

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
СХІДНОУКРАЇНСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВОЛОДИМИРА ДАЛЯ

МЕТОДИЧНІ ВКАЗІВКИ
до виконання контрольної роботи з дисципліни
"Вступ до інтелектуального аналізу даних"
для здобувачів вищої освіти за спеціальністю
"Комп'ютерні науки" заочної форми навчання

(електронне видання)

ЗАТВЕРДЖЕНО
на засіданні кафедри
комп'ютерних наук та інженерії
Протокол № 6 від 03.04.2026 р.

Київ 2026

УДК 004.02

Методичні вказівки до виконання контрольної роботи з дисципліни "Вступ до інтелектуального аналізу даних" для здобувачів вищої освіти за спеціальністю "Комп'ютерні науки" заочної форми навчання / Уклад. : Л.О. Шумова – Київ: вид-во СНУ ім. В. Даля, 2026. – 27 с.

Письмова контрольна робота є основним етапом вивчення курсу "Вступ до інтелектуального аналізу даних". Головними завданнями її виконання є систематизація, закріплення і поглиблення теоретичних знань про використання технологій інтелектуального аналізу даних для задач бізнес-аналітики. Методичні вказівки та завдання для виконання контрольної роботи розроблені у відповідності до робочої програми курсу "Вступ до інтелектуального аналізу даних" і містять вивчення основних завдань та методів збору та аналізу даних у різних предметних додатках.

Укладач:

Л.О. Шумова, к.т.н., доцент;

Рецензент:

Є.В. Щербаков, к.т.н., доцент

ЗМІСТ

ВСТУП	4
1 ОСНОВНІ ВИМОГИ ДО КОНТРОЛЬНОЇ РОБОТИ.....	5
2 ВИМОГИ ДО ЗМІСТУ ТА ЕТАПІВ ВИКОНАННЯ РОБОТИ.....	6
3 ПРИКЛАД ВИКОНАННЯ КОНТРОЛЬНОЇ РОБОТИ	9
ПЕРЕЛІК РЕКОМЕНДОВАНОЇ ЛІТЕРАТУРИ.....	25
ДОДАТОК А. Питання теоретичного завдання	26
ДОДАТОК Б. Візуалізація вхідних даних	27

ВСТУП

У найзагальнішому сенсі аналіз даних є процесом дослідження, фільтрації, перетворення й моделювання даних з метою отримання корисної інформації. Основною метою курсу "Вступ до інтелектуального аналізу даних" є придбання теоретичних і практичних знань, умінь і навичок, орієнтованих на ефективне професійне використання сучасних технологій інтелектуального аналізу даних.

Сучасна концепція аналізу даних отримала назву Data Mining, що перекладається як «видобуток» або «розкопка даних». Нерідко поруч з Data Mining зустрічаються означення «виявлення знань в базах даних» (knowledge discovery in databases) і «інтелектуальний аналіз даних», які можна вважати синонімами Data Mining, хоча кожне з них концентрує увагу на певних аспектах загального процесу інтелектуального аналізу даних.

Вирішуються наступні завдання:

- дослідження механізмів збору та підготовки даних,
- аналіз інструментальних засобів візуалізації,
- дослідження та використання методів інтелектуального аналізу даних.

При виконанні контрольної роботи формуються та закріплюються навички в наступних областях:

- пошук: робота з традиційними носіями і інтернет-евристика;
- критичне мислення: навчання ефективним технологіям осмислення фактів, подій та ін.;
- рішення задач інтелектуального аналізу даних: використання моделей, методів, алгоритмів та програмного забезпечення;
- презентація результатів: написання підсумкових звітів, візуалізація даних.

У перших двох розділах методичних вказівок до контрольної роботи надано порядок виконання і вимоги до структури і змісту основних етапів роботи. У розділі 3 надано приклад виконання та оформлення контрольної роботи.

1 ОСНОВНІ ВИМОГИ ДО КОНТРОЛЬНОЇ РОБОТИ

У межах курсу "Вступ до інтелектуального аналізу даних" студенти мають виконати одну контрольну роботу в міжсесійний період після вивчення відповідних тем навчальної програми.

Номера питань та задач вибираються згідно номеру по журналу. Викладач може замінити варіант контрольної роботи, враховуючи специфіку діяльності студента.

Контрольна робота надається викладачу в електронному вигляді та повинна бути виконана, дотримуючись державних стандартів оформлення технічної документації, які включають вимоги до шрифту, інтервалу та полів. Зазвичай використовується шрифт Times New Roman, розмір 14 пунктів, з інтервалом 1,5 між рядками..

Виконуючи завдання, необхідно переписати зміст питання і задачі, а потім дати на нього конкретну змістовну відповідь з посиланням на відповідні рисунки, літературу тощо.

Після виконання всіх завдань, повинен бути приведений список використаної літератури (автор, назва, рік видання). На титульному аркуші необхідно вказати найменування навчального закладу, назву дисципліни, курс, групу, прізвище, ім'я, по батькові виконавця роботи.

Виконана згідно індивідуального завдання і правильно оформлена контрольна робота подається викладачу на рецензування в термін, встановлений планом-графіком факультету ІТтаЕ. Отримавши рецензію на зараховану контрольну роботу, студент зобов'язаний уважно ознайомитися із зауваженнями викладача, внести необхідні виправлення чи доповнення.

Контрольна робота, яка не відповідає пред'явленим вимогам, виконана з відступом від завдання, не рецензується і не зараховується. Вона виконується студентом повторно і подається викладачу для повторного рецензування. На титульному листі повторно виконаної роботи необхідно зробити помітку "Повторна".

Здобувачи освіти, які не виконали контрольну роботу в установленний термін, до складання іспиту не допускаються.

2 ВИМОГИ ДО ЗМІСТУ ТА ЕТАПІВ ВИКОНАННЯ РОБОТИ

2.1 Зміст контрольної роботи

Контрольна робота складається з чотирьох завдань.

Завдання 1. Теоретичне

Підготуйте відповідь на одне з питань (Додаток А).

Завдання 2

Дослідження механізмів збору та підготовки даних.

Завдання 3

Аналіз інструментальних засобів для візуалізації даних.

Завдання 4

Дослідження методів та алгоритмів інтелектуального аналізу даних. Завдання виконати за рекомендаціями до практичної роботи №6.

2.2 Етапи виконання контрольної роботи

В процесі виконання контрольної роботи виділяються наступні етапи:

1. Отримання завдання.
2. Ознайомлення з теоретичним матеріалом наданим в конспекті лекцій (презентація, додатковими матеріалами).
3. Виконання практичних робіт з дисципліни «Вступ до ІАД».
4. Розрахунки, оформлення і завантаження звітів окремо за кожною з задач на сторінці курсу «Вступ до ІАД» у Moodle.
5. Оформлення заключного звіту по контрольній роботі.

