WRITTEN EXAMINATION

Course: Intro to Data Science A1N

Examination

Course code: IT803A

Credits for written examination: 7.5

Date: 2025-01-15

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

[Course objective: extensively describe and problematize the state of the art of the field of Data Science, and discuss fundamental application areas for Data Science]

1. There's no definitive, commonly agreed upon definition of data science, although several suggestions have been put forward in the literature over the years. Certain fundamental

elements, concepts, principles and themes persist across multiple definitions, albeit with slight variations. What are those elements, concepts, principles and themes?

2. Describe the data science project lifecycle.
3. Name three regression algorithms.
4. Discuss the importance of data preprocessing in achieving high-quality results in Data Science
5. What challenges arise when working with unstructured data in Data Science?
6. Describe the main principles for ensuring *graphical integrity* in the context of visualizing data and results from data analysis tasks, that is, how to make sure that the graphics (plots, graphs, diagrams etc.) give the true picture and do not mislead, disguise or fabricate.
7. What is the purpose of splitting a dataset into training, validation, and test sets, and how do these splits contribute to building a robust machine learning model?
8. What are some real-world applications of supervised learning and unsupervised learning?

Question 2

[Course objective: extensively exemplify and contrast different perspectives on central foundations, principles, methods and theories within the field]

1. Explain the concepts of data *transformation* and data *normalization/standardization* in the context of data science. What are their purposes? What are the similarities/differences between the two concepts? etc.
2. Give two examples of techniques for performing data transformation and give two examples of techniques for performing data normalization, including descriptions of how each technique works.
3. For each technique described in subquestion 2, explain for which data and (subsequent) algorithms the technique is suitable/not suitable.
4. For each technique described in subquestion 2, show how the technique can be implemented in Python.
5. Explain why *categorical* data often needs to be handled differently than numerical data in the context of data science.
6. Give three examples of techniques for handling categorical data in order for it to be used for subsequent analysis tasks, including descriptions of how each technique works and for which types of categorical data the technique is suitable.

7. Explain when to use/not to use a *sequential* color scale, a *diverging* color scale and a *qualitative* color scale, respectively, in the context of data visualization..
8. When visualizing data, explain when (for which data) a *line chart* is preferable over a *bar chart*, and vice versa.

Question 3

[Course objective: present and discuss ethical and societal issues in connection to Data Science and its applications]

1. What are the main ethical concerns in data collection for data science projects?
2. What is data privacy, and why is it crucial in Data Science?
3. What is Natural Language Processing (NLP), and what are some of its key applications?
4. How can you ensure that NLP models maintain fairness and avoid bias?
5. What are Large Language Models (LLMs), and what are their key applications?
6. What is the General Language Understanding Evaluation (GLUE) benchmark, and what is its purpose?
7. List and explain some protection mechanisms used in data privacy.
8. What is Federated Learning? outline some of the key challenges associated with it.

Question 4

[Course objective: account for syntax and semantics for programming languages that are particularly suited for Data Science]

1. What makes Python a popular language for data science?
2. What are the key Python libraries used for data manipulation in data science?
3. How can you perform hyperparameter tuning using the `GridSearchCV` class in Scikit-learn?
4. What is the purpose of `train_test_split` in Scikit-learn, and how do you use it?
5. What is the difference between `fit()` and `predict()` methods in TensorFlow's classification models?
6. What is the purpose of `dropna()` and `fillna()` in Pandas, and when should you use them during EDA?
7. What are the key differences between TensorFlow and scikit-learn
8. How do you calculate summary statistics in Pandas?

Question 5

[Course objective: critically reflect on and describe requirements and issues for programming within the area; and independently develop computer programs within the area]

1. Explain the difference between univariate and multivariate analysis in EDA
2. What are outliers, and how would you identify and handle them during EDA? Provide a Python-based approach.
3. How can MLflow be used to track and manage different versions of regression models?
4. What are the key steps involved in deploying a machine learning model as an API,
5. How can frameworks like Flask or FastAPI be utilized to serve the model?
6. How do you handle model evaluation and validation in your data science programs?
7. What challenges might you face when deploying data science programs to production?
8. Given is a Python `DataFrame` about cars, including columns for a manufacturer, price (in USD), model year, production month, color of the car, whether the car is electric or not and the number of driving wheels. The following shows a small excerpt from the `DataFrame`:

	Manufacturer	Price	Year	Month	Color	Electric	# driving wheels
0	Tesla	50000	2023	January	Red	True	4
1	Ford	30000	2021	March	Blue	False	2
2	BMW	45000	2022	June	Black	True	4
3	Toyota	20000	2019	September	White	False	2
4	Audi	55000	2023	November	Gray	True	4

Show how to use Python to correctly handle the categorical data in the above `DataFrame` for the data to be useful in subsequent data analysis. You can assume that there are no missing values or true outliers in the data.