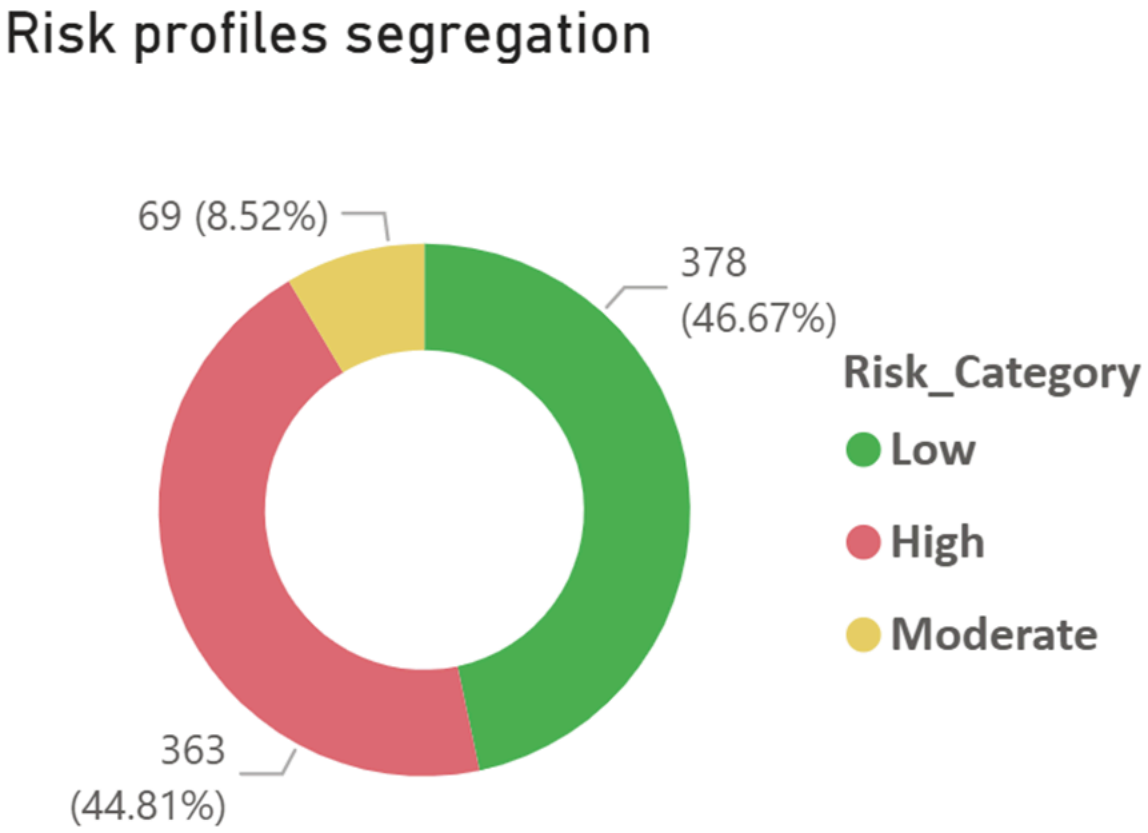


Figure . 11.4.2 Risk Analysis

11.2.2 Individual Student Overview Dashboard

1. Student ID Search

Provides a focused, student-level view:

- A **dropdown slicer** lets users filter by **Student ID**, dynamically updating all associated visuals.

Figure . 11.2.1.1 Student Profile Search

2. Student Profile Card

Summarizes student details in an easy-to-read format:

- Displays **name, email, gender**, and **Student ID**.
- Converts encoded data (e.g., gender: 0/1) into readable labels using DAX.

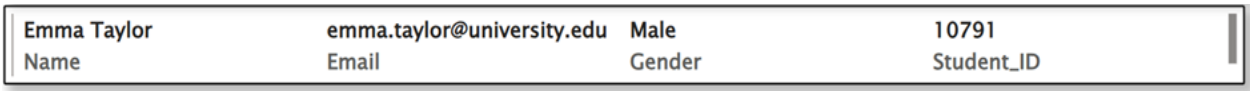


Figure . 11.2.2.1 Student Profile Card

3. Key Metrics: Attendance & Dropout Probability

Visualizes the student's performance and risk:

- **Gauge chart** for Attendance Rate.

