Національний технічний університет України
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

Department of Applied Mathematics

MACHINE LEARNING

Curriculum of the academic discipline (Syllabus)

Course details	
Level of higher education	***Second (Master's)***
Field of knowledge	*05 Social and behavioural sciences*
Specialisation	*054 Sociology*
Educational programme	*Social Data Analytics*
Status of discipline	*Elective*
Form of study	*Full-time (day)*
Year of study, semester	*First year, 2nd semester*
Scope of the discipline	*4 ECTS credits/120 hours:* *18 hours of lectures, 36 hours of practical classes, 66 hours of independent work.*
Semester assessment/assessment measures	Test/ Modular control work
Class schedule	*rozklad.kpi.ua* *1 hour of lectures and 2 hours of computer workshops per week*
Language of instruction	*Ukrainian*
Information about course director/teachers	*Lecturer: Prof. Oleg Romanovich Chertov, Doctor of Technical Sciences, oleg.r.chertov@gmail.com* *Computer workshops: Volodymyr Viktorovych Malchikov, mavr2k@gmail.com*
Course location	*Electronic resource "KPI Campus" https://ecampus.kpi.ua/*

Curriculum

1. Description of the academic discipline, its purpose, subject matter and learning outcomes

The aim *of the course is to develop a modern scientific worldview in students in the field of machine learning methods, to familiarise them with a number of machine learning models and to explain aspects of their application.*

The subject *of the course is methods of supervised and unsupervised learning, cross-validation and feature selection.*

After completing the course, students should demonstrate the following learning outcomes

learning:

- select a set of features (factors) for classification or regression and perform preliminary data processing, select the type of machine learning model depending on the task being solved;

- be able to formulate and solve basic machine learning tasks.

According to the educational and scientific programme, mastering the discipline contributes to the strengthening of the following competencies and programme learning outcomes:

- *LC 01 Ability to think abstractly, analyse and synthesise.*
- *FC 03 Ability to design and conduct sociological research, develop and justify its methodology.*

- *FC 04 Ability to collect and analyse empirical data using modern methods of sociological research.*
- *FC 11 Ability to analyse open source data (OSINT), analyse qualitative information, textual data, and use intellectual analysis for social data.*
- *PRN 01 Analyse social phenomena and processes using empirical data and modern concepts and theories of sociology.*
- *PRN 04 Apply scientific knowledge, sociological and statistical methods, digital technologies, and specialised software to solve complex problems in sociology and related fields of knowledge.*
- *PRN 05 Search for, analyse and evaluate necessary information in scientific literature, databases and other sources.*
- *PRN 09 Plan and carry out scientific research in the field of sociology, analyse results, and justify conclusions.*
- *PRN 12 Analyse open source intelligence (OSINT), analyse qualitative information and text data, and use intelligent analysis for social data.*

2. Prerequisites and post-requisites of the discipline (place in the structural-logical scheme of training under the relevant educational programme)

Prerequisites: mastery of university courses in higher mathematics (algebra, mathematical analysis), probability theory, mathematical statistics, and programming is required.

Post-requisites: studying the discipline contributes to the deepening of students' analytical and research training, forms skills in building, evaluating and interpreting models for data analysis, which can be used when mastering professional disciplines, performing practical tasks and preparing a master's thesis.

3. Contents of the course

Section 1. Controlled learning

Topic 1.1. Introduction
Topic 1.2 Classification metrics
Topic 1.3 Decision trees
Topic 1.4. Metric classification and regression methods

Section 2. Unsupervised learning. Complex models

Topic 2.1. Clustering
Topic 2.2. Feature engineering
Topic 2.3. Model Combination

4. Teaching materials and resources

Basic literature

1. *Kononova K. Yu. Machine learning: methods and models. – Kharkiv: V. N. Karazin Kharkiv National University, 2020. – 301 p. (https://github.com/katerynakononova/ML/blob/master/ML.pdf)*
2. *Intellectual Data Analysis and Machine Learning. Part 1: Basic Methods and Means of Data Analysis / Y. V. Ivanchuk et al. / Vinnitsa National Technical University. – Vinnitsa: VNTU, 2021. - 68 p. (http://pdf.lib.vntu.edu.ua/books/2022/Ivanchuk_P1_2021_69.pdf).*

Supplementary literature

1. *Burkov, A. (2019) The Hundred-Page Machine Learning Book (https://themlbook.com/).*
2. *James, G.M., Witten, D.M., Hastie, T.J., & Tibshirani, R. (2021). An Introduction to Statistical Learning. Springer Texts in Statistics (https://hastie.su.domains/ISLR2/ISLRv2_website.pdf).*

Educational content

5. Methodology for mastering the academic discipline (educational component)

5.1. Lectures

Lecture 1 Introduction to machine learning

Definition of machine learning (according to Tom Mitchell). Types of learning and classes of machine learning tasks. Basic machine learning scheme.

Lectures 2-3 Classification metrics

Classification metrics: confusion matrix and first- and second-order errors, accuracy, precision, and recall, F-scores, Matthews correlation coefficient (MCC), precision-recall curve, ROC (receiver operating characteristic) curve, area under curve (AUC). Binary and multi-class cases.

