MSc Project (CMP060L050H)

Machine Learning-Based Student Behaviour Analysis and Prediction on Online Learning Platforms

Student Name: Geng Rui

Supervisor: SURYANTO

Second Marker: Dr Chandrashekhar Kumbhar

Date: 2025.11.12

Outline

01 Context & Motivation (WHY)

02 Problem & Objectives (WHAT)

03 Literature Snapshot & Gap

04 Data, Ethics & Scope

05 End-to-End Architecture (HOW)

06 Modules & Responsibilities

07 Dashboard (2 Page)

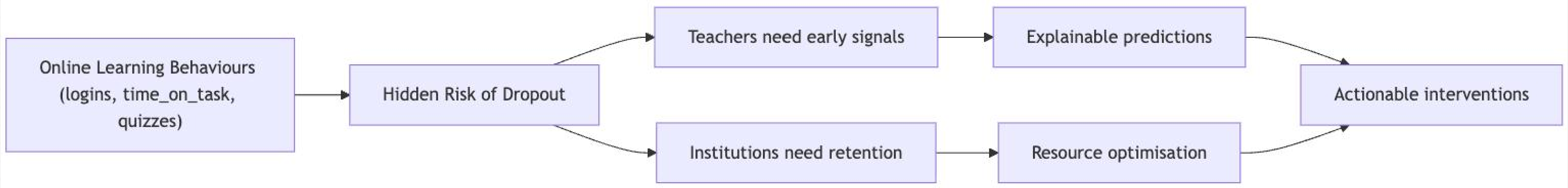
08 Evaluation Design

09 Evaluation Results (2 Page)

10 Project Management & Actual Timeline

11 Contributions & Conclusion

Context & Motivation (WHY)



Online Learning Behaviours (logins, time_on_task, quizzes)

Hidden Risk of Dropout

Teachers need early signals

Explainable predictions

Institutions need retention

Resource optimisation

Actionable interventions

江西应用工程职业学院
JIANGXI APPLIED ENGINEERING VOCATIONAL INSTITUTE
姓名：耿 锐
岗位：教 师
江西应用工程职业学院 | 100周年
Jiangxi Applied Engineering Vocational Institute
The 100th Anniversary

- Vocational learners show sparse, irregular engagement; tools must be interpretable.
- Teachers need actionable & timely risk signals, not opaque scores.
- False Negative cost > False Positive cost in classrooms → recall-aware design.

Problem & Objectives (WHAT)

Problem: off-the-shelf models transfer poorly; opaque predictions are hard to act on.

Objective 1 Interpretable end-to-end pipeline (DB → dashboard).

Objective 2 Transparent evaluation (Accuracy, Precision, Recall, F1, AUC) on held-out test.

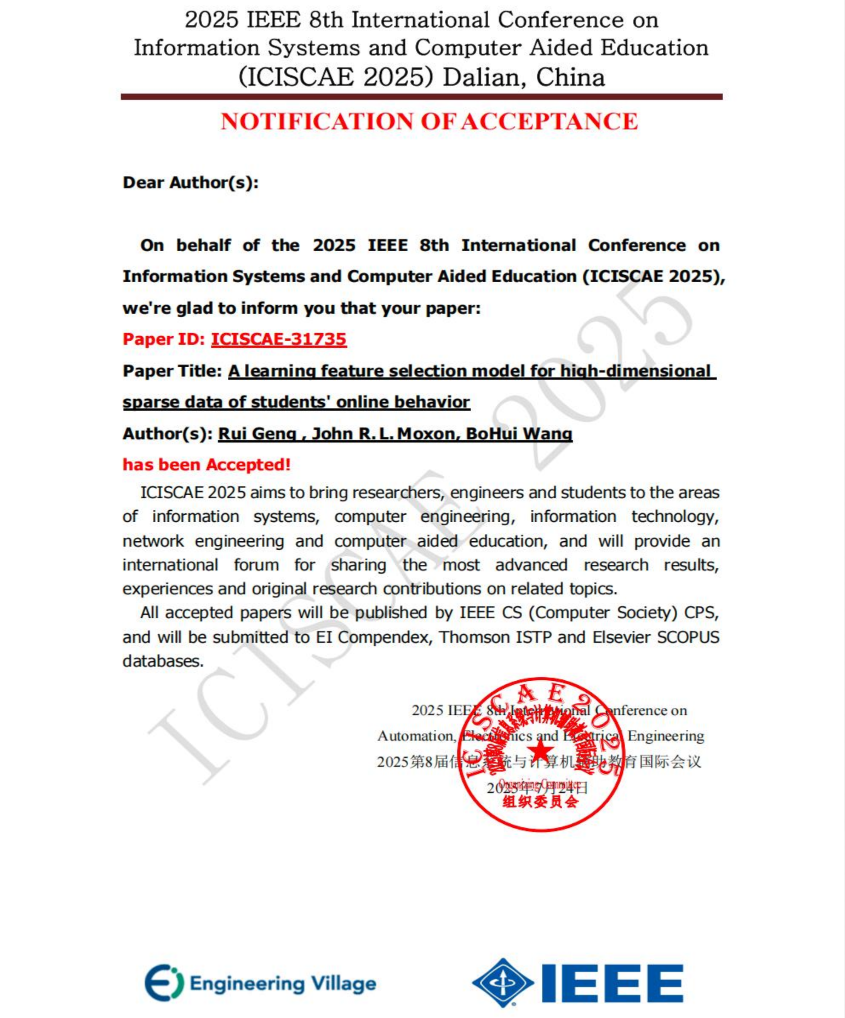
Objective 3 Minimal UI — high-risk table, a few visuals, single-student form.

Literature Snapshot & Gap

Transparent (LR/DT) vs high-capacity (RF/GB/Deep): accuracy vs readability.

For vocational cohorts: start with an interpretable baseline; extend complexity later.

Earlier short paper (ICISCAE 2025) guided feature choices and readability stance.



2025 IEEE 8th International Conference on
Information Systems and Computer Aided Education
(ICISCAE 2025) Dalian, China

NOTIFICATION OF ACCEPTANCE

Dear Author(s):

On behalf of the 2025 IEEE 8th International Conference on Information Systems and Computer Aided Education (ICISCAE 2025), we're glad to inform you that your paper:

Paper ID: ICISCAE-31735

Paper Title: A learning feature selection model for high-dimensional sparse data of students' online behavior

Author(s): Rui Geng , John R. L. Moxon, BoHui Wang

has been Accepted!

ICISCAE 2025 aims to bring researchers, engineers and students to the areas of information systems, computer engineering, information technology, network engineering and computer aided education, and will provide an international forum for sharing the most advanced research results, experiences and original research contributions on related topics.

All accepted papers will be published by IEEE CS (Computer Society) CPS, and will be submitted to EI Compendex, Thomson ISTP and Elsevier SCOPUS databases.

2025 IEEE 8th International Conference on
Automation, Electronics and Electrical Engineering
2025第8届信息系统与计算机辅助教育国际会议
2025年7月24日

ICISCAE 2025
Organizing Committee
组织委员会

Engineering Village IEEE

Data, Ethics & Scope

Data & scope:

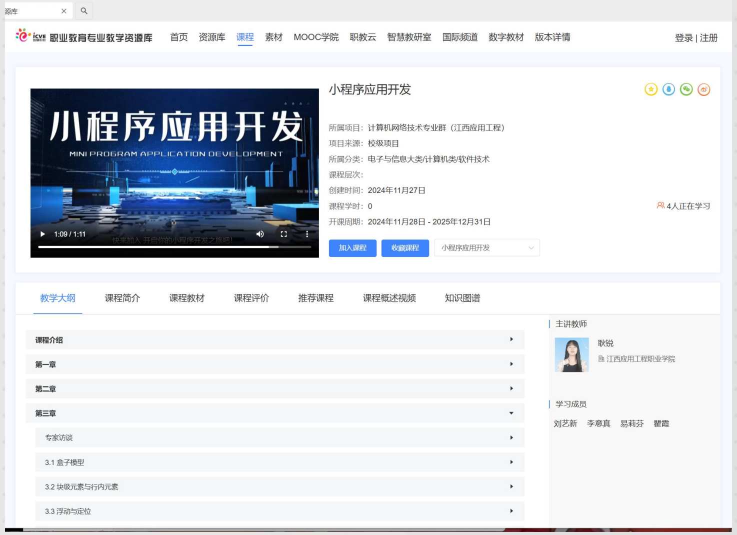
Synthetic behavioural data; minimal features (logins, study time (hours), quiz attempts); no real ICVE distributions; generated via parametric draws within realistic bounds.

