КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА

Факультет комп'ютерних наук та кібернетики
Кафедра прикладної статистики

«ЗАТВЕРДЖУЮ»
Заступник декана
з навчальної роботи
_____ Олена КАШПУР
«12» лютого 2021 року

РОБОЧА ПРОГРАМА НАВЧАЛЬНОЇ ДИСЦИПЛІНИ
Математичні методи обробки інформації

для студентів

галузь знань	**12 – Інформаційні технології**
спеціальність	**122 - Комп'ютерні науки**
освітній рівень	**бакалавр**
освітня програма	**Інформатика**
вид дисципліни	**вибіркова**

Форма навчання	**денна**
Навчальний рік	**2022/2023**
Семестр	**5**
Кількість кредитів ECTS	**3**
Мова викладання, навчання та оцінювання	**українська**
Форма заключного контролю	**залік**

Викладачі: **к.ф.-м.н, доц. Лівінська Ганна Володимирівна**

Пролонговано: на 20__/20__ н.р. _____(_____) «__»____20__р.

на 20__/20__ н.р. _____(_____) «__»____20__р.

КИЇВ – 2021

Розробник: **Лівінська Ганна Володимирівна** канд. фіз.-мат. н., доцент кафедри прикладної статистики

ЗАТВЕРДЖЕНО
Зав. кафедри Прикладної статистики

_____ (Євген ЛЕБЄДЄВ)

Протокол № 7 від «11» лютого 2021 р.

Схвалено гарантом освітньо-професійної програми «Інформатика»

_____ Людмила ОМЕЛЬЧУК «11» лютого 2021 рік
(підпис) (прізвище та ініціали)

Схвалено науково-методичною комісією факультету комп'ютерних наук та кібернетики

Протокол від «11» лютого 2021 року № 7

Голова науково-методичної комісії _____ Людмила ОМЕЛЬЧУК
(підпис) (прізвище та ініціали)

1. Мета дисципліни – засвоєння основних методів та засобів розв'язання задач аналізу та обробки даних незалежно від їх природи, а також навичок по їх використанню за допомогою мови програмування R.

2. Попередні вимоги до опанування або вибору навчальної дисципліни

– *Знати*: основні поняття та методи теорії ймовірностей та математичної статистики, математичного аналізу та алгебри, основи програмування.

– *Вміти*: застосовувати знання з теорії ймовірностей та математичної статистики, розробляти програми на базовому рівні.

– *Володіти елементарними навичками*: розв'язувати задачі з теорії ймовірностей та математичної статистики, навичками програмування.

3. Анотація навчальної дисципліни

Дисципліна має такі розділи: Попередня обробка даних. Кореляційний аналіз. Регресійний аналіз. Дисперсійний аналіз. Коваріаційний аналіз. Аналіз часових рядів. Задачі класифікації. Факторний аналіз. Дисципліна є дисципліною вільного вибору студента. Використовує поняття з теорії ймовірностей та математичної статистики, математичного аналізу та алгебри. Виступає базовою для дисциплін: основи розпізнавання образів, інтелектуальна обробка даних, проблеми штучного інтелекту, нейромережі та нейрообчислення, розпізнавання образів та аналіз сцен, основи Data Mining, ряду дисциплін вільного вибору студента (за блоками), а також буде корисна при написанні випускних кваліфікаційних робіт бакалаврів та магістрів. Викладається в 5-му семестрі, обсяг 90 год. (3 кредити ECTS), з них лекції – 28 год., практичні – 14 годин, консультації – 2 год., самостійна робота – 46 год. Передбачено 6 лабораторних робіт та залік.

В результаті вивчення навчальної дисципліни студент повинен

– **знати**: основні визначення, формули, моделі, методи та засоби розв'язання задач по всім основним розділам аналізу даних.

– **вміти**: користуватися усім спектром методів аналізу даних при розв'язанні прикладних проблем, в тому числі з використанням мови програмування R.

