Національний технічний університет України
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

Кафедра інженерії
програмного забезпечення
в енергетиці (ІПЗЕ)

ОСНОВИ BIG DATA АНАЛІТИКИ
Робоча програма навчальної дисципліни (Силабус)

Реквізити навчальної дисципліни

Рівень вищої освіти	**Другий (магістерський)**
Галузь знань	12 Інформаційні технології
Спеціальність	121 Інженерія програмного забезпечення
Освітня програма	Інженерія програмного забезпечення інтелектуальних кібер-фізичних систем в енергетиці
Статус дисципліни	Вибіркова
Форма навчання	очна(денна)
Рік підготовки, семестр	4 курс, осінній
Обсяг дисципліни	4 кред/120 год.(лекцій 36 год., практ. 18 год., СРС 66 год.)
Семестровий контроль/ контрольні заходи	Залік, мкр
Розклад занять	http://rozklad.kpi.ua/
Мова викладання	Українська/Англійська/Німецька / Французька
Інформація про керівника курсу / викладачів	Лектор: д.т.н., Федорова Наталія Володимирівна, Natasha_f@ukr.net, telegram, viber, Zoom session Практичні: д.т.н., Федорова Наталія Володимирівна, Natasha_f@ukr.net, telegram, viber, Zoom session
Розміщення курсу	Кампус

Програма навчальної дисципліни

1. Опис навчальної дисципліни, її мета, предмет вивчання та результати навчання

Метою дисципліни «Основи Big Data аналітики» є **в**ивчення даної дисципліни, що надасть важливі знання та практичні навички, щоб можна впевнено було працювати з даними в будь-якій сфері, будь то бізнес, наука або громадський сектор. Головний фокус в даній дисципліни спрямовано саме на енергетиці для створення нових можливостей IEC – інтелектуальних електроенергетичних мереж (Smart Grid).

Предметом дисципліни «Основи Big Data аналітики» є великі дані (Big Data), як основний інструмент, а також аналіз даних, тенденцій, закономірностей та розробки різних систем аналітики даних, класифікацій й прогнозування з подальшою інтерпретацією, візуалізацією результатів.

Основні завдання кредитного модуля.

Згідно з вимогами програми навчальної дисципліни студенти після засвоєння кредитного модуля мають продемонструвати такі результати навчання:

Загальні компетенції:

- здатність проводити дослідження на відповідному рівні (ЗК 3);
- здатність генерувати нові ідеї (креативність) (ЗК 5).

Фахові компетентності:

- здатність проектувати та розробляти програмні системи з використанням методів інтелектуального аналізу даних (ФК-11);
- здатність проектувати та розробляти програмне забезпечення для роботи в хмарі (ФК-14).
- здатність застосувати принципи обробки Big Data до задач електроенергетики.

- здатність визначати типи та характеристики наявного електрообладнання та обирати найбільш ефективну реалізацію залежно від обраних характеристик;

- здатність створення цифрових двійників в енергетиці та smart-рішень.

Згідно з вимогами ОПП/ОНП Інженерія програмного забезпечення інтелектуальних кібер-фізичних систем в енергетиці, студенти після засвоєння навчальної дисципліни мають продемонструвати такі результати навчання.

Програмні результати навчання:

- розробляти, аналізувати, обґрунтовувати та систематизувати вимоги до програмного забезпечення (ПРН 5);

- збирати, аналізувати, оцінювати необхідну для розв'язання наукових і прикладних задач інформацію, використовуючи науково-технічну літературу, бази даних та інші джерела (ПРН 17);

- вміти проектувати та розробляти програмні системи з використанням методів інтелектуального аналізу даних (ПРН 19);

- вміти проектувати та розробляти програмне забезпечення для роботи в хмарі (ПРН 22).

2. Пререквізити та постреквізити дисципліни (місце в структурно-логічній схемі навчання за відповідною освітньою програмою)

Дисципліна містить п'ять кредитів.

Вивчення дисципліни спирається на знання, отримані за програмою попередніх років навчання за спеціальністю 121 Інженерія програмного забезпечення. Студенти мають досвід у імперативному, об'єктно-орієнтованому і функціональному програмуванні.

Викладений матеріал може бути інструментальною основою для підготовки магістерських дисертацій.

Міждисциплінарні зв'язки забезпечуються дисциплінами: «Фізичні основи кібер-фізичних систем», «Проектування кібер-фізичних систем» «Основи Інтернету речей».

Зміст навчальної дисципліни

Розділ 1. Техніки та технології великих даних

Тема 1. Поняття «Великих даних».

