МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
СХІДНОУКРАЇНСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВОЛОДИМИРА ДАЛЯ
ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ ТА ЕЛЕКТРОНІКИ
КАФЕДРА КОМП'ЮТЕРНИХ НАУК ТА ІНЖЕНЕРІЇ

УДК 004.32.2

До захисту допускається
В.о. завідувача кафедри
комп'ютерних наук та інженерії
д.т.н., проф. Рязанцев О. І.
_____
«____»_____ 2021 р.

МАГІСТЕРСЬКА РОБОТА

НА ТЕМУ:

«Аналіз ефективності методів машинного навчання для систем розпізнавання письмових символів»

Освітньо-кваліфікаційний рівень «Магістр»
Спеціальність 123 «Комп'ютерна інженерія»

Науковий керівник роботи:	_____	Скарга-Бандурова І.С.
	(підпис)	(ініціали, прізвище)
Консультант з охорони праці:	_____	Критська Я. О.
	(підпис)	(ініціали, прізвище)
Студент:	_____	Доброжан З.Т.
	(підпис)	(ініціали, прізвище)
Група:		КІ-19 дм

Сєвєродонецьк – 2021

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
СХІДНОУКРАЇНСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВОЛОДИМИРА ДАЛЯ

Факультет інформаційних технологій та електроніки

Кафедра комп'ютерних наук та інженерії

Освітньо-кваліфікаційний рівень магістр

Спеціальність 123 «Комп'ютерна інженерія»

«ЗАТВЕРДЖУЮ»
Т.в.о. завідувача кафедри
комп'ютерних наук та інженерії
к.т.н., доц. Кардашук В. С.

"____" _____ 2020 року

ЗАВДАННЯ
НА МАГІСТЕРСЬКУ РОБОТУ СТУДЕНТУ

Доброжану Захару Тарасовичу

(прізвище, ім'я, по-батькові)

1. **Тема проекту (роботи):** «Аналіз ефективності методів машинного навчання для систем розпізнавання письмових символів» затверджена наказом по університету № 140/15.15 від «05» жовтня 2020 р.

2. Строк здачі студентом закінченого проекту (роботи): 11.01.2021 р.

3. Вихідні дані проекту (роботи): матеріали переддипломної практики

4. Зміст розрахунково-пояснювальної записки (перелік питань, які необхідно розробити):

1. Аналіз методів машинного навчання для розпізнавання рукописних текстів;
2. Огляд та застосування технологій зменшення розмірності даних;
3. Використання методів машинного навчання для розпізнавання рукописних символів;
4. Охорона праці та безпека в надзвичайних ситуаціях.

5. Перелік графічного матеріалу (з точною назвою обов'язкових креслень):

електронні плакати

6. Консультанти роботи, з вказівкою розділів, що до них відносяться

Розділ	Консультант	Підпис, дата	
		Завдання видав	Завдання прийняв
Охорона праці та безпека в надзвичайних ситуаціях	Критська Я. О.		

7. Дата видачі завдання: 11.10.2020 р.

Керівник Скарга-Бандурова І.С.

(підпис)

Завдання до виконання прийняв Доброжан З.Т.

(підпис)

КАЛЕНДАРНИЙ ПЛАН

№ п/п	Найменування етапів дипломного проекту (роботи)	Строк виконання етапів проекту (роботи)	Примітки
1.	Отримання завдання, збір матеріалів.	11.10.20- 24.10.20	
2.	Дослідження та аналіз сучасних методів машинного навчання.	25.10.20–28.10.20	
3.	Аналіз існуючих технологій зменшення розмірності.	29.10.20– 28.11.20	
4.	Програмна реалізація та дослідження ефективності методів машинного навчання для розпізнавання рукописних символів.	28.11.20–31.12.20	
5.	Аналіз результатів досліджень.	03.01.21 – 04.01.21	
6.	Оформлення пояснювальної записки.	05.01.21 – 08.01.21	
7.	Підготовка та подання магістерської роботи до захисту.	09.01.21 – 10.01.21	

Магістр _____ Доброжан З.Т.
(підпис) (прізвище та ініціали)

Науковий керівник _____ Скарга-Бандурова І.С.
(підпис) (прізвище та ініціали)

АНОТАЦІЯ

Доброжан З.Т. Аналіз ефективності методів машинного навчання для систем розпізнавання письмових символів.

В роботі представлено вирішення задачі підвищення ефективності розпізнавання письмових символів з використанням методів машинного навчання. Метою роботи було визначення який із методів машинного навчання найбільш ефективний для систем розпізнавання письмових символів та має найменшу вірогідність помилки. Проведено аналіз задачі, огляд сучасних методів і технологій машинного навчання, окрема увага приділяється технологіям зменшення розмірності. Використаними алгоритмами є метод k-найближчих сусідів та випадковий ліс, в якості методу оцінки використано точність та матриці плутанини. Представлено програмну реалізацію та виконано дослідження ефективності методів машинного навчання для розпізнавання рукописних символів з бази даних MNIST.

Рис.: 24 Табл. 8 Бібліогр.: 22.

Ключові слова: машинне навчання, MNIST, метод головних компонент, метод найближчих сусідів, Random Forest

АННОТАЦИЯ

Доброжан З.Т. Анализ эффективности методов машинного обучения для систем распознавания письменных символов.

В работе представлены решения задачи повышения эффективности распознавания письменных символов с использованием методов машинного обучения. Целью работы было определение какой из методов машинного обучения наиболее эффективен для систем распознавания письменных символов и имеет наименьшую вероятность ошибки. Проведен анализ задачи, сделан обзор современных методов и технологий машинного обучения, особое внимание уделяется технологиям

уменьшения размерности. Использованными методами являются метод k-ближайших соседей и случайный лес, в качестве метода оценки использовано точность и матрица путаницы. Представлена программная реализация и выполнено исследование эффективности методов машинного обучения для распознавания рукописных символов из базы данных MNIST.

Рис .: 24 Табл. 8 Библиогр .: 22.

Ключевые слова: машинное обучение, MNIST, метод главных компонент, метод ближайших соседей, Random Forest

ABSTRACT

Zakhar Dobrojan Analysis of the effectiveness of machine learning methods for written character recognition.

The paper presents a solution to the problem of improving the efficiency of recognition of written characters using machine learning methods. The aim of the study was to determine which of the machine learning methods is most effective for written character recognition systems and has the highest accuracy. The analysis of the problem, the review of modern methods and technologies of machine learning are carried out, the special attention is paid to technologies of dimensionality reduction. The algorithms used are the method of k-nearest neighbors and the random forest, the accuracy and confusion matrix are used as a method of estimation. The software implementation is presented and the research of the efficiency of machine learning methods for the recognition of handwritten symbols from the MNIST database is performed.

Figures: 24 Tables: 8 Bibliography: 22.

Keywords: machine learning, MNIST, principal component analysis, nearest neighbors method, Random Forest

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧОК І СКОРОЧЕНЬ

ГК головні компоненти

ПК персональний комп’ютер

kNN k-near neibours (метод найближчих сусідів)

MNIST Modified National Institute of Standards and Technology database

NIST National Institute of Standards and Technology (Національний інститут стандартів і технологій США)

OOB Out Of Bag (помилка поза мішком)

PCA Principal Component Analysis (метод головних компонент)

PC Principal Component (головна компонента)

RF Random Forest (випадковий ліс)

ВСТУП

Поступово, системи, які слідкують за нашою поведінкою, впроваджуються в наше життя. Найбільш помітним є системи, які аналізують наші музичні смаки, вивчають наші вподобання в кіно та одязі, наші інтереси та бажання, згідно чого і формують попит на рекламний продукт згідно з нашими вподобаннями. Галузь, яка називається машинним навчанням, є підгалуззю штучного інтелекту в інформатиці та вивчає методи навчання штучних обчислювальних систем на підставі даних без програмування їх дії або поведінки. Щоденно класичні статистичні алгоритми розв'язують питання, пов'язані з прийняттям рішень на основі даних.

