Last name: Wang ID number: K5196286

First name: Hegao

Data Processes Exam

Referred Exam

30 june-2026

General instructions

- The exam must be done individually. If any non-conformity with this criterion is detected (copy, plagiarism, etc.), actions will be taken according to UPM internal rules.
- You need to answer questions in the given space.

Question 1:

You are starting a project to predict student dropouts in the 3 months after enrollment.

1. Write a plan (bullet points) including:

- Data that you would need,
- Artifacts you would request (≥5),
- Usability checks you would run (eg. variable meaning, time validity, join feasibility, completeness/coverage, and basic quality (formats/ranges/duplicates).

I'm assuming school refers to primary & secondary school, not universities
all available, historical or currently in use, records of students containing fields like name, student number, gender, age, enrolment year, latest grade class performance, exam grades, medical condition (whether have special needs), detention, detention reason, attendance, absence, late for school, with timestamps
request access to data, teachers' reviews, family background, parents' profession, student satisfaction survey, the common id used in different records (student number preferred), e.g. past dropout # and check for categorical variables the cardinality, missingness, rate in the first 90 attendance, grades etc should be non-negative days after school start in each year in the past X years
dates in valid format, students who already dropped out should have whether there's no records after the dropout date. if dropout reason is very incomplete understand school's intervention cost & capacity, what they think are the main reasons of previous dropouts) in trend or unavailable, data quality is low
know the earliest possible date
check if coverage differs in 1-year observation window vs longer windows

Last name: _____ ID number: _____

First name: _____

Since dropout is not like voluntary attrition that can happen more frequently motivated by personal career goals, better offers ..., we do not need a rolling window design & there's no major change in management & staff structure, curriculum structure, admission criteria in the past few years. Set a reference year eg 2024 enrolment. get all records of dropouts first. count # with complete info in the most important fields. We cannot 'blindly' take school's report of dropout # and rate; some may have zero or very incomplete records due to system updates / integration. Say rate reported by school is 5%. School has avg of 500 students per cohort, assuming 4 cohorts in total, 2000 in total. 10-year record would only give us 150 dropouts. This is a very severe sample-size constraint. Assume school's earliest valid record is 15 years ago?

We have maximum 225 posities. the usable ones will be lower.

Question 2

List the minimum deliverables you produce at the end of Data Understanding so another team can understand the data:

State clearly the reference date

list of all sources ranked by priority, relevance, completeness provide access, common id. create crosswalk if none exists

usable feature list, leakage list, dictionaries of table → field, categorical field → unique categories, completeness for all relevant fields.

State clearly assumptions made

include challenge/limitation: eg poor data quality
bias by class / tutor
uncooperative departments at school
sample-size constraint

Last name: _____ ID number: _____

First name: _____

Question 3:

If you are building a classification model, which techniques would you use? How would you plan the validation, and which metrics would you use for evaluating the generated model? Justify your answers:

use regularized logistic regression & boosting tree-based models
↓ allow class weight, handle class imbalance
↓ handle NA natively
handle interaction among features

. date. only keep records before that date

Assume 160 positives.
preprocess data. drop duplicates. remove highly correlated fields.
fine-tune C, L1_ratio. max_iteration. 20%
5-fold validation. test on held-out test
exactly one

metrics: recall@topk - better at identifying dropouts
precision@topk - avoid wasting school intervention costs
PR-AUC - since dropout is rare event

choose best metric on oof performance

select MMC threshold

test on 20% held out test

Last name: _____ ID number: _____

First name: _____

Question 4:

Dropout rate is around 10-30%, how would this affect if you are building a classification model?

My assumption was too conservative.

I can build model of slightly more complexity.

as I can allow more negative samples.

more iterations.

I expect higher statistical significance of evaluated metrics.

Question 5:

missing dropout reason

Provide Python/pandas code for at least two of your cases showing:

- how you create a missing-indicator feature, and
- how you apply the chosen policy (e.g., group median imputation, "Unknown" category, or 0 with justification):

```
df["feature_NA"] = df.feature.isnull().astype(int)
df.loc[df.cat.isnull(), "cat"] = "Unknown"
                ↳ student's class
df.loc[df.num.isnull(), "num"] = df.loc[df.num.notnull(), "num"].
                ↳ student's entrance exam score
    df.group = "group"                    df.group="group"
                                          median()
```

Last name: _____ ID number: _____

First name: _____

Last name: _____ ID number: _____

First name: _____

Question 6:

Propose a small experiment matrix that varies:

1. model family + key hyperparameters,
2. one missing-value policy choice,
3. one decision policy (threshold or top-k),
   and state what metrics you will report.

tree-based n_estimators = [150, 300]

logreg model l1_ratio = [0, 0.1, 0.5, 0.9, 1.0]
class_weight = ["balanced", default]
C [0.001, 0.01, ... 100]

if missing impt fields, drop the student.

top-k.

report lift

Last name: _____ ID number: _____

First name: _____

Question 7:

You are given a dataset containing numerical and categorical variables describing a set of students. No target variable is available now.

a) Explain which type of Data Science task can be addressed with this dataset and why.

b) Describe the main steps of this type of analysis (just enumerate them as we learnt in class).

c) Explain how you would interpret and validate the obtained groups, and how the results could be useful from a decision-making perspective

unsupervised. no label.
see distributions, ranges for numerical variables
unique categories for categorical variables identify any outliers
PCA to obtain clustering.
examine % variance explained by top N PCs.
5-fold validation, see if 80% of them still form similar groups.
hold indicator of missingness and see if result changes.

they may represent dropout Yes | No. of PC. = 2

Check with school if true