

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Львівський національний університет імені Івана Франка
Механіко-математичний факультет
Кафедра математичної статистики і диференціальних рівнянь



Затверджено

На засіданні
кафедри математичної статистики і
диференціальних рівнянь
механіко-математичного факультету
Львівського національного університету
імені Івана Франка
(протокол № 1 від 22.06.2023 р.)

Завідувач кафедри: Олег БУГРІЙ

Силабус з навчальної дисципліни
“Інженерія даних в Python”,
що викладається в межах ОПШ “ Статистичний аналіз даних ”
першого (бакалаврського) рівня вищої освіти для здобувачів з
спеціальності 112 - Статистика

Львів 2023 р.

Назва дисципліни	Інженерія даних в Python
Адреса викладання дисципліни	Головний корпус ЛНУ ім. І. Франка м. Львів, вул. Університетська 1, 79000
Факультет та кафедра, за якою закріплена дисципліна	Механіко-математичний факультет Кафедра математичної статистики і диференціальних рівнянь
Галузь знань, шифр та назва спеціальності	11 - Математика та статистика 112 - Статистика
Викладачі дисципліни	Бокало Т.М., кандидат фізико-математичних наук, асистент кафедри математичної статистики і диференціальних рівнянь
Контактна інформація викладачів	taras.bokalo@lnu.edu.ua , https://new.mmf.lnu.edu.ua/employee/bokalo_t_m ;
Консультації з питань навчання по дисципліні відбуваються	Консультації в день проведення лекцій/практичних занять (за попередньою домовленістю). Головний корпус ЛНУ ім. І. Франка, каб. 267. м. Львів, вул. Університетська, 1
Сторінка курсу	https://new.mmf.lnu.edu.ua/course/inzdp-112-bak23
Інформація про дисципліну	Дисципліна “Інженерія даних в Python” є дисципліною вільного вибору студента з спеціальності 112 – Статистика для освітньої програми “Статистичний аналіз даних”, яка викладається в 4-му семестрі в обсязі 4-ох кредитів (за Європейською Кредитно-Трансферною Системою ECTS).
Коротка анотація дисципліни	У курсі розглядаються основні принципи збирання й оброблення даних, основні етапи та підпроцеси циклу інженерії даних. Визначаються ключові підходи до видобування, зберігання, перетворення та подачі даних. Вивчаються основи мови програмування Python. Розглядаються інструменти мови програмування Python щодо первинної обробки даних, реалізації методів описової статистики, групування та очищення даних, інтерактивних методів аналізу.
Мета та цілі дисципліни	<i>Мета:</i> формування у студентів теоретичних знань і практичних здатностей застосовувати методи та засоби обробки та подання даних з використанням мови програмування Python. <i>Цілі:</i> викласти основні можливості мови програмування Python для обробки, аналізу та візуалізації даних та базові методи автоматизації процесу підготовки аналітичних звітів та інфографічних матеріалів; сформуванню вміння збирати та проводити первинне оброблення великих обсягів даних з різних джерел походження, проводити аналітичну обробку інформації різної структури та змісту.
Література для вивчення дисципліни	Основна: 1) Paul Crickard. <i>Data Engineering with Python: Work with massive datasets to design data models and automate data pipelines using Python</i> . Packt Publishing, 2020. 2) Bas P. Harenslak and Julian Rutger de Ruitter. <i>Data Pipelines with Apache Airflow</i> . Manning, 2021.

- 3) Sandy Ryza, Uri Laserson, Sean Owen, Josh Wills. *Advanced Analytics with Spark: Patterns for Learning from Data at Scale 2nd Edition*. O'Reilly, 2017.
- 4) Bill Chambers, Matei Zaharia. *Spark: The Definitive Guide: Big Data Processing Made Simple 1st Edition*. O'Reilly Media, 2018.
- 5) Jules Damji, Denny Lee, Brooke Wenig, Tathagata Das. *Learning Spark 2nd Edition*. O'Reilly, 2020.
- 6) David Mngadi. *Apache Spark 3 for Data Engineering and Analytics with Python*. Packt Publishing, 2021.
- 7) Benjamin V. Root *Interactive Applications Using Matplotlib*. Packt Publishing, 2015.
- 8) Dr. Pooja. *Data Visualization with Python: Exploring Matplotlib, Seaborn, and Bokeh for Interactive Visualizations*. BPB Publications, 2023.
- 9) Ashwin Pajankar. *Hands-on Matplotlib. Learn Plotting and Visualizations with Python 3. 1st Ed*. Apress, 2022.
- 10) AI Publishing. *Data Visualization with Python for Beginners: Visualize Your Data using Pandas, Matplotlib and Seaborn (Machine Learning & Data Science for Beginners)*. AI Publishing LLC, 2020.
- 11) Adam Schroeder, Christian Mayer, Ann Marie Ward. *The Book of Dash: Build Dashboards with Python and Plotly*. Packt Publishing, 2021.
- 12) Elias Dabbas. *Interactive Dashboards and Data Apps with Plotly and Dash: Harness the power of a fully fledged frontend web framework in Python – no JavaScript required*. No Starch Press, 2022.
- 13) Скотт Берінато. *Хороші діаграми. Поради, інструменти та вправи для кращої візуалізації даних*. ArtHuss, 2022.
- 14) Офіційна документація бібліотек.