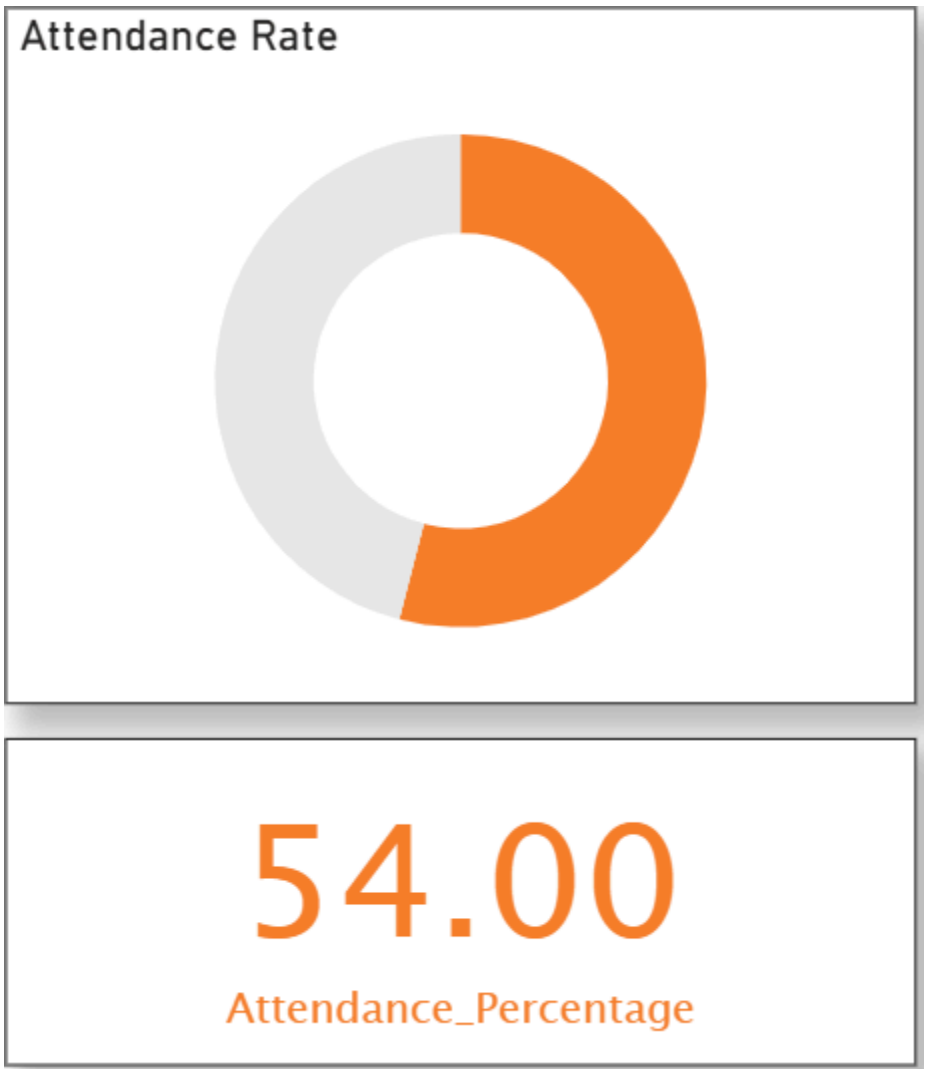


Figure . 11.2.3.1 Attendance Rate

- **Donut chart** for Dropout Probability with traffic light color-coding:
  - **Red** for High Risk (>60%)
  - **Yellow** for Moderate Risk (30–60%)
  - **Green** for Low Risk (<30%)

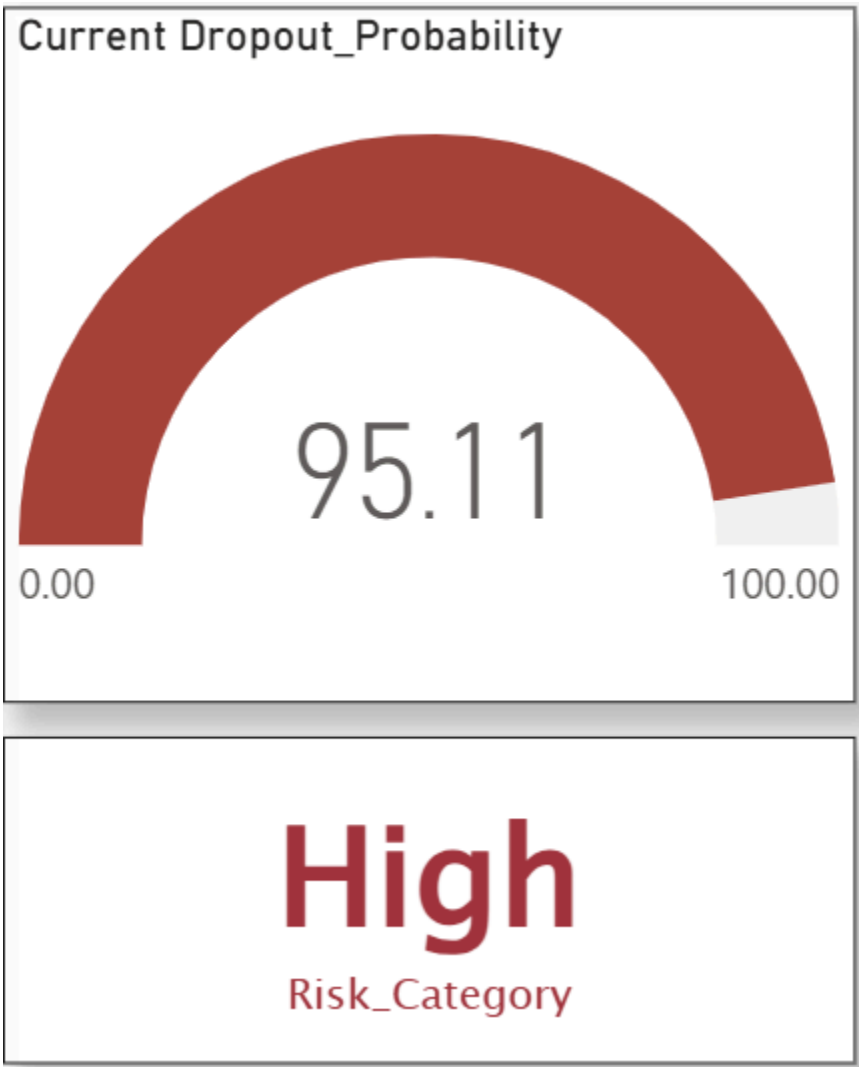


Figure 11.2.3.2 Dropout Risk Score

4. Risk Category Indicator

Highlights the student's risk level clearly:

- **Risk Category** is displayed in bold, with consistent color-coding from the overview dashboard.

5. Dropout Probability Trend

This scatter plot reflects **discrete intervention checkpoints**, where each point represents a student's predicted dropout probability at a specific moment in time.

Intervention Tracking:

- Each plotted point corresponds to a **post-intervention snapshot** (e.g., after academic advising, financial aid disbursement, or counseling sessions).
- Enables stakeholders to assess whether institutional actions are effectively reducing dropout risk.