Governance:

Anonymisation-by-design; training artefacts (trained_model.pkl, scaler.pkl) kept outside the UI layer.

Protocol:

Stratified 80/20 split (seed=42); StandardScaler fit on train only; label = completion-rate threshold with median fallback (≈50/50 balance).



小程序应用开发

所属项目：计算机网络技术专业群（江西应用工程）

项目来源：校级项目

所属分类：电子与信息大类/计算机类/软件技术

课程层次：

创建时间：2024年11月27日

课程学时：0

开课周期：2024年11月28日 - 2025年12月31日

4人正在学习

课程介绍

第一章

第二章

第三章

专家访谈

3.1 盒子模型

3.2 块级元素与行内元素

3.3 浮动与定位

主讲教师

耿锐

江西应用工程职业学院

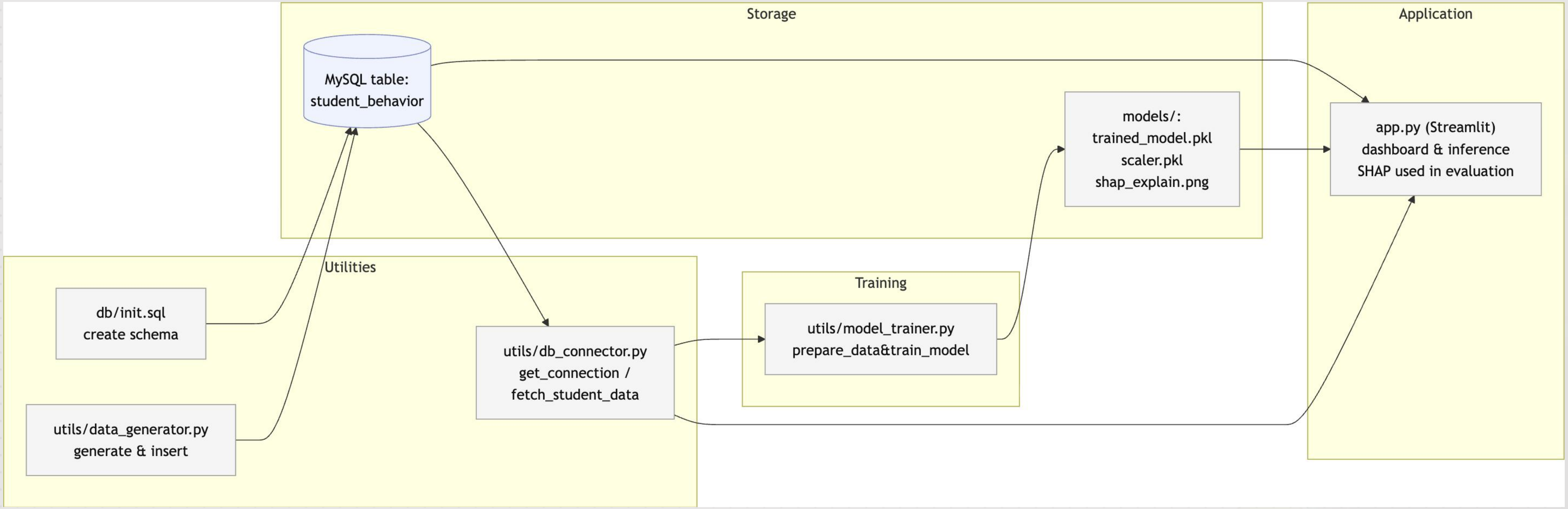
学习成员

刘艺新 李意真 易莉芬 瞿霞

Quality & next:

Realism checks (descriptives + correlation) show activity ↑ → High Risk ↓ (expected); see “Evaluation Results – Synthetic Data Realism”; Post-MSc validation, calibration & A/B trials.

End-to-End Architecture (HOW)



- DB init → feature prep → offline train & select → persist artefacts → explain → UI render.
- 80/20 stratified split (seed=42); scaler fit on train only; LR/RF/GB trained.
- Persisted artefacts (scaler.pkl, trained_model.pkl) decouple training from UI (no retrain).

Modules & Responsibilities

db_connector.py

- connection handling; fetch_student_data()

data_generator.py

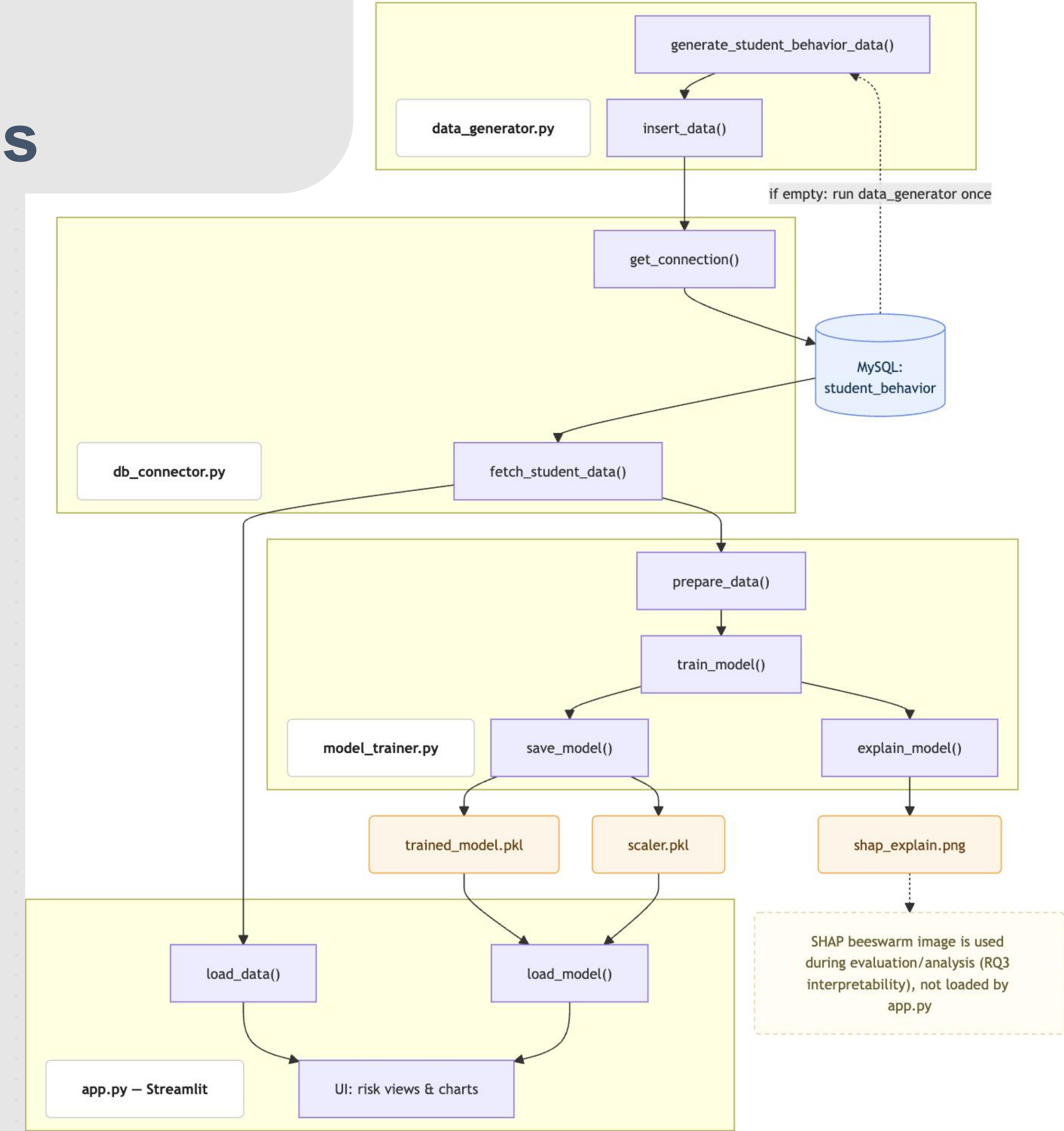
- reproducible cohort synthesis (fixed seed)