Однією з основних задач машинного навчання є класифікація. Вона передбачає собою визначення категорій та розподілу об'єктів згідно із заданими ознаками. Система сортує об'єкти та надає можливість працювати з об'єктами, які або підпадають під певну ознаку, або ні.

Класифікація - це підклас навчальних проблем, що контролюються; контрольовані навчальні проблеми включають дані навчання та відповідні їм цільові вихідні значення, які алгоритму пропонується передбачити на основі даних, які вони отримують, використовуючи певну функцію. Коли цільовим вихідним значенням є категоріальна змінна, перед нами постає проблема класифікації. Мета цих завдань - навчити алгоритм визначати справжній клас об'єкта, наприклад, чи є на знімку собака чи кішка.

Кожен набір проблем класифікації містить точки даних, які слід класифікувати, інформацію про них та їх справжні позначки. Потім цей набір даних слід розділити на два окремі набори даних, навчальний та тестовий. Набір навчальних даних представлений алгоритму разом з вектором його справжніх міток, щоб алгоритм міг «дізнатися», які

особливості точок даних відповідають якій мітці. Потім набір даних тестування використовується для оцінки ефективності алгоритму, коли прогнозовані мітки даних тестування порівнюються з їх справжніми мітками.

Об'єкт дослідження: системи автоматичного розпізнавання рукописних письмових символів

Предмет дослідження: технології розпізнавання письмових символів набору MNIST

Мета і завдання дослідження: Метою дослідження є підвищення точності оцінки розпізнавання письмових символів за рахунок аналізу методів машинного навчання придатних до використання в системах автоматичного розпізнавання.

Для досягнення мети дослідження необхідно вирішити такі **завдання**:

- Аналіз особливостей набору даних MNIST.
- Аналіз моделей машинного навчання для розпізнавання рукописних символів.
- Програмна реалізація методу змешення розмірності та оцінка ефективності виконаного перетворення.
- Програмна реалізація, тестування та аналіз ефективності методів машинного навчання. Використаними алгоритмами є k-найближчі сусіди та випадковий ліс (Random Forest)
- Розробка заходів з охорони праці.

Методи дослідження. Проведені в роботі дослідження основані на технологіях інтелектуального аналізу даних, зокрема методах машинного навчання. Перевірка результатів дослідження ґрунтувалась на методах статистичного аналізу, методах експерименту та порівняння, які використовувались при розробленні практичної частини магістерської роботи.

Особистий внесок здобувача полягає у аналізі літератури, виборі алгоритмів, проведенні експериментів, розробці програмного забезпечення для системи розпізнавання рукописних символів, що дозволяють вирішити поставлені в роботі задачі. Усі основні результати отримані автором особисто.

Апробація матеріалів магістерської роботи. Основні положення, ідеї, та висновки магістерської роботи доповідалися та обговорювалися на IV регіональному форумі IT-Ідея (Сєвєродонецьк, Україна, 7 грудня 2019).

Практичне значення отриманих результатів.

– запропоновано технологію обробки даних з використанням різних методів машинного навчання;

– програмна реалізація структурних елементів системи розпізнавання рукописних символів пройшла тестування та готова до подальшого використання;

– матеріали досліджень плануються до використання для розробки навчального курсу «Машинне навчання».

Публікації. За темою магістерської роботи з викладенням її результатів опубліковано тези доповіді на конференції.

Структура та обсяг магістерської роботи. Кваліфікаційна магістерська робота складається із вступу, чотирьох розділів, висновків, переліку посилань.

Загальний обсяг складає 78 сторінок, з яких основний текст на 61 сторінці, список використаних джерел із 22 найменувань та 2 додатків на 18 сторінках. Робота містить 8 таблиць, 24 рисунки.

1 АНАЛІЗ МЕТОДІВ МАШИННОГО НАВЧАННЯ ДЛЯ РОЗПІЗНАВАННЯ РУКОПИСНИХ ТЕКСТІВ

1.1 Опис набору даних MNIST

Набір даних MNIST – база даних, яка містить зразки рукописних варіантів написання чисел від 0 до 9. MNIST є стандартом, який був запропонований Національним інститутом стандартів і технологій США (NIST), для калібрування і зіставлення методів розпізнавання зображень за допомогою машинного навчання.

Набір даних представляє собою 60000 зображень для навчання і 10000 зображень для тестування. Кожне зображення є прикладом рукописного фрагменту (рис.1.1), який отримано шляхом обробки чорно-білих зразків символів NIST розміром 20X20 пікселів. Оскільки NIST, в свою чергу, використовували набір зразків з Бюро перепису населення США і рукописних робіт студентів американських університетів, база отримала назву Multi-NIST або MNIST.

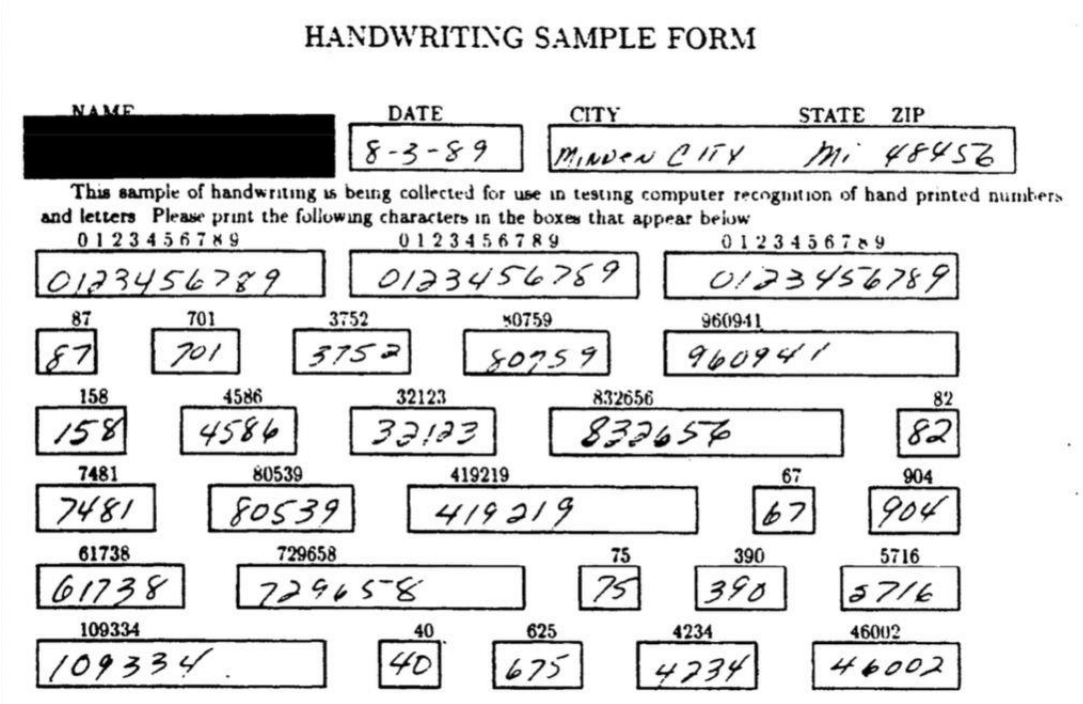


Рисунок 1.1 – Оригінальні символи з набору даних MNIST

Набір є мультикласовим, оскільки містить 10 класів (0-9), з лише однією вихідною міткою для кожного класу (рис.1.2).