2.3 Рекомендації щодо виконання завдань

Завдання 2. Дослідження механізмів збору та підготовки даних.

Завдання виконати за рекомендаціями до практичної роботи №4.

Для проведення аналізу механізмів збору та підготовки даних необхідне виконання наступних кроків:

1. Визначитися з напрямом дослідження, запропонувати власний варіант проекту або отримати у викладача типові завдання і посилання на БД з вихідними даними для проектування.
2. У відповідності з обраним варіантом провести збір даних. Збір даних містить в собі відбір випадків або прикладів об'єктів з різних типів або категорій (класів), для подальшого використання для виявлення груп або подібності в даних (у неконтрольованому навчанні) або прогнозування класу нових об'єктів (у випадку з навчанням з учителем).

3. Провести підготовку даних. Якщо введення даних проводиться вручну, необхідно об'єднати їх в лист таблиці MS Excel і зберегти для подальшого перетворення в формат CSV, один з форматів, який використовується в програмному забезпеченні Weka. CSV (Comma-Separated Values) - текстовий формат, призначений для представлення табличних даних. Кожен рядок файлу - це один рядок таблиці. Значення окремих колонок розділяються розділовим символом - комою.

4. Отриманий файл треба конвертувати в сумісний з програмним забезпеченням Weka формат. Ця дія може бути виконана в MS Excel в опції «Зберегти як» вибрати «Інші формати» і визначити тип файлу CSV (роздільники - коми). Однак, таке збереження не завжди коректно сприймається Weka, тому рекомендується конвертація за допомогою вільного он-лайн ресурсу <http://www.zamzar.com>.

5. Сформувати матрицю даних в Weka.

Завдання 3. Аналіз інструментальних засобів для візуалізації даних, подання найважливіших особливостей набору даних у графічному вигляді. Завдання виконати за рекомендаціями до практичної роботи №5.

1. У відповідності з обраним варіантом, визначитися для якого етапу аналізу даних буде проводиться візуалізація.

2. Визначити об'єм даних, для яких на цьому етапі заплановано проводити візуалізацію (всі дані або тільки їх частина).

3. Обрати метод, який найбільш наочно відображає закономірності, притаманні досліджуваному набору даних.

4. Виконати візуалізацію вихідного набору даних за допомогою обраного варіанту програмного забезпечення.

5. Провести аналіз та надати пояснення щодо отриманих результатів візуалізації даних.

Завдання 4. Дослідження методів та алгоритмів інтелектуального аналізу даних для вирішення практичних задач прийняття рішень. Завдання виконати за рекомендаціями до практичної роботи №6.

В цій частині роботи необхідно провести дослідження якості алгоритмів регресії або класифікації даних, отриманих при виконанні завдання 2. Аналіз даних рекомендовано проводити за допомогою вбудованих алгоритмів:

1. Запустити Weka, вибрати програму Explorer.

2. На вкладці Preprocess завантажити навчальний набір даних у Weka, кількість даних якого складає 60-70 % від тестового набору даних.

3. Перейти на вкладку Classify (рис.3.6). За допомогою кнопки Choose вибрати необхідний для роботи алгоритм.

4. Провести аналіз даних. Умовно його можна поділити на декілька частин:

а) опис вхідних даних (кількість атрибутів, наборів даних, обраного варіанту аналізу).

б) опис класифікованої моделі,

в) прогноз для набору даних, що аналізується.

г) аналіз інформації про побудовану модель, по якій ми зможемо судити про її якість та коректність.

5. Виконати візуалізацію отриманої моделі, якщо для використовуваного алгоритму доступна ця функція.

6. Провести аналіз помилок для свого завдання.

7. Виконати візуалізацію помилково класифікованих результатів.

8. Отримані показники, що характеризують точність та якість методів класифікації або регресії досліджуваних моделей для навчального набору даних, представити у вигляді таблиці, шаблон якої надано у табл. 2.2.

Таблиця 2.2 – Порівняльний аналіз показників точності та якості методів класифікації.

Алгоритм	Показники точності та якості моделі	
	Correctly Classified Instances, %	ROC Area
Naïve Bayes		
Bayes Net		
Random Forest		
Alternating decision tree		

9. Виходячи з порівняльного аналізу показників, зробити висновок, щодо точності використовуваних моделей.

10. На підставі висновку, щодо точності використовуваних моделей, визначитись з найкращою моделлю, що може бути використана для подальшої роботи та класифікації нових даних, для яких клас ще невідомий.

Завдання 5. Заключний звіт з результатами виконання роботи.

1. Проаналізувати результати попередніх завдань.

2. Використовуючи Приклад у п. 3 методичних вказівок скласти заключний звіт з виконання проекту.

3. Заключний звіт з результатами виконання контрольної роботи завантажити на сторінці курсу «Вступ до ІАД» у Moodle.

3 ПРИКЛАД ВИКОНАННЯ КОНТРОЛЬНОЇ РОБОТИ

Завдання 2. Збір та підготовка даних

Прогнозування кількості орендованих велосипедів з метою підтримки прийняття рішень при плануванні навантаження пунктів прокату.

Короткий опис задачі:

Система спільного використання велосипедів - система прокату, зазвичай створена на некомерційній основі, що дозволяє орендувати велосипед на одній з автоматизованих станцій, зробити поїздку і повернути велосипед у будь-який пункт прокату, встановлений в цьому ж місті. На сьогоднішній день існує близько 500 пунктів вело-обміну в усьому світі, які складаються з більше 500 тис. велосипедів. Інтерес до цих систем обумовлений бажанням поліпшити транспортну та екологічну обстановку в містах, а також покращити стан здоров'я населення.

Весь процес роботи систем оренди велосипедів, членство, оренда та повернення, - автоматизований. Всі дані фіксуються. Прогнозування завантаженості пунктів прокату велосипедів допоможе оптимізувати їх роботу.

Реалізація задачі.

Кількість орендованих велосипедів напряму залежить від екологічних та сезонних умов. Наприклад, погодні умови, опади, день тижня, сезон, час дня, і т.д. можуть вплинути на кількість орендованих велосипедів.

В роботі використовувалися дані за 2011 рік та за 11 місяців 2012 року, отримані в Capital Bikeshare system, міста Вашингтон, округу Колумбія, США, які доступні на сайті системи за посиланням <http://capitalbikeshare.com/system-data>. Також додані відповідні дані про погодні умови та сезонна інформація, доступна за посиланням <http://www.freemeteo.com>.

На підставі цих даних ми зробимо прогнозування щодо кількості орендованих велосипедів на наступний місяць.