model_trainer.py

- prepare features; train LR/RF/GB; pick F1-best; save artefacts

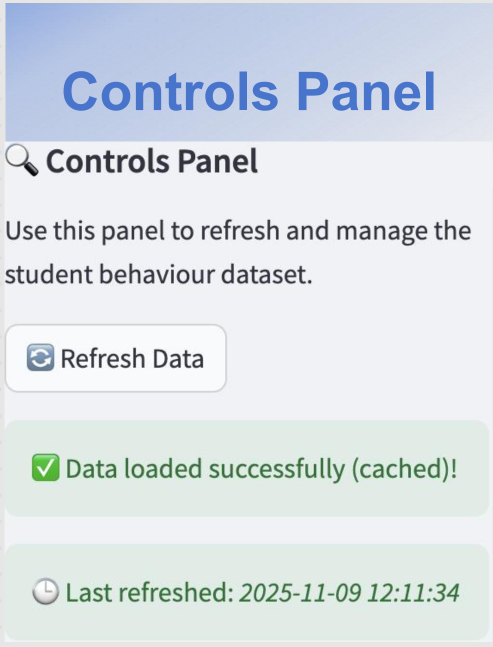
app.py

- cached loaders; cohort views; single-student form; visuals



SHAP beeswarm image is used during evaluation/analysis (RQ3 interpretability), not loaded by app.py

Dashboard - Student Risk & Visualisation



Controls Panel

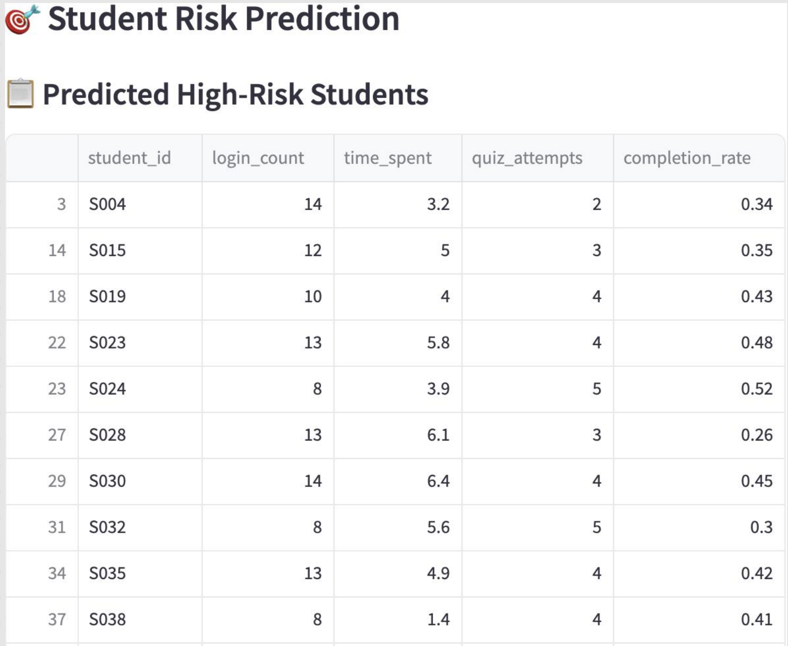
🔍 Controls Panel

Use this panel to refresh and manage the student behaviour dataset.

🔄 Refresh Data

✅ Data loaded successfully (cached)!

🕒 Last refreshed: *2025-11-09 12:11:34*

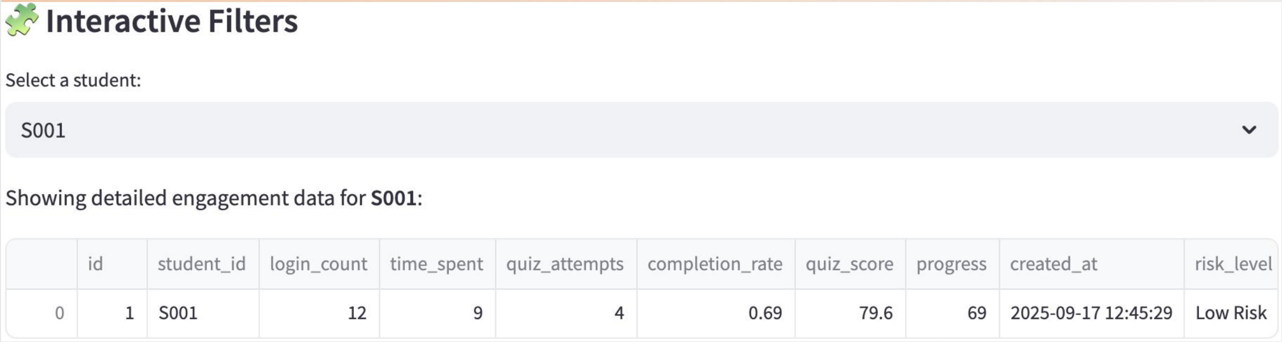


🎯 Student Risk Prediction

📋 Predicted High-Risk Students

	student_id	login_count	time_spent	quiz_attempts	completion_rate
3	S004	14	3.2	2	0.34
14	S015	12	5	3	0.35
18	S019	10	4	4	0.43
22	S023	13	5.8	4	0.48
23	S024	8	3.9	5	0.52
27	S028	13	6.1	3	0.26
29	S030	14	6.4	4	0.45
31	S032	8	5.6	5	0.3
34	S035	13	4.9	4	0.42
37	S038	8	1.4	4	0.41

High-risk table(↑) + quick filters(↓) support weekly teacher review.