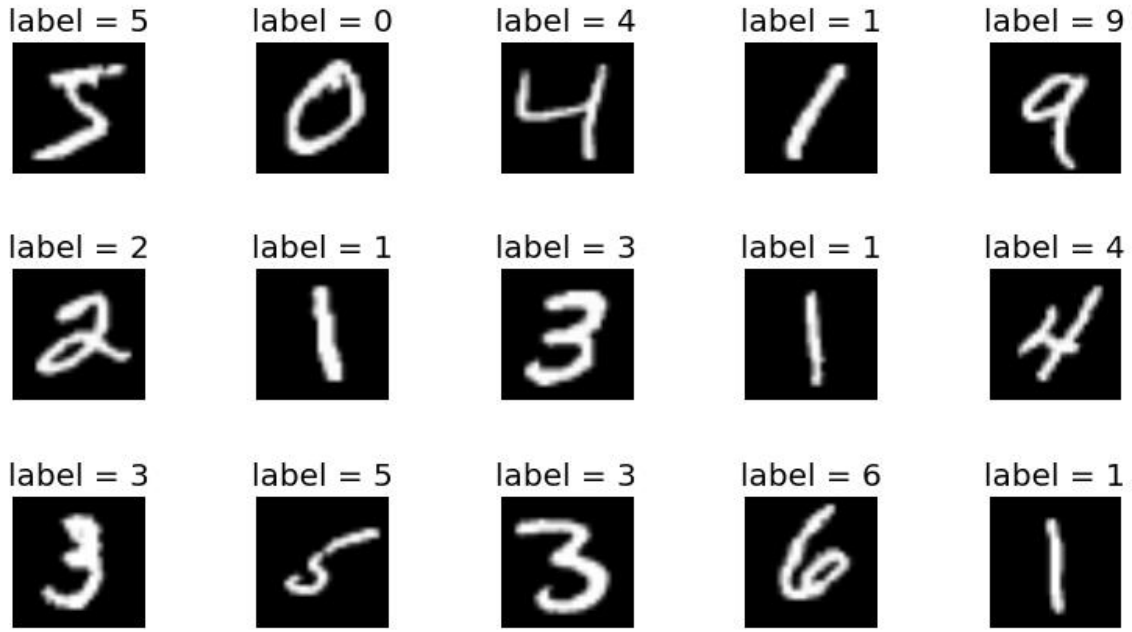


Рисунок 1.2 – Приклади цифр в наборі MNIST

Кожна цифра набору представлена цілочисельною матрицею розміром 28×28, значення кожної комірки матриці можуть приймати значення від 0 до 255 (рис.1.3).

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]	[,15]	[,16]	[,17]	[,18]	[,19]	[,20]	[,21]	[,22]	[,23]	[,24]	[,25]	[,26]	[,27]	[,28]
[1,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[2,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[3,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[4,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[5,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[6,]	0	0	0	0	0	0	0	0	0	0	0	0	3	18	18	18	126	136	175	26	166	255	247	127	0	0	0	0
[7,]	0	0	0	0	0	0	0	0	30	36	94	154	170	253	253	253	253	253	225	172	253	242	195	64	0	0	0	0
[8,]	0	0	0	0	0	0	0	49	238	253	253	253	253	253	253	253	253	251	93	82	82	56	39	0	0	0	0	0
[9,]	0	0	0	0	0	0	0	18	219	253	253	253	253	253	198	182	247	241	0	0	0	0	0	0	0	0	0	0
[10,]	0	0	0	0	0	0	0	0	80	156	107	253	253	205	11	0	43	154	0	0	0	0	0	0	0	0	0	0
[11,]	0	0	0	0	0	0	0	0	0	14	1	154	253	90	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[12,]	0	0	0	0	0	0	0	0	0	0	0	139	253	190	2	0	0	0	0	0	0	0	0	0	0	0	0	0
[13,]	0	0	0	0	0	0	0	0	0	0	0	11	190	253	70	0	0	0	0	0	0	0	0	0	0	0	0	0
[14,]	0	0	0	0	0	0	0	0	0	0	0	0	35	241	225	160	108	1	0	0	0	0	0	0	0	0	0	0
[15,]	0	0	0	0	0	0	0	0	0	0	0	0	0	81	240	253	253	119	25	0	0	0	0	0	0	0	0	0
[16,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	45	186	253	253	150	27	0	0	0	0	0	0	0	0
[17,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	93	252	253	187	0	0	0	0	0	0	0	0
[18,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	249	253	249	64	0	0	0	0	0	0	0
[19,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	46	130	183	253	253	207	2	0	0	0	0	0	0	0
[20,]	0	0	0	0	0	0	0	0	0	0	0	0	39	148	229	253	253	253	250	182	0	0	0	0	0	0	0	0
[21,]	0	0	0	0	0	0	0	0	0	0	24	114	221	253	253	253	253	201	78	0	0	0	0	0	0	0	0	0
[22,]	0	0	0	0	0	0	0	0	23	66	213	253	253	253	253	198	81	2	0	0	0	0	0	0	0	0	0	0
[23,]	0	0	0	0	0	0	18	171	219	253	253	253	253	195	80	9	0	0	0	0	0	0	0	0	0	0	0	0
[24,]	0	0	0	0	55	172	226	253	253	253	253	244	133	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[25,]	0	0	0	0	136	253	253	253	212	135	132	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[26,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[27,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[28,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Рисунок 1.3 – Приклад цифри «5»

Чисельні спроби досягти мінімальної помилки після навчання по базі даних MNIST. Рівень помилки було доведено до 0,23%, завдяки використанню найбільш потужних нейронних мереж. Самі творці бази даних передбачили кілька методів тестування, та відомо, що використання методу опорного вектору показував результати на рівні помилки 0,8%.

1.2 Аналіз моделей машинного навчання для розпізнавання рукописних символів

1.2.1 Метод найближчих сусідів

Метод найближчих сусідів – метричний класифікатор, який базується на оцінюванні подібності об’єктів. Оцінюваний об’єкт відноситься до того класу, якому належать сусідні до нього об’єкти в навчальній виборці.

Метод k-найближчих сусідів для підвищення своєї надійності, відносить об’єкт до того класу, до якого належать більшість його сусідів з навчальної вибірки (рис.1.4). Для виключення ситуацій, коли однакова кількість сусідів належать різним класам, беруть непарне число сусідів.

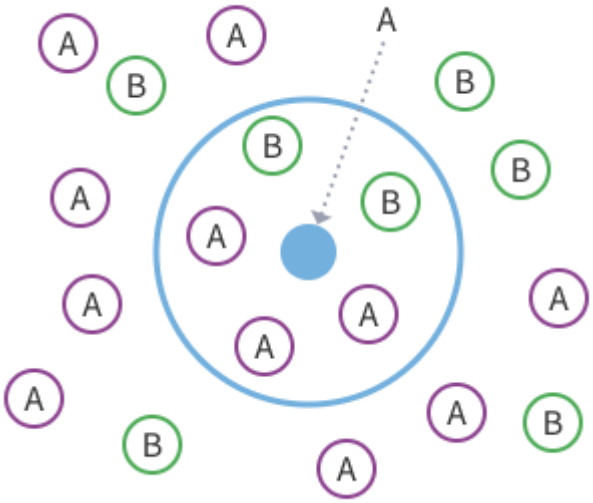


Рисунок 1.4 – Метод «найближчих сусідів»

В ситуації коли в завданнях беруть участь об’єкти з трьома класами і більше ситуації неоднозначності дають необ’єктивні результати. За для запобігання помилкам, об’єктам приписується так звана вага, регресивна до зросту рангу сусіда та оцінюваний об’єкт відноситься до того класу, який набирає більшу сумарну вагу серед k-найближчих сусідів.

Попри всі його переваги, такі як алгоритмічна простота, головним його недоліком є висока обчислювальна трудомісткість, яка збільшується в квадраті, згідно з ростом числа навчальних прикладів.