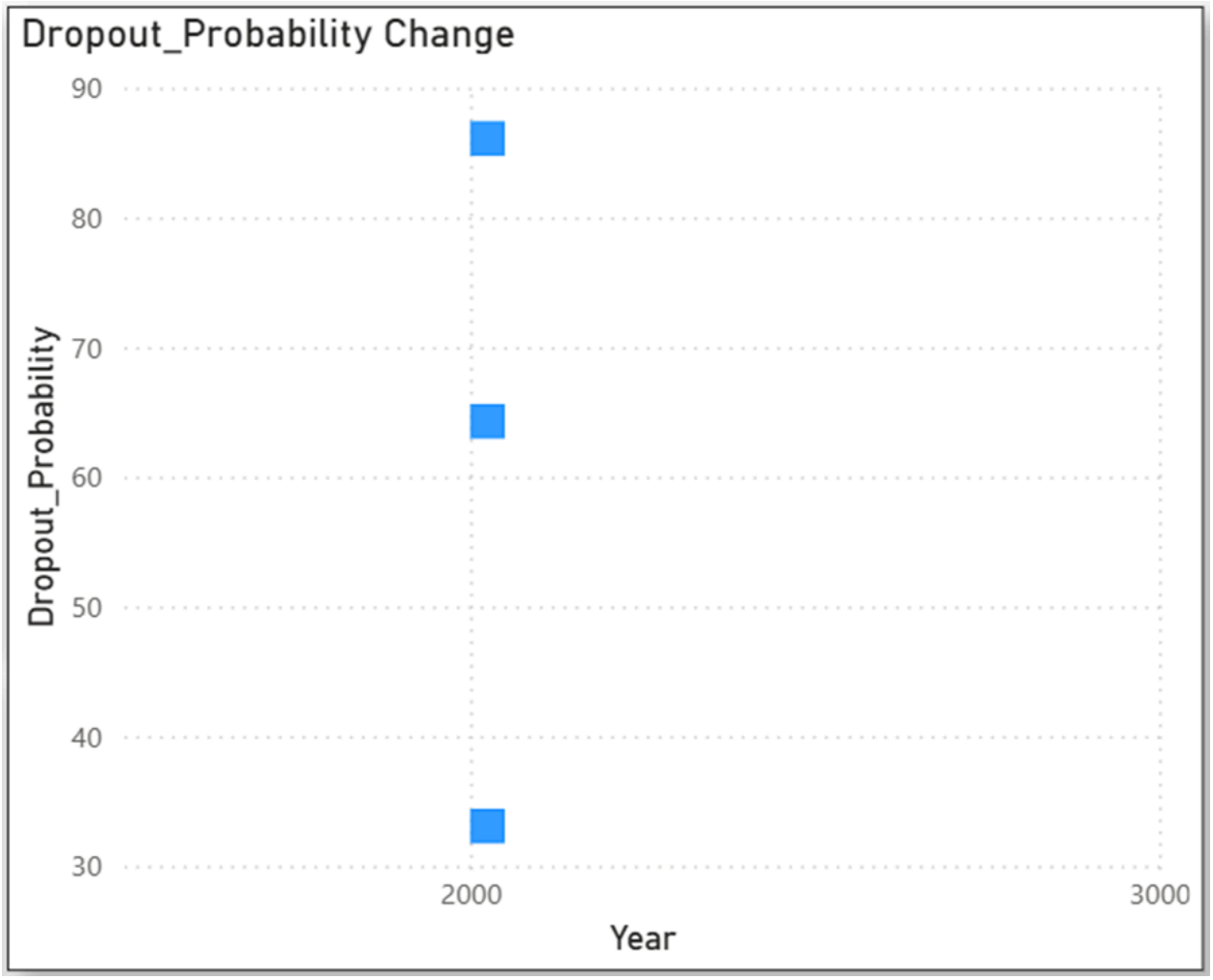


Figure 11.2.5.1 Dropout Probability Trend

12. Technology Architecture Diagram

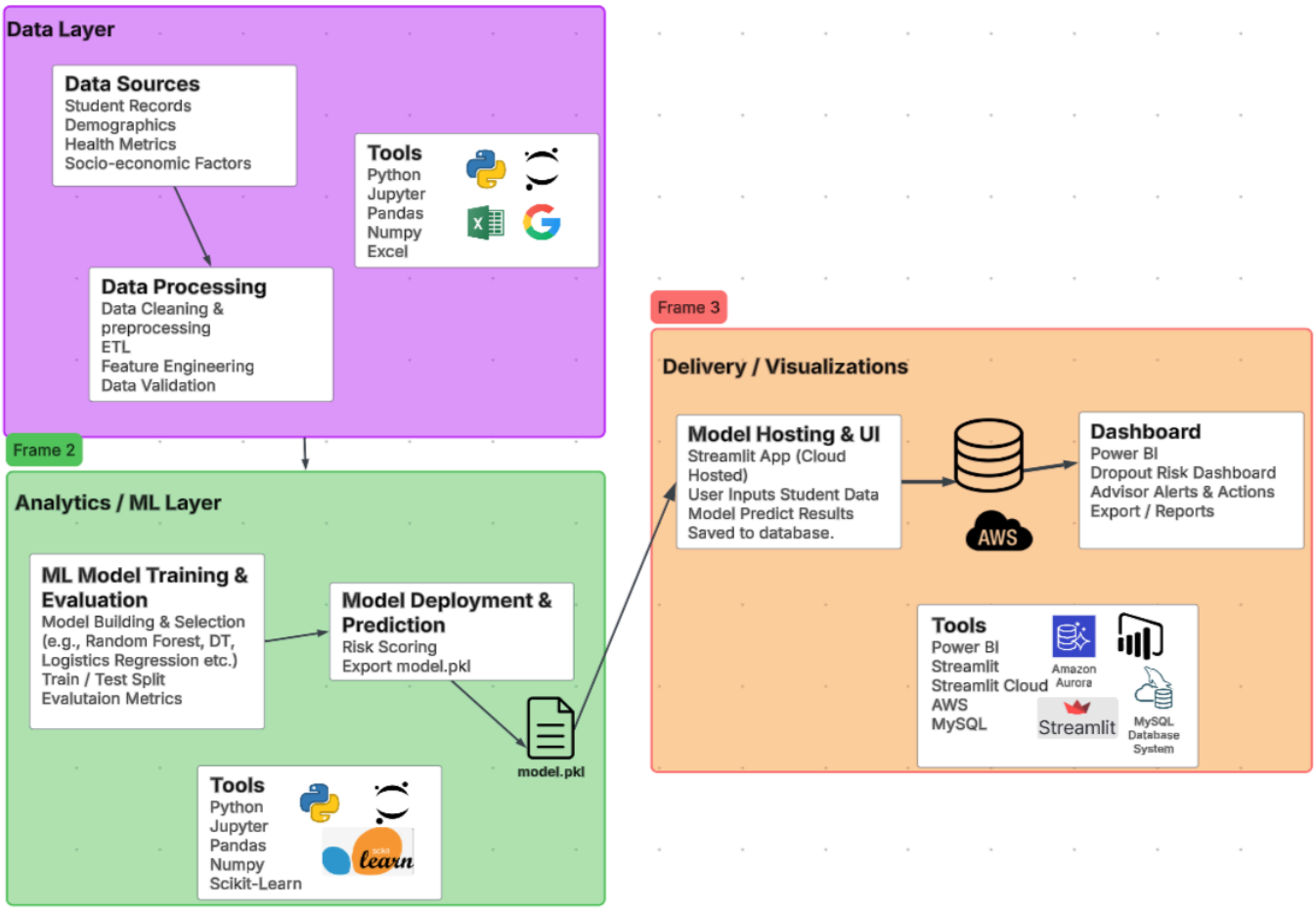


Figure 12: Technology_Architecture_Diagram

The Technology Architecture (figure 12) for the Student Dropout Prediction follows a three-layer structure aligned with the DVM stages. The **Data Layer** handles sourcing, preprocessing, and transformation of raw student data using tools such as Python, Pandas, and Excel.

The processed data feeds into the **Analytics Layer**, where the models are trained and evaluated to select the best-performing one, which is then exported for deployment.

The **Delivery Layer** provides user access to the model through a Streamlit web application, hosted using **Streamlit Cloud** and **Github**. University staff can input student data and receive real-time dropout risk predictions, which