Основні аспекти та складові елементи трактування поняття «Великі дані». Сфери застосування надвеликих масивів даних. Характеристики великих даних. Проблема масштабування. Базові компоненти аналізу даних. Роль великих даних в техніці, економіці та житті.

Тема 2. Техніки великих даних: Консолідація даних. Візуалізація. Класифікація. Кластеризація. Регресійний аналіз. Аналіз асоціативних правил. Нейронні мережі.

Тема 3. Технології та інструменти великих даних. Apache Hadoop. Storm – система потокової обробки даних. Мова програмування R.

Тема 4. Аналітика даних як корпоративний проєкт. Життєвий цикл проєкту аналітики великих даних. Дослідження. Підготовка даних. Планування моделі. Побудова моделі. Обговорення результатів. Використання.

Розділ 2. Великі дані в енергетиці

Тема 1. Інтелектуальна електроенергетика

Тема 2. Системи обробки даних в інтелектуальних енергомережах.

Тема 3. Огляд проєктів Soft Grid, що орієнтовані на аналітику в розподілених системах.

Тема 4. Системи керування енергоспоживанням.

Тема 5 Великі дані перетворюються в енергію: віртуальні електростанції.

Тема 6. Керування в мережах з багатьма накопичувачами енергії.

Тема 7. Системи моніторингу та керування ресурсами в нормі та в аварійній ситуації.

3. Навчальні матеріали та ресурси

Базова література

1. Miner, Donald, and Adam Shook. Mapreduce Design Patterns. Beijing: O'Reilly, 2012.
2. Lam, Chuck. Hadoop in Action. Greenwich, Conn: Manning Publications, 2011.
3. Matei Zaharia, Holden Karau, Andy Konwinski, Patrick Wendell Learning Spark Lightning-Fast Big Data Analysis: O'Reilly Media, 2015, 276 p.
4. Holden Karau, Rachel Warren High Performance Spark: Best Practices for Scaling and Optimizing Apache Spark : O'Reilly Media, 2017, 358 p.
5. Sandy Ryza, Uri Laserson, Josh Wills, Sean Owen Advanced Analytics with Spark, 2nd Edition Patterns for Learning from Data at Scale, O'Reilly Media, 2017, 280p.
6. Stephen Marsland Machine Learning: An Algorithmic Perspective, Chapman & Hall/CRC, 2009m 390 p.

Додаткова література

7. Holmes, Alex. Hadoop in Practice. Shelter Island, NY: Manning, 2012.
8. Alpaydin, Ethem. Introduction to Machine Learning., 2014.
9. Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar. Foundations of Machine Learning. Cambridge, MA: MIT Press, 2012.
10. Shalev-Shwartz, Shai, and Shai Ben-David. Understanding Machine Learning: From Theory to Algorithms., 2014.

Навчальний контент

4. Методика опанування навчальної дисципліни (освітнього компонента)

Лекційні заняття

№ з/п	Назва теми лекції та перелік основних питань
1	Поняття «Великих даних». Основні аспекти та складові елементи трактування поняття «Великі дані». Сфери застосування надвеликих масивів даних. Характеристики великих даних. Проблема масштабування. Базові компоненти аналізу даних. Роль великих даних в техніці, економіці та житті.
2	Консолідація даних - набір технік, що направленні на вилучення даних з різних джерел, забезпечення їх якості, перетворення в єдиний формат та завантаження в сховище даних – "аналітичну пісочницю" (analytic sandbox) або «озеро даних» (data lake).
3	Візуалізація. Техніка візуалізації є потужним методом інтелектуального аналізу даних. Як правило, її використовують для перегляду та верифікації даних перед створенням моделі, а також після створення прогнозів. Візуалізація – це перетворення чисельних даних на деякий візуальний образ, з метою спрощення сприйняття великих масивів інформації.
4	Методи кластеризації Задача кластеризації. Типи методі кластеризації. Алгоритм к-середніх. Змішені моделі. Максимізація очікування. Ієрархічна кластеризація. Застосування методів кластеризації.
5	Навчання з учителем Постановка задачі навчання з учителем. Основні класи задач. Задача регресії Лінійна регресія. Градієнтний спуск. Метод нормальних рівнянь. Задача класифікації. Логістична регресія.

