Federal State Autonomous Educational Institution of Higher Education
Peoples' Friendship University of Russia named after Patrice Lumumba

Academy of Engineering
(name of the main educational unit (MEU) that developed the educational program of higher education)

WORKING PROGRAM OF THE DISCIPLINE

MACHINE LEARNING AND BIG DATA MINING
(name of discipline/module)

Recommended for the field of study/specialty:

27.04.04 CONTROL IN TECHNICAL SYSTEMS
(code and name of the field of study/specialty)

The discipline is mastered within the framework of the implementation of the main professional educational program of higher education (EP HE):

Artificial Intelligence, Machine Learning, and Space Science
(name (profile/specialization) of the educational institution of higher education)

2026

1. THE GOAL OF MASTERING THE DISCIPLINE

The course "Machine Learning and Big Data Mining" is part of the Master's program "Artificial Intelligence, Machine Learning, and Space Sciences" in the 27.04.04 "Control in Technical Systems" program and is studied in the first semester of the first year. The course is offered by the Department of Mechanics and Control Processes. It consists of five sections and 20 topics and focuses on the foundations of modern technological civilization and global trends in the changing scientific worldview, types of scientific rationality, value systems that guide scientists, an analysis of the main ideological and methodological issues arising in science at the current stage of its development, and an analysis of the main methods for solving typical problems and an introduction to their application in professional activities.

The purpose of mastering the discipline is to develop fundamental knowledge and skills in applying problem-solving methods necessary for professional activity, to increase the general level of literacy of students in the discipline of history and methodology of science, to form an understanding of the trends in the historical development of science, as well as a modern understanding of the organization of research activities in the chosen field.

2. REQUIREMENTS FOR THE RESULTS OF MASTERING THE DISCIPLINE

Mastering the discipline "Machine Learning and Big Data Analysis" aimed at developing the following competencies (parts of competencies) in students:

Table 2.1. List of competencies developed in students while mastering the discipline (results of mastering the discipline)

Cipher	Competence	Indicators of Competency Achievement (within this discipline)
GPC-5	Capable of conducting patent research, determining forms and methods of legal protection and defense of rights to the results of intellectual activity, and managing rights to them to solve problems in the development of science, engineering, and technology	GPC-5.1 Knows the methods and approaches to conducting patent research, forms and methods of legal protection and defense of rights to the results of intellectual activity; GPC-5.2 Able to manage rights to the results of intellectual activity to solve problems in the field of development of science, engineering and technology; GPC-5.3 Has knowledge of methods and approaches to conducting patent research, knows the methods of legal protection and defense of rights to the results of intellectual activity.
GPC-6	Capable of collecting and analyzing scientific and technical information, generalizing domestic and foreign experience in the field of automation and control equipment	GPC-6.1 Knows the basic methods of collecting and analyzing scientific and technical information; GPC-6.2 Able to analyze and generalize domestic and foreign experience in the field of automation and control equipment; GPC-6.3 Has mastered the methods of collecting and analyzing scientific and technical information, and can also generalize domestic and foreign experience in the professional field.

3. PLACE OF THE DISCIPLINE IN THE STRUCTURE OF THE EDUCATIONAL EDUCATIONAL INSTITUTION

Machine Learning and Big Data Analysis refers to the mandatory part of block 1 “Disciplines (modules)” of the educational program of higher education.

As part of the higher education program, students also master other disciplines and/or practices that contribute to the achievement of the planned results of mastering the discipline "Machine Learning and Big Data Analysis".

Table 3.1. List of components of the educational program of higher education that contribute to the achievement of the planned results of mastering the discipline

Cipher	Name of competence	Previous courses/modules, practical training*	Subsequent disciplines/modules, practices*
GPC-6	Capable of collecting and analyzing scientific and technical information, generalizing domestic and foreign experience in the field of automation and control equipment		Research work / Scientific research work; Undergraduate Training; Advanced Methods of Earth Remote Sensing;
GPC-5	Capable of conducting patent research, determining forms and methods of legal protection and defense of rights to the results of intellectual activity, and managing rights to them to solve problems in the development of science, engineering, and technology		Research work / Scientific research work; Undergraduate Training; Dynamics and Control of Space Systems;

* - filled in accordance with the competency matrix and the SUP EP HE

** - elective courses/practices

4. SCOPE OF THE DISCIPLINE AND TYPES OF EDUCATIONAL WORK

The total workload of the course "Machine Learning and Big Data Analysis" is 5 credits.