1.2.2 Кластерний аналіз

Кластерний аналіз полягає в розбитті заданої вибірки об’єктів на підмножини, групи об’єктів, які схожі між собою. Такі групи звуться кластерами. При істотній або абсолютній схожості об’єктів в кластері, вони повинні істотно відрізнятися від об’єктів в інших кластерах.

Методів кластерізації, попри те, що поняття «кластеру» не може бути точно визначено, існує велика кількість. Всі вони мають різні алгоритми але виконують схожу задачу – це об’єднання схожих об’єктів у групи. Типовими кластерними моделями є:

— «Моделі зв’язності», яка будується на основі відстані між вузлами;

— «Центроїдні моделі» представляють кожен кластер єдиним усередненим вектором;

— «Статистичні моделі», кластери в якому ґрунтуються на статистичних розподілах, таких як багатовимірний нормальний розподіл;

— «Моделі засновані на щільності» в якій кластери визначаються як зобов’язані області відповідної щільності у просторі даних;

— «Групові моделі», де алгоритми описують групування об’єктів;

— «Графові моделі» основою яких є кліки у графі, тобто частина ребер може бути відсутня та будуються за алгоритмом HCS;

— «Нейронні моделі», можна охарактеризувати як схожі на одну або подібні якійсь з наведених вище моделей, коли нейронні мережі реалізують метод «головних компонент».

Основна мета кластерного аналізу – знаходження груп схожих об’єктів у вибірці. Кластерний аналіз виконує завдання, такі як:

— дослідження корисних схем групування об’єктів;

— перевірка гіпотез або дослідження груп, виділених певним способом, присутні в наявних даних;

— розробка класифікації або типології;

— породження гіпотез на основі дослідження даних.

Універсальність методу привела до появи великої кількості методів і підходів, які змінюють розуміння кластерного аналізу та ускладнюють однозначне використання.

1.2.3 Дерево ухвалення рішень

Дерево ухвалення рішень містить такі елементи як «листок» та «гілки». Гілки поєднують між собою листки та мають атрибути, відповіді, які розрізняють випадки. В листках записано значення цільової функції. Щоб класифікувати новий випадок, треба поступово, відповідно до атрибутів, спуститися до листа і видати відповідне до нього значення.

Дерево може бути вивчено поділом вихідних наборів змінних, що засновано на тестуванні значень атрибутів. Рекурсія завершиться тоді, коли значення цільової змінної і підмножини в вузлі будуть однакові.

В інтелектуальному аналізі даних дерева рішень можуть бути використані як обчислювальні методи, щоб допомогти описати, класифікувати та узагальнити набір даних.

Дерево рішень складається з трьох типів вузлів: рішення, імовірнісні та замикаючі. Використовувані в Data Mining дерева рішень, бувають двох основних типів: регресивний аналіз дерева та аналіз дерева класифікації.

Аналіз дерева класифікації, коли прогнозований результат є класом, до якого належать дані (рис.1.5).

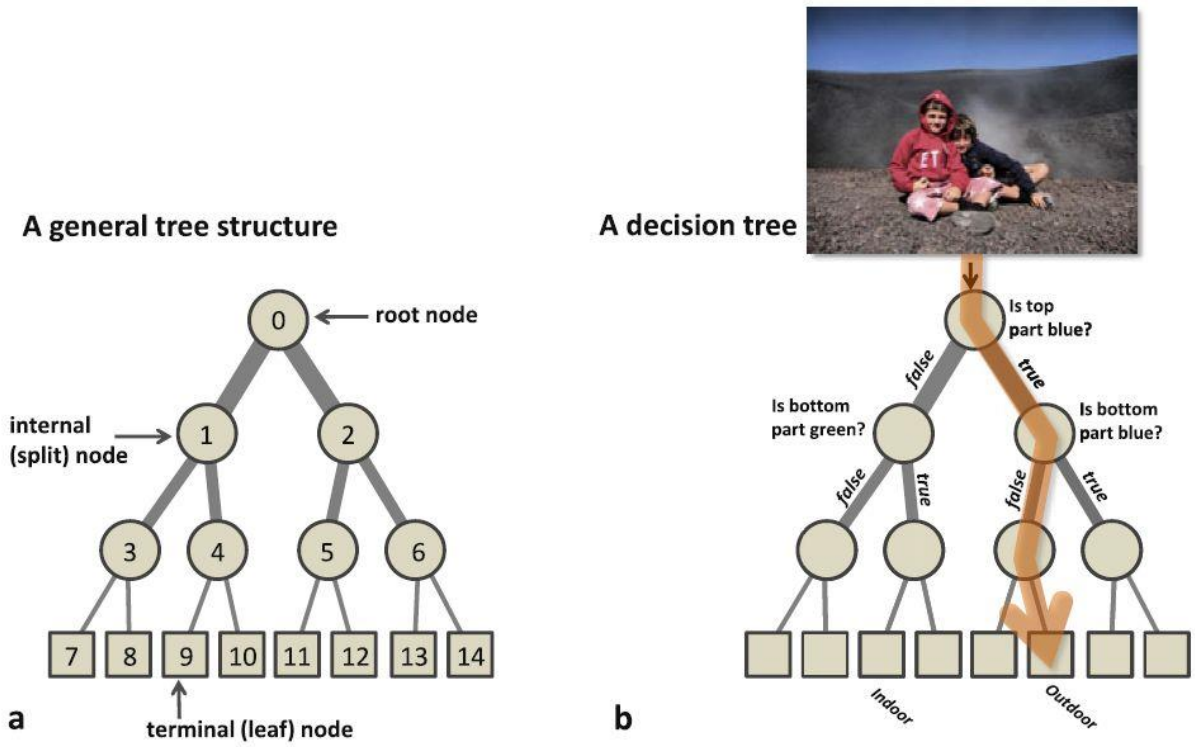


Рисунок 1.5 – Структура (a) та приклад генерації рішення (b) в дерева прийняття рішень

Регресивний аналіз дерева, коли прогнозований результат можна розглядати як дійсний висновок.

1.2.4 Штучні нейронні мережі

Один з видів машинного навчання, є системою взаємодіючих і з'єднаних між собою штучних нейронів, які отримують вхід, змінюють свій внутрішній стан, відповідно до цього входу і виробляють вихід. Вихід формується відповідно до входу та збудження, тобто зміни внутрішнього стану. На рис. 1.6 представлено модель персептрону – базової моделі штучних нейронних мереж.

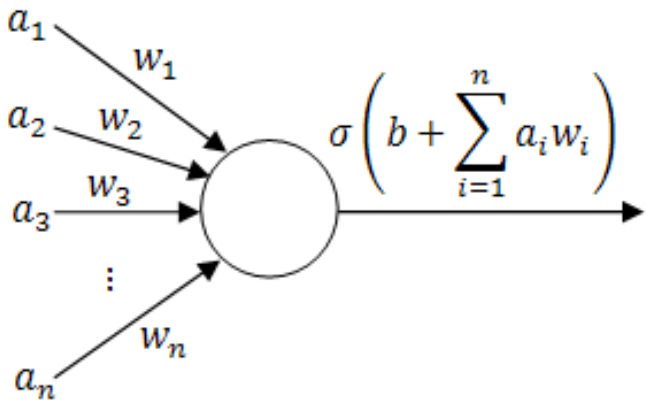


Рисунок 1.6 – Модель персептрону

Мережа представляє собою поєднання входів і виходів нейронів з утворенням орієнтованого зваженого графу. Функції, що визначають збудження, можуть змінюватися процесом, який керується правилом навчання.

Правило навчання – це алгоритм, який змінює параметри нейронної мережі, щоб вхід до мережі визначав придатний вихід. Цей процес полягає в зміні ваг та порогів змінних мережі.

