Auditory exercises and project instructions

Introduction to Data Science – Academic Year 2024/2025

Auditory exercises

Starting with the 2nd week of classes, the student's task will be to follow the materials given in auditory exercises on the course. Students are advised to follow the materials closely and learn gradually, as it may become difficult to learn all the materials right before the final exam term. The auditory exercises will cover the practical topics related to published lecture materials, including some of the Python language basics.

The auditory exercises will be held in room B1 (Monday, 12 AM – 2 PM) or D-273 (Thursday, 3 PM – 5 PM), depending on the detailed course schedule available at the course website. Attending the auditory exercises live is not obligatory, but is advised.

If something is not clear in the auditory exercise, a student may contact an assistant in charge of that auditory exercise (the contact information will be given at the beginning of the published Jupyter Notebook materials at the course website).

Project

At the start of the semester, the students choose one of the offered scientific articles for their project work. A maximum of 2 students can apply for each offered article. Students who do not choose an article by the deadline will be randomly assigned to articles with remaining vacancies. It is recommended that students study all articles before applying for a particular article.

The articles offered (and the assistants in charge) can be seen in this table. In the same table, students select articles by writing their first and last name under the desired article.

Deadline for selecting an article: **October 18, 2024**

After the deadline for article selection, students are required to create a suitable GitHub repository individually and add the assistant in charge as a collaborator (Settings → Collaborators → Add people).

The project will consist of three parts, with three checkpoints. In the first part, students work individually on the preparation of a dataset assigned to them by the assistant in charge. The dataset assigned by the assistant will be tabular and will come from the Kaggle platform. In the second part, students individually prepare a dataset from the scientific article for which they applied. In the third part, the project is done in a group of two to four students (English

students may work with Croatian ones), where the goal is to achieve the replication of the results achieved in the scientific article and to propose an improvement of the results of the article.

In case of possible problems or ambiguities during the project, students can always contact the assistant in charge of the selected article.

First part of the project

The goal of this part of the project is to implement data preparation methods that are described in lectures and auditory exercises on a certain tabular dataset that the assistant selected from Kaggle.
In doing so, data preparation methods are carried out in such a way as to solve pre-prepared tasks in Jupyter Notebook, which will be given to students by their assistant.
Students should independently solve the above-mentioned tasks on that dataset and upload the solution in the form of a Jupyter Notebook to GitHub by the deadline.
The assistant in charge will review and evaluate the submitted notebook and the students will be informed of the results.

Submission deadline: **November 15, 2024**
Deadline for assistant scoring of submitted notebooks: **December 13, 2024**
Maximum number of points: **10**

Second part of the project

The goal of this part of the project is to familiarize yourself with the data from the scientific article. Students should read the selected article and download the data that was used. After that, you need to familiarize yourself with the data. The guidelines on how to do this are:

- upload data
- check which data types exist and display descriptive data statistics
- check for missing values and outliers
- visualize the data in several different ways (e.g. histogram of features, line diagrams of time series, dot diagrams depending on the target class, ...)
- ...

Students are not limited exclusively to the items listed above, they serve as guidelines only, since the articles differ from each other. During this part of the project, you should familiarize yourself with the dataset used in detail and at your own discretion.
The result of this part of the project is also a Jupyter Notebook that should be submitted to the GitHub repository by the deadline and should include data preparation steps for the dataset related to the article.

Submission deadline: **December 20, 2024**
Deadline for assistant scoring of submitted notebooks: **January 8, 2025**
Maximum number of points: **10**

Third part of the project

At the beginning of the third part of the project, the replication and improvement of the project results, a consultation will be held with the assistant in charge at the specified date and time. During the consultation, students can discuss their own solutions for the second part of the project and then discuss what they plan to do in the third part of the project with the assistant. The third part of the project is done in independently formed student groups, whose compositions (names and surnames of students) should be reported to the assistant by the time of the consultation.
In the third part of the project, students first use the approaches from the selected article to replicate the presented results. During the actual implementation, students can use already implemented functions from packages such as numpy, scikit-learn, etc. Once the methods are implemented, they need to be run on the previously prepared data, correctly evaluated, compared with the results from the article and any differences explained.

For example, if you are working on a classification problem, it is advisable to display:

- the value of metrics such as accuracy, precision, responsiveness, etc.
- AUC/ROC curves
- confusion matrices
- . . .

Students are, of course, free to present the results in alternative ways that seem interesting to them. In case the selected article deals with a specific issue, it is advisable to contact the assistant in charge for advice.

Then, the goal of this part of the project is also to improve the results of the scientific article. Some of the ways you can try to improve your results are:

- to correct any shortcomings of the article
- extend the data with feature engineering
- apply some of the methods/algorithms discussed in the lectures
- optimize the parameters
- create your own method
- ...

Any potential improvement must be separately evaluated and compared to the original results of the article. In the event that the improvement is not achieved, it is necessary to comment on what is the cause of this and what else could be tried that is currently beyond your capabilities.

Particularly high-quality improvements to the article (according to the assistant's assessment) will be awarded with a possible up to 5 additional points from the project.

The results of the third part of the project, again in the form of a Jupyter Notebook, must be uploaded to the GitHub repository of all group members. At the end of the third part of the project, it is necessary to make a short video presentation lasting maximally 4 minutes in which

you will present the article and the tested improvements. The video presentation must also be uploaded to the GitHub repository by the submission deadline.

During the review of the results of the third part of the project, the assistant in charge will go through the assigned notebook together with the group and, after the questions asked, will determine the final number of points for all members of the respective group. The distribution of points is the responsibility of the assistant and according to the maximum values listed below.

Finally, several of the best improvements to the articles will be selected by the assistants and proposed for a presentation (play of the recorded video presentation + discussion) during the last week of classes.

Date of consultation with the assistant about the article and deadline for reporting the group: **January 9, 2025**
Final project submission deadline: **January 22, 2025**
Review of the group results: **January 27, 2025**
Maximum number of points: **15** points for replication, **5** points for improvement, additional **5** points for special quality improvement

Additional notes

- You do all of the above using Python and Jupyter Notebook. The final version of the notebook that you will submit (in each part of the project) must contain comments/conclusions of all the steps taken. The notebook must be easy to follow without reading too much code. The notebook is submitted by uploading it to the GitHub repository.
- If you don't have a computer with enough resources to do everything needed in the project, we suggest you use Google Colab.
- In the event that the student is late in submitting the solution for a certain part of the project, that part of the project will be scored with 0 points, and the respective part of the project cannot be compensated later in terms of awarding points.
- If the student does not collect a minimum of 25% points from the project (10 points), the student does not have the right to take the final exam (and other exam terms).
- Both laboratory exercises and the first and second part of the project are considered to be individual student assignements. This means that working in a team to achieve the solutions is not allowed in this course and, if suspected to occur, can result in significant penalties for the involved students.