Lectures 4 Decision trees

Using decision trees for classification. Exponential size of the hypothesis space. Classification of continuous data using trees. Greedy tree construction algorithm. Problems with attributes with many values. Building a decision tree.

ID3 algorithm.

Comparison of splitting criteria:

- *Information Gain*
- *Gini impurity*
- *(Information) Gain ratio (normalised information gain)*
- *Classification error.*

Advantages and disadvantages of decision trees.

Tree pruning (pre-pruning, post-pruning). What is it used for?

Lectures 5 Metric classification methods

The difference between regression and classification tasks. Distance-based vs similarity-based metrics. Greedy/lazy algorithms. Compactness and continuity hypotheses. K-nearest neighbours algorithm. Which value of k is better? Weight selection. Voronoi diagram. Noise sensitivity. Feature normalisation. Examples of metrics (Minkowski metric, cosine measure, Jaccard distance, Hamming distance).

Lectures 6 Clustering

Clustering as an example of unsupervised learning. Types of clustering algorithms. Cluster characteristics. Clustering stopping criteria. K-means method. How to determine the number of clusters? Hierarchical agglomerative clustering.

Lecture 7 Feature engineering

Feature selection: identifying non-informative features using filters; wrapping methods; methods built into the learning model.

Lecture 8 Model combinations

General scheme of model ensembles. What is the advantage of an ensemble over a single classifier? Bagging. Examples of aggregation functions. Simple and weighted voting. Random forest. Boosting.

Lecture 9 A typical approach to solving machine learning problems

Demonstration of typical approaches to data collection and cleaning, filling in missing values, exploratory data analysis, selecting informative features, choosing machine learning models, and optimising their hyperparameters using the example of Jupyter Notebook for predicting the number of passengers who survived after the sinking of the Titanic.

5.2. Practical classes (computer workshops)

1. *Practical class 1-3. Data analysis using Python.*
2. *Practical classes 4-6. Application of classification metrics.*

3. *Practical classes 7-9. Decision trees.*
4. *Practical classes 10-13. Metric classification and regression methods. The k-nearest neighbours algorithm.*
5. *Practical classes 14-16. Clustering.*
6. *Practical session 17. Modular control work.*
7. *Practical class 18. Test.*

6. Independent work

- *Preparation for defending assignments for computer workshops - 56 hours.*
- *Preparation for the modular control work - 4 hours.*
- *Preparation for the test - 6 hours*

Policy and control

7. Academic discipline policy (educational component)

- *Attendance at lectures and computer workshops is highly desirable.*
- *During lectures, students must listen attentively and refrain from making noise. It is permitted (and recommended!) to ask "good" questions to clarify unclear points and to answer the lecturer's questions. Mobile phones, tablets, smart watches and other gadgets must be turned off or set to silent mode; their use is not permitted unless they are used for educational purposes.*
- *Computer workshops are conducted using one of the programming languages (Python is recommended), and the possibility of using third-party libraries is determined separately for each workshop. Computer workshops are defended orally, in person or via conference (for distance learning). During the defence, the student must describe the course of the work, formulate conclusions and answer the teacher's questions. Work must be submitted and defended on time; late submissions will result in a lower grade.*
- *A Modular control work is carried out during one of the practical classes.*
- *While studying the discipline, students must adhere to the rules of academic integrity, which specifically prohibits plagiarism, cheating, and other ways of passing off someone else's work as their own. Failure to comply with these rules will result in punishment, including removal of the offender(s) from the relevant assessment and a grade of "0" for that assessment. If signs of cheating are found during the verification of computer practicals or Modular control work, a grade of "0" will be given to both the person who cheated and the person who allowed the cheating.*

8. Types of control and the rating system for assessing learning outcomes (RSO)

A student's grade for a subject consists of the points they receive:

1) for completing, submitting and defending computer practicals;

2) for completing modular control work (MCW);

RATING POINT SYSTEM

1. Completion, submission and defence of computer practicals

During the semester, the student must prepare, submit and defend 5 computer practicals to the teacher. When evaluating computer practicals, as when solving real practical problems for a specific customer, the performance of the programme, the timeliness of its delivery and the ability to explain the work of the programme and interpret the results obtained are taken into account.

Academic integrity is checked separately – students must prove to the teacher that they understand what exactly is done in the programme and how, and can explain the features of the code and the reasoning behind the decisions made during the programme implementation.

