LOMONOSOV MOSCOW STATE UNIVERSITY

FACULTY OF ECONOMICS

APPROVAL

Dean of the Faculty of economics at Lomonosov MSU

_____/проф. А.А.Аузан/

«_____» _____2025 г.

COURSE SYLLABUS

«Machine Learning and Data Analysis-1»

Level of Higher Education

Master's Degree

Field of Study

38.04.01 Economics

Specialization Profile

Data Analysis in Economics

Form of Study

Full-time

Moscow, 2025

1. **Course name: «Machine Learning and Data Analysis-1»**
   a. Author(s): Иванов Михаил Алоизович, Машин Иван Сергеевич
   b. E-mail: ivanovma1@my.msu.ru,
   c. Level of Higher Education: Master's Degree
   d. Field of Study: Economics
   e. Language of materials: English
2. **Status and Place of the Discipline within the Core Curriculum of the Master's Program**
   a. Status: mandatory
   b. Semester: 2nd
3. **Entry Requirements for Acquiring the Discipline**
   a. probability theory, mathematical statistics, econometrics;
   b. Python programming;
   c. linear algebra.

Competency code	Indicators of achievement	Planned outcomes
ОПК-5. Capable of using modern information technologies and software tools in professional tasks.	ОПК-5.И-1. Applies general or specialized software packages for data processing, visualization, and analysis, including econometric analysis and simulation modeling.	ОПК-5.И-1.У-1. Able to apply general or specialized software packages (e.g., MS Excel, Eviews, Stata, SPSS, AnyLogic, Tableau, etc.) or programming languages (e.g., R, Python, etc.) for data processing, visualization, and analysis, including econometric analysis and simulation modeling, in line with professional goals.
ПК-8. Capable of applying various instrumental methods for calculating and analyzing socio-economic indicators.	ПК-8.И-2. Applies modern instrumental methods for calculating and analyzing socio-economic indicators in solving practical and/or research tasks.	ПК-8.И-2.З-1. Is familiar with data analysis platforms suitable for implementing selected calculation and analysis methods, including business intelligence platforms, BI systems, and data visualization platforms. ПК-8.И-2.У-1. Is able to apply instrumental methods for calculating and analyzing indicators, including utilization of various data analysis platforms and programming languages R and Python. ПК-8.И-2.У-2. Capacity to adapt existing instrumental methods for calculating and analyzing indicators to fit specific tasks, possibly through custom programming in R and Python.

Competency code	Indicators of achievement	Planned outcomes
МПК-1. Capable of developing econometric models and machine learning models for studying economic processes and phenomena, interpreting the results.	МПК-1.И-1. Applies advanced econometric toolsets and machine learning methods to construct models of economic processes and phenomena. Interprets the results of conducted modeling, derives conclusions, and makes recommendations based on them.	МПК-1.И-1.У-1. Proficiency in developing econometric models and models based on machine learning techniques. МПК-1.И-1.У-2. Skill in interpreting modeling results, whether econometric or derived from machine learning applications. МПК-1.И-1.У-3. Competence in drawing conclusions from modeling results—both econometric and machine learning-based—and making recommendations based on these findings.
МПК-1.И-1.У-1. Proficiency in developing econometric models and models based on machine learning techniques. МПК-1.И-1.У-2. Skill in interpreting modeling results, whether econometric or derived from machine learning applications. МПК-1.И-1.У-3. Competence in drawing conclusions from modeling results—both econometric and machine learning-based—and making recommendations based on these findings.	МПК-4.И-1. Analyzes and systematically arranges collected data, creating recommendations for decision-makers based on it.	МПК-4.И-1.У-1. Ability to establish logical connections within organized data. МПК-4.И-1.У-2. Competence in developing recommendations for managers responsible for decision-making in both governmental institutions at various levels and businesses.

4. Competencies Graduates Should Achieve

5. **Volume of the Discipline by Type of Activity**
   The total volume of the discipline is 6 credits, equivalent to 216 academic hours, distributed as follows:

   a. Auditorium Load: 56 hrs (lecture type: 28 hrs, seminar type: 28 hrs)

   b. Group Contact Work: 52 hrs

   c. Individual Contact Work: 0 hrs

   d. Independent Student Work: 108 hrs

6. **Study Format**
   Full-time, using the electronic educational environment of the Faculty of Economics at Lomonosov MSU, «ON.ECON».