4. Завдання (навчальні цілі)

набуття знань, умінь та навичок (компетенцій) на рівні новітніх досягнень у сфері обробки та аналізу різноманітних даних та набуття досвіду програмування мовою R при розв'язанні прикладних задач, відповідно до кваліфікації «фахівець з інформаційних технологій». Зокрема розвивати:

– володіння усім арсеналом методів та засобів аналізу даних

– здатність застосувати знання у практичних ситуаціях;

– здатність оцінювати та забезпечувати якість виконуваних робіт;

– здатність вірно сформулювати задачу та обрати методи обробки даних, проектувати етапи необхідного аналізу, реалізовувати відповідні процедури аналізу на комп'ютері за допомогою мови програмування R та надавати трактовку отриманим результатам аналізу.

5. Результати навчання за дисципліною

Результат навчання (РН) (1. – знати; 2. – вміти; 3. – комунікація; 4. – автономність та відповідальність)		**Форми (та/або методи і технології) викладання та навчання**	**Методи оцінювання та пороговий критерій оцінювання (за необхідності)**	**Відсоток у підсумковій оцінці з дисципліни**
Код	**Результат навчання**			
РН.1.	Знати і розуміти основні розділи і задачі аналізу даних.	Лекції, практичні заняття	Захист лабораторних робіт, поточне оцінювання	18%

РН.2.1	Вміти для реальних даних належним чином сформулювати задачу відповідного розділу аналізу даних, визначити спектр методів для необхідного аналізу, побудувати вірну математичну модель, проаналізувати отриману модель та отримані результати, виявити та виправити недоліки моделі.	Лекції, практичні заняття	Захист лабораторних робіт	22%
РН.2.2	Вміти реалізовувати методи основних напрямків аналізу даних для обробки, аналізу та прогнозу реальних масивів даних за допомогою мови програмування R.	Практичні заняття	Захист лабораторних робіт	24%
РН.3.1	Обґрунтовувати власний погляд на задачу, спілкуватися з колегами з питань проектування та розробки програм, складати письмові звіти	Практичні заняття, самостійна робота	Захист лабораторних робіт, поточне оцінювання	10%
РН.3.2	Демонструвати навички взаємодії з іншими людьми, уміння працювати в команді.	Практичні заняття, самостійна робота	Захист лабораторних робіт, поточне оцінювання	6%
РН.4.1	Уміти організувати самостійну роботу та одержувати результат у рамках обмеженого часу	Самостійна робота	Захист лабораторних робіт, поточне оцінювання	6%
РН.4.2	Виявляти здатність до самонавчання та продовження професійного розвитку.	Самостійна робота	Захист лабораторних робіт, поточне оцінювання	6%
РН.4.3	Відповідально ставитися до виконуваних робіт, нести відповідальність за їх якість	Практичні заняття	Захист лабораторних робіт, поточне оцінювання	8%

6 Співвідношення результатів навчання дисципліни з програмними результатами навчання

Результати навчання дисципліни **Програмні результати навчання**	**РН 1.1**	**РН 1.2**	**РН .3. 1**	**РН .3. 2**	**РН .3. 3**
(з опису освітньої програми)					
ПРН1. Застосовувати знання основних форм і законів абстрактнологічного мислення, основ методології наукового пізнання, форм і методів вилучення, аналізу, обробки та синтезу інформації в предметній області комп'ютерних наук.	+	+			
ПРН2. Використовувати сучасний математичний апарат неперервного та дискретного аналізу, лінійної алгебри, аналітичної геометрії, в професійній діяльності для розв’язання задач теоретичного та прикладного характеру в процесі проектування та реалізації об’єктів інформатизації.		+	+		

ПРН3. Демонструвати знання закономірностей випадкових явищ, їх властивостей та операцій над ними, моделей випадкових процесів та сучасних програмних середовищ для розв'язування задач статистичної обробки експериментальних даних і побудови прогнозних моделей.		+	+	+	+

7. Схема формування оцінки

7.1 Форми оцінювання студентів:

- семестрове оцінювання:

1. Лабораторна робота 1 + тематичний тест: РН 1., РН 2.1, РН 2.2 – 18 балів.

2. Лабораторна робота 2 + тематичний тест: РН 1., РН 2.1, РН 2.2 – 12 балів.

3. Лабораторна робота 3 + тематичний тест: РН 1., РН 2.1, РН 2.2 – 18 балів.

4. Лабораторна робота 4 + тематичний тест: РН 1., РН 2.1, РН 2.2 – 16 балів.

