Data Mining Assignment 2 – Classification

Problem Setting: Also for this assignment you are going to work with the "income.xlsx" dataset. This time you want to train a classification model on this data, to predict whether individuals have a high or a low income. A bank wants to use this label as a proxy of whether individuals have the financial means to pay back a loan, so that they can automatize their loan-allocation process. Since receiving or not receiving a loan can have tremendous impact on peoples' life (think about their ability to buy a house or to start a business) the bank wants to make sure that the classification model behaves *fairly*, regarding the sex of the individuals.
Hence, it will be your job to not only build an accurate, but also a just classifier.

Task 1: Build multiple classification models on the "income.xlsx" data, making use of at least two different types of classifiers and different feature combinations to learn them.
You will see that including or excluding some of the available features can change the behaviour of your models. When choosing features, you can make use of the frequent pattern analysis results from the previous assignment. Describe how you utilize these results to build a more accurate and fairer model.

In your report compare the different models that regarding their performance and their fairness. You should discuss at least one "bad", one "mediocre" and one "good" result, and choose one "good" model that you will use for the second assignment.

Explain and justify which measures you use to assess the models' performance and the fairness. There is not one right set of measures, it is more important that you can motivate your choices.

Task 2: Next to the "income.xlsx" data you have received a file "test.xlsx". This is some very recent data that a bank has collected about their loan applicants, and they want to use your best classification model to decide which people should get a loan. Try to give a performance estimate on how your best model will behave on this data. Can you make any reliable estimations on how accurate will it be, and how many loans will it hand out? On what do you base this performance estimate?

Some Notes on your Code: You may implement your code in any programming language of your choice, but we do recommend Python as this is the most common language used for Data Mining tasks. You might find these libraries useful:

- `pandas` for data storing/preprocessing
- `sklearn` to build and evaluate classification models
- `matplotlib` or `seaborn` for generating figures

Handing in instructions: Put all your findings in a report of about 3-4 pages (there is no hard page limit, but be concise). The deadline for handing in your report and the code is 06-05 . is expected from you!

Grading Rubric:

Below you can find the rubric we will use to grade your assignment. Please consult this rubric to understand what is expected of you.

	Less than 7	**7 to 10**	**10 to 13**	**14 to 17**	**18 to 20**
Layout	Layout of the report is sloppy, report is hard to read, there are many typos		Layout is okay, but at some parts there are typos or section headings, figure titles etc. are missing	The layout of the report is good, there are barely any typos	The layout is excellent
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 explanation behind the methodology/results is 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	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	Complete and thorough analysis was done to

		insufficient argumentation is given for the performance estimate		aspects have been overlooked	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 different tasks can be run