are stored in our database hosted in the AWS RDS environment. These predictions are dynamically visualized in **Power-BI** via an interactive Dropout Risk Dashboard, supporting academic staff in identifying at-risk students and guiding interventions.

13. To-Be Process

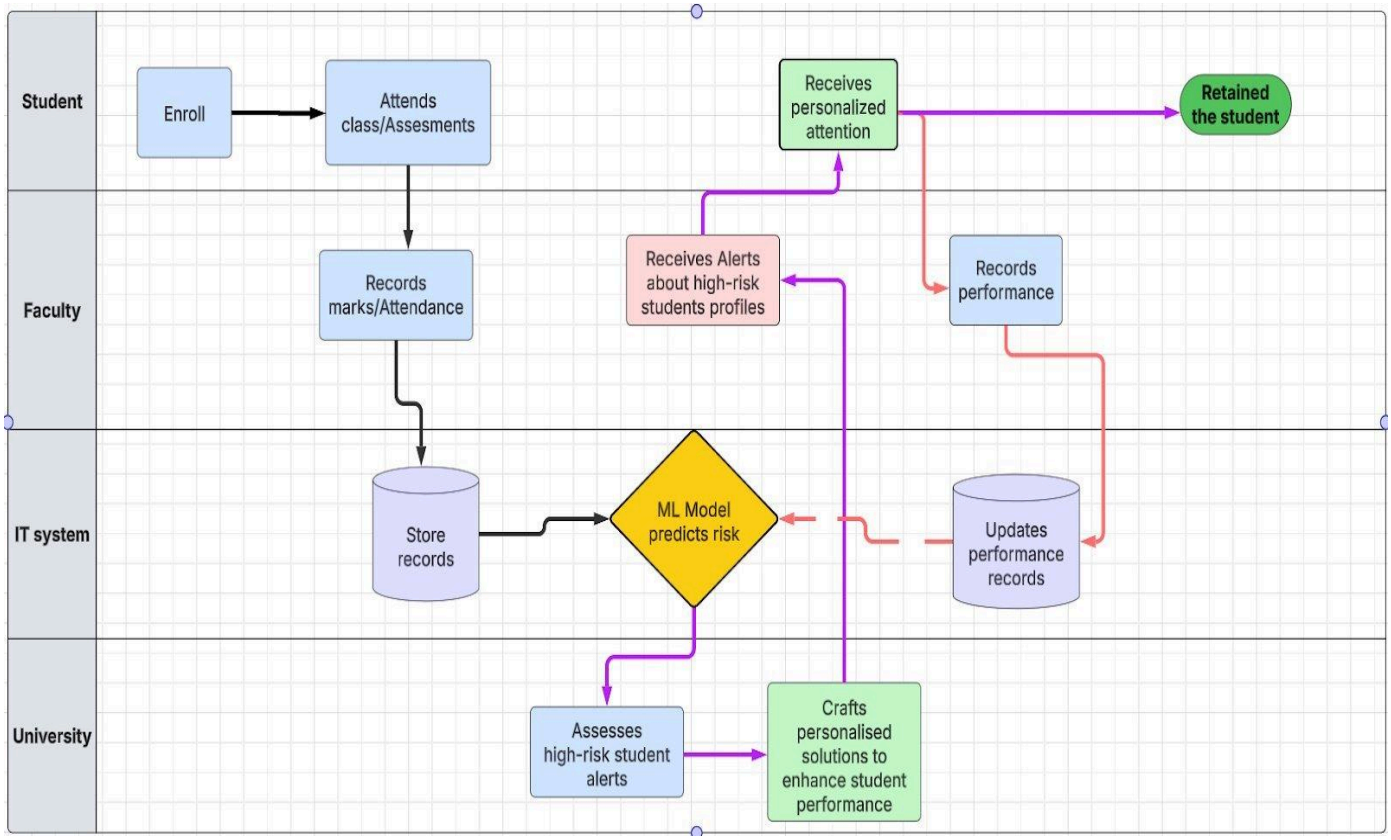


Figure 13.1: To-Be process diagram

In the enhanced student support process, students enroll and attend classes as usual, with faculty recording academic performance and attendance. These records are stored by the IT system and continuously fed into a machine learning model that predicts the likelihood of student dropout.

When a student is identified as high-risk, alerts are sent to faculty and university advisors. The university support team then reviews these alerts and designs personalized intervention strategies tailored to the student's specific needs.