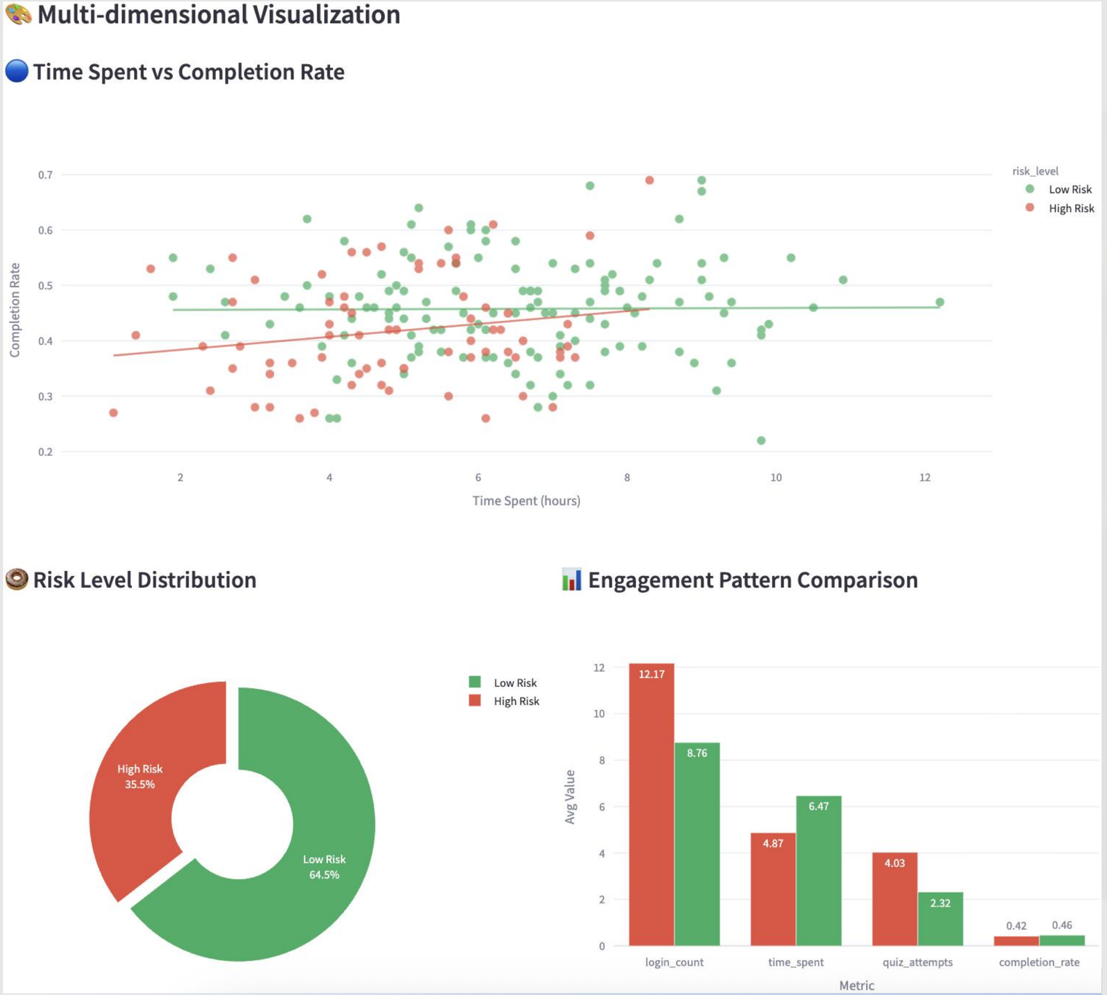
🧩 Interactive Filters

Select a student:

S001

Showing detailed engagement data for **S001**:

	id	student_id	login_count	time_spent	quiz_attempts	completion_rate	quiz_score	progress	created_at	risk_level
0	1	S001	12	9	4	0.69	79.6	69	2025-09-17 12:45:29	Low Risk



🎨 Multi-dimensional Visualization

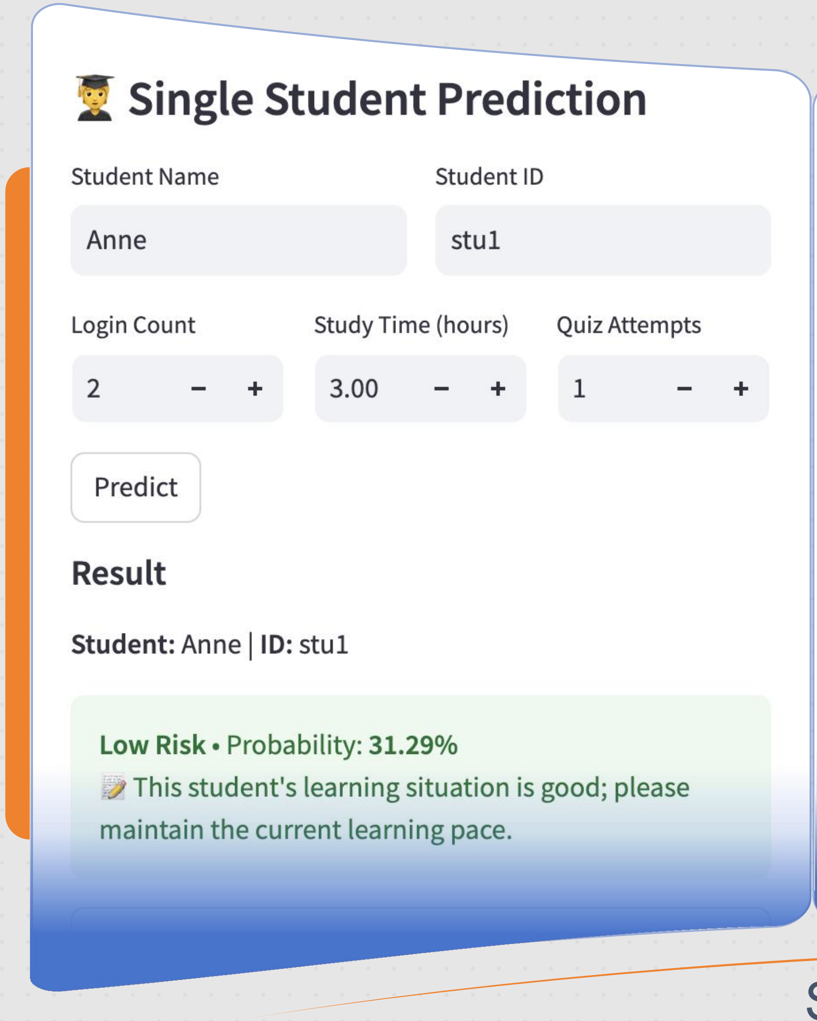
🔵 Time Spent vs Completion Rate

🍩 Risk Level Distribution

📊 Engagement Pattern Comparison

Visuals explain cohort structure: scatter, donut, grouped bars.

Dashboard - Single Student Prediction



🧑‍🎓 Single Student Prediction

Student Name: Anne

Student ID: stu1

Login Count: 2

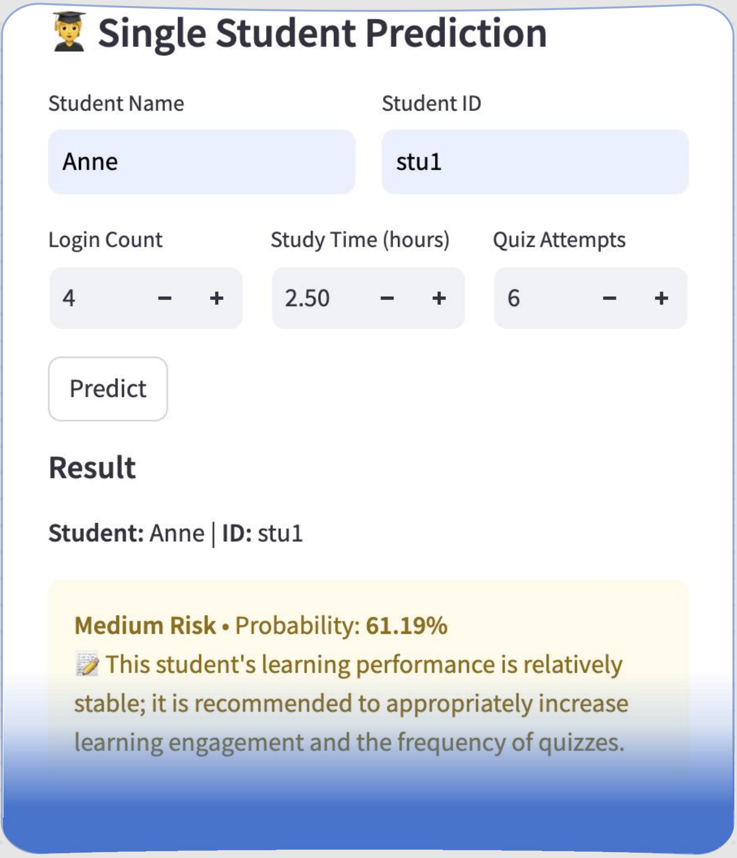
Study Time (hours): 3.00

Quiz Attempts: 1

Result

Student: Anne | **ID:** stu1

Low Risk • Probability: **31.29%**
📝 This student's learning situation is good; please maintain the current learning pace.