1.3 Постановка наукового завдання та обґрунтування методики дослідження

З огляду на вищевикладене, актуальним науковим завданням є оцінити якість класифікаторів для розпізнавання рукописних символів.

Дослідницьке питання: наскільки ефективними є методи машинного навчання для розпізнавання рукописних символів?

Мета: Підвищення якості розпізнавання рукописних символів

Базова гіпотеза: в роботі передбачається що можливо отримати якісну модель, що дозволить відрізняти рукописні символи і використовувати її в системах автоматичного розпізнавання, що працюють зі смартфонів.

Можливий набір моделей та алгоритмів, що будуть використані в якості класифікаторів: k-NN, SVM, NN, кластерний аналіз, Random Tree, Random Forest

Методика дослідження містить наступні етапи:

Частина 1: Підготовка даних

- Завантажити дані
- Перевірити дані (наприклад, перевірити розміри набору даних, перевірити відсоток розподілу кожної цифри у тренувальному наборі даних, візуалізація цифр в даних тощо)
- Нормалізувати дані
- Розділити дані на набір даних для тренування та тестовий

Частина 2: Використання методу головних компонент

- Використати метод головних компонент для зниження обсягів даних
- Побудувати графіки для візуалізації даних
- Зробити фактичне зменшення розмірів (матрицю з 784 стовпців слід перетворити в матрицю з 10, 20, або 40 ... стовпців)

Частина 3: KNN

- Запустити метод KNN на наборах даних із зменшеними розмірами
- Побудувати графік точності для різних k
- Вибирати k, який забезпечує найкращу точність
- Порівняти передбачувані мітки з фактичними мітками

Частина 3: Обрати інші методи класифікації та провести аналогічні тести.

Частина 4: Порівняти якість класифікаторів та зробити висновки.

Частина 5: Запропонувати елементи інформаційної технології для використання обраних класифікаторів в додатках для смартфонів.

Написати короткий висновок про результати та порівняти з результатами, опублікованими для інших алгоритмів на домашній сторінці набору даних.

Висновки до розділу 1

В результаті аналізу існуючих методів машинного навчання, ми дізналися про їх типи та класифікації, а також визначилися з основними поняттями для подальшої розробки.

Згідно з отриманими результатами визначено набір даних, завдання та основні етапи розробки. Основними методами, що будуть використані в роботі визначено метод головних компонент – для зменшення розмірності даних, метод найближчих сусідів та дерева прийняття рішень для проведення експериментів з класифікації даних.

Основною кінцевою метою досліджуваного наукового напрямку, комплексною складовою якого є магістерська робота, є мінімізація помилки класифікації на тестовому наборі даних.

Література до розділу 1

1. Alpaydin, E. (2014) Introduction to machine learning. Third edn. Cambridge, Massachusetts: MIT Press (Adaptive computation and machine learning).

2. Bishop, C. M. (2006) Pattern recognition and machine learning, Springer-Verlag

3. LeCun, Y. (1998) The MNIST database of handwritten digits, LeCun Y. Available at: http://yann.lecun.com/exdb/mnist/

4. LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998) Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), pp.2278-2324.

5. Safavian, S.R. and Landgrebe, D. (1991) A survey of decision tree classifier methodology. IEEE transactions on systems, man, and cybernetics, 21(3), pp.660-674.

2 ТЕХНОЛОГІЯ ЗМЕНШЕННЯ РОЗМІРНОСТІ

Висока розмірність означає, що набір даних має велику кількість вимірів. Основною проблемою, пов'язаною з великою розмірністю в області машинного навчання, є перенавчання моделі, що зменшує здатність до узагальнення, окрім прикладів у навчальному наборі. Річард Беллман описав це явище в 1961 році як прокляття розмірності [1], де "Багато алгоритмів, які чудово працюють у низьких вимірах, стають нерозбірливими, коли вхідні дані є великими". З метою подолання цієї проблеми в цьому розділі наведено задачу зменшення розмірності та основні результати використання методу головних компонент для досліджуваного набору даних рукописних символів.

2.1 Задача зменшення розмірності

Зменшення розмірності – це перетворення даних із високомірного простору в низьковимірний простір, так що другий варіант зберігає деякі основні властивості вихідних даних. Такі перетворення часто використовують в тих галузях, де використовують велику кількість даних або спостережень, таких як розпізнавання мови, обробка сигналів, біоінформатика та нейроінформатика.

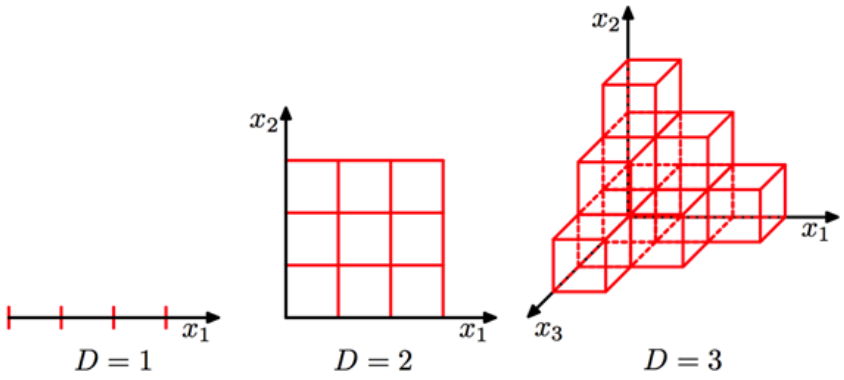


Рисунок 2.1 – Три види розмірностей (Джерело: [2])

Складністю роботи зі зменшенням розмірності є те, що необроблені дані часто є рідкісними і аналіз даних, як правило, є обчислювально нерозв’язним.

Підходи поділяються на вибір об’єктів та їх виділення. Такі задачі можуть бути використані для візуалізації даних, як проміжний крок для полегшення інших аналізів, кластерного аналізу або зменшення шуму.

Обирання ознак – це процес пошуку підмножини первісних змінних. В побудові цієї моделі використовують наступні стратегії:

– фільтрування, використовують для отримання інформації;

– обґрунтування, де необхідний пошук, який керується точністю;

– вкладення, ознаки обираються для видалення або додавання при створенні моделі, яка базується на помилках прогнозування;

– Конструювання ознак може бути лінійним та нелінійними та виділяють наступні методи:

– Розклад невід’ємних матриць розкладає невід’ємну матрицю на добуток двох невід’ємних матриць. За допомогою компонентної бази послідовний розклад невід’ємних матриць здатний зберігати потік при прямому відтворенні навколо зоряних структур в астрономії, як один із способів виявлення екзопланет.

– Автокодувальник використовується для навчання нелінійним функціям зменшення розмірності та кодування разом із оберненою функцією.

– Ядровий метод головних компонент можна використати нелінійним способом за допомогою ядрового трюку. Методика дозволяє побудувати нелінійні відображення, які максимізують дисперсію даних.

– Лінійний розділювальний аналіз – це узагальнення лінійного дискримінанта Фішера, який використовується щоб знайти лінійну комбінацію ознак, які характеризують або відокремлюють два або більше класів об’єктів або подій.

2.2 Метод головних компонент

Метод головних компонент (Principal component analysis або PCA) – це основна лінійна техніка зменшення розмірності, який здійснює лінійне відображення даних в менш вимірний простір таким чином, що максимізується дисперсія даних у маловимірному представленні. Будується матриця коваріації або кореляції даних, і обчислюються власні вектори цієї матриці.

Обчислення головних елементів може бути зведене до обчислення власних чисел і векторів коваріаційної матриці початкових даних або обчислення сингулярного розкладу матриці даних. Це один з основних методів зменшити розмірність даних, втративши найменшу кількість інформації. Застосовується в багатьох галузях, таких як: економетрика, обробка зображень, суспільні науки, біоінформатика та для стиснення даних.