7. Content of the Discipline Structured by Topics with Allocated Academic Hours and Types of Educational Activities

Topic	Workload by Type of Activity (in Academic Hours)				
	Total, hrs	Contact Hours with Instructor			Independent Student Work, Hours
		Lectures	Seminars	Group	
Topic 1. Main tasks of machine learning.		4	4		
Topic 2. Regression (linear models).		2	2	8	
Topic 3. Optimization methods.		2	2	4	
Topic 4. Classification (linear models)		2	2	4	
Topic 5. Model quality estimation.		2	2	4	
Topic 6. Kernel methods (regression).		2	2	4	
Topic 7. Kernel methods (classification).		2	2	4	
Topic 8. Clustering.		2	2	4	
Topic 9. Dimensionality reduction.		2	2	4	
Topic 10. Decision trees.		2	2	4	
Topic 11. Bootstrap and model averaging.		4	4	8	
Topic 12. Boosting.		2	2	4	
Final exam	4	4			
Total hours	216	28	28	52	108

5. Brief Topic Overview

1. **Main tasks of machine learning.** Ontological assumptions (core principles include lawfulness and intelligibility of the world). Classification of ML tasks: supervised (classification, regression) and unsupervised (clustering, dimensionality reduction). Examples of datasets and ML tasks. Data types (continuous, categorical). K nearest neighbors (KNN) for regression and classification. Python libraries for data manipulation, visualization and machine learning. Jupyter Notebook. Course overview.
   Seminar: Python setup, loading datasets, basics of Polars, Matplotlib, NumPy, KNN implementation.
   References: (probML2022, sec. 1, 16.1), (ISL2024/ISL2021, sec. 2)

2. **Regression (linear models).** Model setup, feature matrix, notion of embedding (feature vector). Maximum likelihood estimation with various noise distributions (Gaussian and Laplace).

Likelihood maximization: closed-form solution, gradient descent. Model quality estimation: Mean Squared Error, Mean Absolute Error.
Seminar: regression dataset, sklearn.
References: (probML2023, sec. 11), (ISL2024/2021, sec. 3)

3. **Optimization methods.** Optimization problems, log-likelihood optimization as an optimization problem. Closed-form solutions, iterative methods. Gradient descent and Newton's method. Properties of gradient-based methods, existence of gradient-free methods.
Seminar: dataset and a simple model that requires gradient methods; from-scretch implementation of gradient descent.
References: (probML2022, sec. 8), (EDO2021, sec. 4), (NumOpt2006, sec. 2)

4. **Classification (linear models).** Logistic regression (for 2 classes, for 3+ classes), maximum likelihood estimation (distributions: Bernoulli and categorical). Logistic sigmoid, softmax. Likelihood maximization: gradient ascent. Model quality estimation: accuracy, precision, recall, F-мера, ROC-AUC, …
Seminar: classification dataset (Fisher's Iris), sklearn, confusion matrix.
References: (probML2022, sec. 10)

5. **Model quality estimation.** Overfitting (example with polynomial features), train/test split, cross-validation. Regularization (Ridge, LASSO), regularized estimators as maximum aposteriori estimators.
Seminar: visualization of overfitting, cross-validation-based hyperparameter search.
References: (ISL2024/2021, sec. 6.2), (probML2022, sec. 4.5)

6. **Kernel methods (regression).** Dual problem for least-absolute-values estimation (noise has Laplace distribution) with nonlinear features, kernels in the solution. Quadratic programming. Examples of kernels, their properties, infinite-dimensional feature spaces.
Seminar: regression dataset.
References: (PRML2006/2020, sec. 6)

7. **Kernel methods (classification).** Support vector machine. Classification problem for 2 classes (using linear algebra, margin maximization). Hard SVM for linearly separable data, soft SVM (non-separable data, allows misclassification). SVM loss function as regularized hinge loss. Optimization problem, solution methods: gradient descent, quadratic programming.
Seminar: classification dataset, sklearn, visualization of decision boundaries.
References: (MLPP2012, sec. 14.5), (ISL2024/2021, sec. 9), (PRML2006/2020, sec. 7)

