EXAM DOCUMENTATION.

Dear Sigmaritan,

Hope you enjoyed the learning and exploration of the Artificial Intelligence field, and you found this field as attractive as we all do. But now it's time to put all the knowledge you've accumulated into practice. This examination was created with the aim of testing your knowledge and finding the level of understanding of the Machine Learning and Data Analysis principles.

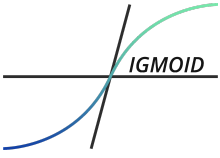The exam will consist of two parts:

1. The exploration of the dataset and the training of a Machine Learning model on it.
2. The presentation of the insights and results obtained from the work you have done on this data set in a PowerPoint presentation in front of the jury.

The first part should be submitted by loading all the notebooks into a GitHub repository and sending the link to that repository to Eduard Balamatiuc by the 9th of September 2025, 23:59.

The second part will be presented individually by each Sigmaritan, that is examined by the jury, consisting of the members of the Castor and Pollux teams. The date of the presentation will be agreed upon personally with each person and the jury.

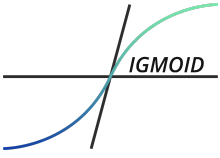More details about the organisational part of the exam can be found on the Pandora page:
https://www.notion.so/vladimirstojoc/Exam-f63df248d45f41a89e329ef90b308040

TASK#1.

In the archive, you have a CSV file. This file is the data that you should use to pass the exam.

The client provided you with a dataset that represents a dynamic tourism route collection offering insights into travel patterns, route optimisation, and tourism behaviour across various destinations. It provides route specifics, such as destination points, travel times, transportation modes, tourist preferences, and seasonal variations, to facilitate a better understanding of dynamic travel planning. Moreover, it contains information about route efficiency, cost factors, attraction popularity, and traveller demographics of tourism experiences. Taking into consideration the travelling preferences and particularities, your task is to predict the age of the tourist. The dataset contains the following fields:

- Route_ID;

- User_ID;

- Sequence;

- Total_Duration;

- Total_Cost;

- Weather;

- Traffic_Level;

- Crowd_Density;

- Event_Impact;

- Optimal_Route_Preference;

- Satisfaction_Score;

- Age;

- Gender;

- Nationality;

- Travel_Companions;

- Budget_Category;

- Preferred_Theme;
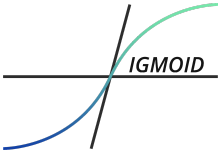
- Preferred_Transport.

In this task, you should:

1. Load and analyse the dataset, and try to extract insights from the data.
2. Prepare the data for a machine learning algorithm-ready format.
3. Handle the NaN values in the dataset.
4. Apply some post-processing and post-analysis techniques, such as Feature Selection, Feature Engineering, or others and see if they can help you.
5. Create a model or a Pipeline that will show the best performance in predicting the target column.

The target column is **Age**. Try to create models that will predict this value.

You MUST show each step of your implementation, even if you consider it "not so important". The goal of this examination is to see what you were able to learn during your study process. The process itself matters more than the final result!

Some more advice:

- You can split different steps of your work into separate notebooks.
- Commend the code and use the markdown cell type to explain what or why you are doing.
- Explain in written form the choices and decisions that you made during the processing of the data.
- Use plots, but don't abuse them.
- Don't chase a perfect model; better show us why you chose a specific model as the best one.

TASK#2.

Your second task is to create and prepare a presentation on the work that you have done. In the presentation that you should explain the work that you have done and why you made the choices that you made during the analysis and Machine Learning model development. Some more pieces of advice for the presentation are:

- KISS (Keep It Simple Stupid) - keep the presentation as simple as possible.
- Give some context about the field on which your data set is about.
- Try to avoid as much as possible having a lot of text on your slides.
- Try to tell a story.
- Don't show the whole process, only the most important findings and decisions.
- Finish with a conclusion.
- Don't make lots of slides.

Good luck!