WRITTEN EXAMINATION

Course: Intro to Data Science A1N

Examination

Course code: IT803A

Credits for written examination: 7.5

Date: 2026-01-12

Examination time: 08:15 – 12:30

Examination responsible: Addi Ait-Mlouk

Teachers concerned:

Aid at the exam/appendices

Other

Instructions

- ☐ Take a new sheet of paper for each teacher.
- ☐ Take a new sheet of paper when starting a new question.
- ☐ Write only on one side of the paper.
- ☒ Write your name and personal ID No. on all pages you hand in.
- ☒ Use page numbering.
- ☒ Don't use a red pen.
- ☒ Mark answered questions with a cross on the cover sheet.

Grade points: Each question is graded 0–10 points. To pass the exam, you need a minimum of 5 points on each question (more details on the next page).

Examination results should be made public within 18 working days

Good luck!

Total number of pages

Questions

- The exam has five questions, one for each course objective.
- Each question has sub-questions (1, 2, 3, ...)
- Each question is graded with up to 10 points.
- To pass a question, you need to have at least 5 points on the question.
- To pass the exam, you need to pass all the questions.
- The maximum number of points on the exam is 50.

Grading

If your score on any question is below 5 points, your grade will be U (Fail). If you have at least 5 points on each question, your grade is determined using the sum of points as follows:

Points	Grade	Percentage
45-50	A	90-100
40-44	B	80-89
35-39	C	70-79
30-34	D	60-69
25-29	E	50-59
0-24	F	0-49

A (Excellent), B (Very good), C (Good), D (Satisfactory), E (Sufficient) or F (Fail)

Don't forget to motivate all your answers!

Good luck!

Question 1 (10 points)

1. The course has centered around a lifecycle model of data science projects with six steps. Briefly describe each of the six steps and how they together form a complete project development process/cycle **(7 points)**

2. In a particular data science project, we need to use the mathematical constant π (pi) in calculations. For that purpose, we include the following Python code in a Jupyter Notebook (cell 1):

```
import math
math.pi
```

getting the expected output (cell 2):

```
3.141592653589793
```

When we try to compute the area of circle given a previously defined radius `r` (cell 3):

```
c_area = pi*r**2
```

we get an error message:

```
---------------------------------------------------------
NameError                 Traceback (most recent call last)
Cell In[3], line 1
----> 1 c_area = pi*r**2
NameError: name 'pi' is not defined
```

A. In the above **`import`** statement, what is `math` an example of? **(1 point)**

B. Why do we get the above error message? **(1 point)**

C. How should the **`import`** statement be changed so we could use `pi` the way we want when computing the circle area in cell 3? **(1 point)**

Question 2 (10 points)

1. What are the two major strengths of Python when it comes to using it for data science projects? Briefly discuss **(3 points)**
2. In the context of data preprocessing, what are **outliers,** and why might they be a problem for data science applications? **(2 points)**
3. One method for identifying outliers in a dataset is based on the **interquartile range (IQR)**. Describe how IQR can be used for finding outliers **(3 points)**

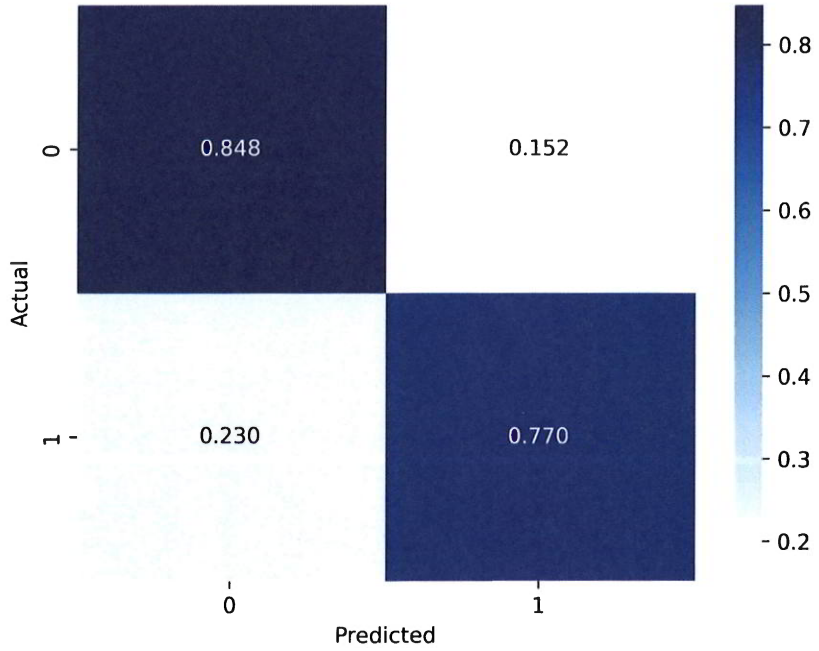
4. How we should handle outliers depends on why we think we have them in the first place. Mention and briefly describe two (2) basic techniques for handling outliers. **(2 points)**

Question 3 (10 points)

On April 15, 1912, RMS Titanic sank after colliding with an iceberg, resulting in the death of 1,502 out of 2,224 passengers and crew. Our task is to build a model that predicts whether a passenger survived (722 passengers or 32.5%) or not (1,502 passengers or 67.5%) using passenger data.

After cleaning the data, handling missing values and outliers, encoding categorical variables, and scaling and normalizing numeric variables, we train a logistic regression model with target "survived" (0 = "died" and 1 = "survived") using a train-test-split of 80/20.

As part of evaluating the classification results on the test set, we inspect the following visualization:



1. What is the above visualization called (what does it depict)? **(1 point)**
2. Briefly explain what is shown in the four quadrants (upper left, upper right, lower right, and lower left). What do they represent? What do they tell us? **(4 points)**
3. What could a possible explanation be for us getting different numbers in the upper left–lower right diagonal (0.848 vs. 0.770), and similarly for the lower left–upper right diagonal (0.230 vs. 0.152)? **(2 points)**

4. The visualization uses a sequential color scale from white to dark blue for representing real numbers ranging between 0 and 1.0. Is this a suitable color scale in this case? Briefly discuss **(2 points)**
5. What information does the confusion matrix provide that overall accuracy does not? **(1 points)**

Question 4 (10 points)

1. What is the fundamental difference between classification and clustering algorithms? **(2 points)**
2. Define overfitting and explain one method to prevent it **(2 points)**
3. What is data leakage, and why is it dangerous in machine learning pipelines? **(2 points)**
4. What is feature engineering, and why can it be more important than model selection? **(2 points)**
5. What is the purpose of cross-validation, and how does k-fold cross-validation work? **(2 points)**

Question 5 (10 points)

1. What is the goal of model distillation regarding model size and performance? **(2 points)**
2. Explain the principle of Federated Learning (FL), contrasting it with traditional centralized Machine Learning **(2 points)**
3. Discuss two major challenges faced by scalable FL systems: system heterogeneity and statistical heterogeneity **(2 points)**
4. In the NLP pipeline, what component is responsible for identifying entities like 'ORG' (organizations) or 'MONEY' within a text? **(2 points)**
5. The GDPR grants individuals several rights to control their data. List and briefly describe at least four distinct rights granted under the GDPR **(2 points)**