Data Analysis with Python

Syllabus

Details of the academic discipline	
Level of higher education	**First (undergraduate)**
Branch of knowledge	12 Information technologies
Specialty	126 Information systems and technologies
Educational program	Integrated information systems, Information management systems and technologies, Information support of robotic systems
Discipline status	Academic discipline of professional and practical training (chosen by students)
Form of education	full-time/correspondence/distance
Year of training, semester	3rd year, autumn semester
Scope of the discipline	120 hours (36 hours – lectures, 18 hours – laboratory, 66 hours – individual work)
Semester control/ control measures	Assessment/assessment work
Lessons schedule	http://rozklad.kpi.ua/Schedules/ScheduleGroupSelection.aspx
Language of teaching	Ukrainian, English
Information about the course leader / teachers	Lecturer: Tymofieieva Yuliia Serhiivna, yulia.s.timofeeva@gmail.com Laboratory: Tymofieieva Yuliia Sergiivna, yulia.s.timofeeva@gmail.com
Placement of the course	https://campus.kpi.ua

Program of educational discipline

1. Description of the educational discipline, its purpose, subject of study and learning outcomes

Description of the discipline. During training, students will get acquainted with the basic concepts, methods and means of preliminary training, statistical processing, visualization and data analysis, the basics of machine learning. In the laboratory classes, you will learn to use the main libraries of the Python language for working with data (NumPy, SciPy, Matplotlib, Seaborn, Scikit-learn, Pandas). The quality control of the acquired knowledge is assesed in the form of a modular test work.

The subject of the academic discipline:methods and means of data analysis and their implementation in the Python language.

Interdisciplinary connections. Basic knowledge of the disciplines: programming, algorithms and data structures, mathematical analysis, discrete mathematics, probability theory.

The purpose of the educational discipline. The goal of the educational discipline is to provide students with theoretical knowledge and practical abilities to apply general methods and tools for data preparation, statistical processing, visualization and analysis, using the main libraries of the Python language for working with data.

The main tasks of the academic discipline

Knowledge:

- methods and means of software presentation of data;
- methods and algorithms for preliminary preparation of data for analysis;
- methods and tools of statistical processing and data visualization;
- basics of machine learning;
- the main functions of the Python language libraries for working with data.

Skills:

- effectively use Python language libraries for data processing, visualization and analysis;
- correctly choose methods and algorithms for data preprocessing;
- correctly choose the most informative ways of presenting data;
- use machine learning algorithms for classification, clustering, regression analysis, etc.
- effectively process, visualize and analyze data obtained as a result of own experiments and research.
- create Python programs to work with data

2. Pre-requisites and post-requisites of the discipline (place in the structural and logical scheme of training according to the relevant educational program)

Prerequisites: be able to use a computer, have basic knowledge of programming, data structures and algorithms, mathematical analysis, mathematical statistics, have experience in writing simple programs in the Python language.

Post-requisites: after completing the discipline, students will be able to effectively process, visualize and analyze data obtained as a result of their own experiments and research, apply basic machine learning algorithms, use NumPu, SciPy, Matplotlib libraries,Seaborn, Scikit-learn, Pandas.

3. Content of the academic discipline

Lecture classes

Chapter 1. Basic concepts of data processing

Chapter 2. Statistical data analysis

Chapter 3. Data structures of NumPy and Pandas libraries

Chapter 4. Data visualization methods

Chapter 5. Data preprocessing methods and algorithms

Chapter 6. Basics of machine learning.

Laboratory classes

1. Basic familiarity with the NumPy library.
2. Statistical data analysis.
3. Pandas data structures.
4. Data visualization using matplotlib and Seaborn.
5. Working with time series in Pandas.
6. Data preprocessing in Pandas.
7. Clustering and regression in scikit-learn.
8. Classification in scikit-learn.

