**Силабус освітнього компонента**
Програма навчальної дисципліни

# Основи Python для Data Science

Шифр та назва спеціальності
122 – Комп'ютерні науки

Інститут
ННІ КНІТ
Навчально науковий інститут комп'ютерних наук та інформаційних технологій

Освітня програма
Комп'ютерні науки. Штучний інтелект та управління проєктами

Кафедра
Кафедра стратегічного управління

Рівень освіти
Бакалавр

Тип дисципліни
Вибіркова

Семестр
5

Мова викладання
Українська

## Викладачі, розробники

**Лисенко Антон Олександрович**

anton.lysenko@khpi.edu.ua
Кандидат технічних наук, асистент кафедри стратегічного управління НТУ «ХПІ»

Досвід роботи – понад 11 років. Автор 26 наукових та навчально-методичних праць. Провідний лектор з дисциплін: «Основи web-технологій», «Основи Python для Data Science», «Стек технологій .NET», «Програмування, бази даних і знань»
Детальніше про викладача на сайті кафедри

## Загальна інформація

### Анотація

Предметом дисципліни є алгоритми обробки даних в умовах їх великого обсягу. Основна практична мета - виявлення закономірностей у даних та отримання знань з даних в узагальненій формі. В курсі вивчається прикладне застосування мови Python для аналізу даних та візуалізації результатів. Розглядаються найбільш поширені алгоритми машинного навчання, такі як алгоритми зменшення розмірності даних, кластеризації та класифікації, пошуку асоціативних правил, тощо... Впродовж курсу студенти знайомляться з популярними бібліотеками, розробленими для обробки даних та візуалізації, такими як numpy, pandas, scikit-learn, matplotlib, seaborn.

### Мета та цілі дисципліни

Мета дисципліни – ознайомлення здобувачів освіти з базовими принципами та методами обробки великих обсягів даних, аналізу даних та побудовою висновків, а також створення осередь студентів достатньо високого усередненого рівня знань і вмінь використання сучасних алгоритмів машинного навчання.

### Формат занять

Лекції, лабораторні роботи, практичні заняття, проміжний модульний контроль, самостійна робота, консультації. Підсумковий контроль – іспит.

### Компетентності

ЗК1. Здатність до абстрактного мислення, аналізу та синтезу.
ЗК2. Здатність застосовувати знання у практичних ситуаціях.
ЗК3. Знання та розуміння предметної області та розуміння професійної діяльності.
ЗК6. Здатність вчитися й оволодівати сучасними знаннями.
ЗК7. Здатність до пошуку, оброблення та аналізу інформації з різних джерел.
ЗК8. Здатність генерувати нові ідеї (креативність).
ЗК11. Здатність приймати обґрунтовані рішення.
СК11. Здатність до інтелектуального аналізу даних на основі методів обчислювального інтелекту включно з великими та слабко структурованими даними, їхньої оперативної обробки та візуалізації результатів аналізу в процесі розв'язування прикладних задач.

### Результати навчання

РН12. Застосовувати методи та алгоритми обчислювального інтелекту та інтелектуального аналізу даних в задачах класифікації, прогнозування, кластерного аналізу, пошуку асоціативних правил з викори-станням програмних інструментів підтримки багатовимірного ана-лізу даних на основі технологій DataMining, TextMining, WebMining.
РНП 1.2. Володіти методами, моделями і алгоритмами для технологій аналізу даних та реалізовувати їх у формі прикладного програмного забезпечення, використовуючи стек технології, знаннями про стан-дарти, методи і засоби забезпечення якості у процесі розробки програмного забезпечення.

### Обсяг дисципліни

Загальний обсяг дисципліни 150 год. (5 кредитів ECTS): лекції 32 год., лабораторні заняття 32 год., практичні заняття 16 год., самостійна робота 70 год.

### Передумови вивчення дисципліни (пререквізити)

Методи та засоби обчислювальної математики, Алгоритми та структури даних,

### Особливості дисципліни, методи та технології навчання

Презентації лекцій з дисципліни, методичне забезпечення для виконання лабораторних робіт в електронному вигляді. Тести для поточного та підсумкового контролю знань і вмінь студентів.

## Програма навчальної дисципліни

### Теми лекційних занять

Тема 1. Введення до вивчення дисципліни.
Що таке Data Science? Історія виникнення та розвитку цього напрямку. Де використовується Data Science? Інструменти та етапи робочого процесу у Data Science. Хто такий дата-саентист і чим він займається? Що таке машинне навчання? Як працює машинне навчання? Складові машинного навчання: дані, ознаки, алгоритми. Які існують типи алгоритмів машинного навчання? Переваги та недоліки моделей машинного навчання. Чому все-таки Python? Огляд найпопулярніших IDE для мови Python. Концепція та філософія Python.

