Exam for Data and Things, spring 2025

Oral exam info:

- There will 20 minutes for each student, including time for assessment and feedback
- At the beginning of the exam, the student draws a random number. Each number will correspond to an exam topic. (see the list of Exam topics below).
- The student then presents on the topic (3-5 min) including potential selected exercises handed in as part of the written product. This is followed by questions about the presentation and the exam question from the examiners (5-10 min).
- At the end, the examiners might relate their questions to some of the other exam topics.
- For each of the exam topics, the student is expected to know the central concepts, methods, theories, and problems discussed in class and be able to explain and exemplify them. Moreover, for those exam topics where there are selected exercises handed in as part of the written product, the student is expected to be able to explain how they solved the exercises and be able to explain their entire code.
- The oral exam is scheduled for the 20-21 of March.

Written product info:

- The written product will consist of answers to the selected exercises – these will be selected from those that have already been done in class.
- The list of selected exercises can be seen below.
- The handed in answers to the selected exercises should be in the format of a single Jupyter Notebook or a zip file containing a notebook for each of the selected exercises.
- The student should be able to run the notebook(s) (i.e. the code) at the exam if requested – the code should work.
- The students can do the selected exercises in self-made groups and either hand in as a group on Eksamen.ruc.dk, or hand in the notebook(s) individually.
- The hand-in must happen through eksamen.ruc.dk before March 16 at 10:00.

Exam topics:

1) Data transformation and exploratory data analysis (EDA)
2) Data engineering
3) Statistics
4) Regression
5) Time series
6) Classification
7) IoT and sensor data
8) Clustering
9) Machine Learning Operations (MLOps)
10) Recommender systems
11) Neural networks and deep learning
12) Generative AI
13) Explainability
14) Ethical reflections on data science

Hand-in exercises:

1) Data transformation and exploratory data analysis (EDA)
    - The hand-in exercise for this topic is Exercise 3 from the notebook “Exercises in DT and EDA.ipynb”. The exercise asks you to do an exploratory data analysis of the adult dataset. There is no change in the exercise and what is required, however, here is an elaboration of what the exercise involves: You need to explain what the data is about, which variables the dataset contains and what their data type is. Moreover, for each individual variable you should investigate/explain its distribution/variation through visualization and descriptive statistics. Finally, you should investigate/explain the variation/correlation between pairs of variables – here it is enough to investigate three pairs of variables, one where both variables are categorical, one where both variables are numeric, and one where one of the variable is categorical and the other is numeric.
2) Data engineering
    - The hand-in exercise for this topic is Task 7 from the notebook “DE_Task.ipynb”. The exercise asks you to wrap the functionality of other tasks in this notebook into an ETL pipeline. Note that the GroupBY tasks (which are extra) are not needed in this ETL pipeline. The data visualization parts are also not needed. The pipeline should contain some of the data cleaning (such as removing duplicates and nulls) and then creation of features which are part of task 4 and 5.
3) Statistics
    - The hand-in exercise for this topic is Exercise 3 from the notebook “Exercises in statistics.ipynb”.
4) Regression
    - The hand-in exercises for this topic is Exercise 1 and 2 from the notebook “Exercises in linear regression.ipynb”.
5) Time series
    - The hand-in exercise for this topic is Task 3,4 and 6 from the notebook “TSA_Task”. This means that you have to do the cleaning of dataset, then create features (at least 5 new features should be created, and you should be able to justify why you created each of the features). And then, you should train an XGBoost model on the dataset. Note that you also need to do relevant train, test, validation split and be able to explain why you chose a certain split. Lastly, you should calculate evaluation metrics: rmse and mae to show performance of your model. The hyperparameter tuning part is not required.
6) Classification
    - The hand-in exercise for this topic is Exercise 2 from the notebook “Exercises in Classification II.ipynb” (It is the same as Exercise 2 from the notebook “Exercises in Classification I.ipynb”).
7) IoT and sensor data
    - There are two hand-in exercises for this topic:
        i. The first one is explained in bullet 2 of slide 16 of the slide deck “IoT & Sensor Data.pptx”. You need to do the entire task described in bullet 2 of slide 16.
        ii. The second one is related to PCA and it is explained in the notebook “pdm_task”. Note that the data simulation parts are already done for you, so you do not need to do those parts. You should start from the cell where tasks are listed and you need to do all 5 tasks.
8) Clustering

- The hand-in exercise for this topic is Exercise 1 from the notebook “Exercises in Clustering.ipynb”.

9) Machine Learning Operations (MLOps)
    - There are two hand in exercises for this topic:
        i. First, one is in the notebook - ‘mlflow_task.ipynb’. Complete all the tasks in this notebook.
        ii. Second, one is in the notebook - ‘MLOps exercises.ipynb’. Do tasks 1 – 6 in this notebook.
10) Recommender systems
    - The hand-in exercise for this topic is Exercise 1 from the notebook “Exercises in Recommender systems.ipynb”.
11) Neural networks and deep learning
    - The hand-in exercise for this topic is Exercise 2 from the notebook “Exercises in neural network and deep learning II.ipynb”.
12) Generative AI
    - The hand-in exercise for this topic is in the notebook named ‘rag_task.ipynb’. Do all 4 tasks within this notebook. For task 2, you should try at least 3 types of chunking such as chunk in paragraphs, sentences or even by punctuation marks – you are welcome to choose your own chunking strategy. For task 4 you should try at least one other type of similarity or distance function to calculate the similarity.
13) Explainability
    - The hand in exercise for this topic is explained in slide 25 in the powerpoint ‘Explainability.pptx’. You need to do all the tasks within bullet 2 of this slide. Regarding the second bullet, you should do at least 3 local plots per model, i.e. explain 3 rows/predictions from your test data and then have one global plot using SHAP. Regarding bullet three, you should again explain 3 rows/predictions using LIME. Global plot with LIME is not required.
14) Ethical reflections on data science
    - The hand-in exercise for this topic is Exercise 1 from the notebook “Exercises in Fairness in Machine Learning.ipynb”.