🧑‍🎓 Single Student Prediction

Student Name: Anne

Student ID: stu1

Login Count: 4

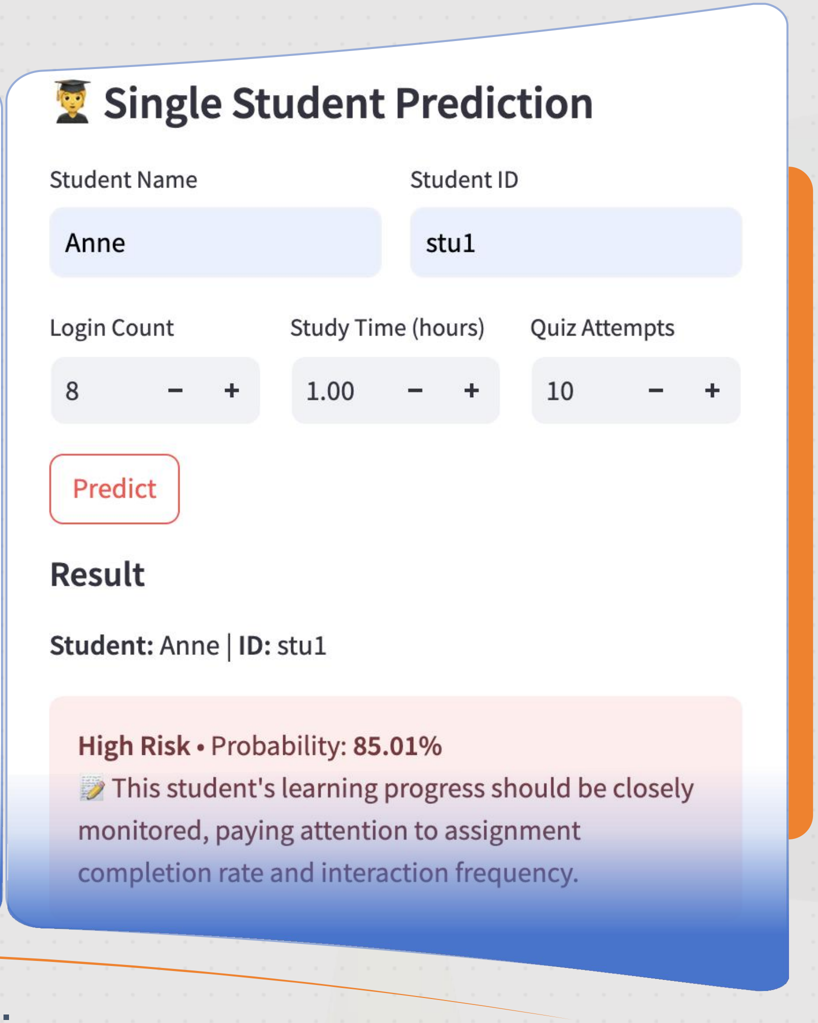
Study Time (hours): 2.50

Quiz Attempts: 6

Result

Student: Anne | **ID:** stu1

Medium Risk • Probability: **61.19%**
📝 This student's learning performance is relatively stable; it is recommended to appropriately increase learning engagement and the frequency of quizzes.



🧑‍🎓 Single Student Prediction

Student Name: Anne

Student ID: stu1

Login Count: 8

Study Time (hours): 1.00

Quiz Attempts: 10

Result

Student: Anne | **ID:** stu1

High Risk • Probability: **85.01%**
📝 This student's learning progress should be closely monitored, paying attention to assignment completion rate and interaction frequency.

Select a student or enter inputs.
Click **Predict** → show label and P(High Risk).
High ≥ 0.66; Medium 0.33–0.6599; Low < 0.33.

Evaluation Design

Split

Stratified 80/20 held-out test (seed = 42)

Model Selection

LR/RF/GB tested; Logistic Regression chosen (F1-best, held-out)

Outputs

Metrics (Accuracy / Precision / Recall / F1 / AUC)

Confusion Matrix (positive class = High Risk)

Classification Report

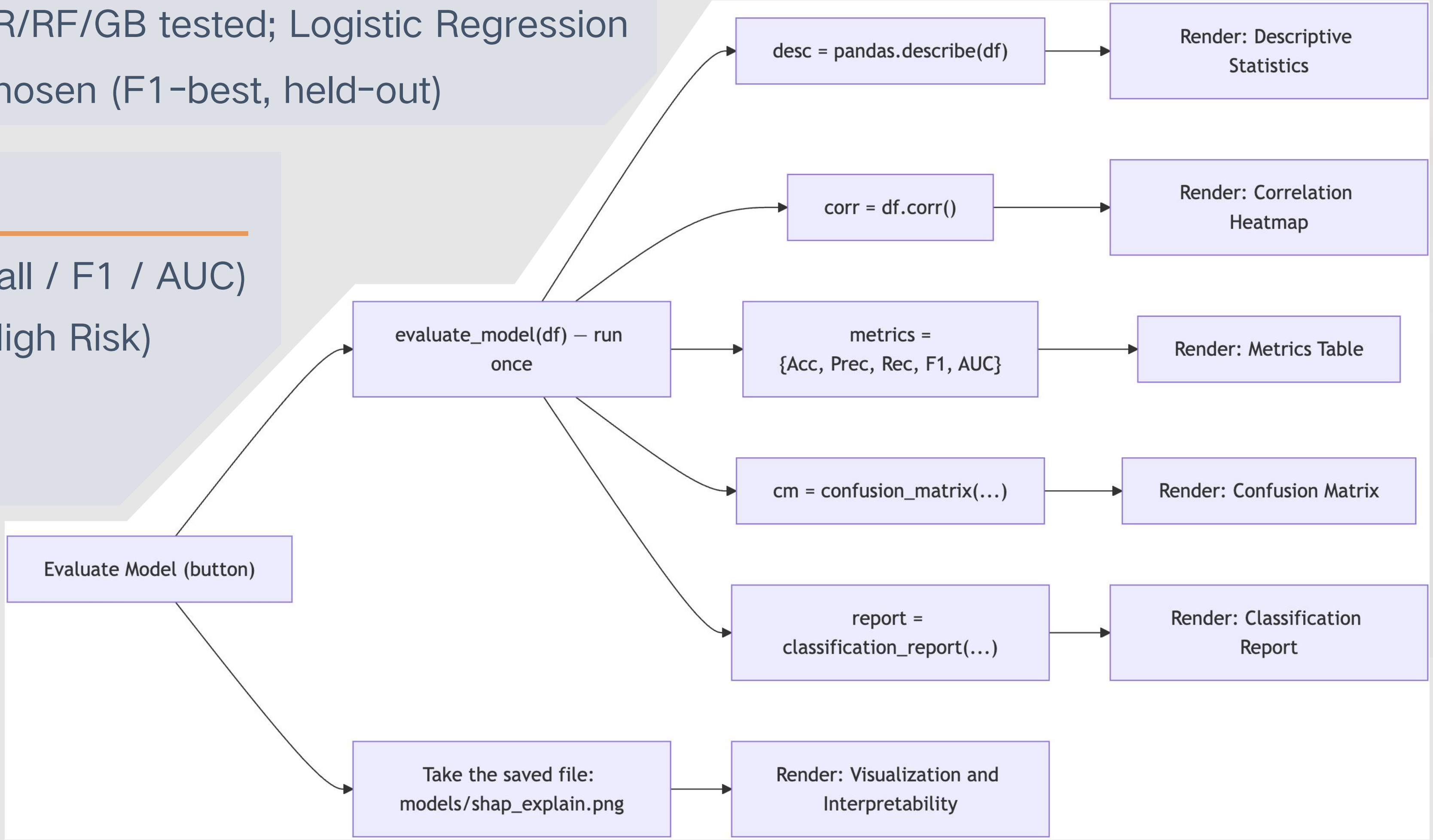
SHAP Beeswarm

Protocol mirrors training

Scaler fit on train only

Identical preprocessing on test

No leakage → Fair comparison



Evaluate Model (button)

evaluate_model(df) — run once

desc = pandas.describe(df)

Render: Descriptive Statistics

corr = df.corr()

Render: Correlation Heatmap

metrics = {Acc, Prec, Rec, F1, AUC}

Render: Metrics Table

cm = confusion_matrix(...)