Для нашого набору даних ми обрали 12 факторів, які можуть впливати на кількість користувачів в точці прокату велосипедів:

instant – випадок;

season – сезон (1 - весна, 2 - літо, 3 - осінь, 4 - зима);

yr – рік (0 - 2011, 1 - 2012);

month – місяць (от 1 до 12);
 holiday – святковий день, (так=1, ні=0);
 weekday – день тижня (0 – неділя, 1 - понеділок, 2 – вівторок, 3 – середа, 4 – четвер, 5 – п'ятниця, 6 - субота);
 workingday – робочий день (якщо не свято та не вихідний, то =1, в протилежному випадку =0);
 weathersit – погодні умови (1- змінна хмарність, декілька хмар, мало хмарно; 2 - туман та ясно, туман та хмарно з проясненням, туман та декілька хмар; 3 – невеликий сніг, дощ, гроза та змінна хмарність, дощ та розсіяні хмари; 4 – ливень, ожеледиця, гроза та туман, сніг та туман);
 temp – нормалізована температура повітря (для кожного випадку розраховується по формулі $(t - t_{\min}) / (t_{\max} - t_{\min})$, $t_{\min} = -8$, $t_{\max} = +39$);
 atemp - нормалізована температура повітря, по відчуттям (для кожного випадку розраховується по формулі $(t - t_{\min}) / (t_{\max} - t_{\min})$, $t_{\min} = -16$, $t_{\max} = +50$);
 hum – нормалізована вологість повітря, істинне значення ділиться на 100 (максимальне значення);
 windspeed – нормалізована швидкість вітру, істинне значення ділиться на 67 (максимальне значення).
 Показник cnt показує загальну кількість користувачів в точці прокату.
 Проведена підготовка даних. Дані об'єднали в лист таблиці MS Excel і зберегли для подальшого перетворення в формат CSV, один з форматів, який використовується в програмному забезпеченні Weka. CSV (Comma-Separated Values) - текстовий формат, призначений для представлення табличних даних. Кожен рядок файлу - це один рядок таблиці. Значення окремих колонок розділяються розділовим символом - коми.
 На наступному етапі було проведено конвертування отриманого файлу в сумісний з програмним забезпеченням Weka формат (рис. 3.1).
 Ця дія виконана в MS Excel в опції «Зберегти як» вибрати «Інші формати» і визначити тип файлу CSV (роздільники - коми).
 З огляду на те, що збереження не завжди коректно сприймається Weka, в даному проекті конвертацію проведено за допомогою вільного он-лайн ресурсу <http://www.zamzar.com>.

	A	B	C	D	E	F	G	H	I
	instant	season	yr	mnth	holiday	weekday	workingday	weathers	temp
2	1	1	0	1	0	6	0	2	0,3441
3	2	1	0	1	0	0	0	2	0,3634
4	3	1	0	1	0	1	1	1	0,1963
5	4	1	0	1	0	2	1	1	
6	5	1	0	1	0	3	1	1	0,2265
7	6	1	0	1	0	4	1	1	0,2043
8	7	1	0	1	0	5	1	2	0,1963
9	8	1	0	1	0	6	0	2	0,1
10	9	1	0	1	0	0	0	1	0,1383
11	10	1	0	1	0	1	1	1	0,1508
12	11	1	0	1	0	2	1	2	0,1690
13	12	1	0	1	0	3	1	1	0,1727
14	13	1	0	1	0	4	1	1	0,1
15	14	1	0	1	0	5	1	1	0,160
16	15	1	0	1	0	6	0	2	0,2333
17	16	1	0	1	0	0	0	1	0,2316
18	17	1	0	1	1	1	0	2	0,1758
19	18	1	0	1	0	2	1	2	0,2166
20	19	1	0	1	0	3	1	2	0,2923
21	20	1	0	1	0	4	1	2	0,2616
22	21	1	0	1	0	5	1	1	0,17
23	22	1	0	1	0	6	0	1	0,055

Рисунок 3.1 – Фрагмент об'єднаних даних в MS Excel

Оброблені дані представлені у форматі .csv - текстовому форматі, призначеному для представлення табличних даних (рис. 3.2).

Далі, ПЗ Weka було запущено в режимі Explorer (рис 3.3).

Сформовано матрицю даних в Weka. За допомогою кнопки "Open file" в режимі "Preprocess" відкрито файл з набором даних, що аналізується. Після успішного завантаження з'являється можливість переглянути деякі дані про кількість атрибутів, їх типи (numeric, nominal і т.д.) і кількість випадків (рядків у сегменті даних або документів).

Для перегляду, редагування або копіювання завантажених даних в табличному режимі натисніть кнопку «Edit». Після створення в Weka таблиці її можна зберегти ARFF(Attribute-Relation File Format) форматі за допомогою опції "Save". На рис. 3.4 представлений скріншот частини результуючого документа.

	A	B	C	D	E	F	G	H	I
1	instant,season,yr,mnth,holiday,weekday,workingday,weathersit,temp,atemp,hum,winds								
2	1,1,0,1,0,6,0,2,0.344167,0.363625,0.805833,0.160446,985								
3	2,1,0,1,0,0,0,2,0.363478,0.353739,0.696087,0.248539,801								
4	3,1,0,1,0,1,1,1,0.196364,0.189405,0.437273,0.248309,1349								
5	4,1,0,1,0,2,1,1,0.2,0.212122,0.590435,0.160296,1562								
6	5,1,0,1,0,3,1,1,0.226957,0.22927,0.436957,0.1869,1600								
7	6,1,0,1,0,4,1,1,0.204348,0.233209,0.518261,0.0895652,1606								
8	7,1,0,1,0,5,1,2,0.196522,0.208839,0.498696,0.168726,1510								
9	8,1,0,1,0,6,0,2,0.165,0.162254,0.535833,0.266804,959								
10	9,1,0,1,0,0,0,0,0.138333,0.116175,0.434167,0.36195,822								
11	10,1,0,1,0,1,1,1,0.150833,0.150888,0.482917,0.223267,1321								
12	11,1,0,1,0,2,1,2,0.169091,0.191464,0.686364,0.122132,1263								
13	12,1,0,1,0,3,1,1,0.172727,0.160473,0.599545,0.304627,1162								
14	13,1,0,1,0,4,1,1,0.165,0.150883,0.470417,0.301,1406								
15	14,1,0,1,0,5,1,1,0.16087,0.188413,0.537826,0.126548,1421								
16	15,1,0,1,0,6,0,2,0.233333,0.248112,0.49875,0.157963,1248								
17	16,1,0,1,0,0,0,0,0.231667,0.234217,0.48375,0.188433,1204								
18	17,1,0,1,1,1,0,2,0.175833,0.176771,0.5375,0.194017,1000								
19	18,1,0,1,0,2,1,2,0.216667,0.232333,0.861667,0.146775,683								
20	19,1,0,1,0,3,1,2,0.292174,0.298422,0.741739,0.208317,1650								
21	20,1,0,1,0,4,1,2,0.261667,0.25505,0.538333,0.195904,1927								
22	21,1,0,1,0,5,1,1,0.1775,0.157833,0.457083,0.353242,1543								
23	22,1,0,1,0,6,0,1,0.0591304,0.0790696,0.4,0.17197,981								