Тема 2. Введення в розробку програм з використанням мови Python.
Основи синтаксису: оператори, вирази, керуючі конструкції, коментарі. Типи та структури даних Python. Особливості зберігання даних у пам'яті при динамічній типизації. Рекомендації щодо стилю оформлення сирцевого коду.

### Тема 3. Основи попередньої обробки та очищення даних.
Знайомство з поняттями «ознака» та «спостереження». Стандартизація та масштабування даних, застосування можливостей бібліотеки sklearn.preprocessing. Робота із масивами. Застосування бібліотеки pandas для завантаження та отримання описових статистик даних. Побудова гістограм ознак та їхня візуальна інтерпретація.

### Тема 4. Бібліотека NumPy.
Вступ. Методи автозаповнення та формування числових діапазонів. Методи створення матриць. Методи формування масивів. Основні типи даних бібліотеки numpy. Властивості та уявлення numpy-масивів. Операції із масивами. Транспонування матриць та векторів. Робота з розмірностями. Індексація, зрізи, ітерування по масивах. Спискова та маскова індексація (fancy & masking indexing). Базові математичні операції з масивами. Поняття "векторизації". Методи генерації псевдовипадкових чисел та перемішування елементів масиву. Методи математичної статистики. Елементи лінійної алгебри, операції з матрицями та векторами. Транслювання масивів (array broadcasting).

### Тема 5. Методи зниження розмірності даних.
Мета зниження розмірності даних та огляд найпопулярніших алгоритмів. Поняття «прокляття розмірності». Проекція як спосіб зменшення розмірності даних, розгляд методу головних компонент (Principal Component Analysis - PCA). Коефіцієнт поясненої дисперсії, критерії Кайзера та Кеттелема. Нелінійні проекції, алгоритми машиного навчання на основі різноманітностей (manifold learning), модифікаці алгоритму PCA. Огляд інших алгоритмів зниження розмірності даних. Метод аналізу факторів, відмінність методу головних компонент від методу аналізу факторів. Факторний аналіз як метод класифікації.

### Тема 6. Бібліотека Pandas.
Очищення та первинна оцінка даних. Розгляд основних можливостей бібліотеки. Об'єкти Series та Dataframe. Індексація. Вирівнювання даних за мітками індексу. Виконання відбору елементів за умовою. Відбір стовпців в об'єкті DataFrame. Відбір рядків в об'єкті DataFrame. Створення зрізів. Відбір значень за допомогою індексу. Ієрархічна індексація. Робота з категоріальними даними. Виконання операцій із ковзним вікном. Чисельні та статистичні методи. Мова запитів. Графічні можливості бібліотеки Pandas.

### Тема 7. Функціональне програмування в Python.
Функції та їх параметри. Область видимості змінних. Оператори розпакування. Лямбда-вирази. Елементи функціонального програмування. Ітератори, спискові включення (list comprehension syntax), замикання (closures), анотації типів, generic-типи. Підпрограми як метод підвищення рівня абстракції. Віртуальне оточення Python. Модулі.

### Тема 8. Навчання на асоціативних правилах (Associations Rules Learning — ARL).
Введення у теорію – основні поняття (support, confidence, lift, conviction). Розгляд алгоритмів Apriori, еквівалентного перетворення (ECLAT), FP-Growth (Frequent Pattern Growth). Ефективність та підбір оптимальних параметрів.

### Тема 9. Об'єктно-орієнтоване програмування в Python.
Огляд концепцій ООП. Класи та об’єкти. Особливості інкапсуляції в Python. Особливості спадкування в Python як в мові з динамічною типизацією. Атрибути класу, статичні методи, строкове уявлення об'єкту. Поліморфізм у Python. Перевантаження операторів. Модульність. Ієрархія та абстракціонізм, помилки (exceptions), контексти, контейнери, ітератори та генератори. Метакласи.

### Тема 10. Кластеризація.
Загальний огляд алгоритмів кластеризації. Практичне застосування алгоритмів кластеризації. Відстань між об’єктами, функція відстані та її різноманіття. Відстань між кластерами, варіанти її обчислення, метод Варда. Ієрархічна кластеризація, побудова дендрограми, визначення оптимальної кількості кластерів. Розгляд алгоритму K-MEANS, переваги, недоліки та обмеження цього алгоритму. Початкове розташування центрів кластерів. Підбір параметрів, оцінка

результату кластеризації. Модифікаці алгоритму K-MEANS. Розгляд інших алгоритмів кластеризації – DBSCAN, OPTICS. Методи оцінки якості кластеризації. Практичний приклад використання кластеризації - зменшення кількості кольорів графічного зображення.

Тема 11. Регресія.
Визначення, основні поняття, мета алгоритму. Функція втрат, метод найменших квадратів, сума квадратів відхилень. Коефіцієнт детермінації та його обмеження («Квартет Анскомба»). Поняття колінеарності та перенавчання. Критерій якості: метод тестової вибірки, метод валідації через виключені спостереження (leave one out validation), метод k-кратної валідації. Регуляризація.