Даний метод дає можливість за х-числом початкових ознак виділити у головних узагальнених ознак. Математична модель методу головних компонент базується на логічному припущенні, що значення множини взаємозалежних ознак породжують деякий загальний результат. Простір головних компонент ортогональний.

Задачі аналізу головних компонент мають чотири базові версії:

– апроксимувати дані лінійними многовидами меншої розмірності;

– пошук підпросторів меншої розмірності, в ортогональній проекції, на яке середньоквадратичне відхилення від середнього значення – максимальне;

– знайти підпростори меншої розмірності, в ортогональній проекції на які середньоквадратична відстань між точками максимальна;

– побудувати ортогональне перетворення координат для даної багатовимірної випадкової величини, внаслідок якого кореляції між окремими координатами перетворяться в нуль.

Матриця A перетворення даних до головних компонент складається з векторів головних компонент, розташованих у порядку спадання власних значень:

$A = \{a_1, \ldots, a_n\}^T$, де T означає транспортування, при чому $AA^T = 1$. Тобто, матриця A є ортогональною.

Велика частина варіації даних буде зосереджена в перших координатах, що дозволяє перейти до простору меншої розмірності.

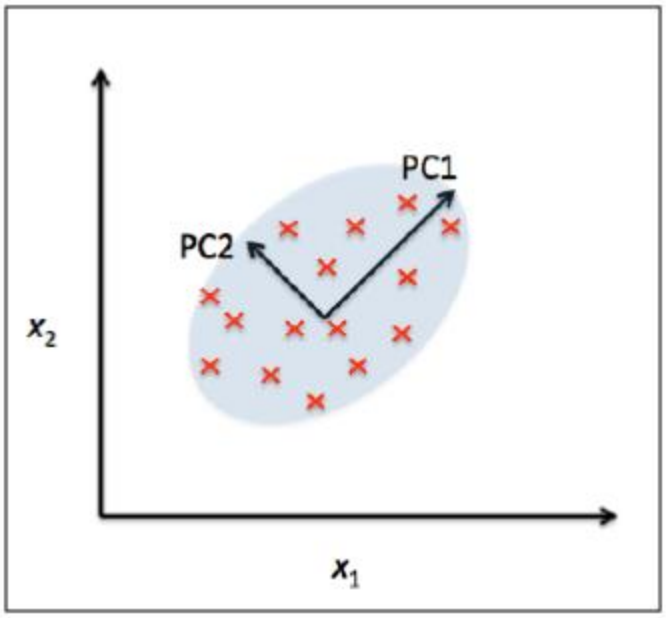


Рисунок 2.2 – Перший та другий головні компоненти

На графіку вище перший з головних компонентів (PC1) є синтетичною змінною, побудованою як лінійна комбінація для визначення величини та напрямку максимальної дисперсії в наборі даних. Цей компонент має найвищу мінливість серед усіх компонентів і, отже, найбільше інформації. Другий головний компонент (PC2) - це також синтетична лінійна комбінація, яка фіксує залишок дисперсії в наборі даних і не корелює з PC1. Наступні основні компоненти аналогічним чином фіксують решту змін, не корелюючи з попереднім компонентом.

PCA - це керований алгоритм навчання, оскільки напрямки цих компонентів обчислюються виключно на основі пояснювальних функцій без будь-якого посилання на змінні відповіді.

Кількість комбінацій ознак дорівнює кількості розмірів набору даних і загалом встановлює максимальну кількість PCA, яку можна побудувати. На рис. 2.3., кожна синя точка відповідає спостереженню, і кожна основна складова зменшує три виміри до двох. Алгоритм знаходить пару ортогональних векторів (червоні стрілки), які визначають простір нижчого розміру (сіра площина), щоб захопити якомога більшу дисперсію від вихідного набору даних.

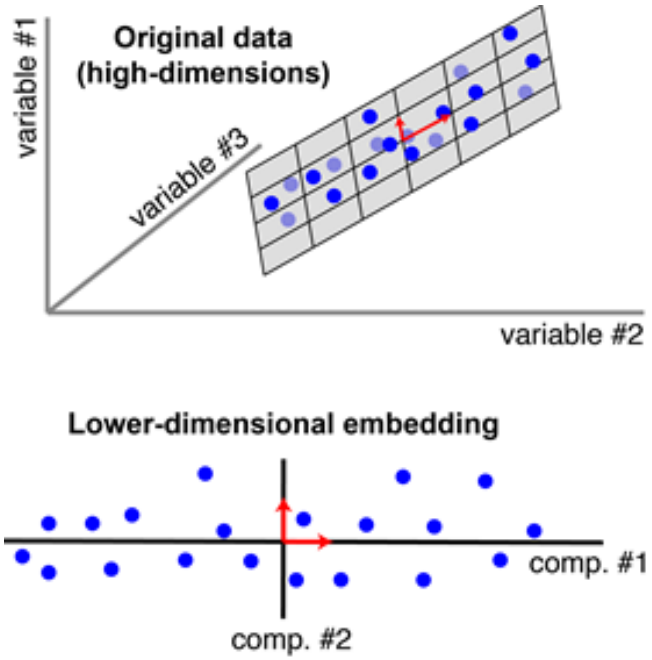


Рисунок 2.3 – Перетворення трьохвимірного простору у двохвимірне (Джерело: [3])

Для кількісної оцінки напрямку та величини варіації використовуються власні вектори та власні значення (рис. 2.4). Власний вектор описує кут або напрямок осі через простір даних, а власне значення кількісно визначає величину дисперсії даних на осі.

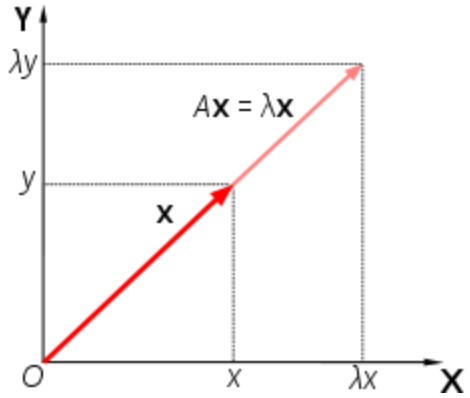


Рисунок 2.4 – Виміри що використовуються в методі головних компонент (A - матриця x n, ƛ - власне значення, а X - власний вектор).

Кількість комбінацій ознак дорівнює кількості розмірів набору даних. Наприклад, набір даних з десятьма вимірами матиме десять комбінацій власних значень / власних векторів.

Кореляція між кожною основною складовою повинна бути нульовою, оскільки наступні компоненти фіксують решту дисперсії. Кореляція між будь-якою парою власного значення / власного вектора дорівнює нулю, так що осі є ортогональними, тобто перпендикулярними одна одній у просторі даних.

Лінія, яка максимізує дисперсію даних, коли її проеціюють у простір даних, еквівалентна пошуку шляху, що мінімізує відстань найменших квадратів проекції (рис. 2.5).

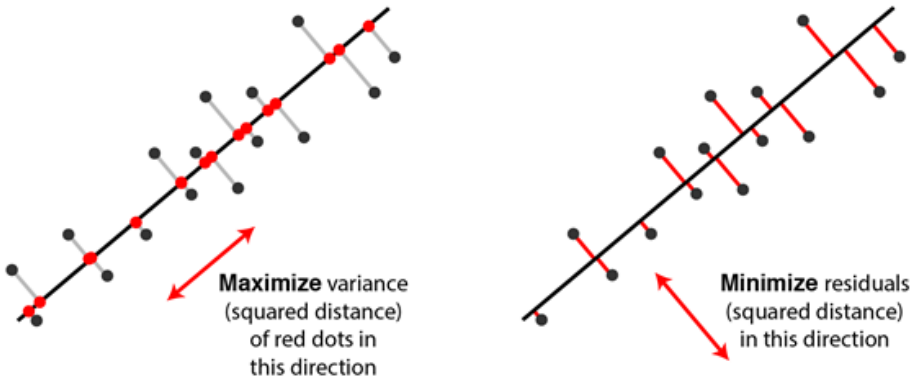


Рисунок 2.5 – Проекція значень

При цьому PCA базується на основі коефіцієнта кореляції Пірсона і до використання методу висуваються наступні припущення

1. Розмір вибірки: Мінімум 150 спостережень та ідеально співвідношення спостереження 5: 1 до особливостей.

