# СИЛАБУС НАВЧАЛЬНОЇ ДИСЦИПЛІНИ
# «Аналіз великих даних в Python»

| | |
|---|---|
| Ступінь освіти | Магістр |
| Галузь знань | 12 Інформаційні технології |
| Спеціальність | всі спеціальності галузі |
| Тривалість викладання | 1 семестр |
| Заняття | Весняний семестр, 3 та 4 чверть |
| лекції | 2 год /тижд |
| практичні роботи | 2 год /тижд |
| Мова викладання | українська |

Передумови для вивчення: вивчення дисципліни «Аналіз великих даних в Python» у встановлених відповідною робочою програмою обсягах передбачає розуміння основ програмування на Python, основ аналізу даних.

**Сторінка курсу в СДО НТУ «ДП»:**
https://do.nmu.org.ua/course/view.php?id=6719
**Консультації:** за окремим розкладом, що попередньо погоджений зі здобувачами освіти.
**Онлайн-консультації:** MS Teams, електронна пошта**.**

**Інформація про викладача:**


Викладач:

Хабарлак Костянтин Сергійович

доктор філософії, доцент каф. САУ

Посилання на профілі:

Сторінка кафедри САУ

Orcid ID

Scopus ID

Google Scholar


## 1. Анотація курсу

Вміння аналізувати великі даних є затребуваною навичкою, адже людина, пристрої Інтернету речей постійно генерують неосяжні об’єми даних. В необробленому вигляді вони мають невисоку цінність, адже їх складно зберігати, структурувати та аналізувати, тому все більший розвиток отримують підходи **Аналізу великих даних в Python.** В даному курсі здобувач отримає практичні навички роботи з інструментарієм обробки та аналізу великих даних: бібліотеками Pandas, Matplotlib, Apache Spark і PySpark, нереляційною базою даних MongoDB, мовою програмування Python для аналізу великих даних. Здобувач навчиться навчитися збирати, зберігати, оброблювати та аналізувати великі масиви даних; будувати моделі регресії та класифікації, використовуючи великі набори даних, та робити передбачення на нових, невідомих вхідних значеннях; опанує нереляційні бази даних для зберігання та обробки великих даних.

## 2. Мета та завдання навчальної дисципліни

**Мета дисципліни** – сформувати у здобувачів вищої освіти навички збору, обробки та аналізу великих із використанням сучасних бібліотек Pandas, Matplotlib, Apache Spark і MongoDB та мови програмування Python. Знання та навички, отримані в курсі, будуть корисними для подальшого працевлаштування здобувача.

**Завдання курсу:**

- навчитися збирати, зберігати, оброблювати та аналізувати великі масиви даних;
- навчитися будувати моделі регресії та класифікації, використовуючи великі набори даних, та робити передбачення на нових, невідомих вхідних значеннях;
- опанувати нереляційні бази даних для зберігання та обробки великих даних;
- отримати практичні навички роботи з бібліотеками Pandas, Matplotlib, Apache Spark, PySpark, базою даних MongoDB, мовою програмування Python для обробки та аналізу великих даних.

## 3. Результати навчання

1. Вміти здійснювати пошук, агрегацію та візуалізацію великих даних для аналізу поведінки досліджуваної системи.
2. Вміти застосовувати методи побудови регресії на великих даних. Проводити аналіз та прогнозування.
3. Вміти застосовувати методи класифікації (лінійних, метричних моделей та на основі дерев рішень) із використанням інструментарію обробки великих даних мови програмування Python.
4. Розуміти способи підгонки параметрів моделей машинного навчання на великих даних.
5. Вміти проводити первинну обробку великих масивів даних різної природи, виявляти закономірності та візуалізувати дані.
6. Розуміти підходи щодо обробки та аналізу слабко структурованих даних для зберігання в нереляційних базах даних, вміти проводити їх подальший аналіз.

## 4. Структура курсу

| Види та тематика навчальних занять | Внесок в загальну оцінку, % |
|---|---|
| **ЛЕКЦІЇ** | **40** |
| **1. Первинний аналіз даних.** Знайомство із мовою програмування Python для аналізу даних. Використання бібліотеки Pandas для табличних даних та розрахунку статистичних оцінок. Типи даних. Бібліотека SciPy. | 3 |
| **2. Візуалізація даних.** Основні види графіків, діаграм. Їх побудова, доречність використання для наявних даних, особливості. | 3 |
| **3. Бутстрап.** Поняття популяції та вибірки даних, зміщеної вибірки, повторного та статифікованого відбору. Алгоритм бутстрапу, розрахунок бутстрапівських оцінок. | 3 |
| **4. Обробка великих даних в Python.** Знайомство із мовою програмування Python для аналізу великих даних. Джерела великих даних. Типи даних. Актуальність. Проблематика. | 3 |
| **5. Apache Spark для обробки та аналізу великих даних в Python.** Знайомство з бібліотекою Apache Spark для аналізу великих даних та її інтерфейсом до мови програмування Python PySpark. Підключення до | 3 |