8. **Clustering.** Problem statement, K-means, DBSCAN algorithms. Finite Gaussian mixtures, EM-algorithm for maximum likelihood estimation.
Seminar: datasets for clustering and mixtures (Old Faithful?), from-scratch k-means implementation, k-means and mixtures in sklearn.
References: (ISL2024/2021, sec. 12.4), (PRML2006/2020, sec. 9), (probML2022, sec. 21), (MLPP2012, sec. 11)

9. **Dimensionality reduction.** Problem statement, algorithms: Principal Component Analysis, matrix factorization. Using resulting lower-dimensional embeddings in downstream models.
Seminar: high-dimensional dataset.
References: (ISL2024/2021, sec. 12.2), (probML2022, sec. 20)

10. **Decision trees.** Binary trees as sequence of if/else instructions. Infeasibility of gradient methods for tree estimation. Algorithms for building locally optimal trees, CART. Interpreting tree-based models.
Seminar: tree construction, visualization, interpretation, hyperparameter tuning.
References: (ISL2024/2021, sec. 8), (probML2022, sec. 18).

11. **Bootstrap and model averaging.** Bootstrap for approximating estimators' sampling distributions, alternative to asymptotic approaches. Bootstrap confidence intervals, bootstrap forecasts, using the bootstrap to decrease estimates' variance (Bootstrap AGGregatING). Stacking. Random forest.
Seminar: bootstrap confidence intervals for logistic regression parameters, building random forests.
References: (PRML2006/2020, sec. 14), (probML2022, sec. 4.7.3, 18).

12. **Boosting.** Boosting for arbitrary models, including decision trees. Boosting for the mean squared error loss. Connection between boosting and gradient descent.
Seminar: sklearn, CatBoost.
Литература: (probML2022, sec. 18.5).

Evaluation Measures for Learning Results

Learning outcomes	Assessment tools
ОПК-5.И-1.У-1. Can apply general or specialized application software packages (such as MS Excel, Eviews, Stata, SPSS, AnyLogic, Tableau, etc.) or programming languages (like R, Python, etc.) designed for data processing, visualization, and analysis, including econometric analysis and simulation modeling, according to one's professional responsibilities.	Homework assignments, project assignments, exam
ПК-8.И-2.У-1. Can use instrumental methods for calculating and analyzing metrics, including those involving various data analysis platforms and applying R and Python programming languages.	Homework assignments, project assignments, exam
ПК-8.И-2.У-2. Can adapt existing instrumentation methods for calculating and analyzing metrics to meet specific task requirements, including writing code in R and Python.	Homework assignments, project assignments, exam
МПК-1.И-1.У-1. Can develop econometric models and models based on machine learning techniques.	Homework assignments, project assignments, exam
МПК-1.И-1.У-2. Can interpret the results of both econometric and machine learning-based modeling.	Homework assignments, project assignments, exam
МПК-1.И-1.У-3. Can draw conclusions from the results of both econometric and machine learning-based modeling and provide recommendations based on these findings.	Homework assignments, project assignments, exam
МПК-4.И-1.У-1. Can establish logical interconnections among collected information.	Homework assignments, project assignments, exam
МПК-4.И-1.У-2. Can create recommendations for decision-making bodies involved in public administration at various levels and in business sectors.	Homework assignments, project assignments, exam

8. **Grading System**
Maximum points per assessment form (current and intermediate evaluation):

Forms of current and intermediate evaluation	Points
Homework	140
Course project	100
Exam	60
Total	300

Final grade criteria:

Grade	Minimum score	Maximum score
Excellent	255	300
Good	195	254.9
Satisfactory	120	194.9
Fail	60	119.9

Note: If a student earns less than 20% of the possible score during the semester, they can only receive a passing grade ("Satisfactory") provided they earn at least 85% of the available points on the final exam covering all course material.

Typical tasks and other materials necessary to assess the learning outcomes:

Example of home assignment.
The data used is in the price.csv file. Select the resulting metric. Regarding empirical distribution for indicators, check with appropriate tests that it is statistically different from normal. Based on the result, conduct tests for comparing the means in 2 groups for the selected metric. Write a short report with the final decision for the pricing policy of company N. Submit this work in the Jupiter Notebook.