The student receives targeted support such as academic counselling, mentoring, or learning resources which are monitored and documented by faculty. Updated performance data is looped back into the system, helping to retrain and improve the model over time.

This proactive, data-driven process enables earlier identification and support of struggling students, shifting the institution's approach from reactive to preventive, and ultimately improving student retention and academic outcomes.

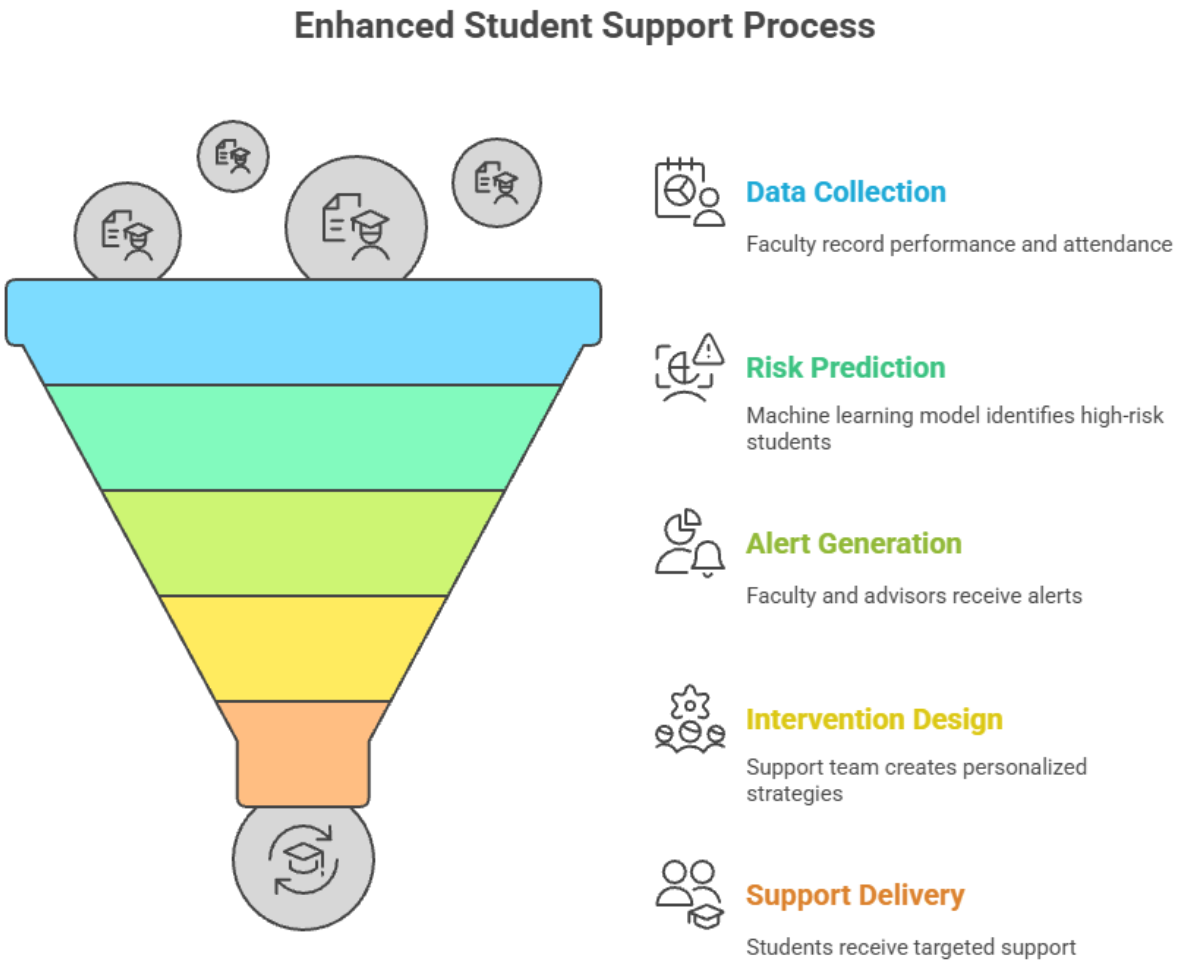


Figure 13.2: Illustration of the enhanced process after implementing the solution

14. Mapping Business Value Through the Mission Model Canvas

The mission of this artefact is to proactively reduce student dropout through predictive analytics. Our solution mobilizes partnerships across faculty, IT teams, advisors, and administrators. Activities span data engineering, ML development, dashboard design, and iterative testing.

The artefact delivers value by supporting earlier and more effective interventions, improving retention, enhancing student wellbeing, and reducing financial losses.