Тема 12. Класифікація.
Загальний огляд алгоритмів. Дерева рішень (Classification And Regression Trees – CART) – розгляд алгоритму, визначення оптимальніх параметрів моделі, критерії зупинки, перенавчання моделі. Розгляд інших алгоритмів, таких як метод наівного Байєсу, лінійний дискримінантний аналіз (Linear Discriminant Analysis - LDA), метод опорних векторів (Support Vector Machine - SVM).

Тема 13. Візуалізація.
Основи бібліотеки matplotlib, побудова графиків, стовпчатих діаграм, малювання у 3D просторі. Знайомство з бібліотекою Seaborn.
.

**Теми практичних занять**

Тема 1. Анализ данных с использованием статистических методов.
Описові статистики, імовірність, двовимірна статистика, нормальний розподіл, візуалізація.
Мета: Оволодіти основами статистичного аналізу даних та статистичного моделювання.

Тема 2. Аналіз даних та обчислення з використанням матриць і ядер.
Застосування матричних методів аналізу даних, обчислення коваріаційної матриці, власних чисел і векторів, а також аналіз залишкової дисперсії та візуалізацію головних компонентів.

Тема 3. Алгоритми асоціативного аналізу даних.
Вивчення алгоритмів асоціативного аналізу, роботу з мінімальним рівнем підтримки та визначення областей пошуку для наборів елементів у структурованих даних.

Тема 4. Аналіз набору даних та виявлення мінімальних генераторів.
Вивчення методів аналізу даних, виявлення замкнутих наборів, підрахунок можливих послідовностей та визначення мінімальних генераторів з використанням різних алгоритмів та понять, що стосуються асоціативного аналізу даних.

Тема 5. Кластерний аналіз і метод k-середніх.
Вивчення та практичне застосування методів кластерного аналізу, моделювання ймовірності та обробки категоріальних даних з метою аналізу та визначення зв'язків у наборах даних різної природи.

Тема 6. Кластерний аналіз з використанням алгоритмів щільнісної кластеризації.
Вивчення та практичне застосування методів кластерного аналізу, включаючи алгоритм DBSCAN, метрики відстаней та визначення основних понять, пов'язаних із кластеризацією даних у просторі.

Тема 7. Використання методів машинного навчання для класифікації даних.
Вивчення та практичне використання методів класифікації даних за допомогою машинного навчання, включаючи наївний байєсівський класифікатор, метод Байесового виведення та побудову дерева рішень.

Тема 8. Методи класифікації та розділення даних.
Вивчення та практичне застосування методів класифікації та розділення даних, включаючи аналіз дисперсії, метод опорних векторів та визначення гіперплощини для класифікації.

### Теми лабораторних робіт

Тема 1. Передобробка та очистка даних.
Знайомство з методами обробки даних з бібліотеки Scikit Learn. Завантаження даних з файлу, побудова гістограм для загальної візуальної оцінки ознак. Приведення до діапазону та стандартизація даних. Нелінійні перетворення та дискретизація ознак.
.
Тема 2. Зниження розмірності даних
Знайомство з методами зниження розмірності даних із бібліотеки Scikit Learn. Завантаження та нормування даних, побудова діаграми розсіювання. Використання методу головних компонент (PCA) та його модифікацій для зниження розмірності даних. Порівняння методу PCA з методом аналізу факторів.
.
Тема 3. Знайомство з методами частотного аналізу з бібліотеки MLxtend.
Завантаження та передобробка даних, побудова бінарної матриці ознак. Проведення аналізу з використанням алгоритму Apriori, знайомство з основними поняттями частотного аналізу. Практичне використання бібліотеки pandas для оцінки результатів.

Тема 4. Знайомство з методами асоціативного аналізу з бібліотеки MLxtend.
Завантаження та передобробка даних, побудова бінарної матриці ознак. Використання алгоритмів FPGrowth та FPMax, побудова графиків та гістограм для візуальної оцінки результатів роботи алгоритмів. Пошук асоціативних правил на основі різних метрік. Використання можливостей бібліотеки NetworkX для відображення знайдених асоціативних правил у вигляді графа.

Тема 5. Знайомство з алгоритмами кластеризації модуля sklearn.
Використання алгоритмів K-Means та ієрархічної кластеризації, знайомство з особливостями та обмеженнями цих алгоритмів. Графічне відображення результатів.

Тема 6. Подальше знайомство із методами кластеризації модуля sklearn.
Використання алгоритмів DBSCAN та OPTICS. Визначення оптимальних параметрів алгоритмів, графічне відображення результатів. Знайомство з особливостями та обмеженнями цих алгоритмів у порівнянні з алгоритмами, які розглядалися у попередній лабораторній роботі.