6	Нейронні мережі. Основні поняття Біологічний нейрон, модель Маккаллок-Питтса як лінійний класифікатор. Функції активації. Проблема повноти. Завдання виключного або. Типи нейронних мереж. Класи задач. Сфери застосування.
7	Навчання нейронних мереж Алгоритм зворотного розширення похибки. Проблеми навчання нейронних мереж. Типи оптимізаторів. Швидкі методи стохастичного градієнта. Метод випадкових відключень нейронів (Dropout). Інтерпретації Dropout.
8	Навчання без учителя. Зниження розмірності Постановка задачі навчання без учителя. Задача зниження розмірності. Сингулярний розклад матриці. Метод головних компонент. T-SNE алгоритм.
9	Програмний засіб Apache Hadoop Проблема масштабування. Базові компоненти аналізу даних. Поняття розподіленої файлової системи. Програмні моделі “великих даних”. Hadoop екосистема. Storm – система потокової обробки даних. Мова програмування R.
10	Аналітика даних як корпоративний проєкт (частина 1): Життєвий цикл проєкту аналітики великих даних. Дослідження.
11	Аналітика даних як корпоративний проєкт (частина 2): Підготовка даних. Планування моделі. Побудова моделі. Обговорення результатів. Використання.
12	Інтелектуальна електроенергетика
13	Системи обробки даних в інтелектуальних енергомережах
14	Огляд проєктів Soft Grid, що орієнтовані на аналітику в розподілених системах
15	Системи керування енергоспоживанням
16	Великі дані перетворюються в енергію: віртуальні електростанції
17	Керування в мережах з багатьма накопичувачами енергії.
18	Системи моніторингу та керування ресурсами в нормі та в аварійній ситуації.

Лабораторні заняття

Основні завдання циклу лабораторних занять полягають у набутті студентами практичних навичок з використання спеціалізованого програмного забезпечення обробки надвеликих масивів даних та реалізації алгоритмів машинного навчання.

№ з/п	Назва теми заняття
1	Активне керування навантажувальними характеристиками споживачів.
2	Ефективне управління розподіленими системами генерації з великою кількістю джерел.
3	Ефективне управління та моніторинг для численних динамічних нестабільних систем генерації.
4	Прогнозування навантаження та ефективне управління елементами мережі з метою енергозбереження та запобігання перевантаженням.
5	Гнучка тарифікація та детектування витоків та розкрадань.
6	Управління майном та технічне обслуговування.
7	Вирішення завдань Grid Analytics (аналітика на мережах).
8	Вирішення завдань Customer Data (аналітика для клієнтів).
9	Завдання Customer Data: - захист доходів; - «тонке» передбачення навантаження; - детальна сегментація користувачів.

5. Самостійна робота студента/аспіранта

№ з/п	Назви тем і питань, що виноситься на самостійне опрацювання та посилання на навчальну літературу
1	Базові програмні засоби роботи з надвеликими масивами даних. Програмна екосистема Hadoop. Практичне застосування Hadoop в задачах Data Science. Огляд HiveQL, Pig, Yarn, Hbase. Аналітика даних з використанням Spark. [2, 4, 5, 7].
	Програмна платформа розподілених даних Spark Модель паралельних обчислень Spark. Spark Job Scheduling. DataFrame API. Представлення даних в DataFrames and Datasets. Core Spark Joins. Ефективні трансформації [5].
2	Технології великих даних на сучасне програмне забезпечення [10]. Програмна реалізація шаблонів проектування Mapreduce [1].
3	Основи машинного навчання Теоретичні засади машинного навчання [9]. Кодування та підготовка даних Mllib. Масштабування та вибір характеристик. MLlib Model Training. Оцінка моделі машинного навчання [5].
4	Нейронні мережі. Основні поняття Біологічний нейрон, модель Маккаллок-Питтса як лінійний класифікатор. Функції активації. Проблема повноти. Завдання виключного або. Типи нейронних мереж. Класи задач. Сфери застосування.
5	Навчання нейронних мереж Алгоритм зворотного розширення похибки. Проблеми навчання нейронних мереж. Типи оптимізаторів. Швидкі методи стохастичного градієнта. Метод випадкових відключень нейронів (Dropout). Інтерпретації Dropout.
6	Навчання з учителем (Supervised Learning). Метод стохастичного градієнта SG. Логістична регресія. Принцип максимуму правдоподібності і логарифмічна функція втрат. Метод стохастичного градієнта для логарифмічною функції втрат. Математичні основи методу опорних векторів. Завдання квадратичного програмування і двоїста задача. Побудова ядер для методу опорних векторів. Критерії якості класифікації: чутливість і специфічність, ROC-крива і AUC, точність і повнота, AUC-PR. [6, 9].
7	Навчання без вчителя (Unsupervised Learning). Постановка завдання кластеризації. Постановка завдання Semisupervised Learning, приклади застосування. Алгоритм k-середніх і EM-алгоритм для поділу Гауссовской суміші. Алгоритм Ланса-Вільямса та його окремі випадки. Алгоритм побудови дендрограми. Визначення числа кластерів. Сингулярний розклад в задачі зниження розмірності. Alternating Least Squares з використанням Spark [6, 9].
8	Перспективні напрямки розвитку програмних засобів обробки надвеликих даних. Ризики, пов'язані з застосуванням надвеликих даних. Проблема конфіденційності. Датифікація [10].