Допоміжна:

- 15) Jake VanderPlas. *Python Data Science Handbook*. O'Reilly, 2016.
- 16) Васильєв О. *Програмування мовою Python*. Навчальна книга – Богдан, 2019.
- 17) Пол Беррі *Head First Python. Легкий для сприйняття довідник*. Фабула, 2021.
- 18) Ерік Маттес. *Пришивидишений курс Python*. Видавництво Старого Лева, 2021.
- 19) Адіт'я Бхаргава. *Грокаємо алгоритми. Ілюстрований посібник для програмістів і допитливих*. ArtHuss, 2023.
- 20) Mark Lutz. *Programming Python: Powerful Object-Oriented Programming 4th Edition*. O'Reilly Media, 2011.
- 21) [A Byte of Python](#).
- 22) Michael Dawson. *Python Programming for the Absolute Beginner, 3rd Edition*. Course Technology, 2010.
- 23) Tony Gaddis. *Starting Out with Python 4th Edition*. Pearson, 2017.
- 24) Daniel Zingaro. *Learn to Code by Solving Problems: A Python Programming Primer*. No Starch Press, 2021.
- 25) Dan Bader. *Python Tricks: A Buffet of Awesome Python Features*. 2017.
- 26) David Beazley, Brian K. Jones. *Python Cookbook: Recipes for Mastering Python 3*. O'Reilly Media, 2013.
- 27) David Beazley. *Python Essential Reference 4th Edition*. Addison-Wesley Professional, 2009.
- 28) Patrick Viafore. *Robust Python: Write Clean and Maintainable Code 1st*

	<p><i>Edition. O'Reilly Media, 2021.</i></p> <p>29) Al Sweigart. <i>Beyond the Basic Stuff with Python: Best Practices for Writing Clean Code.</i> No Starch Press, 2020.</p> <p>30) Luciano Ramalho. <i>Fluent Python: Clear, Concise, and Effective Programming 1st Edition.</i> O'Reilly Media, 2015.</p>
Обсяг курсу	Загальний обсяг: 120 годин. Аудиторних занять: 48 год., з них 32 год. лекційних та 16 год. лабораторних занять. Самостійної роботи: 72 год.
Очікувані результати навчання	<p>У результаті вивчення даного курсу студент буде:</p> <p>знати:</p> <p>поняття інженерії даних, основні концепції та принципи; роль і важливість інженерії даних у сучасному аналізі даних; синтаксис мови програмування Python та базові конструкції програмування; використання Python для зчитування та запису даних; основи роботи з бібліотеками Numpy та Pandas; особливості візуалізації даних з використанням Python; методи перетворення та подання даних; основи роботи з конвеєрами обробки даних в Apache Airflow; змінні та параметри в Apache Airflow; принципи розподіленого обчислення з використанням Apache Spark.</p> <p>вміти:</p> <p>використовувати Python для ефективної обробки та аналізу даних; вирішувати завдання з обробки даних за допомогою Python; застосовувати бібліотеки Numpy та Pandas для обробки та аналізу даних; створювати графіки та діаграми з використанням Matplotlib, Seaborn, Plotly; застосовувати методи Pandas для перетворення структури даних; використовувати Apache Airflow для автоматизації процесів обробки даних; створювати та виконувати розподілені обчислення з використанням Apache Spark.</p>
Ключові слова	Мова програмування Python, типи даних, логічні змінні, оператори управління, функції, бібліотеки Matplotlib, Seaborn, Plotly, Pandas, NumPy, Apache Airflow, Apache Spark, інженер по даних, інженер аналітик, інженер по знаннях.
Формат курсу	Очний
Теми	Див. Схема курсу
Підсумковий контроль, форма	Залік у кінці семестру
Пререквізити	<p>Для вивчення даного курсу студенту потрібні базові знання з:</p> <ul style="list-style-type: none"> - лінійної алгебри та аналітичної геометрії; - основ теорії ймовірностей; - інформатики та програмування.
Навчальні методи та техніки, які будуть використовуватися під час викладання курсу	Інформаційні методи (лекція, презентація, бесіда, ілюстрація, демонстрація); інтерактивні методи (дискусія); індивідуальні завдання.
Необхідне обладнання	Для проведення лекційних занять: комп'ютер (мінімальні характеристики: процесор Intel Core i3, 4ГБ оперативної пам'яті), доступ до мережі Internet, засоби мультимедіа (в т.ч. проектор).