Table 4.1. Types of educational work by periods of mastering the educational program of higher education for full-time education.

Type of academic work	**TOTAL,**academic hours		**Semester(s)**
			1
Contact work, academic hours	*34*		*34*
Lectures (LC)	17		17
Laboratory work (LW)	17		17
Practical/seminar classes (SC)	0		0
Independent work of students, academic hours	*110*		*110*
Control (exam/test with assessment), academic hours	*36*		*36*
Total complexity of the discipline	**academic hours**	***180***	*180*
	credit	***5***	*5*

5. CONTENT OF THE DISCIPLINE

Table 5.1. Content of the discipline (module) by types of academic work

Section number	Name of the discipline section	Topic Title		Topic Contents	Type of academic work*
Section 1	Introduction to Machine Learning and Data Processing. Software Tools for Data Mining and Machine Learning.	1.1	Introduction to Machine Learning and Data Processing. Formulation of Main Problem Classes in Machine Learning.	Machine learning as a scientific direction studying methods for constructing algorithms capable of learning from empirical data without strictly following predefined rules. Main components: features as independent variables, target variable as dependent variable, training sample. Classification of machine learning problems: supervised learning using labeled data, unsupervised learning in the absence of a target variable, reinforcement learning for behavior formation based on interaction with the environment.	LC, LW
		1.2	Regression and Classification; Clustering, Dimensionality Reduction.	Regression task as prediction of a continuous numerical value. Classification task as assigning an object to one of discrete categories. Clustering task as grouping objects based on their similarity in the absence of class labels. Dimensionality reduction task as transforming the feature space while preserving essential information and reducing the number of variables.	LC, LW
		1.3	Text Processing; Image Processing.	Methods for processing textual data: representing textual information numerically through the bag-of-words model and TF-IDF method. Main tasks of text analysis: document classification, sentiment analysis, entity extraction. Methods for image processing: representing visual information as pixel matrices. Main tasks of computer vision: object recognition, image segmentation, edge detection.	LC, LW
Section 2	Regression Analysis and Data Compression.	2.1	Regression Task. Minimizing the Squared Deviation. Regression Function: Conditional Mathematical Expectation.	Regression task as predicting the numerical value of a target variable based on input features. Minimizing the squared deviation as a standard approach to assessing regression quality. Regression function as the conditional mathematical expectation of the target variable given feature values. Relationship between the quadratic loss function and conditional mathematical expectation.	LC, LW
		2.2	Linear Regression and k-Nearest Neighbors Method. Overfitting and Underfitting.	Linear regression as a model assuming a linear relationship between features and the target variable, finding coefficients by minimizing mean squared error. k-nearest neighbors method as a non-parametric regression and classification method predicting based on the nearest objects in the training sample. Overfitting as excessive adjustment of the model to training data with loss of	LC, LW

Section number	Name of the discipline section	Topic Title		Topic Contents	Type of academic work*
				generalization ability. Underfitting as excessive simplification of a model unable to capture patterns in the data.	
		2.3	Decomposition of Error into Noise, Bias, and Variance.	Decomposition of total prediction error into three components: noise as irreducible error due to data randomness, bias as systematic error due to incorrect model assumptions about the data, variance as error due to model sensitivity to small fluctuations in the training sample. Bias-variance tradeoff: simple models have high bias and low variance, complex models have the opposite.	LC, LW
Section 3	Outlier and Anomaly Detection. Data Cleaning and Regularization Technologies.	3.1	Outlier and Anomaly Detection. What Are Outliers, Types of Outliers.	Outliers as data objects that differ significantly from the main body of observations. Types of outliers: point outliers with an anomalous value of one feature, contextual outliers anomalous under specific conditions, collective outliers as groups of objects deviating from the overall distribution. Causes of outliers: measurement errors, rare events, equipment failures.	LC, LW
		3.2	Outlier Detection Methods. Anomaly Search.	Statistical outlier detection methods based on the three-sigma rule and interquartile range. Distance-based methods: Mahalanobis distance, nearest neighbor method. Density-based methods: local outlier factor. Machine learning for anomaly search: one-class classification, isolation forest, autoencoders.	LC, LW
		3.3	Sample Censoring. Outlier Screening, Outlier Removal.	Sample censoring as the process of limiting extreme values to reduce the influence of outliers. Outlier screening as complete removal of anomalous records from the dataset. Criteria for outlier removal: exceeding threshold values, inconsistency with logical constraints, impossibility of verification. Risks when removing outliers: loss of useful information when mistakenly discarding rare but significant events.	LC, LW
		3.4	Data Cleaning and Regularization Technologies. Main Types of Regularization.	Data cleaning as a set of measures to identify and correct errors, omissions, and inconsistencies in data. Regularization as a technology for combating overfitting by adding a penalty term to the loss function. Main types of regularization: L1-regularization lasso leading to sparse solutions, L2-regularization ridge uniformly reducing weights, Elastic Net as a combination of L1 and L2.	LC, LW
		3.5	Dimensionality Reduction Method. Feature Selection Methods.	Dimensionality reduction as the process of reducing the number of features while preserving essential information. Feature selection methods: filter methods evaluating the importance of each feature independently, wrapper methods searching feature subsets based	LC, LW