Політика та контроль

6. Політика навчальної дисципліни (освітнього компонента)

Застосовуються стратегії активного і колективного навчання, які визначаються наступними методами і технологіями:

1) методи проблемного навчання (проблемний виклад, частково-пошуковий (евристична бесіда) і дослідницький метод);

2) особистісно-орієнтовані (розвиваючі) технології, засновані на активних формах і методах навчання («мозковий штурм», «аналіз ситуацій» дискусія, експрес-конференція);

3) інформаційно-комунікаційні технології, що забезпечують проблемно-дослідницький характер процесу навчання та активізацію самостійної роботи студентів (електронні презентації для лекційних занять, використання аудіо-, відео-підтримки навчальних занять).

4) лекційні та лабораторні заняття відносяться до аудиторних занять. Відвідування аудиторних занять є обов'язковим;

5) правила поведінки на заняттях: активність, підготовка коротких доповідей чи текстів, відключення телефонів, використання засобів зв'язку для пошуку інформації на гугл-диску викладача чи в інтернеті тощо;

6) правила захисту лабораторних робіт. На лабораторних заняттях студенти під керівництвом викладача вивчають методику експериментальних досліджень. На кожній лабораторній роботі студенти оформляють звіт у письмовому вигляді. До звіту заноситься перебіг досліду, його результати і даються пояснення отриманих результатів з урахуванням похибок експерименту.

7) індивідуальні завдання з дисципліни (реферати, розрахункової, графічні, тощо) видаються студентам в терміни, передбачені вищим навчальним закладом. Індивідуальні завдання виконуються студентом самостійно при консультуванні викладачем. Допускаються випадки виконання комплексної тематики кількома студентами.

8) правила призначення заохочувальних балів: своєчасне виконання та здача лабораторних, індивідуальних завдань, нестандартний підхід до вирішення певного завдання;

правила призначення штрафних балів: несвоєчасне виконання лабораторних та індивідуальних завдань, а також користування допоміжними засобами (наприклад, мобільний телефон, конспект лекцій) під час виконання контрольної роботи.

9) політика дедлайнів та перескладань: невчасно виконані та здані лабораторні роботи оцінюються нижчою оцінкою (-10-15% від загальної підсумкової оцінки).

10) політика щодо академічної доброчесності: письмові роботи можуть перевірятися на наявність плагіату і допускаються до захисту із коректними текстовими запозиченнями не більше 40%. Списування під час контрольних робіт та екзаменів заборонені.

11) інші вимоги, що не суперечать законодавству України та нормативним документам Університету:

- політика щодо відвідування: відвідування занять є обов'язковим компонентом оцінювання, за яке нараховуються бали. За об'єктивних причин (підтверджених документально) дозволяється перескладання пропущених тем курсу.
- політика щодо виконання завдань: позитивно оцінюється відповідальність, старанність, креативність, фундаментальність.

7. Види контролю та рейтингова система оцінювання результатів навчання (РСО)

1. Оцінка з дисципліни виставляється за багатобальною системою, з подальшим перерахуванням у 4-бальну.

2. Максимальна кількість балів з дисципліни дорівнює 100.

3. Нарахування балів по окремих видах робіт:

Рейтинг студента з кредитного модуля складається з балів, що він отримав за:

- виконання практичних робіт;
- написання контрольної роботи (МКР).