Рисунок 3.2 – Фрагмент оброблених даних у форматі .csv

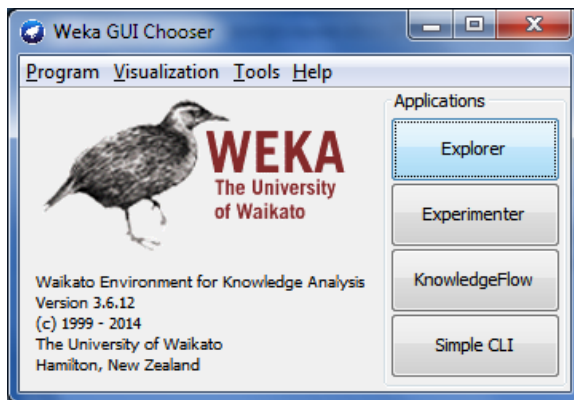


Рисунок 3.3 – Вікно вибору додатка в Weka

Relation: day1

No.	instant	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	cnt
	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
1	1.0	1.0	0.0	1.0	0.0	6.0	0.0	2.0	0.344...	0.363...	0.805...	0.160446	985.0
2	2.0	1.0	0.0	1.0	0.0	0.0	0.0	2.0	0.363...	0.353...	0.696...	0.248539	801.0
3	3.0	1.0	0.0	1.0	0.0	1.0	1.0	1.0	0.196...	0.189...	0.437...	0.248309	1349.0
4	4.0	1.0	0.0	1.0	0.0	2.0	1.0	1.0	0.2	0.212...	0.590...	0.160296	1562.0
5	5.0	1.0	0.0	1.0	0.0	3.0	1.0	1.0	0.226...	0.22927	0.436...	0.1869	1600.0
6	6.0	1.0	0.0	1.0	0.0	4.0	1.0	1.0	0.204...	0.233...	0.518...	0.089562	1606.0
7	7.0	1.0	0.0	1.0	0.0	5.0	1.0	2.0	0.196...	0.208...	0.498...	0.168726	1510.0
8	8.0	1.0	0.0	1.0	0.0	6.0	0.0	2.0	0.165	0.162...	0.535...	0.266804	959.0
9	9.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	0.138...	0.116...	0.434...	0.36195	822.0
10	10.0	1.0	0.0	1.0	0.0	1.0	1.0	1.0	0.150...	0.150...	0.482...	0.223267	1321.0
11	11.0	1.0	0.0	1.0	0.0	2.0	1.0	2.0	0.169...	0.191...	0.686...	0.122132	1263.0
12	12.0	1.0	0.0	1.0	0.0	3.0	1.0	1.0	0.172...	0.160...	0.599...	0.304627	1162.0
13	13.0	1.0	0.0	1.0	0.0	4.0	1.0	1.0	0.165	0.150...	0.470...	0.301	1406.0
14	14.0	1.0	0.0	1.0	0.0	5.0	1.0	1.0	0.16087	0.188...	0.537...	0.126548	1421.0
15	15.0	1.0	0.0	1.0	0.0	6.0	0.0	2.0	0.233...	0.248...	0.49875	0.157963	1248.0
16	16.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	0.231...	0.234...	0.48375	0.188433	1204.0
17	17.0	1.0	0.0	1.0	1.0	1.0	0.0	2.0	0.175...	0.176...	0.5375	0.194017	1000.0
18	18.0	1.0	0.0	1.0	0.0	2.0	1.0	2.0	0.216...	0.232...	0.861...	0.146775	683.0
19	19.0	1.0	0.0	1.0	0.0	3.0	1.0	2.0	0.292...	0.298...	0.741...	0.208317	1650.0
20	20.0	1.0	0.0	1.0	0.0	4.0	1.0	2.0	0.261...	0.25505	0.538...	0.195904	1927.0
21	21.0	1.0	0.0	1.0	0.0	5.0	1.0	1.0	0.1775	0.157...	0.457...	0.353242	1543.0
22	22.0	1.0	0.0	1.0	0.0	6.0	0.0	1.0	0.059...	0.079...	0.4	0.17197	981.0
23	23.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	0.096...	0.098...	0.436...	0.2466	986.0
24	24.0	1.0	0.0	1.0	0.0	1.0	1.0	1.0	0.097...	0.11793	0.491...	0.15833	1416.0

Рисунок 3.4 – Фрагмент частини завантаженого документу

Вже на цьому етапі, за допомогою ПО Weka, ми можемо отримати візуальну інформацію про деякі статистичні дані про атрибути завдяки візуалізації вхідних даних.

Завдання 3. Візуалізація даних

Використане програмне забезпечення - Weka 3.6.12.

Візуалізація призначена для наочності побудованої моделі, її кращому розумінню, кращій інтерпретації отриманого результату, оцінки якості побудованої моделі. Існує безліч видів візуалізації, що мають різне призначення: графіки, діаграми, таблиці, структурні схеми, карти, інфографіка і т.д. Деревоподібна візуалізація, наприклад, використовується для інтерпретації результатів. В залежності від кількості вимірів візуалізація даних може бути одномірною, двовимірною, тривимірною, багатовимірною.

Найчастіше, одну і ту ж інформацію можна представити за допомогою різних способів візуалізації.

Візуалізація за допомогою ПО Weka надає нам кількісну інформацію про розподіл атрибутів. Для перегляду необхідно виділити досліджуваний атрибут у вікні Attributes зліва (рис.3.5).

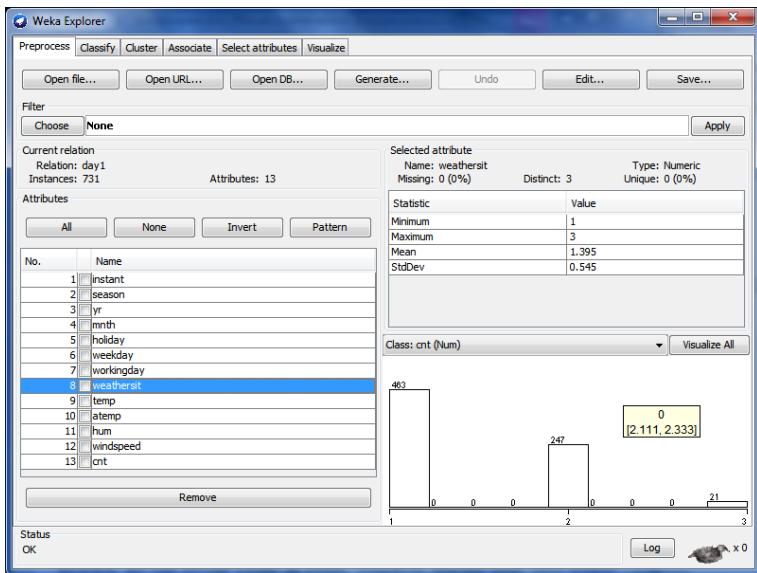


Рисунок 3.5 – Візуалізація вихідних даних за допомогою ПЗ Weka

З представленої на рис. візуалізації даних по погодним умовам ми можемо зробити висновок, що такі погодні умови, як змінна хмарність, декілька хмар, мало хмарно переважають у обраному пункті прокату велосипедів на протязі 2011-2012 років. Для візуалізації всіх відношень одразу треба натиснути кнопку «Visualize all» та отримати розподіл даних по всіх атрибутах.