	<p>Для проведення практичних/лабораторних занять: комп'ютер (мінімальні характеристики: процесор Intel Core i3, 4ГБ оперативної пам'яті), доступ до мережі Internet.</p> <p>Необхідне програмне забезпечення включає в себе ОС Windows 10, програмні додатки (MS Teams, MS Excel, Jupyter Notebook з вбудованим компілятором мови програмування Python).</p>
<p>Критерії оцінювання (окремо для кожного виду навчальної діяльності)</p>	<p>Оцінювання проводиться за 100-бальною шкалою. Бали нараховуються за наступним співвідношенням:</p> <ul style="list-style-type: none"> ● 6 практичних робіт, кожна оцінюється у максимум 10 балів; становить 60% семестрової оцінки; ● 3 контрольні роботи, кожна оцінюється у максимум 10 балів; становить 30% семестрової оцінки; ● активна участь на лекційних та практичних заняттях оцінюється у максимум 10 балів, що становить 10% семестрової оцінки. <p>Підсумкова максимальна кількість балів 100.</p> <p>Академічна доброчесність: Очікується, що роботи студентів будуть оригінальними дослідженнями чи міркуваннями. Списування та втручання в роботу інших студентів становлять, але не обмежують, приклади можливої академічної недоброчесності. Виявлення ознак академічної недоброчесності в написанні завдань є підставою для її незарахування викладачем, незалежно від масштабів плагіату чи обману. Жодні форми порушення академічної доброчесності не толеруються.</p> <p>Відвідування занять є важливою складовою навчання. Очікується, що всі студенти відвідають усі лекції та практичні/лабораторні заняття курсу. Студенти повинні інформувати викладача про неможливість відвідати заняття. У будь-якому випадку студенти зобов'язані дотримуватися термінів визначених для виконання всіх видів робіт, передбачених курсом.</p> <p>Література. Уся література, яку студенти не зможуть знайти самостійно, буде надана викладачем виключно в освітніх цілях без права її передачі третім особам. Студенти заохочуються до використання також й іншої літератури та джерел, яких немає серед рекомендованих.</p> <p>Політика виставлення балів. Враховуються бали, набрані при поточному контролі та бали підсумкового тестування. При цьому обов'язково враховуються присутність на заняттях та активність студента під час практичного заняття; недопустимість пропусків та запізнь на заняття; користування мобільним телефоном, планшетом чи іншими мобільними пристроями під час заняття в цілях не пов'язаних з навчанням; списування та плагіат; несвоєчасне виконання поставленого завдання і т. ін.</p> <p>Оцінювання практичних робіт (6 практичних робіт, максимальна кількість балів: 60) відбувається шляхом оцінки захисту написаної студентом практичної роботи (0-10 балів за одну роботу).</p> <p>Бали оцінювання аудиторного виконання практичних робіт нараховуються за наступним співвідношенням:</p> <p>9-10 – студент в повному обсязі володіє навчальним матеріалом, має повне розуміння розглянутої теми, надає правильні відповіді на запитання по темі, код програми функціонує відповідно до завдання;</p> <p>7-8 – студент достатньо розуміє розглянутий матеріал та принципи написаного ним коду програми, присутні неточності та незначні помилки</p>

