Data Mining Assignment 2 – Classification

Problem Setting: Predicting Student Graduation Success

For this assignment, you will work with the **"graduation_train.csv"** dataset. The goal is to train a classification model that predicts whether students will successfully graduate based on various academic, demographic, and behavioral factors.

Universities are interested in using this model to identify students who might be at risk of dropping out so that they can provide additional support, such as tutoring or counseling. Since educational opportunities can significantly impact students' future careers and lives, the university wants to ensure that the classification model is not only accurate but also fair, especially with regard to gender and socioeconomic background.

Task 1: Building and Evaluating Classification Models

Just like the last assignment, it is important to start with preprocessing the dataset. This may include handling missing values, encoding categorical variables, and normalizing numerical features.

Your first task is to train multiple classification models on the graduation dataset. Specifically you must experiment with at least the following models:

- Decision Tree
- k-Nearest Neighbors
- Naïve Bayes
- One ensemble method

During your experiments, consider different feature selection techniques and discuss their impact on performance. Ensure careful model evaluation by avoiding overfitting, optimizing hyperparameters where applicable, and designing a proper experimental setup to ensure valid conclusions. Report your findings, which must include at least the following metrics: accuracy, Area Under the Curve (AUC), and precision and recall for both classes.

Task 2: Applying the Model to New Student Data

Select the best performing model from the previous task and apply it to a new dataset, **"graduation_test.csv"**, which contains recently admitted students. The university wants to use your model to predict which students are at risk of not graduating and may need additional support. Your task is to generate predictions on this new data and provide an estimate of how well your model is likely to perform in practice.

Can you make a reliable estimate of how accurate the model will be and how many students are predicted to graduate? What factors do you base this prediction on? How does your model perform in terms of gender fairness? Discuss briefly how you could improve fairness in your model.

Fill in the **predictions_template.csv** with a prediction for each new student.

Additional Notes on Code and Submission:

- Implement your code in any language (Python recommended). Use libraries like:
    - **pandas** for data manipulation
    - **sklearn** to build and evaluate classification models
    - **matplotlib** or **seaborn** for visualizations
- Include your findings in a concise report of maximum 4 pages.
- Submit both the code, your results and the report(pdf) as a **.zip** file via Blackboard by **25/04/2025**.
- The ZIP file should include your full name like, **Lastname_Firstname-studentNumber.zip.**

Questions: In case you have any questions specific to the assignment, please send an email to nick.wils@uantwerpen.be

Evaluation Criteria

Your submission will be evaluated based on the following grading rubric.

	Less than 7	7 to 10	10 to 13	14 to 17	18 to 20
Writing Style	The report is very confusing; the writing style is below average.	At places the report is not very clear; there is a lack of clear structure.	The text is overall clear although some parts could be improved.	Most of the text is easy to follow, findings are explained in a clear way.	The text is very clearly and concisely written, illustrative examples and figures are not overused, but added where needed.
Task 1	Task was not completed.	Task was only completed in a minimal way. Incorrect or insufficient argumentation is given for the choices that were made. The result analysis is not clear.	Task was completed. Some of the explanations behind the methodology/results were okay, though some aspects are missing.	Task was completed and the motivation behind the methodology as well as the result analysis are clear. Some interesting points are made.	The motivation behind the methodology and the result analysis are excellent. Most decisions are backed either by numerical analyses, scientific papers or convincing arguments.
Task 2	Task was not completed.	Task was only completed in a minimal way. Incorrect or insufficient argumentation is given for the performance estimate.	Effort was done to complete the task but some important aspects are missing.	Most of the task execution was done and motivated well. Only minor aspects have been overlooked.	Complete and thorough analysis was done to complete the task.
Code	Code raises many errors.	Code raises some errors or is very unclear.	Code runs but lacks clear structure and readability, only little documentation is given.	Code is readable and sufficiently documented.	Code is very readable and well documented. It is structured in a way that only by (un)commenting single lines, the code for the different tasks can be run.

Data Description

The dataset consists of various attributes related to students' academic backgrounds, personal information, and socioeconomic factors. Below is a brief description of each attribute:

- **Student_id:** Unique id for each student.
- **Marital Status**: Indicates whether the student is single, married, or in another marital situation.
- **Application Mode**: The method through which the student applied to the university.
- **Application Order**: The ranking of the chosen course in the student's application preferences.
- **Course**: The specific academic program the student is enrolled in.
- **Daytime/Evening Attendance**: Specifies whether the student is enrolled in daytime or evening classes.
- **Previous Qualification**: The highest academic qualification obtained before enrolling in the university.
- **Nationality**: The student's country of origin.
- **Mother's Qualification**: The highest educational level attained by the student's mother.
- **Father's Qualification**: The highest educational level attained by the student's father.
- **Mother's Occupation**: The profession of the student's mother.
- **Father's Occupation**: The profession of the student's father.
- **Displaced**: Indicates whether the student had to relocate for their studies.
- **Educational Special Needs**: Identifies if the student requires special educational support.
- **Debtor**: Whether the student has outstanding tuition-related debts.
- **Tuition Fees Up to Date**: Indicates if the student has paid their tuition fees on time.
- **Gender**: The gender of the student.
- **Scholarship Holder**: Whether the student receives a scholarship.
- **Age at Enrollment**: The age of the student when they first enrolled in the university.
- **International**: Specifies if the student is an international student.

Academic Performance Indicators

- **Curricular Units 1st Semester (Credited, Enrolled, Evaluations, Approved, Grade, Without Evaluations)**:
  Information on the student's course load in the first semester, including the number of credited, enrolled, evaluated, and approved units, as well as their grades and any units without evaluations.
- **Curricular Units 2nd Semester (Credited, Enrolled, Evaluations, Approved, Grade, Without Evaluations)**:
  Similar information as above, but for the second semester.

Economic Indicators

- **Unemployment Rate**: The national unemployment rate at the time of enrollment.
- **Inflation Rate**: The inflation rate during the student's enrollment period.
- **GDP**: The gross domestic product at the time of enrollment.

Target Variable

- **Target**: The outcome variable, indicating whether the student successfully graduates(1) or drops out(0).