Render: Confusion Matrix

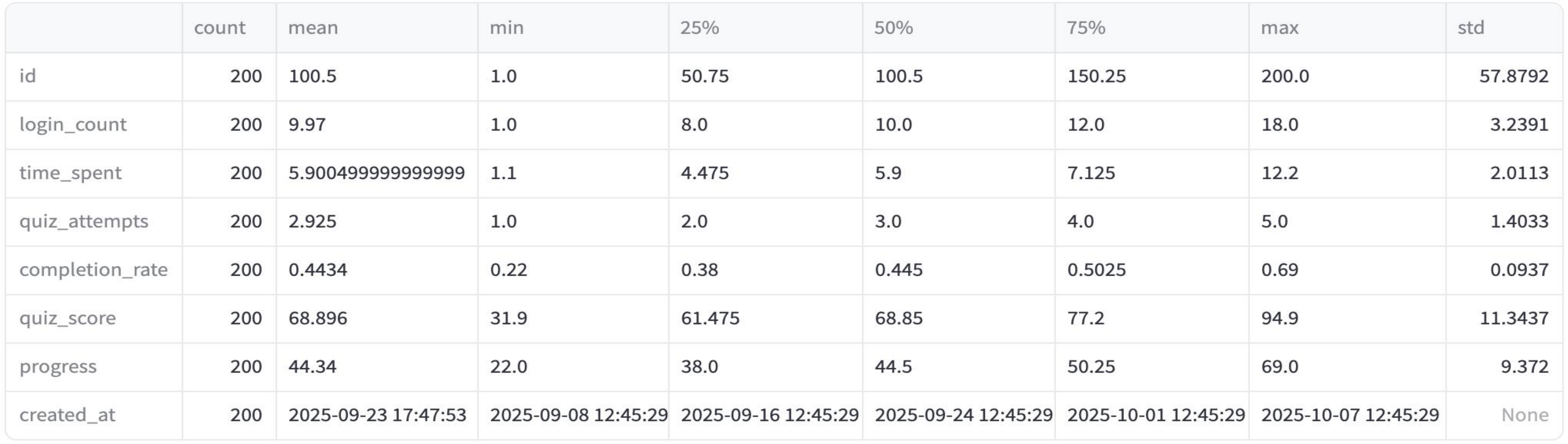
report = classification_report(...)

Render: Classification Report

Take the saved file: models/shap_explain.png

Render: Visualization and Interpretability

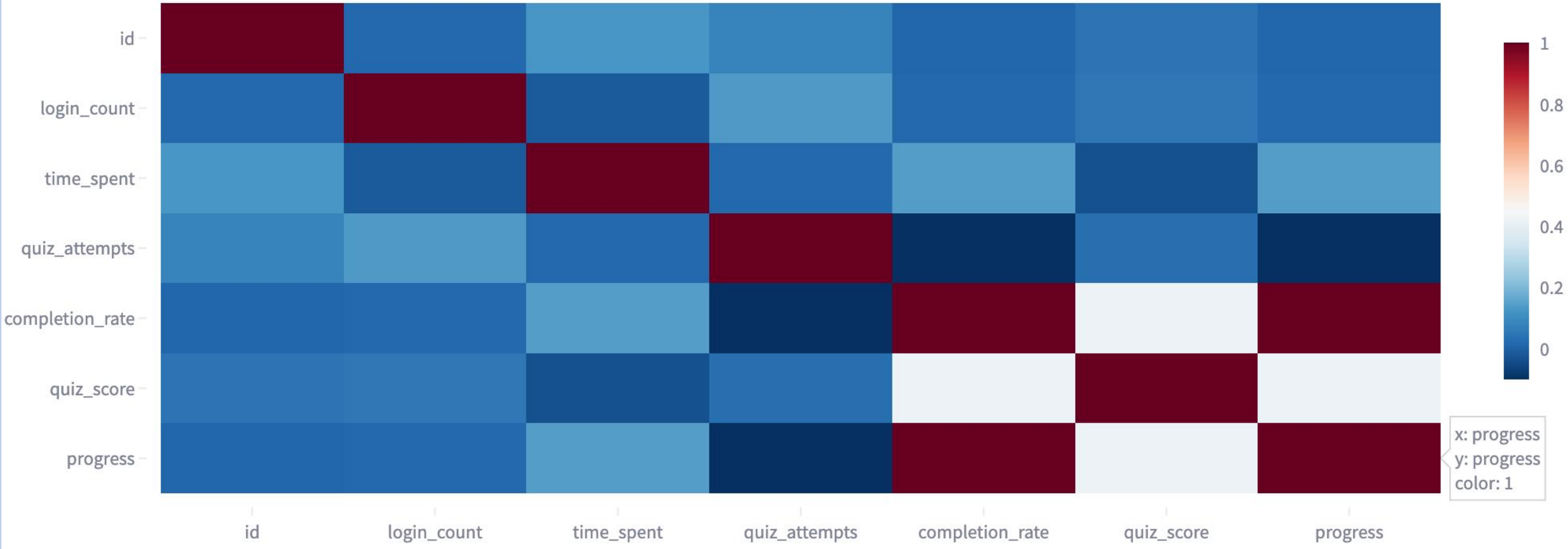
Evaluation Results - Synthetic Data Realism



	count	mean	min	25%	50%	75%	max	std
id	200	100.5	1.0	50.75	100.5	150.25	200.0	57.8792
login_count	200	9.97	1.0	8.0	10.0	12.0	18.0	3.2391
time_spent	200	5.900499999999999	1.1	4.475	5.9	7.125	12.2	2.0113
quiz_attempts	200	2.925	1.0	2.0	3.0	4.0	5.0	1.4033
completion_rate	200	0.4434	0.22	0.38	0.445	0.5025	0.69	0.0937
quiz_score	200	68.896	31.9	61.475	68.85	77.2	94.9	11.3437
progress	200	44.34	22.0	38.0	44.5	50.25	69.0	9.372
created_at	200	2025-09-23 17:47:53	2025-09-08 12:45:29	2025-09-16 12:45:29	2025-09-24 12:45:29	2025-10-01 12:45:29	2025-10-07 12:45:29	None