| | |
|---|---|
| джерел даних. Поняття кадру даних. Pandas і PySpark. Який інструмент є більш доречним для конкретної задачі? | |
| **6. Операції над кадрами даних. Інструменти візуалізації**<br>Перегляд, вибір та агрегація даних. Розрахунок статистичних оцінок. Робота із неповними даними. Проведення первинного аналізу даних за допомогою інструментів PySpark. Обробка та аналіз даних, отриманих з експериментів. Функції візуалізація даних. | 3 |
| **7. Інструменти машинного навчання в PySpark. Регресія.**<br>Методи регресії із використанням інструментарію обробки великих даних мови програмування Python. | 3 |
| **8. Машинне навчання в PySpark. Класифікація.**<br>Класифікація та регресія в PySpark. Побудова моделі та пошук її параметрів на великих масивах вхідних даних. | 3 |
| **9. Методи зменшення розмірності даних.**<br>Проблема обробки даних із великою розмірністю та методи її зменшення. Метод головних компонент. | 3 |
| **10. Нереляційні БД. Переваги та недоліки над реляційними БД.**<br>Причини популярності нереляційних БД в епоху великих даних. Особливості структурування та збереження даних в нереляційних БД. Чи варто завжди використовувати нереляційні БД? | 3 |
| **11. MongoDB – відкрита документна нереляційна БД.**<br>Модель зберігання даних. Встановлення та налаштування БД. Користувацький інтерфейс до MongoDB. | 4 |
| **12. Засоби роботи з MongoDB в Python.**<br>Інтерфейс до мови програмування Python. Створення таблиць, зберігання, обробка та агрегування даних. | 3 |
| **13. Ланка між PySpark та MongoDB.**<br>Інтелектуальний аналіз, машинне навчання та прогнозування даних інструментами PySpark, використовуючи дані з нереляційної БД. | 3 |
| **ПРАКТИЧНІ ЗАНЯТТЯ** | **60** |
| **1: Первинний аналіз даних.**<br>Мета: закріпити теоретичні знання і розвинути практичні навички роботи з таблицями в Pandas, розрахунку статистик та візуалізації даних. | 12 |
| **2: Джерела великих даних. Обробка слабко-структурованої інформації різного походження.**<br>Мета: закріплення навичок роботи із кадрами даних PySpark, початкових аналіз даних, візуалізація великих даних. | 12 |
| **3: Прогнозування на великих даних.**<br>Мета: закріплення навичок застосування різних методів регресії та класифікації на великих даних, підгонки моделей, прогнозування невідомих значень. | 12 |
| **4: Використання нереляційної БД для зберігання та обробки великих масивів даних.**<br>Мета: закріплення навичок роботи із нереляційною БД MongoDB. Створення таблиць, зберігання, обробка та агрегування даних в нереляційній БД в Python. | 12 |
| **5: Інтеграція PySpark і MongoDB.**<br>Мета: узагальнення навичок, отриманих в курсі. Поєднання різних технологій для обробки, структурування, аналізу та передбачення на даних. | 12 |
| | |
| **РАЗОМ** | **100** |

## 5. Технічне обладнання та/або програмне забезпечення

Технічні засоби навчання: мультимедійні та комп'ютерні пристрої.

Засоби дистанційної освіти: Moodle, MS Teams.

Пакети прикладних програм: Python, бібліотеки: Pandas, Matplotlib, Apache Spark, PySpark, MongoDB (безкоштовні).

## 6. Система оцінювання та вимоги

6.1. Навчальні досягнення здобувачів вищої освіти за результатами вивчення курсу оцінюватимуться за шкалою, що наведена нижче:

| Рейтингова шкала | Інституційна шкала |
|---|---|
| 90 – 100 | відмінно |
| 75-89 | добре |
| 60-74 | задовільно |
| 0-59 | незадовільно |

6.2. Здобувач вищої освіти може отримати підсумкову оцінку з навчальної дисципліни на підставі поточного оцінювання знань за умови, якщо набрана кількість балів з поточного тестування та самостійної роботи складатиме не менше 60 балів. Поточна успішність складається з успішності за теоретичну частину курсу (максимум – 40 балів) та оцінок за виконання практичних робіт (кожна робота оцінюється по 100-бальній шкалі. Оцінка за практику є середньою оцінкою за індивідуальні роботи, що переводиться в 60-бальну шкалу). Отримані бали за теоретичну частину курсу та практичні роботи додаються і є підсумковою оцінкою за вивчення навчальної дисципліни. Максимально за поточною успішністю здобувач вищої освіти може набрати 100 балів.