2. Кореляції: Набір функцій корелює, тому зменшений набір функцій ефективно представляє вихідний простір даних.

3. Лінійність: Усі змінні мають постійний багатовимірний нормальний зв'язок, а основні компоненти є лінійною комбінацією вихідних ознак.

4. Оцінки: Немає значних відхилень у даних, оскільки вони можуть мати непропорційний вплив на результати.

5. Велика дисперсія передбачає більшу структуру: осі з великою дисперсією розглядаються як основні компоненти, тоді як осі з низькою дисперсією розглядаються як шум і відкидаються.

2.3 Реалізація методу головних компонент

2.3.1 Підготовка набору даних

На рисунку 2.6 зображено приклад початкового вигляду однієї з карток набору даних MNIST.

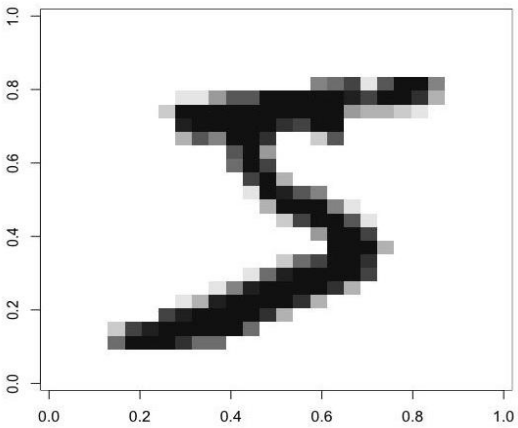


Рисунок 2.6 - Початковий вигляд першої цифри з набору даних

Щоб перевести всі значення числових стовпців у стандартний масштаб в наборі даних Було використано нормалізацію.

Нормалізуємо числові дані:

```
# Normalize the numeric data
train_normalized <- as.data.frame(scale(train$x,scale = FALSE, center = TRUE))
test_normalized <- as.data.frame(scale(test$x,scale = FALSE, center = TRUE))
```

Коваріація – це міра того, наскільки дві величини змінюються разом. Розрахунок матриці коваріації може бути виражений як:

$$C = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T$$

```
digits_covMatrix <- cov(train_normalized)
```

2.3.2 Реалізація методу головних компонент для набору даних MNIST

Загальна ідея методу головних компонент – зменшити розміри набору даних MNIST, одночасно вибираючи основні функції, які все ще представляють набір даних, зберігаючи ті, що мають більшу дисперсію, та відкидати ті, які мають меншу дисперсію. За менших розмірів візуалізація може бути більш життєздатною.

Запускаємо PCA, використовуючи навчальну матрицю коваріації

```
# Run PCA using the training covariance matrix
pca_train <- prcomp(train_normalized)
```

Дисперсія отримується діленням суми квадратів між дисперсіями на всі 784 компоненти.

Нижче наведено результат 50-ї дисперсії

```
# Look at 50th variance
```

NmbrPCs	CumVar
50	0.8246469

За евристикою, сукупний відсоток пояснювальної дисперсії становить понад 80%. Ми вибираємо 50 ГК, які з 50 ГК містять 82,5% відхилень даних.

Побудуємо графік даних (рис.2.7):

```
par(mfrow=c(2,2))
plot (variance_explained$NmbrPCs, variance_explained$CumVar,
    xlab = "Number of Factors", ylab = "Proportion of Variance Explained",
    type = "l", col = "red")
plot (pca_train$sdev^2/sum(pca_train$sdev^2), xlab = "Principal Component",
   ylab = "Proportion of Variance Explained", type = "b" )
plot(pca_train, type = "l", main = "Scree plot")
plot(pca_train, type = "barplot", main = "Scree plot")
```

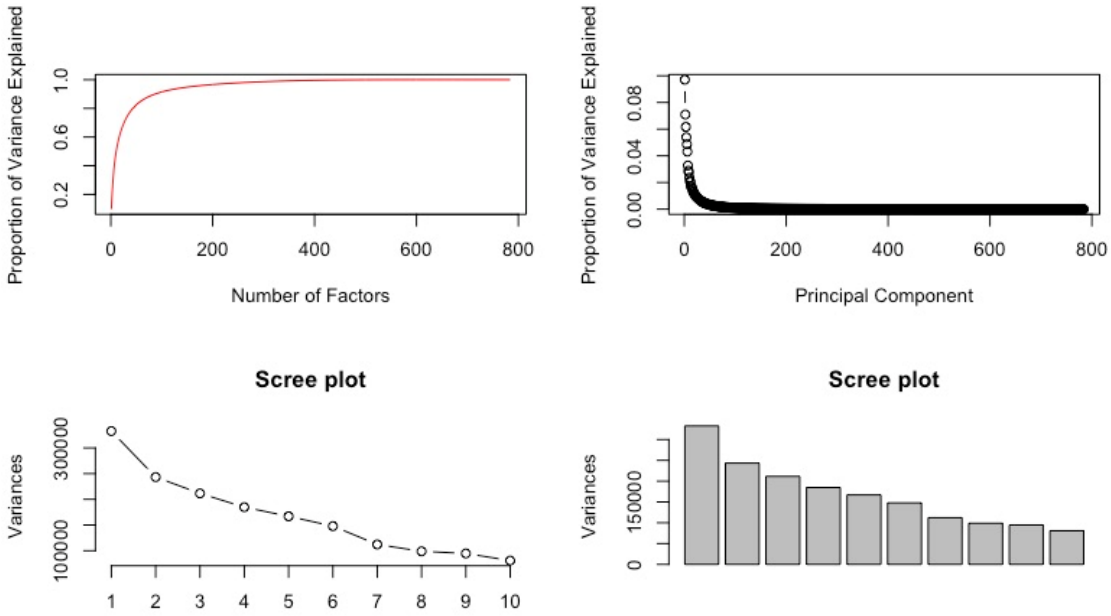


Рисунок 2.7 – Графіки даних

Розрахуємо важливість перших 5 компонент (рис.2.8)

```
summary(pca_train)$importance[,1:5]
```

	PC1	PC2	PC3	PC4	PC5
Standard deviation	576.82291	493.23822	459.89930	429.85624	408.56680
Proportion of Variance	0.09705	0.07096	0.06169	0.05389	0.04869
Cumulative Proportion	0.09705	0.16801	0.22970	0.28359	0.33228

Рисунок 2.8 – Характеристики перших 5 компонент

Згідно результатів (рис. 2.8), бачимо, що п’ять перших головних компонент пояснюють лише 33,288% змінності, що загалом непогано, але недостатньо для проведення якісної класифікації.

Графік на рис. 2.10 вказує на наявність великої кількості кластерів у наборі даних які важко розрізнити, що означає, що двокомпонентний PCA не дає нам бажаних шаблонів даних. Навіть тривимірний графік не дає жодних уявлень про можливий розподіл даних.