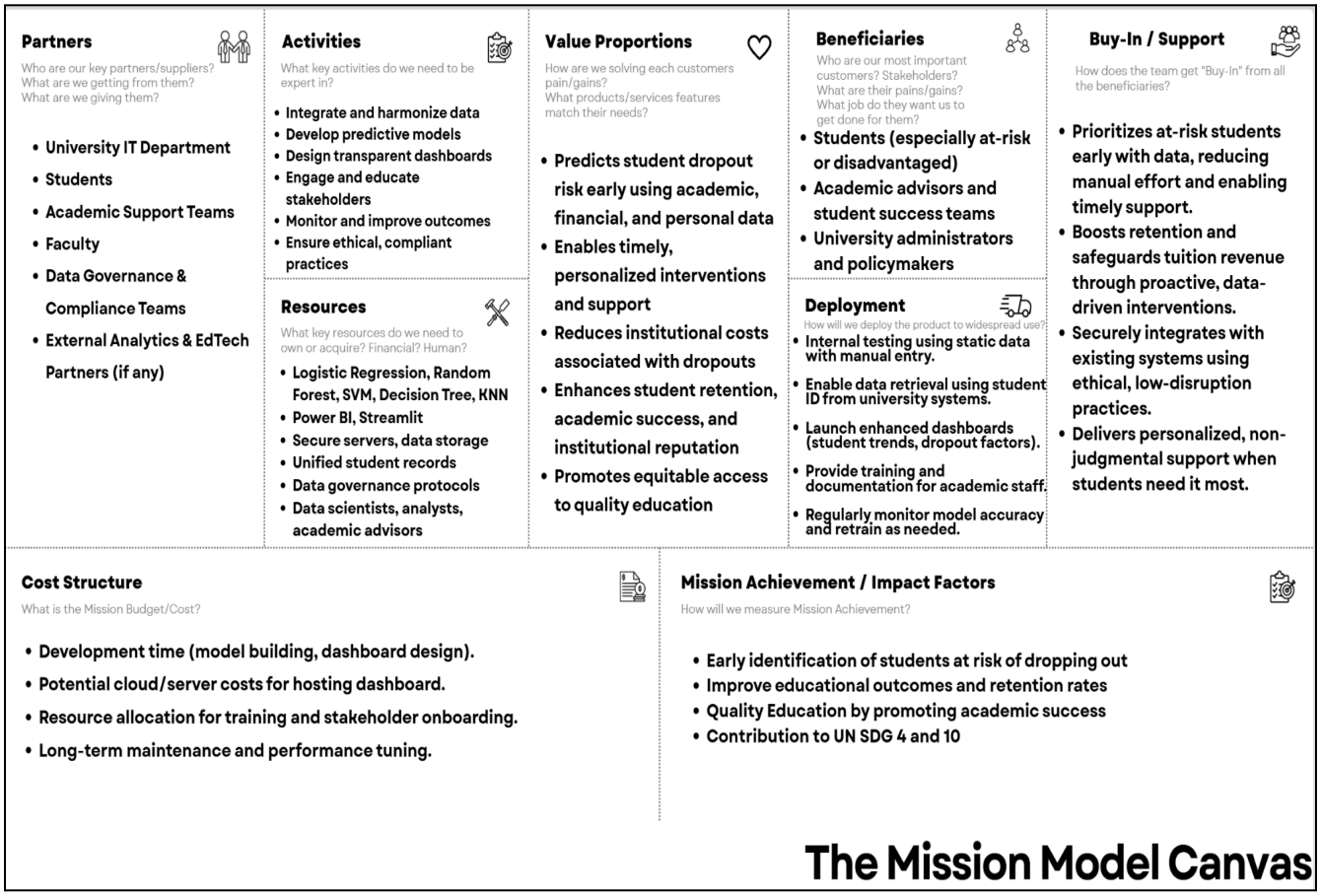


Figure 14.1: Mission Model Canvas

The value propositions of the system are multifaceted as it enables early detection of dropout risks, supports timely and personalized student interventions, reduces institutional costs related to late dropouts, and enhances both student retention and academic success. Furthermore, it promotes equitable access to quality education, especially for disadvantaged or marginalized students. These benefits directly translate to the system's key beneficiaries such as students, academic support staff, and university decision-makers.

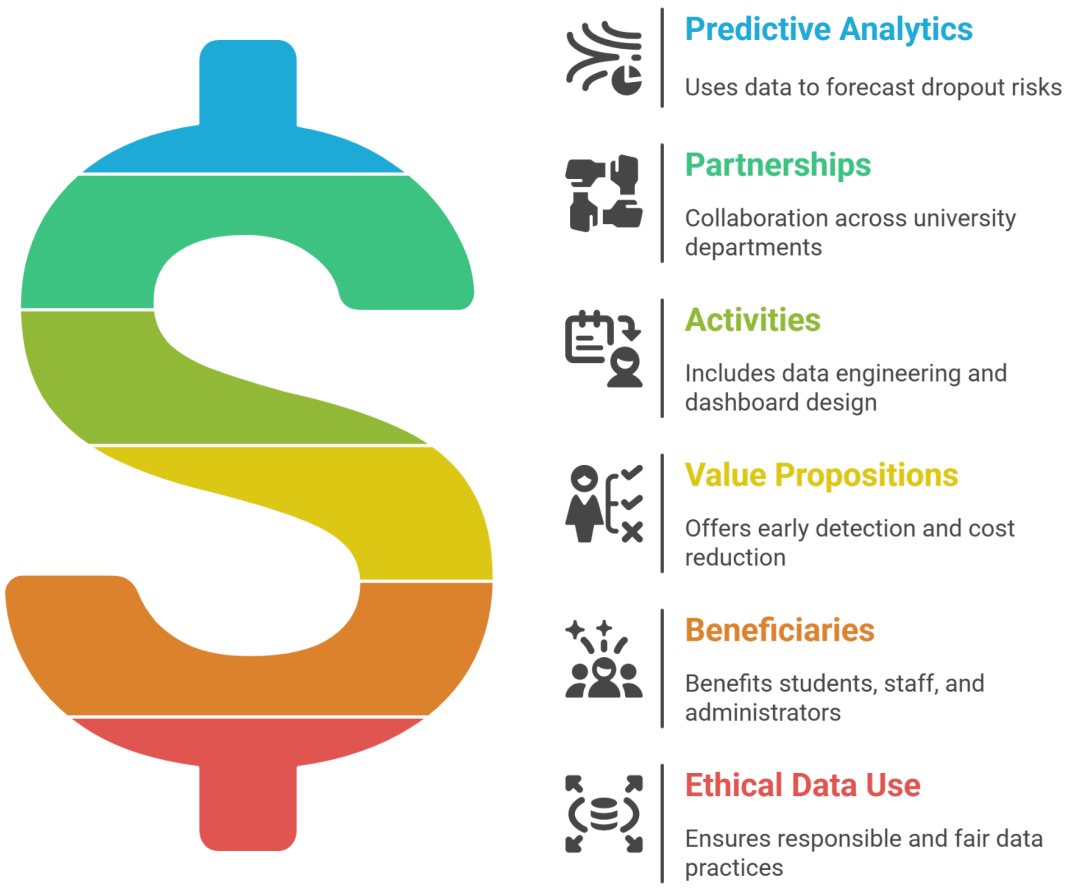


Figure 14.2: Enhancing Student Retention Through Data-Driven Solutions

For students, especially those at higher risk, this initiative delivers non-judgmental, data-informed support that can help them overcome challenges before they escalate. Academic advisors and student success teams benefit from enhanced visibility into student performance trends and actionable recommendations for intervention. For university administrators and policymakers, the system boosts institutional reputation and helps safeguard tuition revenue by proactively managing student outcomes through ethical, low-disruption data use.