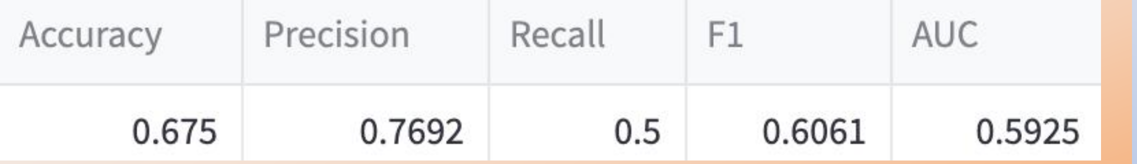
Descriptive Statistics

Correlation Heatmap



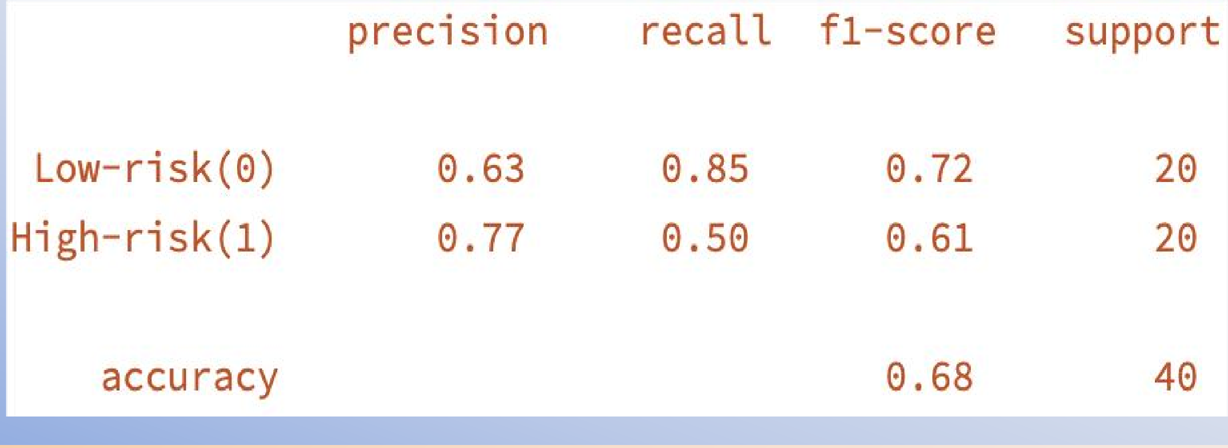
Evaluation Results - Performance and Interpretability

Model Metrics (Positive class = High Risk)



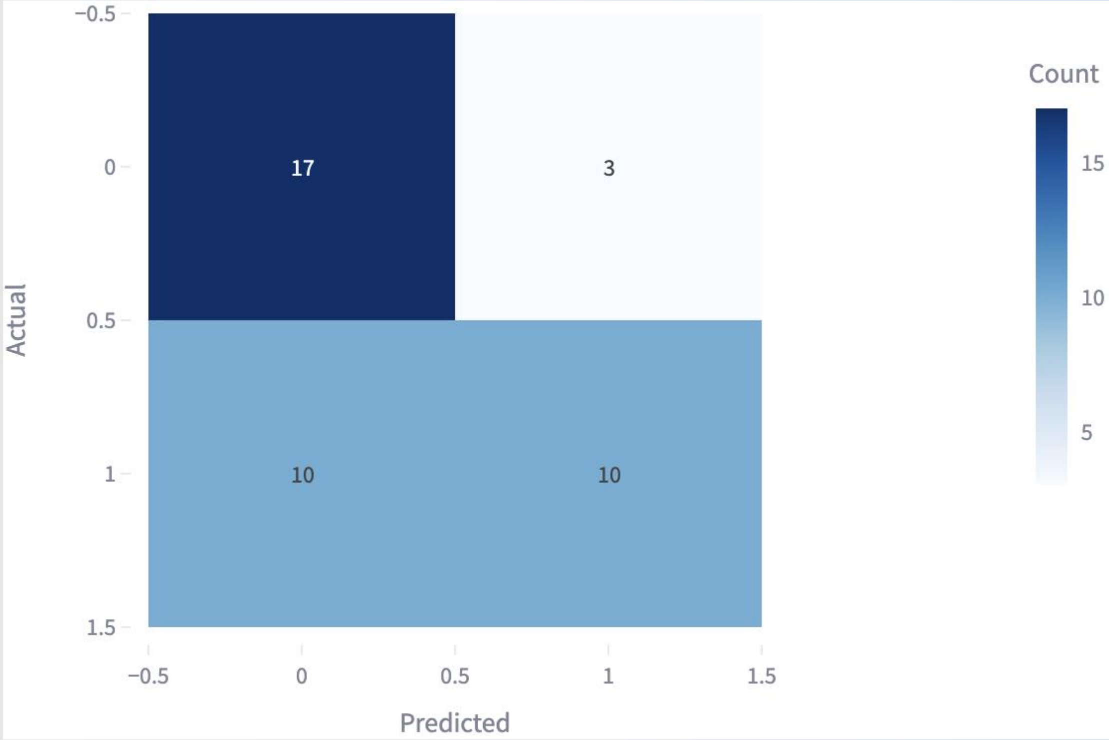
Accuracy	Precision	Recall	F1	AUC
0.675	0.7692	0.5	0.6061	0.5925

Classification Report

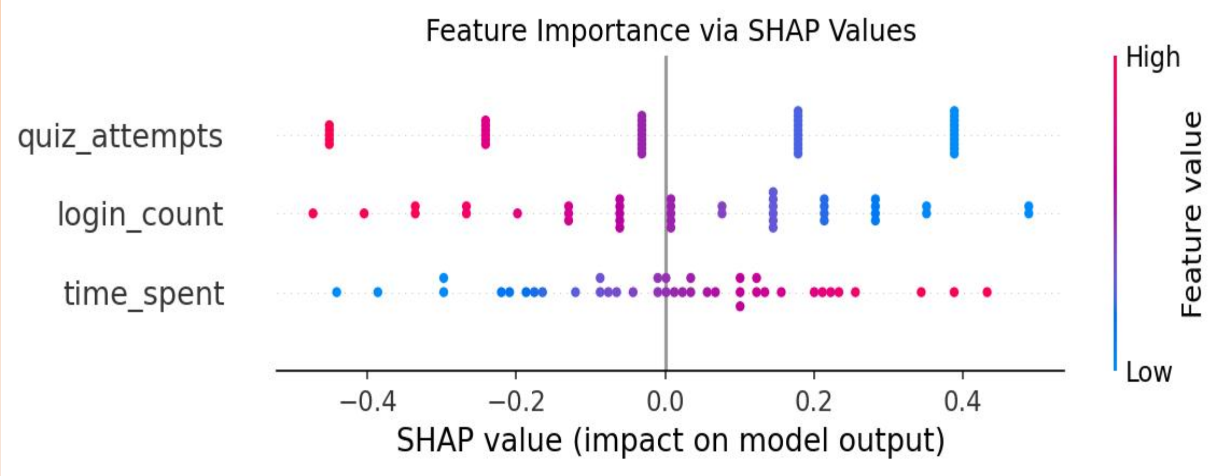


```
              precision    recall  f1-score   support

 Low-risk(0)       0.63      0.85      0.72        20
High-risk(1)       0.77      0.50      0.61        20

    accuracy                           0.68        40
```