Приклад візуалізації вихідних даних за допомогою інфографіки можна подивитись у Додатку Б.

Завдання 4. Дослідження методів та алгоритмів інтелектуального аналізу даних для вирішення практичних задач прийняття рішень

Використане програмне забезпечення - Weka 3.6.12

Продовжимо аналіз даних для вирішення практичних задач прийняття рішень. Поставлена задача має на увазі використання алгоритмів класифікації або регресії. Як класифікаційна, так і регресійна моделі знаходять залежності між вхідними і вихідними змінними. Вхідні дані нашого набору даних та вихідна змінна моделі безперервні, тобто можуть приймати будь-яке значення в межах певного інтервалу, значить

перед нами задача регресії. Виходячи з цього в роботі використовуємо алгоритми регресії та для оцінки ефективності моделі використовуємо показники ефективності моделей регресії.

Проводимо дослідження якості та ефективності алгоритмів регресії. Для дослідження будемо використовувати такі алгоритми: Linear Regression, SMOReg, Multiplayer Perceptron та Isotonic Regression.

Після завантаження досліджуваного набору даних переходимо на вкладку Classify (рис.3.6). За допомогою кнопки Choose вибрати необхідний для роботи алгоритм.

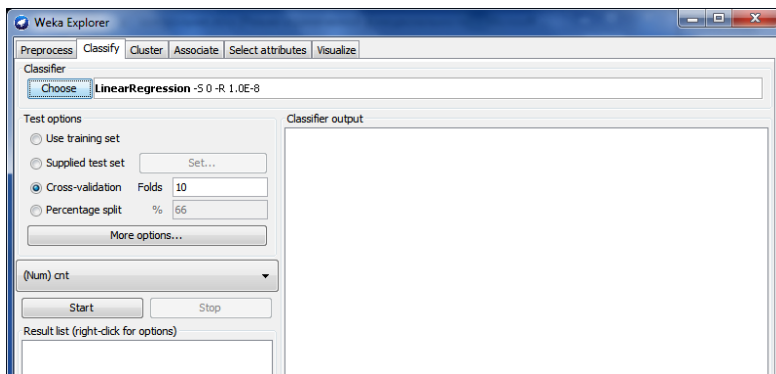


Рисунок 3.6 – Вибір алгоритму на вкладці Classify

Обираємо варіант аналізу. Це може бути Cross-validation (перехресна перевірка), можна використовувати аналіз даних за допомогою навчального та тестового наборів даних, де спочатку для навчання моделі проводиться обробка даних навчального набору, потім для тестового. При цьому варіанті аналізу розмір навчального набору даних становить 60-70 відсотків від тестового, а для оцінки ефективності моделі використовують результати обробки тестового набору даних. Ми обираємо Cross-validation. При цьому варіанті аналізу модель будується на підмножинах вихідних даних і для кінцевої моделі використовується середнє значення. Кратність перевірки можна брати будь-яку, ми оставимо 10-кратну перевірку. При такому варіанті аналізу кожна величина використовується один раз для тестування і дев'ять разів для навчання. Для перегляду подробиць прогнозування активуємо функцію Output Predictions в More options (рис.3.7).

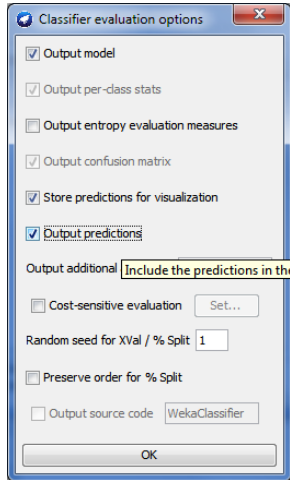


Рисунок 3.7 – Активування функції Output Predictions

Далі, ми обираємо змінну, для якої будемо будувати модель. Для нашого прогнозування це змінна *cnt*. Натискаємо Start і через декілька секунд у вікні справа бачимо результат.

Проводимо аналіз даних.

Умовно його можна поділити на декілька частин, які можуть трохи відрізнятися в залежності від використовуваного методу або алгоритму:

a) опис вхідних даних (кількість атрибутів, наборів даних, обраного варіанту аналізу).

=== Run information ===

Scheme: weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8

Relation: day1

Instances: 700

Attributes: 13

instant

season

yr

mnth

holiday

weekday

workingday

weathersit

temp

atemp

hum

windspeed

cnt

Test mode:10-fold cross-validation

Цей розділ надає нам інформацію про кількість випадків (рядків у сегменті даних або документів), в нашому випадку 700, кількість атрибутів та їх назву, інформацію про метод аналізу, що використовується.

б) *опис класифікованої моделі*, наприклад докладну формулу розрахунку, побудованої за допомогою алгоритму Linear Regression моделі, прогнозованого результату для обраного параметру cnt, яку наведено нижче.

=== Classifier model (full training set) ===

Linear Regression Model

cnt =

-5.2622 * instant +
324.8122 * season +
4041.6293 * yr +
188.0184 * mnth +
-587.692 * holiday +
64.0853 * weekday +
-590.4561 * weathersit +
2343.6995 * temp +
3075.4806 * atemp +
-955.27 * hum +
-2587.018 * windspeed +
1526.7364

в) *прогноз для набору даних, що аналізується*.

Фрагмент прогнозу наведено нижче.

=== Predictions on test data ===

inst#, actual, predicted, error

1	683	846.765	163.765
2	3429	3640.264	211.264
3	4400	3612.669	-787.331
4	7290	5998.577	-1291.423
5	5202	5165.369	-36.631
6	3767	3603.104	-163.896
7	2423	2361.258	-61.742
8	4068	3974.644	-93.356

Де *actual* – актуальний клас набору даних, *predicted* – прогнозований моделлю клас, *error* – помилки моделі (різниця між актуальним та прогнозованим значенням).

Так, для першого випадку актуальна кількість орендованих велосипедів становить 683, в той час як модель прогнозує 846, помилка моделі дорівнює 163. Прогноз моделі та її помилка розраховуються по формулі, що складена побудованою моделлю за допомогою алгоритму Linear Regression, яка наведена вище в описі класифікованої моделі.