5. Лабораторна робота 5 + тематичний тест: РН 1., РН 2.1, РН 2.2 – 18 балів.

6. Лабораторна робота 6 + тематичний тест: РН 1., РН 2.1, РН 2.2 – 18 балів.

Максимальна кількість балів які можуть бути отримані студентом: 100 балів.

Для отримання загальної позитивної оцінки з дисципліни кількість балів, набраних студентом протягом навчального семестру, має бути не меншою за 60.

7.2 Організація оцінювання

Терміни проведення оцінювання

1. Лабораторна робота 1 + тематичний тест: не пізніше 4 тижня семестру.

2. Лабораторна робота 2 + тематичний тест: не пізніше 6 тижня семестру.

3. Лабораторна робота 3 + тематичний тест: не пізніше 8 тижня семестру.

4. Лабораторна робота 4 + тематичний тест: не пізніше 10 тижня семестру.

5. Лабораторна робота 5 + тематичний тест: не пізніше 12 тижня семестру.

6. Лабораторна робота 6 + тематичний тест: не пізніше 14 тижня семестру.

За відсутності студента з поважних причин перездача лабораторних робіт здійснюється відповідно до «Положення про порядок оцінювання знань студентів при кредитно-модульній системі організації навчального процесу» від 1 жовтня 2010 року.
У разі неякісного виконання лабораторної роботи, викладач має право не зарахувати лабораторну роботу, або знизити за неї бали.
Студент має право здавати лабораторні роботи після закінчення визначеного для них терміну, але з втратою одного балу за кожен тиждень, який пройшов з моменту закінчення терміну її здачі.

7.3 Шкала відповідності оцінок

Зараховано / Passed	60-100
Незараховано / Failed	0-59

8. Структура навчальної дисципліни. Тематичний план лекцій і практичних занять

№ п/п	Назва лекції	Кількість годин		
		лекції	практичні	с/р
	Частина 1 **«Попередня обробка даних. Кореляційний та регресійний аналізи.»**			
1	**Тема 1. Вступ.** Збір даних. Методи та задачі аналізу даних. Основні розділи аналізу даних. Мова програмування R. *Самостійна робота:* Ознайомлення з R, R-studio, бібліотеками та пакетами прикладних програм аналізу даних CRAN.	2	2	4
2	**Тема 2. Попередня обробка даних.** Типи даних. Аномальні спостереження. Групування даних. Описові статистики: характеристики положення центру, характеристики розсіювання, характеристики форми розподілу. Характеристики випадкових векторів. Перевірка стохастичності вибірки. *Самостійна робота:* Ознайомлення з R, R-studio, бібліотеками та пакетами прикладних програм аналізу даних CRAN.	2		4
3	**Тема 3. Розвідувальний аналіз.** Графічні методи аналізу вибірки Визначення виду розподілу спостережуваної величини. Перевірка нормальності вибірки. *Самостійна робота:* Ознайомлення з R, R-studio, бібліотеками та пакетами прикладних програм аналізу даних CRAN. Виконання лабораторної роботи №1.	2	2	4
4	**Тема 4. Кореляційний аналіз кількісних даних.** Кореляційний та причинно-наслідковий зв'язок між змінними. Аналіз наявності статистичного зв'язку між кількісними змінними. Перевірка статистичної значущості показників статистичного зв'язку. Показники наявності нелінійного статистичного зв'язку. Аналіз множинних статистичних зв'язків. *Самостійна робота:* Реалізація методів кореляційного аналізу в R. Виконання лабораторної роботи №2.	2	2	4
5	**Тема 5. Кореляційний аналіз інших типів даних. Візуалізація кореляцій.** Кореляційний аналіз ординальних змінних, рангова кореляція. Виявлення статистичного зв'язку між номінальними змінними, таблиці спряженості. Візуалізація кореляцій: діаграма розсіювання, карта кореляцій, граф кореляцій. *Самостійна робота:* Реалізація методів кореляційного аналізу в R. Виконання лабораторної роботи №2.	2		2
6	**Тема 6. Регресійний аналіз. Побудова моделі.** Поняття регресії. Основні етапи побудови та верифікації регресійної моделі. Проста лінійна регресія. Метод найменших квадратів. Метод найменших модулів. Метод повторних медіан. Загальна лінійна модель. Умови Гаусса-Маркова. МНК-оцінки для загальної лінійної моделі. *Самостійна робота:* Реалізація методів регресійного аналізу в R. Виконання лабораторної роботи №3.	2		4
7	**Тема 7. Регресійний аналіз. Тестування та корекція моделі.** Оцінювання точності оцінок коефіцієнтів регресії. Оцінка точності моделі. Загальна техніка підгонки регресійних моделей з використанням графічних методів. Типові недоліки регресійної моделі. Зважений МНК. Розширення лінійної моделі. *Самостійна робота:* Реалізація методів регресійного аналізу в R. Виконання лабораторної роботи №3.	2	2	2
	Частина 2 **«Дисперсійний та коваріаційний аналізи. Аналіз часових рядів. Задачі класифікації.»**			
8	**Тема 8. Однорідність вибірок.** Однорідність двох незалежних вибірок. Альтернативи однорідності. Повторні вибірки. Методи перевірки однорідності двох незалежних вибірок. Порівняння залежних (повторних) вибірок. Множинні порівняння груп. *Самостійна робота:* Тести для перевірки однорідності в R.	2		2