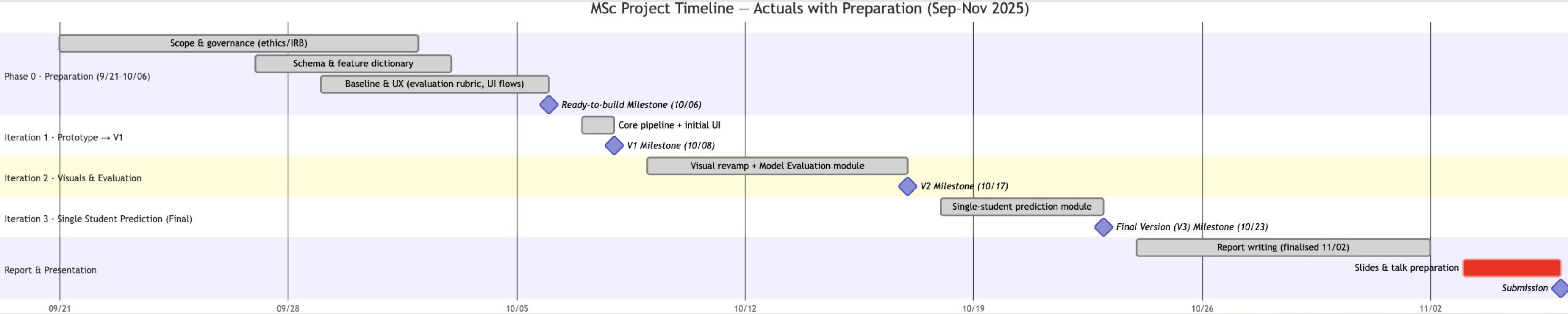
Confusion Matrix

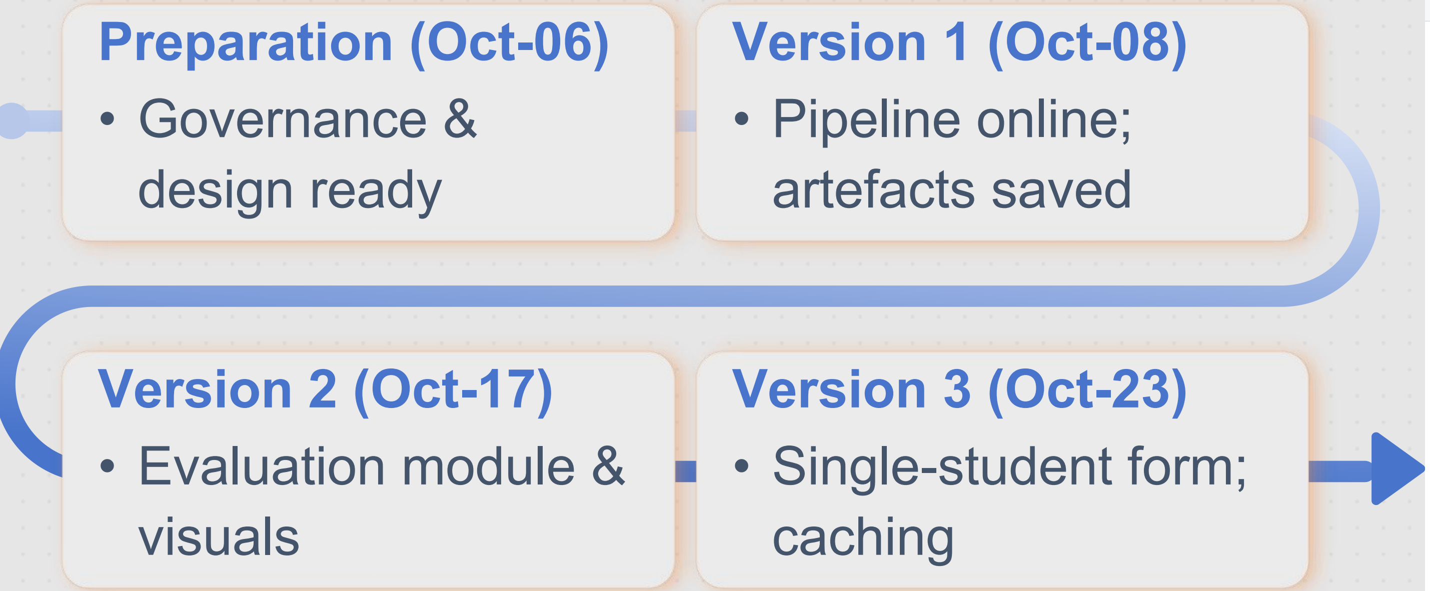


SHAP Beeswarm w.r.t. P(High Risk)



Project Management & Actual Timeline





Preparation (Oct-06)

- Governance & design ready

Version 1 (Oct-08)

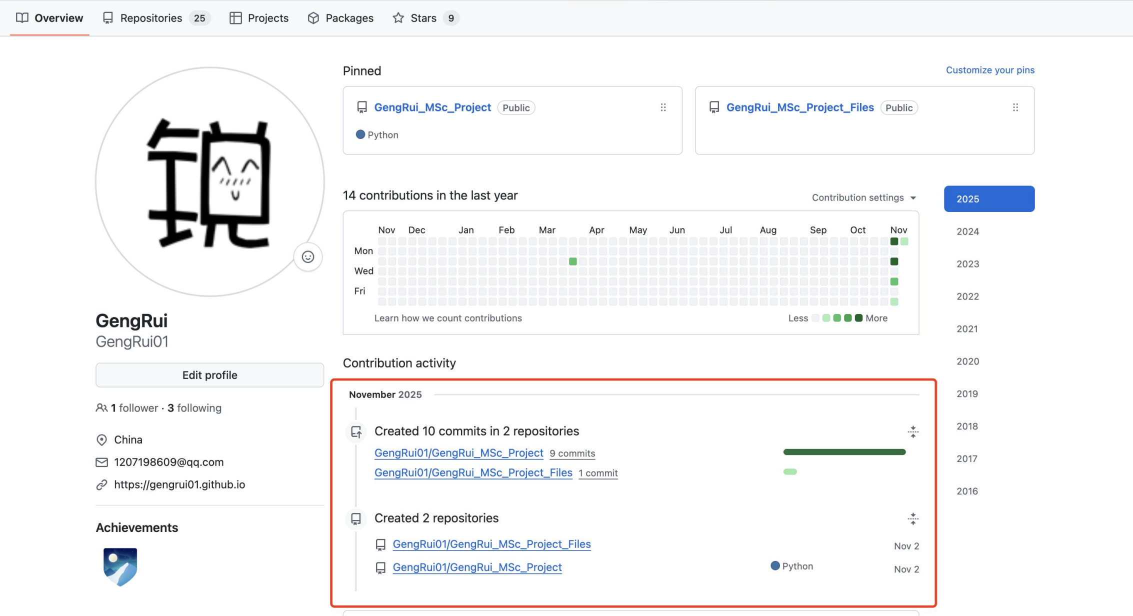
- Pipeline online; artefacts saved

Version 2 (Oct-17)

- Evaluation module & visuals

Version 3 (Oct-23)

- Single-student form; caching



GengRui
GengRui01

1 follower · 3 following

China
1207198609@qq.com
https://gengrui01.github.io

Achievements

Pinned

GengRui_MSc_Project Public
Python

GengRui_MSc_Project_Files Public

14 contributions in the last year

Contribution activity

November 2025

Created 10 commits in 2 repositories
GengRui01/GengRui_MSc_Project 9 commits
GengRui01/GengRui_MSc_Project_Files 1 commit

Created 2 repositories
GengRui01/GengRui_MSc_Project_Files Nov 2
GengRui01/GengRui_MSc_Project Python Nov 2

Contributions & Conclusion

Contributions

explainability-first pipeline (DB → model → dashboard); persisted artefacts (models/*.pkl); recall-prioritised evaluation.

Results P(H)

Acc 0.6750

Prec 0.7692

Recall 0.500

F1 0.6061

AUC 0.5925

Limitations

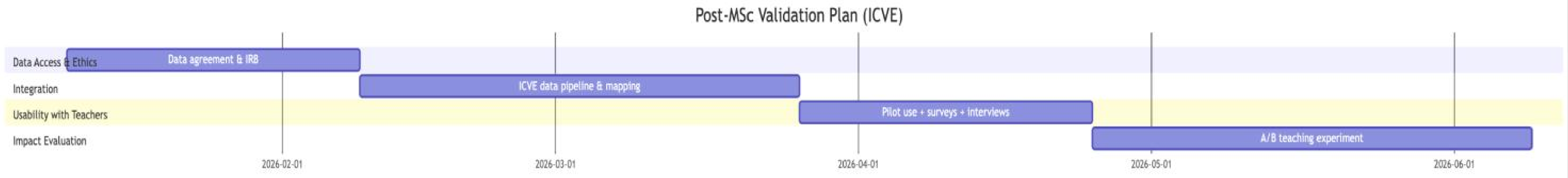
synthetic cohorts; baseline tuned for recall; data-sharing constraints.

Conclusion

A teacher-actionable, reproducible, and fairness-oriented, recall-prioritised prototype that supports early identification and intervention.

Future Work

data-sharing(ICVE) → schema mapping → cohort ingestion → threshold calibration → subgroup-parity checks → A/B teaching trials → ethics & governance loop.


Post-MSc Validation Plan (ICVE)
Data Access & Ethics
Data agreement & IRB
Integration
ICVE data pipeline & mapping
Usability with Teachers
Pilot use + surveys + interviews
Impact Evaluation
A/B teaching experiment