Fundamentals of Data Science and Data Analytics

Lecturer: Vahan Sargsyan, Ph.D

Course Description:

This course offers a practical and accessible introduction to data science for upper-level undergraduate and graduate students seeking to build genuine analytical skills in a structured, applied setting. Students are guided through the full data science workflow — from understanding and cleaning data, to building predictive models and evaluating their performance with rigor and responsibility.

Bridging programming, statistics, and problem-solving in an AI-driven world, the course emphasizes the data tasks most relevant to applied research, business analytics, and introductory machine learning. Students work extensively with Python-based materials and real-world examples throughout the semester, leveraging modern AI tools as an integral part of the programming workflow.

Beyond core foundations- — including data structures, key model types (OLS, logistic regression, random forests, and neural networks), and supervised versus unsupervised learning paradigms — the course develops essential practical competencies: data preparation, feature engineering, train-test splitting, classification metrics, and ROC analysis. The course also situates neural networks within the broader landscape of modern AI and large language models, while maintaining a clear emphasis on classical data science

Course Prerequisites:

Introductory statistics or econometrics is recommended.
Prior programming experience is helpful but not required; the course materials are designed to support beginners in Python.

Course Objectives:

By the end of the course, students will be able to:

- Explain the role of data science in research and applied decision-making;
- Distinguish between major types of learning, data, models, and AI systems;
- Load, explore, clean, and restructure data for downstream analysis using Python;
- Create simple derived variables and prepare a modeling dataset;
- Estimate and interpret regression models for classification tasks;

- Understand the intuition and practical use of random forest and other ensemble models;
- Compare models using confusion matrices, accuracy, precision, recall, F1 score, ROC curves, and ROC AUC;
- Understand the basic logic of a multilayer perceptron for tabular prediction;
- Develop basic technical understanding of Neural Networks and Large Language Models.

Software and Materials:

- Python environment: students should install a working Python setup before the first practical session (VS Code or equivalent, with the required scientific packages – numpy, pandas, sklearn, etc.).
- AI configuration: students should be capable to use some sort of AI code assistant tool (CODEX, Claude Code, OCA or equivalent).
- Core reading: Burkov, A. (2019). The Hundred-Page Machine Learning Book.
- Course materials: lecture slides, practice scripts, datasets, and supplementary notes distributed through the course page.

Assessment:

- Participation and attendance: 15%
- Midterm assessment: 25%
- Data science project: 25%
- Final exam: 35%

Mandatory Completion Policy
Note that all mandatory assignments and exams must be completed to the best of your ability in order for your final grade to be issued. Failure to complete a mandatory assignment or exam may result in a failing grade.

Practice Component:

The practical materials are aligned with the lecture progression. Students first learn to load, explore, and clean a dataset; then implement ML models to reveal data patterns (including neural-network models) and interpret the results of different models; then make predictions and finally evaluate multiple models with different evaluation methods, including ROC-based tools.

Letter Grade	Percentage	Description
A+	97-100	Excellent Work
A	93-97	Outstanding Work
A-	90-92	
B+	87-89	Good work
B	83-86	
B-	80-82	
C+	77-79	Acceptable Work
C	73-76	
C-	70-72	
D+	67-69	Work that is significantly below average
D	63-66	
D-	60-62	
F	0-59	Work that does not meet the minimum standards for passing the course

AEP Academic Integrity Policy
Plagiarism and other forms of academic dishonesty are not tolerated. The use of Artificial Intelligence (AI) for the development of knowledge and learning is encouraged at many stages of the learning process. While we value technology for educational purposes, we also value originality and the retainment of knowledge, and thus using AI for assignments and examinations, even if rephrased, is strictly prohibited and considered an academic integrity violation, unless the instructor explicitly allows for it in the context of evaluated work

AEP Non-Discrimination/Harassment Policy
The AEP program in Prague promotes a diverse learning environment where the dignity, worth, and differences of each individual are valued and respected. Discrimination and harassment, whether based on a person's race, gender, sexual orientation, color, religion, national origin, age, disability, or other legally protected characteristics, are repugnant and completely inconsistent with our objectives. Retaliation against individuals for raising good faith claims of harassment and/or discrimination is prohibited.

AEP Diversity Policy
AEP is committed to fostering an inclusive and welcoming community that values diversity in all its forms. We believe that one of the most meaningful lessons of studying abroad is learning to navigate and appreciate differences with curiosity and an open mind. While engaging across differences can sometimes be challenging or uncomfortable, these moments are essential for growth and learning. We recognize that every member of our community, even with the best intentions, may occasionally make missteps. Our commitment is to provide a supportive environment where respectful and honest dialogue helps us learn from these experiences, ensuring that every student has the opportunity to thrive and broaden their perspective.

Weekly Schedule

Week 1
CEE Introductory Lecture Series

AEP Introductory Lecture Series

Week 2
Introduction to Data Science and Course Orientation

Lecture
- Data science as a profession
- Machine learning as coding plus statistics
- Overview of the course workflow and expectations

Practice
- Orientation to course tools and working structure

Week 3
Programming Setup and Python Workflow

Lecture
- Python environments, IDE setup, scripts, and packages
- Reproducible working habits for data analysis

Practice
- Environment setup and basic workflow in Python

Week 4
Key Notations, Definitions, and Learning Paradigms

Lecture
- Structured vs. unstructured data
- Supervised, unsupervised, semi-supervised, and reinforcement learning
- Classical AI vs. generative AI

Practice
- Discussion of data and model categories through examples

Academic Experience in Prague

Week 5
Working with Data and Data Ethics

Lecture
- Data sources, privacy, confidentiality, copyright, and security
- Data quality and the role of judgment in applied analytics

Practice
- Examples of data quality issues and practical discussion

Week 6
Data Cleaning and Dataset Preparation

Lecture
- Importing data, joins vs. appends, missing values, duplicates
- Renaming, restructuring, and basic feature engineering

Practice
- Practice Session 1: loading, exploring, cleaning, and feature engineering

Week 7
Fundamental Modeling Concepts and OLS

Lecture
- Regression intuition and prediction logic
- OLS as a bridge from statistics to machine learning

Practice
- Discussion of regression-based prediction examples

Week 8
Logistic Regression, Interpretation and Regularization

Lecture
- Probabilities, odds, log-odds, and classification targets
- Interpreting logistic regression output
- Limits of linear decision boundaries and the role of regularization
- Train-test split and classification setup

Practice
- Practice Session 2: Logistic Regression in Python

Week 9
Decision Trees and Random Forest

Lecture
- Tree logic, ensemble intuition, and overfitting
- Why random forest performs well on tabular data

Practice
- Practice Session 3: random forest and comparison with logistic regression

Week 10
Introduction to Neural Networks

Lecture
- Artificial neurons, hidden layers, and multilayer perceptrons
- Strengths, trade-offs, and orientation to modern AI and large language models

Practice
- Introductory neural-network example for tabular data

Week 11
Model Evaluation I

Lecture
- Training, validation, and test sets
- Confusion matrix, accuracy, precision, recall, and F1 score

Practice
- Comparing predictive models using standard classification metrics

Week 12
Model Evaluation II and ROC Analysis

Lecture
- ROC curve, ROC AUC, threshold tradeoffs, and comparative evaluation
- Course wrap-up and synthesis across models

Practice
- Practice Session 4: ROC analysis and final model comparison

Week 13
Students Presentations of Data Science Projects