Section number	Name of the discipline section	Topic Title		Topic Contents	Type of academic work*
				on model quality, embedded methods with feature selection within the learning process. Principal component analysis as a linear transformation of original features into new uncorrelated components.	
Section 4	Clustering and Classification Technologies. Neural Networks. Genetic Algorithms	4.1	Clustering and Classification Technologies. K-means. EM Algorithm.	Clustering as an unsupervised task of grouping objects based on their similarity. K-means method with iterative assignment of objects to nearest cluster centers and recalculation of centers. Choosing the number of clusters using the elbow method and silhouette coefficient. EM algorithm for probabilistic clustering assuming a mixture of Gaussian distributions. Alternating E-step for estimating object membership and M-step for recalculating distribution parameters.	LC, LW
		4.2	Other Clustering Methods. Classification Tasks. Bayesian Classifier.	Agglomerative hierarchical clustering with sequential merging of the nearest objects into clusters. DBSCAN as a density-based clustering method identifying regions of high density. Classification as the task of assigning an object to one of predefined classes. Bayesian classifier based on Bayes' theorem and the assumption of feature independence naive Bayes classifier.	LC, LW
		4.3	Linear Methods for Classification. Logistic Regression, Maximum Likelihood.	Linear classification methods with class separation by a linear boundary in feature space. Logistic regression as a binary classification method predicting class membership probability through a sigmoid function. Maximum likelihood as a learning criterion for logistic regression instead of minimizing the squared error. Interpretation of logistic regression coefficients as the influence of features on the log odds ratio.	LC, LW
		4.4	Neural Networks: General Architecture. Multilayer Networks. Backpropagation.	General architecture of a neural network: input layer of features, one or more hidden layers with non-linear activation functions, output layer with dimensionality matching the task. Multilayer networks as universal approximators of any continuous function. Forward pass for computing network outputs given input data. Backpropagation as an algorithm for computing gradients of the loss function with respect to network weights through sequential application of the chain rule from output layer to input layer.	LC, LW
		4.5	Stochastic Gradient Descent. Genetic Algorithms.	Stochastic gradient descent as an optimization method for training neural networks with weight updates based on a random mini-batch of examples rather than the entire dataset. Advantages of	LC, LW

Section number	Name of the discipline section	Topic Title		Topic Contents	Type of academic work*
				stochastic gradient descent: faster convergence, ability to work with large datasets, escaping local minima. Genetic algorithms as evolutionary optimization methods simulating natural selection processes. Basic operations of a genetic algorithm: selection of the best solutions, crossover for exchanging chromosome parts, mutation for random changes.	
Section 5	Feature Detection; Data Normalization. Fuzzy Sets. Bayesian Networks	5.1	Feature Extraction.	Feature extraction as the process of creating new informative features from raw data. Difference between feature extraction and feature selection: extraction creates new features, selection chooses a subset of existing ones. Feature extraction methods for images: corner and edge detectors, histograms of oriented gradients. Feature extraction methods for texts: n-grams, topic modeling, word vector representations.	LC, LW
		5.2	Feature Transformations. Data Normalization. Data Normalization Methods.	Feature transformations as bringing data to a form convenient for model training. Data normalization as scaling features to a specific range or distribution. Necessity of normalization for methods sensitive to feature scale: gradient descent, principal component analysis, distance-based methods.	LC, LW
		5.3	Min-Max Normalization. Z-score Normalization. Decimal Scaling.	Min-max normalization with linear transformation of features into a given range, usually from zero to one. Preservation of the original distribution shape with min-max normalization. Z-score normalization as centering and scaling to unit variance: subtracting the mean and dividing by the standard deviation. Transforming features to a standard normal distribution. Decimal scaling as dividing all feature values by a power of ten to shift the decimal point. Simplicity and reversibility of decimal scaling.	LC, LW
		5.4	Fuzzy Sets. Bayesian Networks. Bayesian Inference Tasks. Method for Constructing a Fuzzy Bayesian Network.	Fuzzy sets as an extension of classical set theory with element membership specified by a continuous function from zero to one. Membership functions for describing linguistic variables: high, medium, low. Bayesian networks as probabilistic graphical models representing dependencies between variables through directed acyclic graphs. Bayesian inference tasks: computing posterior probabilities of variables given observations. Method for constructing a fuzzy Bayesian network: combining fuzzy logic for working with imprecise data and Bayesian networks for probabilistic inference. Application in decision support systems with incomplete	LC, LW