	<p>у відповідях на запитання по темі, код програми функціонує відповідно до завдання;</p> <p>4-6 – студент не досить добре розуміє розглянутий матеріал та написаний ним код програми, вагається та надає неточні/не конкретні відповіді на запитання по темі, код програми функціонує з помірними недоліками;</p> <p>1-3 – студент погано розуміє розглянутий матеріал та написаний ним код програми, студент в більшості надає помилкові відповіді на питання по темі, код програми не функціонує належним чином;</p> <p>0 - студент зовсім не засвоїв розглянутий матеріал, написаний ним код програми не відповідає темі/не функціонує взагалі.</p> <p>Оцінювання контрольних робіт відбувається шляхом оцінки відповідей на тестові завдання та письмових відповідей студента на поставлені запитання.</p> <p>Відсотки нарахування балів оцінювання відповіді на кожне відкрите запитання нараховуються за наступним співвідношенням:</p> <p>75-100% – тема відтворюється в повному обсязі, правильно, обґрунтовано, логічно;</p> <p>50-75% – відтворюється значна частина розглянутої теми, проте присутні неточності та/або невідповідності;</p> <p>25-50% – виявлено множинні неточності та невідповідності, пояснення відсутні чи частково помилкові;</p> <p>0-25% – тему майже не розкрито, кількість викладеного матеріалу не відповідає загальним нормам обраного виду роботи.</p> <p>Критерії оцінювання результатів неформальної освіти:</p> <p>Нарахування балів відбувається за публікацію студентом тез доповідей на конференціях, наукових статей, за участь студента у діяльності наукових гуртків, семінарів, круглих столів, конкурсів, участь у заходах неформальної освіти, за отримання сертифікатів про проходження навчання на різних освітніх платформах (Coursera, Prometheus тощо), курсах провідних ІТ компаній за тематикою навчальної дисципліни. Кількість балів визначається відсотком покриття результатів відповідної активності до вимог результатів навчання з навчальної дисципліни</p>
<p>Питання для контрольного тестування</p>	<p>Основні концепції і принципи інженерії даних; використання Python в інженерії даних; використання бібліотеки Numpy для обробки даних; використання бібліотеки Pandas для аналізу та маніпуляції даними; візуалізація даних з використанням Matplotlib, Seaborn та Plotly; методи перетворення даних; ефективне представлення даних для аналізу та візуалізації; організація конвеєрів обробки даних за допомогою Apache Airflow; робота зі змінними в Apache Airflow; використання Apache Spark для розподіленого обчислення.</p>
<p>Опитування</p>	<p>Анкету-оцінку з метою оцінювання якості курсу буде надано по завершенню курсу.</p>

**Схема курсу “Інженерія даних в Python”
для студентів спеціальності 112 - Статистика**

Т и ж н і	Лекційний курс		Лабораторні заняття		К - с т ь г о д С Р	Літерату ра
	Назва теми	К - с т ь г о д	Назва теми	К - с т ь г о д		
1	2	3	4	5	6	7
1	Поняття інженерії даних	2	Ознайомлення з Python в інженерії даних	2	5	[1], Сайт курсу
2	Використання Python в інженерії даних	2			4	[15]-[30], Сайт курсу
3	Основи програмування в Python	2	Виконання ПР 1.	2	5	[15]-[30], Сайт курсу
4	Контрольна робота 1	2			4	
5	Використання NumPy	2	Консультація з ПР 1. Виконання ПР 2	2	5	[14], Сайт курсу
6	Використання Pandas	2			4	[10], [14], Сайт курсу
7	Особливості візуалізації даних. Візуалізація засобами Python	2	Здача ПР 1. Консультація з ПР 2. Виконання ПР 3	2	5	[7]-[14], Сайт курсу
8	Контрольна робота 2	2			4	
9	Перетворення даних	2	Здача ПР 2. Консультація з ПР 3. Виконання ПР 4	2	5	[1], Сайт курсу
10	Подання даних	2			4	[1], Сайт курсу
11	Платформи для організації конвеєрів обробки даних	2	Здача ПР 3. Консультація з ПР 4. Виконання ПР 5	2	5	[2], Сайт курсу
12	Конвеєри в Apache Airflow	2			4	[2], Сайт курсу
13	Змінні в Apache Airflow	2	Здача ПР 4. Консультація з ПР 5. Виконання ПР 6	2	5	[2], Сайт курсу
14	Розподілене обчислення з Apache Spark	2			4	[3]-[6], Сайт курсу
15	Контрольна робота 3	2	Здача ПР 5 та ПР6	2	5	
16	Підсумкове заняття	2			4	
	Разом	32		16	72	
	Викладач: Бокало Т.М.		Викладач: Бокало Т.М.			