WRITTEN EXAMINATION

Course: Intro to Data Science A1N

Examination

Course code: IT803A

Credits for written examination: 7.5

Date: 2026-02-24

Examination time: 14:15 – 18:30

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. As data analysis tasks, what are the principal differences between *clustering* and *classification*? **(4 points)**

2. When it comes to algorithms/methods:

   A. Describe and explain one (1) algorithm/method for clustering. **(2 points)**

   B. Describe and explain one (1) algorithm/method for classification. **(2 points)**

3. Concerning model validation/evaluation:

   A. Briefly describe and explain one (1) approach for validating/evaluating clustering models. **(1 point)**

   B. Briefly describe and explain one (1) approach for validating/evaluating classification models. **(1 point)**

Question 2 (10 points)

Solving data science problems using Python typically involves *defining and using functions.*

1. What are the major benefits of functions in this context? Briefly discuss. Note that this is not about how to concretely define and use functions (syntax) but about benefits of functions on a more general level. **(3 points)**

2. Consider the following function definition:

   ```
   def greet(name, msg="How do you do?"):
       print(f"Hello {name}, {msg}")
   ```

   What would be the outputs/results of running (executing) the following function calls? **(0.5 points each)**

   A. `greet("Göran")`

   B. `greet("Göran", "Pleased to meet you!")`

   C. `greet(msg="Long time no see!")`

   D. `greet(msg="What's up?", name="Göran")`

   E. `greet(msg="Good to see you!", "Göran")`

   F. `greet([1, 2, 3])`

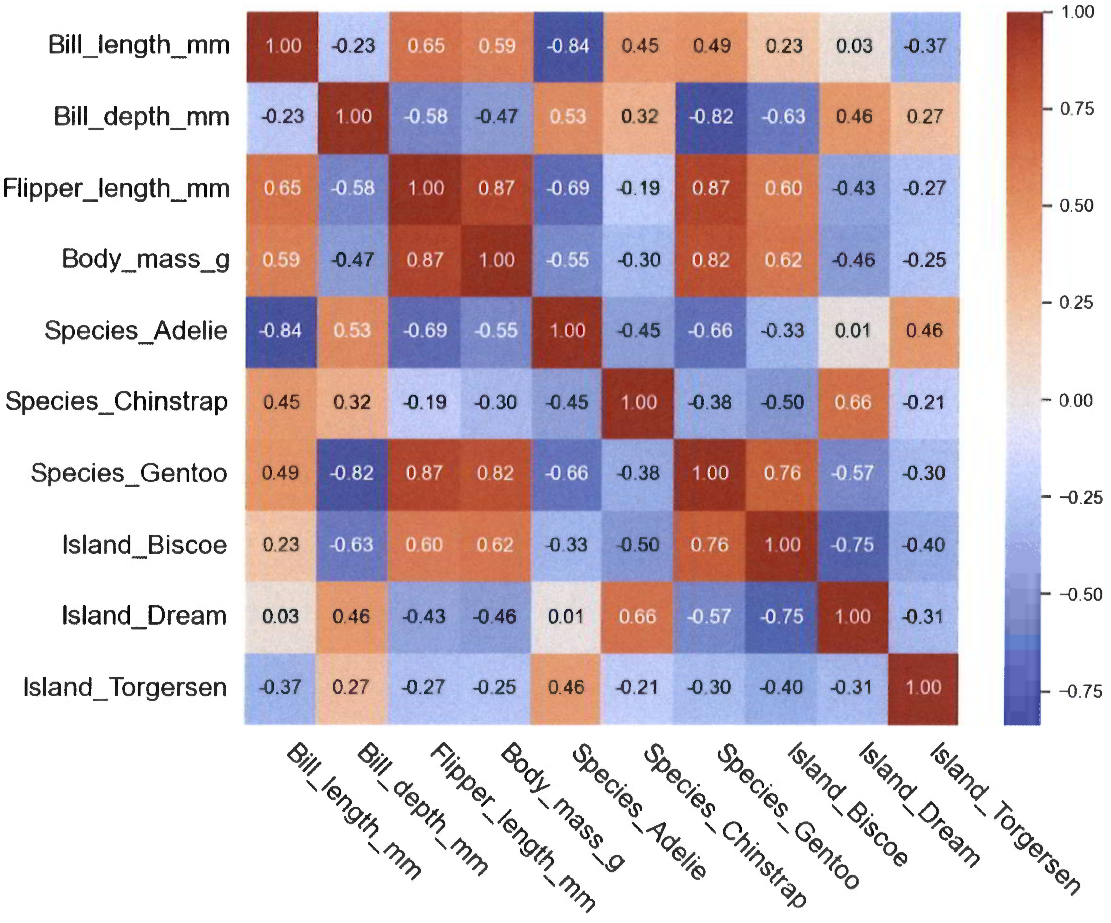
3. Pandas DataFrames (objects of class `pandas.DataFrame`) are commonly used when working with data in data science applications. Because of their flexibility, DataFrames are widely used as a starting point for understanding data and preparing it for further analysis or modeling.