г) *аналіз інформації про побудовану модель*, по якій ми зможемо судити про її якість, коректність та ефективність.

=== Cross-validation ===

=== Summary ===

Correlation coefficient	0.8911
Mean absolute error	656.9508
Root mean squared error	880.2981
Relative absolute error	41.3976 %
Root relative squared error	45.322 %
Total Number of Instances	700

Цей розділ містить інформацію про якість та ефективність алгоритмів регресії, які визначені за допомогою вбудованих у Weka наступних методів оцінки ефективності:

Correlation coefficient - коефіцієнт кореляції (зв'язку між змінними) - показує наскільки сильний зв'язок між вхідними і вихідними значеннями, може приймати значення між -1 і +1, де 1 - сильний зв'язок, а 0 - немає ніякого зв'язку. Якщо коефіцієнт кореляції негативний, це означає наявність протилежного зв'язку: чим вище значення однієї змінної, тим нижче значення іншої. Існує декілька способів представлення та інтерпретації отриманого результату.

Візуально градації коефіцієнта кореляції можна представити наступним чином, як зображено на рис. за допомогою діаграми розсіювання (рис. 3.8).

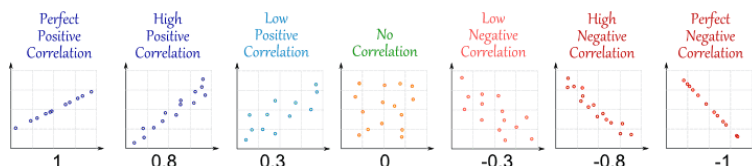


Рисунок 3.8 - Візуалізація градації коефіцієнта кореляції

Для словесного опису величини коефіцієнта кореляції використовуються наступні градації, які представлено в табл.3.1.

Таблиця 3.1 – Інтерпретація коефіцієнту кореляції

Значення коефіцієнта кореляції	Інтерпретація отриманого значення
до 0,2	Дуже слабка кореляція
до 0,5	Слабка кореляція
до 0,7	Середня кореляція
до 0,9	Висока кореляція
вище 0,9	Дуже висока кореляція

В нашому випадку коефіцієнт кореляції дорівнює 0,8911, що вказує на високу кореляцію.

Mean absolute error - середня абсолютна помилка - залежить від значення вихідних даних, та вимірюється в тих же одиницях, що і вихідні дані. MAE підходить для порівняння ефективності моделей тільки з однаковими вихідними даними.

Нехай помилка це різниця:

$$e(i) = a_i - p_i$$

де a_i - фактичне значення (actual target),

p_i – прогнозоване значення (predicted target)

Тоді формулу для середньої абсолютної помилки для n спостережень можна записати в наступному вигляді:

$$MAE = \frac{1}{n} \sum_{i=1}^n |p_i - a_i|$$

Root mean squared error - квадратний корінь з середньоквадратичної помилки - використовується для вимірювання частоти помилок моделі. Цю оцінку можна використовувати для порівняння моделей з однаковими даними, тобто, коли помилки моделей вимірюються в одних і тих же одиницях.

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (p_i - a_i)^2}{n}}$$

Relative absolute error - відносна абсолютна похибка - можна порівняти між моделями, чиї помилки вимірюються в різних одиницях.

$$RAE = \frac{\sum_{i=1}^n |p_i - a_i|}{\sum_{i=1}^n |a - a_i|}$$

Root relative squared error - квадратний корінь з відносного квадрату помилки - прийняті значення від 0 до нескінченності. Ідеальний результат 0.

RAE і RRSE показують нам наскільки актуальне значення параметра відрізняється від його середнього значення, що показує нам наскільки актуальне значення відрізняється від самого себе в цій моделі. Функція візуалізації отриманої моделі для цього алгоритму недоступна.

Проведемо аналіз помилок для свого завдання. Ретельно вивчивши результати аналізу в розділі Predictions on test split, фрагмент якого представлений на рис. 8, де actual – актуальний клас набору даних, predicted – прогнозований моделлю клас, error – помилка моделі (різниця між актуальним та прогнозованим значенням), ми бачимо, що для випадків з невеликою кількістю орендованих велосипедів алгоритм прогнозує набагато більшу кількість, тому в цих випадках найбільша різниця між актуальним та прогнозованим значенням, тобто найбільша помилка (рис.3.9).

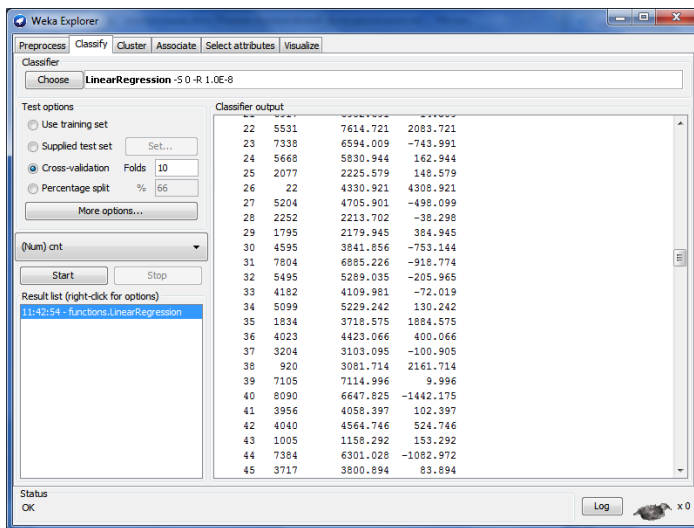


Рисунок 3.9 – Фрагмент результатів аналізу з позначеними помилками моделі

Виконаємо візуалізацію помилково класифікованих результатів.

Для цього обираємо у меню опцію Visualize classifier errors при натисканні у вікні Result-list правою кнопкою миші (рис. 3.10).

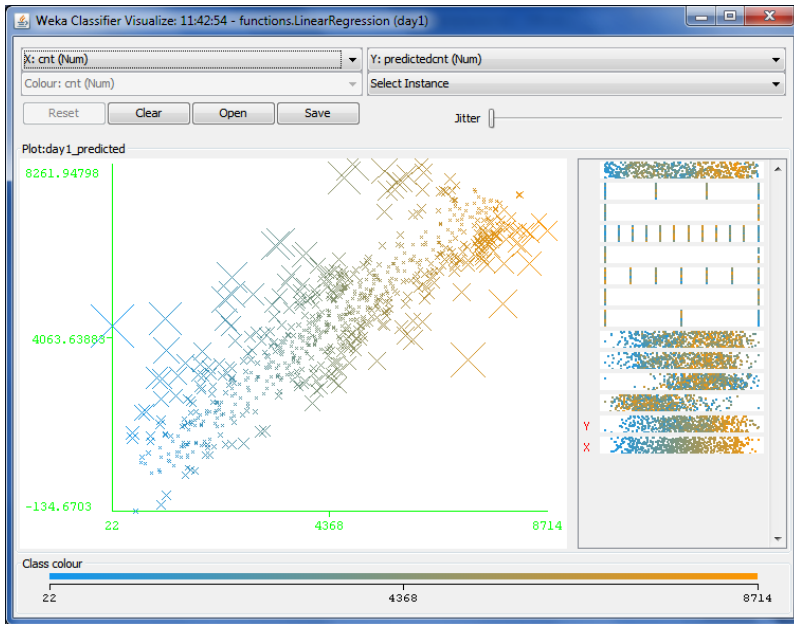


Рисунок 3.10 – Діаграма розсіювання, що відображає значення актуальної та прогнозованої кількості орендованих велосипедів для досліджуваного набору даних

На рис. 3.10 надано графічне представлення зв'язку між актуальним значенням даних та прогнозованим.