Система рейтингових (вагових) балів та критерії оцінювання

1.Виконання практичних робіт

Оцінюються 9 робіт, передбачених робочою програмою. Максимальний ваговий бал rЛР =63

Сума вагових балів практичних робіт:

N л. р.	Назва практичної роботи	Максимальний ваговий бал
1	Активне керування навантажувальними характеристиками споживачів.	7

2	Ефективне управління розподіленими системами генерації з великою кількістю джерел.	7
3	Ефективне управління та моніторинг для численних динамічних нестабільних систем генерації.	7
4	Прогнозування навантаження та ефективне управління елементами мережі з метою енергозбереження та запобігання перевантаженням.	7
5	Гнучка тарифікація та детектування витоків та розкрадань.	7
6	Управління майном та технічне обслуговування.	7
7	Вирішення завдань Grid Analytics (аналітика на мережах).	7
8	Вирішення завдань Customer Data (аналітика для клієнтів).	7
9	Завдання Customer Data: - захист доходів; - «тонке» передбачення навантаження; - детальна сегментація користувачів.	7
Разом		63

Оцінювання лабораторних робіт:

–якщо робота виконана невчасно знімається 10-15% від максимальної кількості балів (кількість процентів залежить від терміну запізнення);

–якщо робота виконана не самостійно та простежується не індивідуальне виконання то знімається 50% від максимальної кількості балів;

–якщо в програмі не витримані основні правила створення програмних продуктів (модульність, дружній інтерфейс, наявність коментарів та т.п.) знімається 5%.

2. Модульний контроль

На одному з лекційних занять проводиться модульна контрольна робота: Максимальний ваговий бал rМКР = 10.

Оцінювання модульної контрольної роботи виконується наступним чином:

–якщо на всі питання дані повні та чітко аргументовані відповіді, контрольна виконана охайно, з дотримання основних правил, то виставляється 9 - 10 балів;

–якщо методика виконання запропонованого завдання розроблена вірно, але допущені непринципові помилки у теоретичному описі або розрахунках, то виставляється 6 - 8 балів;

–від 3 до 5 балів нараховується, якщо методика виконання завдання розроблена в основному вірно, але допущені деякі з наступних помилок: помилки у представлені вихідних даних, не обгрунтовані теоретичні рішення, помилки у методиці розрахунків;

–нижче 3 балів нараховується, якщо завдання не виконане або допущені грубі помилки.

3. Екзамен

Екзамен відбувається у письмовій формі. Максимальна оцінка за екзамен складає rЕК = 27 балів.

Умови позитивної проміжної атестації

Для отримання „зараховано” з першої проміжної атестації студент повинен мати не менше, ніж 12 балів (за умови, що за 8 тижнів згідно з календарним планом контрольних заходів студент повинен отримати 24 бали).

Для отримання „зараховано” з другої проміжної атестації студент повинен мати не менше, ніж 40 балів (за умови, що за 14 тижнів згідно з календарним планом контрольних заходів студент повинен отримати 76 балів).

Розрахунок шкали (R) рейтингу:

Сума вагових балів контрольних заходів протягом семестру складає:

$$R=63+10+27=100 \text{ балів}$$

Таким чином, рейтингова шкала з кредитного модуля складає 100 балів.

Умови допуску до іспиту: зарахування всіх лабораторних робіт, а також стартовий рейтинг $r \geq 40$ балів.

Для отримання студентом відповідних оцінок (ECTS та традиційних) його рейтингова оцінка RD переводиться згідно таблиці:

Таблиця відповідності рейтингових балів оцінкам за університетською шкалою:

Кількість балів	*Оцінка*
100-95	Відмінно
94-85	Дуже добре
84-75	Добре
74-65	Задовільно
64-60	Достатньо
Менше 60	Незадовільно
Не виконані умови допуску	Не допущено

8. Додаткова інформація з дисципліни (освітнього компонента)

Методичні рекомендації

Для кращого засвоєння матеріалу дисципліни рекомендується використовувати на лекціях мультимедійні засоби навчання, які дозволяють інтенсифікувати навчальний процес, стимулювати розвиток мислення та уяви студентів, збільшувати обсяг навчального матеріалу для творчого засвоєння і використання його студентами, викликати зацікавленість та позитивне ставлення до навчання.

Методика побудована таким чином, що матеріал майже кожної лекції закріплюється виконанням завдання комп'ютерного практикуму. Завдання студенти отримують заздалегідь і на аудиторному занятті під керівництвом викладача виправляють помилки в разі їх наявності та відповідають на запитання щодо програмної реалізації та теоретичних засад роботи.
Якість самостійної роботи перевіряється на заняттях комп'ютерного практикуму.

Робочу програму навчальної дисципліни (силабус): «Основи Big Data аналітики»:

Складено професором кафедри ІПЗЕ, д.т.н., доц. Федоровою Наталією Володимирівною

Ухвалено кафедрою ІПЗЕ (протокол № 34 від 10.05.2024 р.)

Погоджено Методичною комісією факультету[1] (протокол № 9 від 31.05.2024 р.)

---

[1] Методичною радою університету – для загальноуніверситетських дисциплін.