   Explain what a DataFrame is and how it supports the representation and handling of data and discuss the key characteristics that make DataFrames central to modern data analysis workflows, including data exploration, transformation, and preparation for modeling. **(4 points)**

Question 3 (10 points)

The *Palmer penguins dataset* contains 342 observations of three species of Antarctic penguins from three islands in the Palmer archipelago. Our task is to predict the species (“Species_Adelie”, “Species_Chinstrap”, or “Species_Gentoo”) from morphological features (“Bill_length_mm”, “Bill_depth_mm”, “Flipper_length_mm”, and “Body_mass_g”) and location (“Island_Biscoe”, “Island_Dream”, or “Island_Torgersen”).

After proper preparation of the data, we are conducting an exploratory data analysis (EDA). As part of the EDA, we have constructed the following *correlation matrix* (correlation heatmap):



	Bill_length_mm	Bill_depth_mm	Flipper_length_mm	Body_mass_g	Species_Adelie	Species_Chinstrap	Species_Gentoo	Island_Biscoe	Island_Dream	Island_Torgersen
Bill_length_mm	1.00	-0.23	0.65	0.59	-0.84	0.45	0.49	0.23	0.03	-0.37
Bill_depth_mm	-0.23	1.00	-0.58	-0.47	0.53	0.32	-0.82	-0.63	0.46	0.27
Flipper_length_mm	0.65	-0.58	1.00	0.87	-0.69	-0.19	0.87	0.60	-0.43	-0.27
Body_mass_g	0.59	-0.47	0.87	1.00	-0.55	-0.30	0.82	0.62	-0.46	-0.25
Species_Adelie	-0.84	0.53	-0.69	-0.55	1.00	-0.45	-0.66	-0.33	0.01	0.46
Species_Chinstrap	0.45	0.32	-0.19	-0.30	-0.45	1.00	-0.38	-0.50	0.66	-0.21
Species_Gentoo	0.49	-0.82	0.87	0.82	-0.66	-0.38	1.00	0.76	-0.57	-0.30
Island_Biscoe	0.23	-0.63	0.60	0.62	-0.33	-0.50	0.76	1.00	-0.75	-0.40
Island_Dream	0.03	0.46	-0.43	-0.46	0.01	0.66	-0.57	-0.75	1.00	-0.31
Island_Torgersen	-0.37	0.27	-0.27	-0.25	0.46	-0.21	-0.30	-0.40	-0.31	1.00

1. What does the Palmer penguins’ correlation matrix tell us? Explain and discuss using concrete examples from the above correlation matrix. **(3 points)**

2. What does the Palmer penguins’ correlation matrix **not** tell us? Explain and discuss. **(1 point)**

3. Is the color scale used in the above correlation matrix appropriate, and why or why not? Briefly explain and discuss. **(1 point)**

4. Explain how the self-attention mechanism in transformers allows NLP models to capture long-range dependencies. **(2 points)**

5. Discuss appropriate evaluation metrics for NLP tasks (classification, summarization, translation) and explain how to detect and mitigate biases in NLP models. **(3 points)**

Question 4 (10 points)

Several hospitals want to train an AI model to detect diseases from medical images. Due to privacy laws, patient data cannot be shared. With federated learning, each hospital trains the model locally on its own data and only sends model updates (not raw data) to a central server, which aggregates them into a better global model. This improves accuracy while preserving patient privacy.

1. Explain how federated learning addresses regulatory constraints (e.g., GDPR, HIPAA) in the hospital use case, and discuss remaining privacy risks. **(1 point)**
2. Analyze the impact of data heterogeneity across hospitals on model convergence and performance in federated learning. **(3 points)**
3. Compare federated learning with secure centralized training using anonymization in the context of medical imaging. When is FL preferable? **(3 points)**
4. Discuss how communication efficiency and system scalability affect federated learning deployment across multiple hospitals. **(3 points)**

Question 5 (10 points)

A bank builds a binary classification model to detect fraudulent credit card transactions in real time. The dataset includes transaction amount, time, location, merchant category, and device features. The system must handle highly imbalanced data, adapt to evolving fraud patterns, and deploy the model with low latency to avoid blocking legitimate transactions.

1. Explain how you would handle class imbalance in the fraud detection dataset and justify your choice of techniques. **(2 points)**
2. Discuss feature engineering strategies that can improve fraud detection while minimizing data leakage. **(3 points)**
3. Propose two classification models suitable for this task in terms of interpretability, performance, and risk. **(3 points)**
4. Analyze the challenges of deploying a fraud classification model in a real-time banking system, including latency, monitoring, and model updates. **(2 points)**