Data Analytics and Mining

Intro and Course overview

Majid Sohrabi
National Research University Higher School of Economics

September 19, 2025

Teaching Team Contacts / Timetable

Majid Sohrabi (lectures)

– Email: msohrabi@hse.ru

Alexandra Kogan (seminars)

– Email: akogan@hse.ru

Lectures	Seminars	Format	Duration
Fridays 6:10 pm	Fridays 6:10 pm	Offline	Modules 1-2

Repository with course material

https://github.com/Majid-Sohrabi/DAM-2025

Course content

Introduction to Data Analytics and Mining

- Overview of data analytics and mining
- Importance and applications
- Types of data, structured vs. unstructured

Data Processing and Cleaning

- Handling missing data, outliers, noise
- Data transformation, normalization, standardization, encoding
- Feature selection and dimensionality reduction

Exploratory Data Analysis (EDA)

- Descriptive Statistics and Data Summarization
- Data Visualization Techniques (Histograms, Scatter Plots, Box Plots)
- Identifying Patterns, Trends, and Anomalies

Course content

Statistical Foundations for Data Analysis

- Probability Theory and Distributions
- Correlation and Causation, introduction to Regression Analysis

Machine Learning Fundamentals

- Supervised vs. Unsupervised Learning
- Linear and Logistic Regression
- Support Vector Machines
- Model Evaluation (accuracy, precision, recall, F1-score), Cross-Validation

Data Mining Techniques

- Classification, decision tree, k-Nearest Neighbors
- Clustering, k-Means

Course content

Advanced-Data Mining and Machine Learning

– Ensemble Methods: Random Forests

Overview

Compulsory course for year 2

Duration: 1st half of the academic year (modules 1 and 2)

Assessment elements:

- Quizzes (50% weight)
- Exam (50% weight), in the form of a project, with **progress tracked during the semester** (topic choice deadline, preliminary results deadline, final result deadline)

Format:

- Offline (lectures & seminars)

Grade Formula

Grade Component	Percentage	Evaluation Criteria
Quizzes	50%	In-class mini tests once a week or once in two weeks, each test consists of several questions, and a single test is a 10-point scale. IMPORTANT: Quizzes cannot be retaken. The only exception is if you are unable to attend due to illness and provide a medical certificate verified by the study office.
Final Project	50%	The final project is a 10-point scale (in groups), choose a dataset, make relevant analysis, write a report, and present their works.

The formula

$$\text{Final grade} = 0.5 \cdot \text{Quizzes} + 0.5 \cdot \text{Final Project}$$

$$0 \leq \text{Quizzes} \leq 10$$

$$0 \leq \text{Final Project score} \leq 10$$

Rounding to the closest integer

Arithmetic rounding. E.g. 3.5 is rounded to 4, 3.49 is rounded to 3.

Quiz

Quiz are conducted during the seminars

Format: Quizzes will be taken in the LMS using the Safe Exam Browser (SEB)

Devices: Please use only your main device for the quiz. Other devices like phone, tablet, including AI tools, are not allowed.

- **Note**: To keep things fair for everyone, any violation of these rules will unfortunately result in a score of 0 for that quiz.

Exam

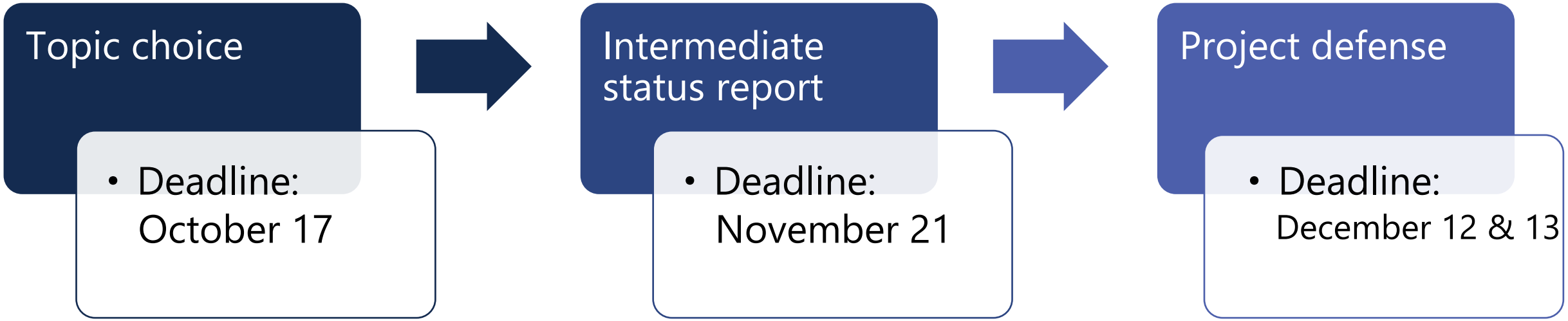
Exam in the form of project defense

The project is:

- Participation in a competition (on www.kaggle.com or similar)
  - Teams of up to 3 people are OK (roles of all members of a team should be clear and significant)

Please discuss your choice with me

Exam project timeline



Missing deadlines for stage 1, 2 will result in -1 point for the project (if both stage 1 and 2 were missed then -2 applies)

Missing the project defense (stage 3) will result in 0 for the whole project.

All members of a group need to attend the final defense.

Motivation

Statistics and data analysis underlie machine learning and artificial intelligence (AI) technologies

Comprehension of its basic principles helps to understand the world we live in

https://www.youtube.com/watch?v=RNnZwvklwa8&ab_channel=AdamEubanks

Thank you!

Majid Sohrabi

msohrabi@hse.ru