4. Educational materials and resources

Literature:

1. Wes McKinney. Python for Data Analysis_ Data Wrangling with Pandas, NumPy, and IPython. - O'Reilly Media, 2017. - 482 p.
2. Gayathri Rajagopalan. A Python Data Analyst's Toolkit. — apress, 2021. — 409 p.
3. Alex Campbell. Data Visualization Guide. — 2021. – 113 p.
4. Sanjeev J. Wagh. Fundamentals of Data Science. — Taylor & Francis Group, LLC, 2022. - 297
5. Avinash Navlani. Python Data Analysis. — Packt Publishing, 2021. - 463 p.
6. Joel Grus. Data Science from Scratch. - O'Reilly Media, Inc., 2019. - 513p.

Educational content

5. Methods of mastering an educational discipline (educational component)

Lecture classes

No. z/p	The name of the topic of the lecture and a list of the main questions (a list of didactic tools, references to the literature and tasks on the SRS)
1	**Lecture 1.**Basic concepts of data processing The structure of the discipline and RSO. Concept of data, types of data, structured data, data sets and their components, main tasks of data processing, features of data processing. Classification and general overview of stages and methods of data processing. Python libraries for working with data. L. (1,3,5) **Tasks on SRS.**Concept of information and knowledge. Data attributes
2	**Lecture 2.**Descriptive statistics with Python The main means of statistics for identifying the characteristics of a data set. Identification of statistical characteristics of data arrays using NumPy and SciPy. Hypothesis testing, general hypothesis testing algorithm, basic statistical tests. L. (1.5) **Tasks for SRS**Installing librariesthe Python language.
3	**Lecture 3.**Statistical hypothesis testing with SciPy Hypothesis testing using the functions of the stats module of the SciPy library. Functions for one sample. Testing hypotheses about the equality of mathematical

	expectations. Use of one-sided alternative hypotheses. Testing hypotheses about the distribution of a random variable, stats.normaltest function. L. (2.5) **Tasks for SRS**. Dispersion analysis of data.
4	**Lecture 4.**Correlation and regression data analysis using Python. Correlation data analysis, corresponding NumPy and SciPy functions. Pearson and Spearman correlation coefficients. A simple linear regression model and its construction using SciPy tools. L. (5) **Tasks for SRS.**Setting up the NumPy library.
5	**Lecture 5.**NumPy library data structures NumPy multidimensional arrays, advantages of their use in data processing. Creating arrays, attributes of arrays, working with different dimensions. Functions for combining arrays. Arithmetic and Boolean operations with arrays. L. (4,5) **Tasks for SRS.**Pandas library settings.
6	**Lecture 6.**Pandas data structures The Series object, its attributes and indices. Two-dimensional DataFrame object, flexible row and column indexes. Add and remove rows and columns. Index object and its types. L. (3,4) **Tasks for SRS**. Data types of the Index object
7	**Lecture 7.**Joining and grouping data in Pandas Joining a DataFrame and a Series using the concat function. The merge function and its parameters, different types of merging. Data grouping using groupby. Methods of grouped objects and their use. L. (2,4) **Tasks for SRS**. Preparation for the first test.
8	**Lecture 8.**Data visualization The first test. Data analysis by visualization methods. Main libraries for data visualization: matplotlib, Pandas, Seaborn. Bar charts. L. (6) **Tasks for SRS**. Additional graph settings in matplotlib.
9	**Lecture 9.**Diagrams for one and two characteristics Histograms and their construction using Python. Scaling charts and their use to detect statistical outliers. Construction of scatter diagrams. Using heatmaps to display the correlation between quantities. L. (6) **Tasks for SRS**. Additional features of Seaborn.