9	**Тема 9. Дисперсійний аналіз (ANOVA). Основні поняття та побудова однофакторної моделі.** Постановка задачі ANOVA, його сутність та переваги. Класифікація видів ANOVA. Побудова математичної моделі однофакторного ANOVA. Математичні припущення ANOVA. Оцінювання параметрів моделі за допомогою МНК. *Самостійна робота:* Реалізація методів дисперсійного аналізу в R. Виконання лабораторної роботи №4.	2	2	2
10	**Тема 10. Дисперсійний аналіз.** Аналіз контрастів. Дисперсійний аналіз по Краскеру-Уоллісу. Побудова математичної моделі двофакторного ANOVA. Багатофакторний дисперсійний аналіз. *Самостійна робота:* Реалізація методів дисперсійного аналізу в R. Виконання лабораторної роботи №4.	2		4
11	**Тема 12. Часові ряди. Методи аналізу.** Визначення, класифікація та приклади ЧР. Цілі аналізу ЧР. Моделі ЧР: стаціонарні ЧР, складові адитивної моделі ЧР, мультиплікативна модель. Згладжування ЧР, метод рухомого середнього. Оцінка компонент адитивної моделі ЧР: оцінка тренду, оцінка сезонної складової. Автокореляція та корелограма. *Самостійна робота:* Реалізація методів аналізу часових рядів в R. Виконання лабораторної роботи №5.	2		4
12	**Тема 13. Часові ряди. Методи прогнозу.** Побудова прогнозу з використанням експоненційного згладжування: метод Брауна; метод Хольта; метод Хольта-Вінтерса. Модель авторегресії рухомого середнього (ARMA) та авторегресії проінтегрованого рухомого середнього (ARIMA). Оцінка якості прогнозу ЧР. *Самостійна робота:* Реалізація методів аналізу часових рядів в R. Виконання лабораторної роботи №5.	2	2	4
13	**Тема 14. Зниження розмірності простору ознак. Метод головних компонент.** Мета та методи зниження розмірності простору ознак. Обґрунтування методу головних компонент. Змістовні обмеження МГК. Властивості головних компонент. Етапи МГК. Зображення та аналіз отриманих результатів. *Самостійна робота:* Реалізація методу головних компонент в R.	2		2
14	**Тема 15. Задачі класифікації. Кластерний аналіз.** Визначення та задачі кластерного аналізу. Типи даних та міра відстані між об'єктами. Методи кластерного аналізу: ієрархічні та неієрархічні. Вибір оптимальної кількості кластерів. Оцінка якості кластеризації. *Самостійна робота:* Реалізація методів кластерного аналізу в R. Виконання лабораторної роботи №6.	2	2	4
	ВСЬОГО	**28**	**14**	**46**

Загальний обсяг 90 *год.,* в тому числі:
Лекцій **– 28** *год.*
Практичні **– 14** *год.*
Самостійна робота **– 46** *год.*
Консультації – **2 *год***.