For completing and submitting one computer workshop, the student can receive points for each of the following three components:

1) (timeliness): if the computer workshop is sent to the teacher no later than the agreed deadline and passes the basic testing without errors (runs without errors and gives the correct result for at least one standard test example), then 4 points are awarded for this;

2) (operability):

- *if the programme runs without errors, 3 points are awarded;*
- *if the programme has minor flaws and sometimes gives incorrect results, 1 to 2 points are awarded;*
- *if the programme works incorrectly in most cases, it is awarded 0 points;*

3) (presentability):

- *if the student answers questions about the functionality of the programme and the features of its implementation correctly or almost correctly, 6 to 7 points are awarded;*
- *if the student is partially confused when answering questions about the functionality of the programme and the features of its implementation, then 3 to 5 points are awarded for this;*
- *if the student is partially confused when answering questions about the functionality of the programme and cannot justify the choice of tools for its implementation, then 1 to 2 points are awarded for this;*
- *if the student cannot explain the basic functionality of the programme and the features of its implementation, i.e. there is a clear violation of academic integrity , then all points for the corresponding computer workshop are reset to zero and, in general, the student receives 0 points for it;*
- *In order to understand how well the student has understood the theoretical material corresponding to the computer workshop, the teacher may ask the student several questions on the relevant theory and, depending on the correctness of the answers, increase or decrease the grade for "presentability".*

Thus, for one computer workshop, a student can receive a maximum of:

*4 points + 3 points + 7 points = **14 points**.*

In total, for a cycle of computer workshops, a student can receive a maximum of:

*14 points × 5 = **70 points (R_{Lab})**.*

The lecturer has the right to reward the student with a certain number of points (maximum – 5 points per semester) for providing an interesting and original algorithmic or software implementation of the computer workshop task.

<u>2. Modular control work</u>

*Weighting score – **30 points (R_{CW})** for a one-hour Modular control work.*

The test includes two types of tasks:

1) perform certain calculations (construct a PR or ROC curve, etc.);

2) select the correct answer(s) from those provided and justify your choice.

Each task of the test is evaluated (depending on its complexity) in two, three or five points.

The student receives the maximum number of points if they provide a complete and correct solution to the problem or make minor errors/mistakes that do not significantly affect the solution.

A student receives less than the maximum number of points (usually half of the maximum number) if the solution provided is correct but incomplete (not all correct answers were given), or if the solution process is correct but the student made mistakes that significantly affected the answer.

The student receives zero points if the problem is not solved at all, or if the solution process contains gross errors, or if only the answer is given without justification.

3. *<u>Bonus points are awarded for:</u>*

- *a presentation at a lecture (by prior agreement with the lecturer) with a report on the course material (or similar) or with a report on personally obtained scientific results in machine learning or its applications, up to 5 bonus points are awarded;*
- *active participation in lectures, i.e.*
  - *answers to the lecturer's questions to the general audience,*
  - *finding typos/errors in lectures,*
  - *asking "the right questions", i.e. questions that demonstrate the student's thoughtful work with the course material*

  a total of up to 5 bonus points are awarded.

4. *<u>Recognition of learning outcomes obtained in informal education, etc.</u>*

For completing a specific element of informal education (e.g., Coursera online courses or similar), for winning or participating in thematic hackathons or other competitions, students may be awarded additional points, credited with completing computer practicals for the entire discipline with the highest possible grade, or, in general, awarded 100 points.

The specific amount of the incentive for the student is determined by the lecturer based on the completeness, importance, and results of the student's completion of the relevant elements of informal education. The recognition of the results of informal education takes place in accordance with the procedure set out in the relevant Regulations of Igor Sikorsky KPI: https://osvita.kpi.ua/node/179

Calculation of the rating scale (R)

The sum of the weighted points for control measures during the semester is:

$$RC = R_{Lab} + R_{CW} = 70 \text{ points} + 30 \text{ points} = 100 \text{ points.}$$

An indispensable condition for admission to the exam is that the student must earn 30 points during the semester. Otherwise, the student will receive a grade of "not admitted" in the first report card and must then fulfil these conditions (earn points for admission). After fulfilling the conditions for admission to the exam, the student writes a test during the first and/or second retake.

If, at the end of the semester, the student has earned at least 60 rating points and has fulfilled the conditions for admission to the exam, they will automatically receive a passing grade in accordance with the table below.

The student may attempt to improve their grade by writing a credit test, in which case their points earned during the semester will be cancelled.

If the total rating points are less than 60, but the conditions for admission to the semester exam have been met, the student takes the exam, and their rating is cancelled, after which points are awarded based on the results of the exam.

The test is conducted as a colloquium with 5 written and oral tasks. Each task is worth 20 points.

Criteria for evaluating answers to each task:

20 points — correct and meaningful answer;

18–19 points — correct, meaningful answer, but with minor flaws;

14–17 points — the answer contains minor errors or is incomplete;

12–13 points — the answer contains significant errors and is incomplete;

0 points — no answer.

To obtain the appropriate grades, the student's rating score R_D *is converted according to the following table:*

Rating points, R_D	***Grade***
95–100	*Excellent*
85–94	*Very good*
75–84	*good*
65–74	*satisfactory*
60–64	*sufficient*
Total score < 60	*unsatisfactory*
Semester rating < 30	*not admitted*

Work programme for the academic discipline (syllabus):

Prepared *by the head of the Department of Applied Mathematics, Doctor of Technical Sciences, Prof. CHERTOV O. R.*

Approved *by the Department of Applied Mathematics (Minutes No. 18 of 10.06.2024).*

Approved by *the Methodological Commission of the Faculty of Applied Mathematics (Minutes No. 12 of 21.06.2024).*