10	**Lecture 10.**Time series in Pandas Concept of time series. Working with time series in Pandas. Using datetime as an index in Pandas objects. Visualization of time series. Analysis of time series by means of Pandas. L. (4,5,6) **Tasks for SRS**. Moving average and other indicators.
11	**Lecture 11.**Working with files in Pandas Basic functions for reading files of various formats typical for representing datasets. Options for reading and writing files. Different json file orientations. Reading datasets from web pages. L. (4) **Tasks for SRS**. Structure of json files.
12	**Lecture 12.**Data pre-processing Basic actions during data preprocessing. Renaming data. Sorting in Pandas. Conversion of data types. Detection and processing of duplicate data. Missing data, their detection, deletion and filling. L. (4,5) **Tasks for SRS**. Preparation for the second test.
13	**Lecture 13.**Data cleaning and transformation The second test. Determination of erroneous and anomalous data. Data filtering. Wide and long formats and transformation between them. L. (4,5) **Tasks for SRS**. Installation and configuration of the scikit-learn library
14	**Lecture 14.**Preparation of data for training Main categories of machine learning models. Getting to know the scikit-learn library. Division of data into training and test samples. Scaling and centering data. Coding of categorical data. Missing data detection and filling in scikit-learn. L. (5) **Tasks for SRS**. Indicator variables.
15	**Lecture 15.**Regression and clustering in scikit-learn Linear regression. Model training and testing. Metrics for regression model evaluation. Hyperparameters of the regression model. Clustering problem, k-means algorithm. Metrics for evaluating the clustering model. L. (2,4,5) **Tasks for SRS.**Visualization of clusters.
16	**Lecture 16.**Classification in scikit-learn Binary classification and logistic regression. Metrics for evaluating the classification model. Balanced and unbalanced classes. Decision tree, ensemble methods and boosting. L. (1.5) **Tasks for SRS.**Gini coefficient.
17	**Lecture 17.**Improvement of models in scikit-learn

	Finding the best hyperparameters using the GridSearch class. Cross validation. Selection, extraction, enhancement and feature creation. Analysis of principal components. L. (2.5) **Tasks for SRS**. Definition of anomalies.
18	**Lecture 18.** Recommendation systems Different types of recommendation systems. Movie recommendation system with scikit-learn. Anomaly detection problem. L. (2.5)

Laboratory classes

No. z/p	The name of the laboratory session	Number of aud. hours
1	A basic introduction to the NumPy library	2
2	Statistical data analysis	2
3	Pandas data structures	2
4	Data visualization using matplotlib and Seaborn	2
5	Working with time series in Pandas	2
6	Data preprocessing in Pandas	2
7	Clustering and regression in scikit-learn	2
8	Classification in scikit-learn	2
9	Final lesson	2

6. Independent work of student

No. z/p	The name of the topic submitted for independent processing	Number of hours of SRS
1	Concept of information and knowledge. Data attributes	3
2	Installing librariesthe Python language	4
3	Dispersion analysis of data	5
4	Setting up the NumPy library	5
5	Pandas library settings	4
6	Data types of the Index object	4
7	Preparation for the first part of the modular test	4
8	Additional graph settings in matplotlib	4
9	Additional features of Seaborn	5
10	Moving average and other indicators	4
11	Structure of json files	3
12	Preparation for the second part of the modular control work	4
13	Installation and configuration of the Skit-learn library	4
14	Indicator variables	3
15	Visualization of clusters	3
16	Gini coefficient	4
17	Definition of anomalies	3

Policy and control

7. Policy of academic discipline (educational component)

The system of requirements for the student:

- attending lectures and laboratory classes is a mandatory component of studying the material;
- the teacher uses his own presentation material at the lecture; uses Google Drive for teaching the material of the current lecture, additional information, tasks for laboratory work, etc.;
- questions at lectures are asked in the time allotted for this purpose;
- laboratory reports are downloaded on the eve of the defense; to defend the laboratory work, it is necessary to demonstrate the operation of the program corresponding to the task and answer the questions about the program and control questions;
- modular test papers are written in lectures without the use of aids (mobile phones, tablets, etc.); the result is downloaded in a file through a Google form to the appropriate directory of Google Drive;
- incentive points are awarded for: participation in faculty and institute olympiads in academic disciplines, participation in work competitions, preparation of reviews of scientific works, etc. The number of encouraged points is no more than 10;
- Penalty points are awarded for: late submission of laboratory work without valid reasons. The number of penalty points is no more than 10.