Typical Task for exam. For a binary classifier with a threshold of 0.5 and a given sample (y, p), where y are the true responses and p are the corresponding estimates of the probabilities of belonging to the positive class, as output by the classifier.
1. Calculate:
accuracy
- *precision*
- *recall*
- *f1-мepy*
- *auc roc*

1. Calculate *roc-curve*

Y	p
1	0.6
0	0.25
1	0.25
0	0.9
0	0.8

1	0.9
0	0.6
0	0.4
0	0.3
1	0.7

Example of project assignment.
The project is a short text (up to 15 pages) containing a statement of the problem, a brief review of the literature, an analysis of the theoretical model (or models) with premises and conclusions, as well as a selection of examples from empirical articles illustrating the conclusions of the model and own calculations. Selection of data and their description is required. Training data should be taken from open sources.
The task is performed in groups (of 2-3) or individually.
Submit this work in the Jupiter Notebook. Yandex DataSphere service is recommended.

Methodological guidelines and assignment requirements:

Home assignments are practical tasks aimed at consolidating the skills of machine learning and data analysis in Python within various topics. The task is submitted in the Jupyter-notebook .ipynb format with the code, comments and answers to the questions of the task. The comments should describe the methods and data used to solve the problem, justify the choice of the algorithm in detail, and present the results of calculations (using tables and figures if necessary).
Example of project assignment.
The project is a short text (up to 15 pages) containing a statement of the problem, a brief review of the literature, an analysis of the theoretical model (or models) with premises and conclusions, as well as a selection of examples from empirical articles illustrating the conclusions of the model and own calculations. Selection of data and their description is required. Training data should be taken from open sources.
The task is performed in groups (of 2-3) or individually.
Submit this work in the Jupiter Notebook. Yandex DataSphere service is recommended.

Information Resources for the Discipline

a. Primary references

i. (ISL2024) Джеймс Г., Уиттен Д., Хасти Т., Тибширани Р., Тейлор Дж. Введение в статистическое обучение с примерами на языке Python / пер. с англ. А.Ю. Гинько. -М.: ДМК Пресс, 2024. - 846 с.: ил.

ii. (PRML2020) Бишоп, Кристофер М. Распознавание образов и машинное обучение. : Пер. с англ. — СПб. : ООО «Диалектика», 2020. — 960 с. : ил. — Парал. тит. англ.

iii. (probML2022) Мэрфи К.П. Вероятностное машинное обучение: введение / пер. с англ. А.А. Слинкина. - М.: ДМК Пресс, 2022. - 940 с.: ил.

iv. (ISL2021) James G., Witten D., Hastie T., Tibshirani R., «An Introduction to Statistical Learning with Applications in R», Springer, 2021. https://doi.org/10.1007/978-1-0716-1418-1

v. (PRML2006) Bishop C.M., «Pattern Recognition and Machine Learning», Springer, 2006. ISBN-10: 0-387-31073-8.

vi. (MLPP2012) Murphy K.P., «Machine Learning: a Probabilistic Perspective», The MIT Press, 2012. ISBN 978-0-262-01802-9.

vii. (probML2022) Murphy K.P., «Probabilistic Machine Learning: an Introduction», The MIT Press, 2022.

viii. (EDO2021) Joaquim R.R.A. Martins and Andrew Ning. Engineering Design Optimization. Cabridge University Press, 2021. ISBN: 9781108833417. https://mdobook.github.io

ix. (NumOpt2006) Nocedal J., Wright S.J. «Numerical Optimization», Springer, 2006. ISBN-10: 0-387-30303-0.

b. Documentation:

i. Polars: https://docs.pola.rs/api/python/stable/reference/index.html

ii. NumPy: https://numpy.org/doc/

iii. Matplotlib: https://matplotlib.org/

iv. scikit-learn: https://scikit-learn.org/stable/index.html

Description of material and technical support

To organize classes in the discipline the following technical training tools are needed: a computer class with a projector and a blackboard.

Language of instruction: *English*

Professor (professors): assistant professor MSU M. Ivanov (М.А. Иванов), Salyutdevices - Head of Development Group I. Mashin (И.С. Машин)

Syllabus authors: assistant professor MSU M. Ivanov (М.А. Иванов), Salyutdevices - Head of Development Group I. Mashin (И.С. Машин)