Тема 7. Знайомство із методами класифікації модуля sklearn.
Завантаження та передобробка даних. Застосування алгоритмів класифікації на основі дерев та методу наївного байєсу. Вивчення особливосей цих алгоритмів, оцінка точності класифікації, графічне відображення результатів. Розгляд параметрів використаних алгоритмів, визначення оптимальних їх значень.

Тема 8. Подальше знайомство із методами класифікації модуля sklearn.
Завантаження та передобробка даних. Застосування алгоритмів лінійного дискримінантного аналізу та методу опорних векторів, розгляд параметрів цих алгоритмів та порівняння з алгоритмами, які використовувались у попередній лабораторній роботі.

### Самостійна робота

Опрацювування лекційного матеріалу. Підготовка до модульного тесту та іспиту.

## Література та навчальні матеріали

Основна література
1. Joel Grus. Data Science from Scratch, 2nd edition – O'Reilly Media, 2019. – 406 p.
2. Michael Heydt. Learning Pandas, 2nd edition – Packt Publishing, 2017. – 446 p.
3. Jake VanderPlas. Python Data Science Handbook, 1st edition – O'Reilly Media, 2017. – 546 p.

4. Davy Cielen, Arno Meysman, Mohamed Ali. Introducing Data Science: Big Data, Machine Learning, and more, using Python tools. – Manning, 2016. – 320 p.
5. Aurelien Geron. Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 1st edition – O'Reilly Media, 2017. – 574 p.
6. Paul Barry. Head First Python: A Brain-Friendly Guide, 1st edition – O'Reilly Media, 2010. – 494 p.
7. Початок роботи з Python і робота з даними. Лабораторний практикум з навчальної дисципліни «Основи програмування Python (дисципліна вибору 02)» Частина перша. / М.М. Козуля, Т.В. Козуля – Харків, НТУ «ХПІ», 2022. – 97 с.

Додаткова література

8. Travis E. Oliphant. Guide to NumPy, 2006. – 370 p.
9. Florent Buisson. Behavioral Data Analysis with R and Python – USA: O'Reilly, 2021 – 336 p.
10. Peter Farrell. The Statistics and Calculus with Python Workshop / P.Farrell, Alvaro Fuentes, Ajinkya Sudhir Kolhe, Quan Nguyen, Alexander Joseph Sarver, Marios Tsatsos – UK: Packt Publishing Ltd. – 705 p.
11. Dr. Patrick Jeff. The advanced python for data analysis, 2020 – 60 p.
12. Mehendi Hzn. Python Tricks And Tips Magazine: Gain Insider Skills : Advanced Guides & Tips — 2021
13. Alan D. Moore Python GUI Programming with Tkinter – Packt Publishing, 2018 – 452 p.
14. Методичні вказівки до лабораторної роботи «Основи роботи в сере-довищі Jupyter Notebook» з курсу «Обробка даних Python» / С.М. Коваленко, С.В. Коваленко, О.В. Шматко – Харків, НТУ «ХПІ», 2021. – 28 с.

## Система оцінювання

### Критерії оцінювання успішності студента та розподіл балів

100% підсумкове оцінювання у вигляді іспиту та поточного оцінювання.
30% іспит, 70% поточне оцінювання:
Модульний контроль 1 (5%)
Модульний контроль 2 (5%)
Лабораторні роботи (40%)
Практичні заняття (20%)

### Шкала оцінювання

| Сума балів | Національна оцінка | ECTS |
|---|---|---|
| 90–100 | Відмінно | A |
| 82–89 | Добре | B |
| 75–81 | Добре | C |
| 64–74 | Задовільно | D |
| 60–63 | Задовільно | E |
| 35–59 | Незадовільно (потрібне додаткове вивчення) | FX |
| 1–34 | Незадовільно (потрібне повторне вивчення) | F |

## Норми академічної етики і політика курсу

Студент повинен дотримуватися «Кодексу етики академічних взаємовідносин та доброчесності НТУ «ХПІ»: виявляти дисциплінованість, вихованість, доброзичливість, чесність, відповідальність. Конфліктні ситуації повинні відкрито обговорюватися в навчальних групах з викладачем, а при неможливості вирішення конфлікту – доводитися до відома співробітників дирекції інституту. Нормативно-правове забезпечення впровадження принципів академічної доброчесності НТУ «ХПІ» розміщено на сайті: http://blogs.kpi.kharkov.ua/v2/nv/akademichna-dobrochesnist/

## Погодження

| | | |
|---|---|---|
| Силабус погоджено | Дата погодження, підпис | Завідувач кафедри<br>Марина ГРИНЧЕНКО |
| | Дата погодження, підпис | Гарант ОП<br>Марина ГРИНЧЕНКО |