This mission-driven solution not only addresses institutional efficiency but also upholds educational equity and student well-being, ensuring that those who need support the most receive it through data-driven, ethically governed practices. By aligning with institutional goals and SDGs, our artefact not only enhances student outcomes but also reinforces the university's strategic vision for equity and excellence in education.

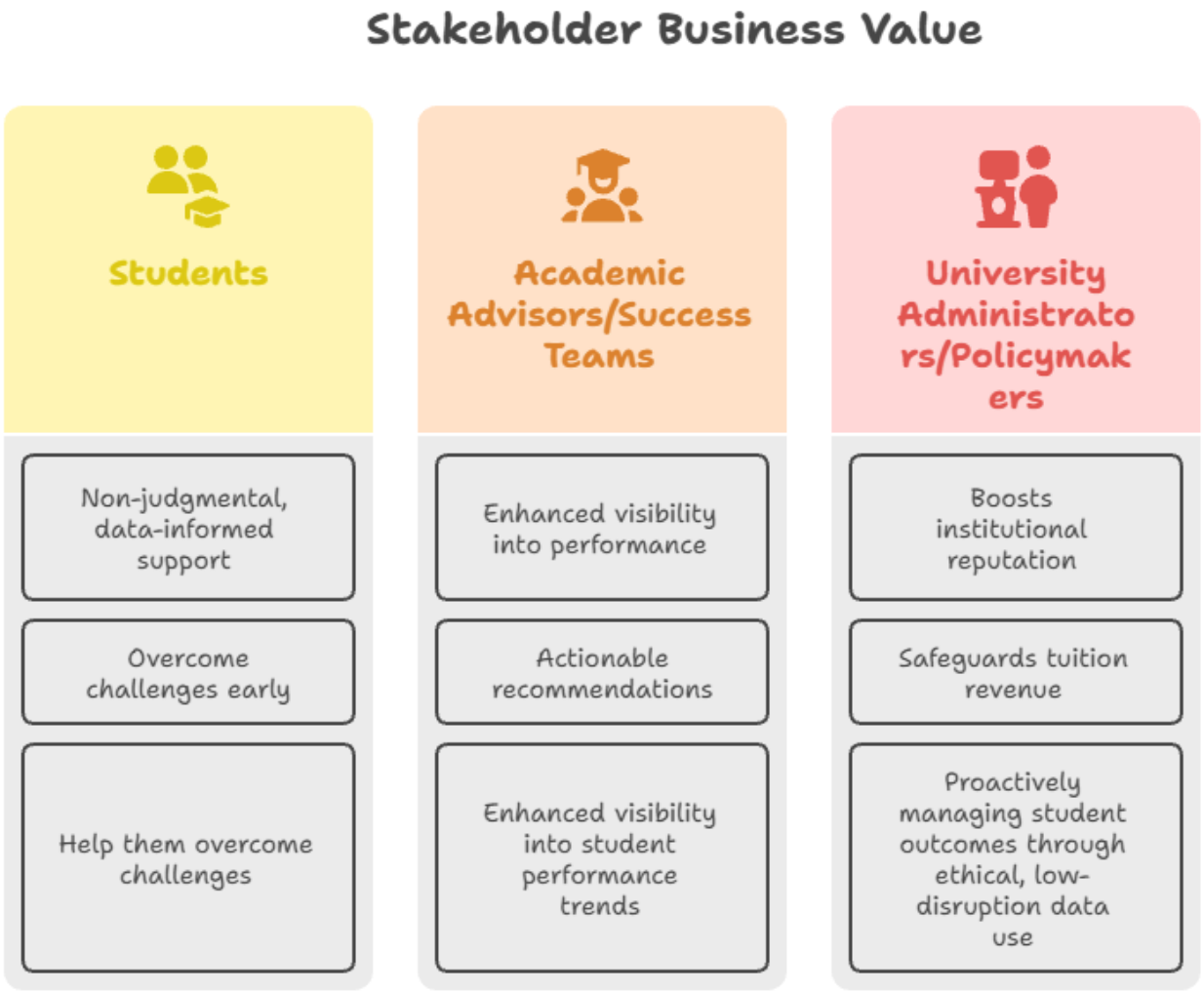


Figure 14.3: Illustration of business values to the core stakeholders

15. Team Collaboration & Development Process

Our project was driven by collaboration, adaptability, and shared vision, with all team members contributing cohesively. We adopted an agile, iterative approach, holding regular brainstorming sessions, design reviews, and model evaluations to ensure inclusive decision-making.

Key aspects of our teamwork:

- Flexible Roles: Tasks like data selection, model refinement, and dashboard design were tackled collectively, leveraging diverse expertise.

- Feedback Integration: Continuous input from advisors and peers shaped improvements across Power BI visuals, classification models, and reporting.
- Unified Output: The final artefact and report reflect harmonized language and narrative flow, blending academic rigor with practical teamwork.
- Used tools like Trello, Github, Google Collab, Google Docs and MS Teams for effective collaboration.

This experience strengthened both our technical skills and collaborative problem-solving for real-world analytics challenges.