Умови лабораторних робіт:

Лабораторна робота 1: Попередня обробка даних. Описові статистики. Розвідувальний аналіз.
Лабораторна робота 2: Виявлення кореляційного зв'язку між кількома змінними. Перевірка на значущість показників кореляційного зв'язку. Зображення кореляцій.
Лабораторна робота 3: Побудова регресійної моделі. Перевірка якості та корекція моделі.
Лабораторна робота 4: Проведення дисперсійного аналізу для відповідної задачі. Аналіз контрастів.
Лабораторна робота 5: Аналіз та прогноз часового ряду. Перевірка якості побудованого прогнозу.
Лабораторна робота 6: Кластеризація даних з використанням різних методів. Обґрунтування вибору кількості кластерів. Оцінка якості кластеризації.
Деталізовані умови лабораторних робіт даються студентам на першому практичному занятті.

9. Рекомендовані джерела:

Основні:

1) Каримов Р. Н. *Основы дискриминантного анализа: Учебно-методическое пособие.* — Саратов: СГТУ, 2002. -108ст.
2) Лагутин М.Б., *Наглядная математическая статистика*, М.: БИНОМ, 2007.
3) Майборода Р.Є. *Комп'ютерна статистика – професійний старт*, 2018, 482 ст. http://probability.univ.kiev.ua/userfiles/mre/compsta1.pdf
4) Майборода Р.Є., Сугакова О.В. *Аналіз даних за допомогою пакета R* http://matphys.rpd.univ.kiev.ua/downloads/courses/mmatstat/Statistics_with_R.pdf
5) Слабоспицький О.С., *Аналіз даних. Попередня обробка*, ВПЦ "Київський університет" (2001).
6) Слабоспицький О.С., *Основи кореляційного аналізу даних*, ВПЦ "Київський університет" (2006).
7) Слабоспицький О.С., *Основи дисперсійного аналізу даних*, ВПЦ "Київський університет" (2006).
8) Brockwell, P.J., Davis, R.A., *Introduction to Time Series and Forecasting* (3-d edition, 2016).
9) A. Coghlan, *A Little Book of R For Multivariate Analysis* (2017).
10) P. Dalgaard, *Introductory Statistics with R*, 2-nd edition, Springer (2008).
11) Br. S. Everitt, S. Landau, M. Leese, D. Stahl, *Cluster Analysis*, 5-th edition, Wiley series in probability and statistics (2011).
12) A. Field, *Discovering Statistics Using SPSS*, third edition, SAGE Publications Ltd (2009).
13) W. Hardle, L. Simar, *Applied Multivariate Statistical Analysis*, second edition, Springer (2007).
14) D. W. Hoster, St. Lemeshow, *Applied Logistic Regression*, second edition, Wiley series in probability and statistics (2000).
15) G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer (2013).
16) J. Maindonald, W.John Braun, *Data Analysis and Graphics Using R – an Example-Based Approach*, Cambridge Univ. Press.(2010).
17) R.H. Shumway and D.S. Stoffer, *Time Series Analysis and Its Applications. With R Examples.* 2nd edition. Springer (2006).
18) N. H. Timm, *Applied Multivariate Analysis*, Springer (2002).
19) http://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/

Додаткові:

1) Айвазян С.А., Енюков И.С., Мешалкин Л.Д., *Прикладная статистика. Основы моделирования и первичная обработка данных.* М.: Финансы и статистика, 1983.
2) Бокс, Дж., Дженкинс, Г., *Анализ временных рядов. Прогноз и управление.* (1974)
3) Бриллинджер, Д., *Временные ряды. Обработка данных и теория.* (1980)
4) Ивченко Г.И., Медведев Ю.И. *Математическая статистика.* – М.: Высш. шк., 1984.
5) Кендэлл, М., *Временные ряды.* (1981)
6) Майборода Р.Є., *Регресія: лінійні моделі.* ВПЦ "Київський університет", 296 р. – 2007. http://probability.univ.kiev.ua/userfiles/mre/ora0.pdf
7) Тьюки Дж., Анализ результатов наблюдений. Разведочный анализ. – М.: Мир, 1981.

Datasets

https://www.kaggle.com/datasets

https://vincentarelbundock.github.io/Rdatasets/datasets.html