Section number	Name of the discipline section	Topic Title		Topic Contents	Type of academic work*
				or fuzzy information.	

* - to be completed only for FULL-TIME education: LC – lectures; LW – laboratory work; SC – practical/seminar classes.

6. LOGISTIC AND TECHNICAL SUPPORT OF DISCIPLINE

Table 6.1. Material and technical support for the discipline

Audience type	Equipment of the auditorium	Specialized educational/laboratory equipment, software and materials for mastering the discipline (if necessary)
Lecture	A lecture hall equipped with specialized furniture, a whiteboard (screen), and multimedia presentation equipment.	
Computer class	A computer room for conducting classes, group and individual consultations, ongoing monitoring and midterm assessment, equipped with personal computers (in the amount of ____ units), a board (screen) and technical means for multimedia presentations.	
For independent work	A classroom for independent student work (can be used for seminars and consultations), equipped with a set of specialized furniture and computers with access to the Electronic Information System.	

* - the classroom for independent work of students MUST be indicated!

7. EDUCATIONAL, METHODOLOGICAL AND INFORMATIONAL SUPPORT OF THE DISCIPLINE

Main literature:

1. James, G. et al. An introduction to statistical learning. – Springer, 2013. – 426 pp

2. Trevor Hastie, Robert Tibshirani, et al., The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edition, 2017

3. Vyugin, V. V. Mathematical foundations of machine learning and forecasting: a tutorial / V. V. Vyugin. - Moscow: MCNO, 2014. - 304 p.

Further reading:

1. Bruce, P. C., & Bruce, A. (2017). Practical Statistics for Data Scientists: 50 Essential Concepts (Vol. First edition). Sebastopol, CA: O'Reilly Media

2. Molnar, C. (2018). iml: An R package for Interpretable Machine Learning

3. Explainable and interpretable models in computer vision and machine learning. (2018)

4. Combinatorics and probability theory, textbook, 99 pp., Raigorodsky, A. M., 2013

Resources of the information and telecommunications network "Internet":

1. RUDN University Electronic Library System and third-party electronic library systems to which university students have access based on concluded agreements

- Electronic library system of RUDN - ELS RUDN http://lib.rudn.ru/MegaPro/Web
- Electronic Library System "University Library Online" http://www.biblioclub.ru
- EBS Yurayt http://www.biblio-online.ru
- Electronic Library System "Student Consultant" www.studentlibrary.ru
- Electronic Library System "Troitsky Bridge"

2. Databases and search engines

- electronic fund of legal and regulatory documentation http://docs.cntd.ru/
- Yandex search engine https://www.yandex.ru/
- Google search engine https://www.google.ru/
- SCOPUS abstract database http://www.elsevierscience.ru/products/scopus/

Educational and methodological materials for independent work of students in mastering a discipline/module:*

1. Lecture course on the subject "Machine learning and big data analysis".

* - all teaching and methodological materials for independent work of students are posted in accordance with the current procedure on the discipline page in TUIS!

DEVELOPER:

Associate Professor		Saltykova Olga Alexandrovna
Position, DEPARTMENT	*Signature*	*Surname I.O.*

HEAD OF THE DEPART-MENT:

Head of Department		Razumny Yuri Nikolaevich
Position of the DEPARTMENT	*Signature*	*Surname I.O.*

HEAD OF THE EP HE:

Professor		Razumny Yuri Nikolaevich
Position, DEPARTMENT	*Signature*	*Surname I.O.*