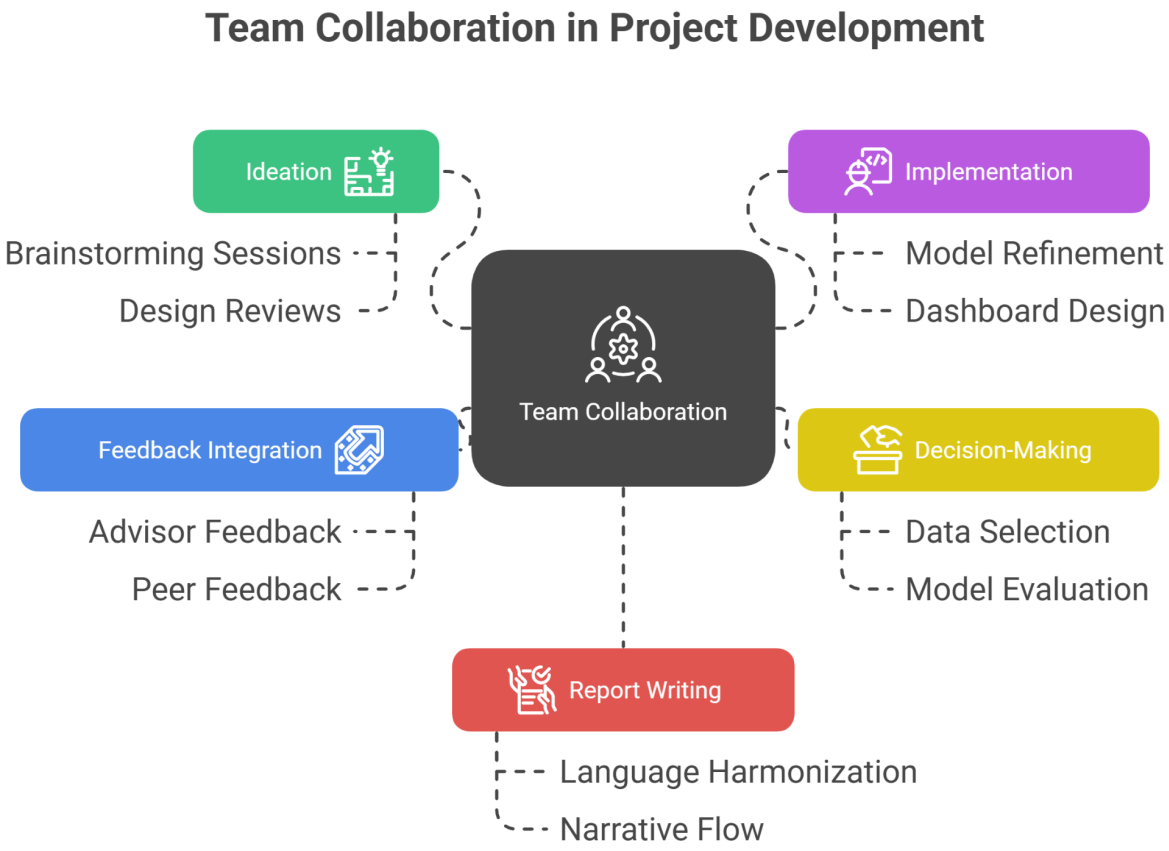


Figure 15: Team Collaboration and Learnings

References

Ali, M. (2024, October). *Synthetic data generation: A hands-on guide in python*. Datacamp.com; DataCamp. https://www.datacamp.com/tutorial/synthetic-data-generation

Atlas. (2024, September). *The direct and indirect costs of high college attrition rates*. AtlasRTX. https://atlasrtx.com/the-direct-and-indirect-costs-of-high-college-attrition-rates/

Cornett, A., Fletcher, C., & Ashton, B. (2024). Student financial wellness survey student financial wellness survey fall 2023 semester results. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4896870

Delgado, K., Origgi, J., Hasanpoor, T., Yu, H., Allessio, D., Arroyo, I., Lee, W., Betke, M., Woolf, B., & Bargal, S. (n.d.). *Student engagement dataset*. Retrieved July 2025, from https://openaccess.thecvf.com/content/ICCV2021W/ABAW/papers/Delgado_Student_Engagement_Dataset_ICCVW_2021_paper.pdf

Explore our data. (n.d.). Healthy Minds Network. https://healthymindsnetwork.org/research/data-for-researchers

Hanson, M. (2024, August). *College dropout rates*. Education Data Initiative. https://educationdata.org/college-dropout-rates

Kassim, M., ZH. Azizul, & AAH. Ahmad Fuaad. (2025). Student engagement dataset (SED): An online learning activity dataset. *IEEE Access*, *PP*, 1. https://doi.org/10.1109/ACCESS.2025.3531102

Kemper, L., Vorhoff, G., & Wigger, B. U. (2020). Predicting student dropout: A machine learning approach. *European Journal of Higher Education*, *10*, 28–47. https://doi.org/10.1080/21568235.2020.1718520

Medhat, M. (2024). *Unveiling the comprehensive student scores dataset: Insights into academic performance and student behavior | kaggle*. Kaggle.com. https://www.kaggle.com/discussions/accomplishments/509324

NCES. (2017). *Indicator 1: Population Distribution*. Ed.gov; NCES. https://nces.ed.gov/programs/raceindicators/indicator_raa.asp

NCES. (2022, May). *COE - Undergraduate Retention and Graduation Rates*. Nces.ed.gov; NCES. https://nces.ed.gov/programs/coe/indicator/ctr/undergrad-retention-graduation

Raisman, N. (2013). *The Cost of College Attrition at Four-Year Colleges & Universities Educational Policy Institute*. Education Policy Institute. https://files.eric.ed.gov/fulltext/ED562625.pdf

Student wellbeing statistics statistics. (n.d.). Inspire Student Hub. https://www.inspiresupporthub.org/students/student-wellbeing/

Vardishvili, O. (2020). The Macroeconomic Cost of College Dropouts. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3755800

Whistle, W. (2019a, January). *Ripple effect: The cost of the college dropout rate – third way*. Www.thirdway.org. https://www.thirdway.org/report/ripple-effect-the-cost-of-the-college-dropout-rate

Whistle, W. (2019b, January 28). *Ripple Effect: The Cost of the College Dropout Rate – Third Way*. Www.thirdway.org. https://www.thirdway.org/report/ripple-effect-the-cost-of-the-college-dropout-rate

Wikipedia. (2023, July 2). *National Survey of Student Engagement*. Wikipedia. https://en.wikipedia.org/wiki/National_Survey_of_Student_Engagement

Wine, J., Siegel, P., Stollberg, R., & Hunt-White, T. (2018). https://nces.ed.gov/pubs2018/2018482.pdf

World Health Organization. (2025). *World health organization*. Who.int; World Health Organization. https://www.who.int/