Якщо ми порівняємо отриманий результат з візуальною градацією коефіцієнта кореляції, то можемо зробити висновок про високу кореляцію. Перехід від одного кольору до іншого вказує на збільшення або зменшення значення параметрів.

В цьому вікні ми можемо відслідковувати залежність кількості орендованих велосипедів від будь-якого атрибуту, наприклад від швидкості вітру, як представлено на рис. 3.11.

По цій діаграмі ми можемо зробити висновок про слабку кореляцію, тобто про дуже низьку або відсутню залежність оренди велосипедів від швидкості вітру.

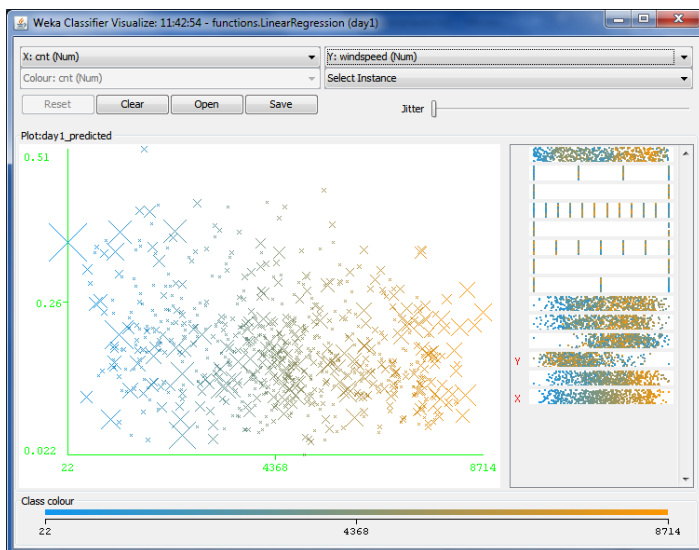


Рисунок 3.11 – Діаграма розсіювання, що відображає значення актуальної кількості орендованих велосипедів та швидкості вітру для досліджуваного набору даних

Натиснувши на будь який випадок правою кнопкою миші, можна отримати детальну інформацію про конкретний випадок.

Проводимо обробку даних за допомогою всіх обраних алгоритмів.

Отримані показники, що характеризують точність, якість та ефективність досліджуваних моделей, представлено у табл. 3.2.

Таблиця 3.2 – Порівняльний аналіз показників точності, якості та ефективності методів регресії.

Алгоритм	Показники точності, якості та ефективності моделі		
	Correlation coefficient	Mean absolute error	Relative absolute error, %
Linear Regression	0.8911	656.9508	41.3976
SMOreg	0.8906	652.2465	41.1012
Multiplayer Perceptron	0.8756	646.4922	40.7385
Isotonic Regression	0.8295	804.9378	50.723

Ні підставі даних табл. 3.2 можемо зробити висновок, що найкращі результати, тобто найвищий коефіцієнт кореляції, найменшу середню абсолютну помилку та найменшу відносну абсолютну похибку, показав алгоритм Linear Regression. Зберігаємо цю модель для подальшої обробки даних. Для цього натискаємо правою кнопкою миші на назві алгоритму у вікні Result list та обираємо опцію Save model.

Проведемо прогнозування кількості орендованих велосипедів для наступного місяця за допомогою збереженої моделі.

Створюємо тестовий набір, який містить існуючий набір даних до яких додаємо відомості про наступний місяць, які відповідають вимогам до атрибутів набору.

Вибираємо опцію Supplied test set та завантажуюмо тестовий файл з відомостями про наступний місяць. У вікні Result list навести курсор на назву моделі, що показала найкращі результати Linear Regression, натискаємо правою кнопкою миші, і вибираємо опцію Re-evaluate model on current test set. Прогнози можна робити на тестовому наборі, при умові, що набір містить допустимі значення класу. Результат буде містити як відомі так і прогнозовані дані.

Також можна використати попередньо збережену модель. Для цього клацнемо правою кнопкою миші у вікні Result list, вибираємо опцію Load model та завантажуюмо збережену раніше модель. Клацаємо правою кнопкою миші на назві моделі і вибираємо опцію Re-evaluate model on current test set. У вікні справа з'являється результат прогнозування.

Результат прогнозування кількості орендованих велосипедів на наступний місяць представлено нижче.

inst#,	actual,	predicted,	error
701	?	5387.442	
702	?	5059.77	
703	?	6413.299	
704	?	6363.348	
705	?	6057.908	
706	?	5530.234	
707	?	5217.25	
708	?	5561.49	
709	?	5038.695	
710	?	5255.971	
711	?	4862.589	
712	?	5064.441	
713	?	5721.323	
714	?	5707.883	

715	?	6059.42
716	?	5100.677
717	?	5269.981
718	?	5896.999
719	?	5700.03
720	?	5234.145
721	?	3688.267
722	?	4018.031
723	?	4291.894
724	?	3605.398
725	?	3143.823
726	?	2392.458
727	?	3162.436
728	?	3870.357
729	?	3815.083
730	?	3660.616
731	?	3426.839

Модель, побудована за допомогою алгоритму Linear Regression зробила своє прогнозування, передбачивши клас для всіх невідомих результатів. Результат прогнозування ми бачимо в Predictions on test split, в колонці Predicted.

Таким чином ми отримуємо підтримку прийняття рішення при плануванні оптимальної кількості необхідних велосипедів на конкретному пункті оренди велосипедів.