Шкала оцінювання (зазначено максимально можливі бали):

| **Теоретична частина** | **Практичні роботи** | | **Разом** |
|---|---|---|---|
| | При своєчасному складанні | При несвоєчасному складанні | |
| 40 | 60 | 40 | 100 |

6.3 Критерії оцінювання поточного та підсумкового контролю:

– підсумкове оцінювання відбувається у формі диференційованого заліку у форматі тесту, який складається з 20 індивідуальних завдань (кожне з яких оцінюється по 5 балів, 17 завдань із однією правильною відповіддю, 3 – з декількома правильними відповідями);

– поточне оцінювання теоретичних робіт відбувається аналогічно у форматі тесту: 20 індивідуальних завдань (кожне з яких оцінюється по 5 балів, 17 завдань із однією правильною відповіддю, 3 – з декількома правильними відповідями);

– поточне оцінювання практичних робіт відбувається шляхом перевірки звіту з відповідної роботи на порталі дистанційної освіти Moodle (максимальний бал – 100, загальна оцінка рівномірно розподіляється між пунктами індивідуального завдання. Оцінюється відповідність завданню, повнота програмного коду, правильність роботи розробленої програми, проведений здобувачем аналіз та обґрунтованість висновків щодо отриманих результатів).

## 7. Політика курсу

**7.1. Політика щодо академічної доброчесності.** Академічна доброчесність студентів є важливою умовою для опанування результатами навчання за дисципліною і отримання задовільної оцінки з поточного та підсумкового контролів. Академічна доброчесність базується на засудженні практик списування (виконання письмових робіт із залученням зовнішніх джерел інформації, крім дозволених для використання), плагіату

(відтворення опублікованих текстів інших авторів без зазначення авторства), фабрикації (вигадування даних чи фактів, що використовуються в освітньому процесі). У НТУ «Дніпровська політехніка» політика щодо академічної доброчесності регламентується положенням "Положення про систему запобігання та виявлення плагіату у Національному технічному університеті "Дніпровська політехніка": "Положення про систему запобігання та виявлення плагіату у Національному технічному університеті "Дніпровська політехніка".

У разі порушення студентом академічної доброчесності (списування, плагіат, фабрикація), робота оцінюється незадовільно та має бути виконана повторно. При цьому викладач залишає за собою право змінити тему завдання.

**7.2. Комунікаційна політика.** Студенти повинні мати активовану університетську (корпоративну на домені @nmu.one) пошту. Усі письмові запитання до викладачів стосовно курсу мають надсилатися на університетську електронну пошту.

**7.3. Політика щодо перескладання.** Роботи, які здаються із порушенням термінів без поважних причин оцінюються на нижчу оцінку. Перескладання відбувається із дозволу деканату за наявності поважних причин (наприклад, лікарняний).

**7.4. Відвідування занять.** Для студентів денної форми відвідування занять є обов'язковим. Поважними причинами для неявки на заняття є хвороба, участь в університетських заходах, відрядження, які необхідно підтверджувати документами у разі тривалої (два тижні) відсутності. Про відсутність на занятті та причини відсутності студент має повідомити викладача або особисто, або через старосту. Якщо студент захворів, ми рекомендуємо залишатися вдома і навчатися за допомогою дистанційної платформи. Студентам, чий стан здоров'я є незадовільним і може вплинути на здоров'я інших студентів, буде пропонуватися залишити заняття (така відсутність вважатиметься пропуском з причини хвороби). Практичні заняття не проводяться повторно, ці оцінки неможливо отримати під час консультації. **За об'єктивних причин (наприклад, міжнародна мобільність) навчання може відбуватись в он-лайн формі за погодженням з керівником курсу.**

**7.5. Участь в анкетуванні**. Наприкінці вивчення курсу та перед початком сесії студентам буде запропоновано анонімно заповнити електронні анкети (MS Office 365), які буде розіслано на ваші університетські поштові скриньки. Заповнення анкет є важливою складовою вашої навчальної активності, що дозволить оцінити дієвість застосованих методів викладання та врахувати ваші пропозиції стосовно покращення змісту навчальної дисципліни.

## 8. Рекомендовані джерела інформації


**Базова:**

1. Pedro Duarte Faria. Introduction to pyspark. – 2024. [Online] URL: https://pedropark99.github.io/Introd-pyspark/
2. Apache Spark documentation [Online]. URL: https://spark.apache.org/docs/latest/index.html
3. MongoDB documentation [Online]. URL: https://www.mongodb.com/docs/
4. Niall O'Higgins. MongoDB & Python. O'Reilly Media – 2011. – 66 p.
5. Математичні методи інтелектуального аналізу даних: [навчальний посібник для здобувачів першого рівня вищої освіти спеціальності 124 Системний аналіз] / Т. Шабельник, О. Дяченко. – Маріуполь: МДУ, 2021. – 163 с.

**Додаткова:**

1. Кононова К. Ю. Машинне навчання: методи та моделі / К. Ю. Кононова. – Харків: ХНУ імені В. Н. Каразіна, 2020. – 301 с.
2. Practical Statistics for Data Scientists / P. Bruce, A. Bruce, P. Gedeck. – O'Reilly Media, 2020.