8. Types of control and rating system for evaluating learning outcomes (RSE)

A student's rating consists of the points he receives for:

1. performance and protection of laboratory work;
2. execution of modular control work;
3. incentive and penalty points.

System of rating points and evaluation criteria

Laboratory tasks:
"excellent", complete answer to the question during the defense (at least 90% of the required information), complete completion of the laboratory task - 9 points;
"good", sufficiently complete answer to the question during the defense (at least 75% of the required information), complete completion of the laboratory task - 7-8 points;
"satisfactory", incomplete answer to the questions during the defense (at least 60% of the required information), minor errors in the performance of the laboratory work task - 5-6 points;
"unsatisfactory", an unsatisfactory answer and/or significant errors in the performance of the laboratory task - 0 points.

Modular control works:
"excellent", a complete answer (at least 90% of the required information), the task was completed without errors, the actions were justified - 14 points;
"good", a sufficiently complete answer (at least 75% of the required information), the task was completed without significant errors - 10-13 points;
"satisfactory", incomplete answer, significant errors may be present in some tasks, but at least 60% are completed correctly - 7-9 points;
"unsatisfactory", an unsatisfactory answer (incorrect performance of tasks), requires mandatory rewriting at the end of the semester - 0 points.

Incentive points
for the performance of creative works from the credit module (for example, participation in faculty and institute olympiads in academic disciplines, participation in work competitions, preparation of reviews of scientific works, etc.) 1-2 points, but not more than 10 in total.

Intersessional certification
According to the results of the educational work for the first 8 weeks, the maximum possible number of points is 41 points (3 laboratory tasks, the first part of the modular control work). At the first certification (8th week), the student receives "passed" if his current rating is not less than 21 points.

According to the results of 13 weeks of training, the maximum possible number of points is 82 points (6 laboratory tasks, a modular control work). At the second certification (14th week), the student receives "passed" if his current rating is not less than 41 points.

The maximum sum of weighted points of control measures during the semester is:
RD = 8*rhl.r.+2*rmkr + (rz - rsh)=8*9+2*14+ (rz - rsh)=100 + (rz - rsh),

where rl.r. – score for the laboratory task (0...9);
rmkr – score for writing part of the MKR (0...14);
rz – incentive points (0...10);
rzsh - penalty points (0...10).

Test:

Students who have fulfilled all the conditions of admission to the semester certification (written all module tests, completed and defended all laboratory work) and scored the required number of points during the semester (RD ≥ 60), receive a credit score (credit) automatically in accordance with the obtained rating (table. 1 below). In this case, RD points and corresponding grades are entered into the credit and examination information.

Students who scored less than 60 points during the semester and do not have debts are required to complete the credit control work.

The final test contains five questions. Each question is worth 20 points.

Students who scored more than 60 points during the semester and fulfilled all admission conditions are given the opportunity to perform a credit test to improve their grade. In this case, the student's previous rating from the credit module is canceled and he receives an assessment taking into account the results of the credit test (Table 1 below).

Question evaluation system:
"excellent", complete answer (at least 90% of the required information) - 18-20 points;
"good", sufficiently complete answer (at least 75% of the required information, or minor inaccuracies) - 15-17 points;
"satisfactory", incomplete answer (at least 60% of the required information and some errors) - 12-14 points;
"unsatisfactory", unsatisfactory answer - 0-11 points.
The sum of RD points or points for credit work is transferred to the credit score according to the table:

Table 1 — Conversion of rating points to grades on the university scale

Scores	Rating
100-95	Perfectly
94-85	Very good
84-75	Fine
74-65	Satisfactorily
64-60	Enough
Less than 60	Unsatisfactorily
Admission conditions not met	Not allowed

9. Additional information on the discipline (educational component)

Working program of the academic discipline (Syllabus):

Folded Tymofieieva Yuliia Serhiivna

ApprovedDepartment of ICT (protocol No. 13 dated 15.06.2022)

Agreedby the methodical commission of the faculty (protocol No. 11 dated 07.07.2022)