Перелік рекомендованої літератури

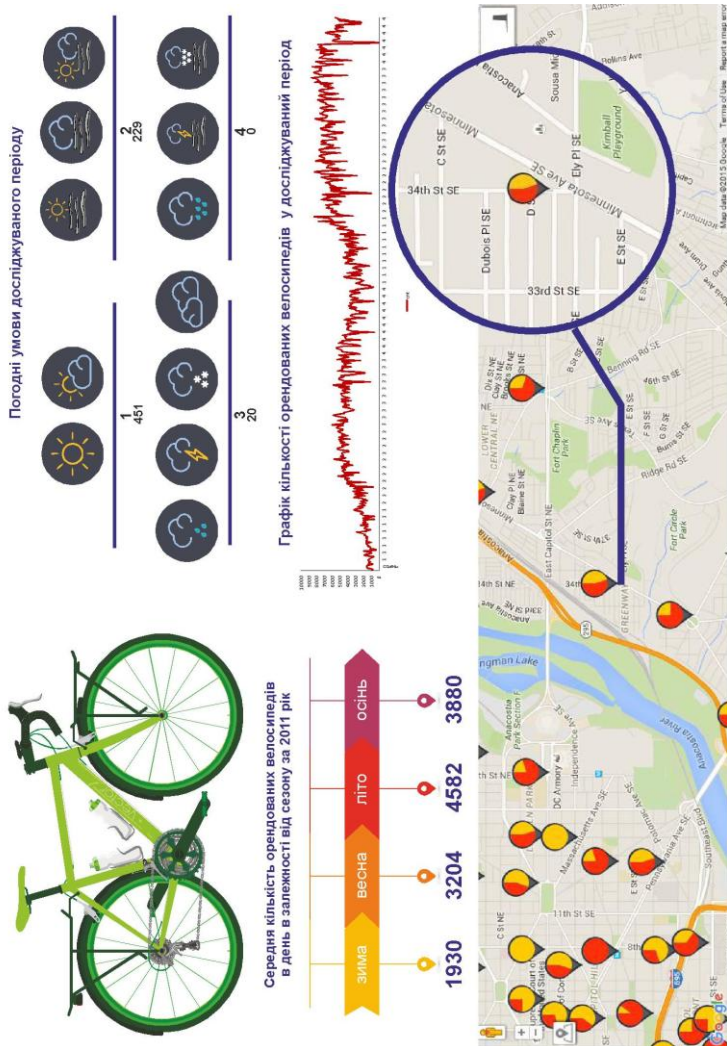
1. Інтелектуальний аналіз даних [Електронний ресурс] : навчальний посібник / А. О. Олійник, С. О. Субботін, О. О. Олійник. – Запоріжжя : ЗНТУ, 2012. – Режим доступу: <http://eir.zntu.edu.ua/bitstream/123456789/2155/1/> – Назва з екрана.
2. Інтелектуальний аналіз даних. Комп'ютерний практикум [Електронний ресурс] : навч. посіб. для студ. спеціальності 122 «Комп'ютерні науки та інформаційні технології» / О. О. Сергєєв-Горчинський, Г. В. Іщенко 2012. – Режим доступу: https://ela.kpi.ua/bitstream/123456789/24971/1/Komp_prakt.pdf
3. Конспект лекцій з дисципліни «Вступ до інтелектуального аналізу даних» (для студентів денної та заочної форми навчання спеціальності 122 "Комп'ютерні науки") Уклад: О.К. Лифар. – Северодонецьк: Вид-во СНУ ім. В. Даля, 2019. – 213с. A periodic table of visualization methods. – URL: http://www.visual-literacy.org/periodic_table/periodic_table.html
4. Марченко О.О., Россада Т.В. Актуальні проблеми Data Mining: Навчальний посібник для студентів факультету комп'ютерних наук та кібернетики. — Київ. — 2017. — 150 с.
5. Фісун М.Т., Кравець І.О., Казмірчук П.П., Ніколенко С.Г. Інтелектуальний аналіз даних: практикум. Навчальний посібник для освітніх напрямів підготовки «Комп'ютерні науки» та «Системний аналіз» — Львів «Новий Світ -2000», 2016 – 161 с. Murphy K. A Brief Introduction to Graphical Models and Bayesian Networks [Електронний ресурс] / K. Murphy – Режим доступу [www. URL: http://www.cs.ubc.ca/~murphyk/Bayes/bayes.html](http://www.cs.ubc.ca/~murphyk/Bayes/bayes.html)
6. Черняк О.І. Інтелектуальний аналіз даних: Підручник / О.І. Черняк, П.В. Захарченко ; Київ. нац. ун-т ім. Т. Шевченка. — К. : Знання, 2014. — 599 с.
7. Few S. Show me the numbers / S. Few ; Analytics Press, 2012. – 371 p.
8. Few S. Information Dashboard Design: Displaying Data for At-a-Glance Monitoring / S. Few ; Analytics Press, 2013. – 260 p.
9. Few S. Data visualization for enlightening communication [Електронний ресурс] / S. Few – Режим доступу [www. URL: https://nces.ed.gov/programs/slds/pdf/08_F_06.pdf](https://nces.ed.gov/programs/slds/pdf/08_F_06.pdf) - 30.09.2015 p.
10. Saed Sayad An Introduction to Data Mining [Електронний ресурс] – Режим доступу [www. URL: http://www.saedsayad.com/data_mining_map.htm](http://www.saedsayad.com/data_mining_map.htm) - 19.10.2015 p.

Додаток А

Питання теоретичного завдання

1. Алгоритм Байєса. Застосування даного алгоритму на вирішення завдань інтелектуального аналізу даних.
2. Алгоритми регресії. Застосування даного алгоритму на вирішення завдань інтелектуального аналізу даних.
3. Алгоритми кластеризації. Застосування даного алгоритму на вирішення завдань інтелектуального аналізу даних.
4. Генетичні алгоритми. Застосування даного алгоритму на вирішення завдань інтелектуального аналізу даних.
5. Аналіз часових рядів. Застосування даного алгоритму на вирішення завдань інтелектуального аналізу даних.
6. Алгоритм пошуку асоціативних правил. Застосування даного алгоритму на вирішення завдань інтелектуального аналізу даних.
7. Нейронні мережі. Застосування на вирішення завдань інтелектуального аналізу даних.
8. Нейронні мережі. Застосування на вирішення завдань розпізнавання образів.

Візуалізація вхідних даних до прикладу виконання контрольної роботи п.4



Навчальне видання

МЕТОДИЧНІ ВКАЗІВКИ

до виконання контрольної роботи з дисципліни

"Вступ до інтелектуального аналізу даних"

для здобувачів вищої освіти за спеціальністю

"Комп'ютерні науки" заочної форми навчання
(електронне видання)

Укладач:

Шумова Лариса Олександрівна;

Редактор

Л.О. Шумова

Техн. редактор

Л.О. Шумова

Оригінал – макет

Л.О. Шумова

Підписано до друку _____

Формат 16 60 × 84 1 . Папір типограф. Гарнітура Times.

Друк офсетний. Ум. друк. арк. _____. Обл.-вид.арк. _____.

Тираж ____ прим. Вид. № _____. Замовл. № _____. Ціна договірна.

Видавництво Східноукраїнського національного університету імені
Володимира Даля

Адреса університету: вул. Іоанна Павла II, 17, м. Київ, 01042, Україна

e-mail: vidavnictvosnu.ua@gmail.com