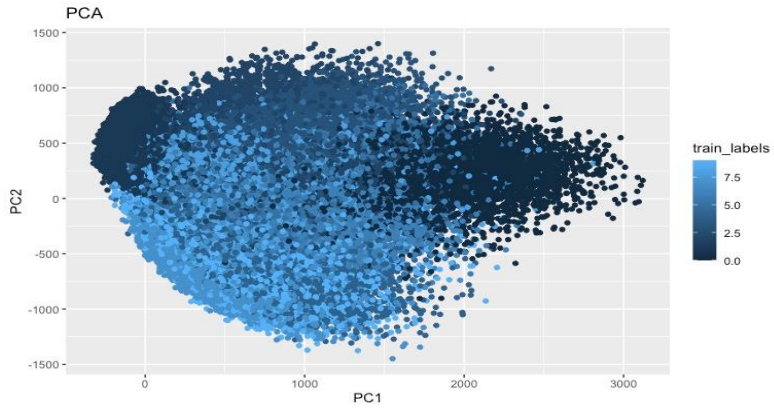


Рисунок 2.9 – Двовимірне представлення PCA у наборі навчальних даних MNIST

Для визначення достатньої кількості компонент скористуємося масштабованим зображенням (рис. 2.10).

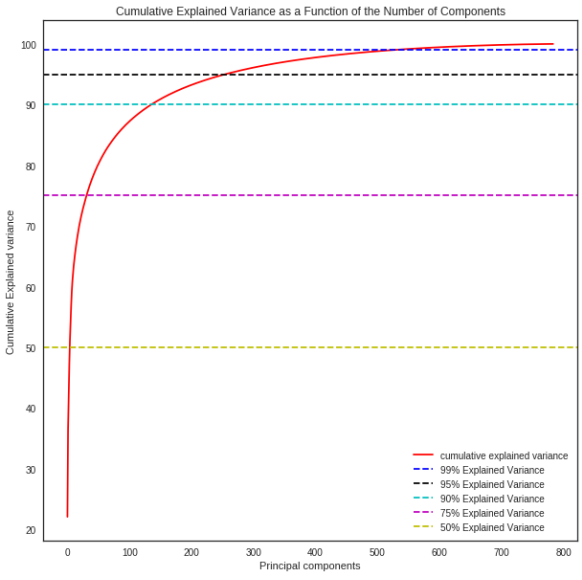


Рисунок 2.10 – Збільшене зображення розподілу головних компонент до кумулятивної дисперсії.

З рис. 2.10 видно, що 50 компонент описують майже 80% даних, отже можливо використовувати найвпливовіші ГК, які враховують початкові впливові характеристики і дозволяють зменшити простір з 748 до 50. Для перевірки цієї гіпотези, проведено реконструкцію рукописних символів для різних значень головних компонент. Нижче наведено приклад реконструкції цифр для PCA = (1, 10, 50, 784).

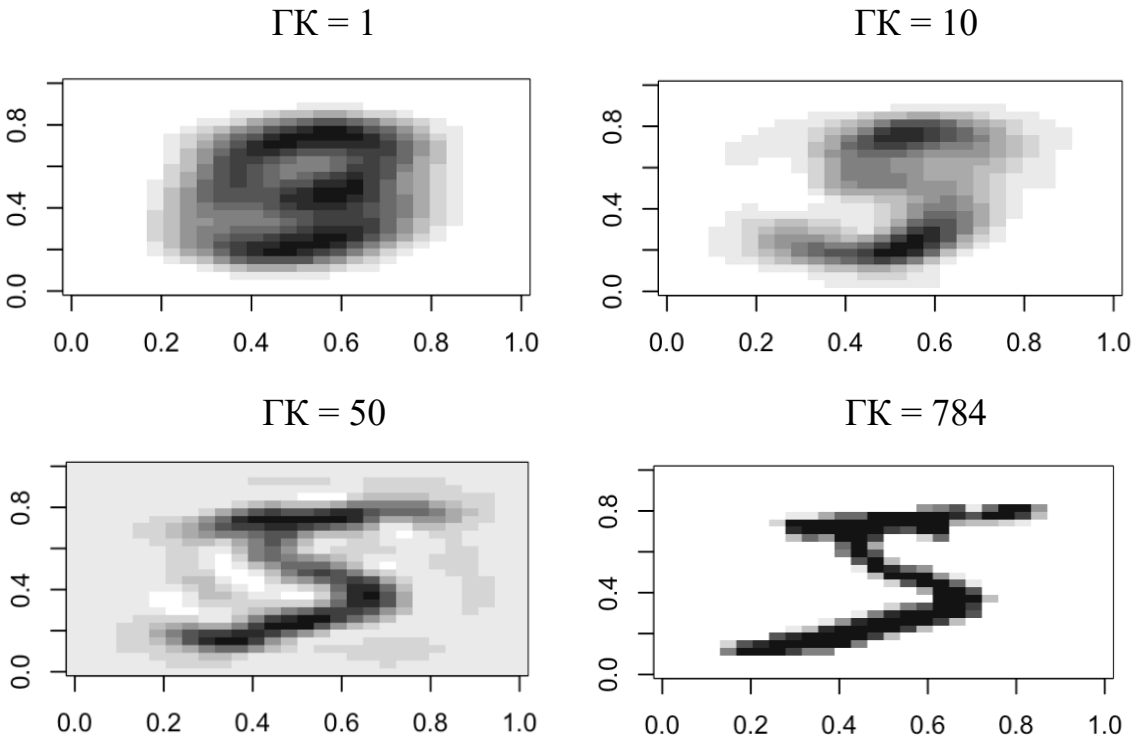


Рисунок 2.11 – Реконструкція цифри «5» з використанням різної кількості головних компонент (ГК)

З рис. 2.11 видно, що більша кількість ГК покращує якість реконструкції. При ГК=10 зображення можуть бути незрозумілими, деякі цифри легше ідентифікувати, ніж інші. З ГК=50 зображення зазвичай були достатньо чіткими, щоб людське око могло ідентифікувати та відокремлювати цифри. При ГК=784 – отримуємо оригінальний нарис цифр.

Отже значення PCA = 50 є найменшим значенням для прийнятного перегляду зображення цифр.

Висновки до розділу 2

В розділі проведено експерименти зі зменшення розмірності набору даних MNIST з використанням методу головних компонент.

За результатами проведеного аналізу, вибір був зроблений на користь 50 головних компонентів.

Спираючись на рішення щодо кількості основних компонентів, створюється остаточний набір даних.

Новий набір даних значно менший за розміром, але втрачає дуже мало з точки зору дисперсії, це означає, що швидкість та ефективність покращуються без значних втрат у дисперсії.

Цей набір даних буде використаний при застосуванні алгоритмів класифікації kNN та Random Forest.

Література до розділу 2

1. Bellman, Richard Ernest (1961). Adaptive control processes: a guided tour. Princeton University Press.

2. Goonewardana H. PCA: Application in Machine Learning https://medium.com/apprentice-journal/pca-application-in-machine-learning-4827c07a61db#:~:text=Principal%20Component%20Analysis%20(PCA)%20is,a%20large%20number%20of%20features.&text=PCA%20can%20also%20be%20used,datasets%2C%20such%20as%20image%20compression.

3. Williams A. Everything you did and didn't know about PCA http://alexhwilliams.info/itsneuronalblog/2016/03/27/pca/#f3b

4. Udell M. Generalized low rank models https://people.orie.cornell.edu/mru8/doc/udell15_thesis.pdf

3 ВИКОРИСТАННЯ МЕТОДІВ МАШИННОГО НАВЧАННЯ ДЛЯ РОЗПІЗНАВАННЯ РУКОПИСНИХ СИМВОЛІВ

3.1 Метод k-найближчих сусідів

Метод найближчих сусідів (kNN - k-Nearest Neighbors) – це алгоритм класифікації, який в процесі роботи зберігає тренувальні дані. Активний класифікатор створює класифікаційну модель в процесі навчання. Коли вводяться нові дані, класифікатор поглинає нові дані класифікаційної моделі.

Він почне процес класифікації лише тоді, коли з’являться нові непозначені дані, kNN спочатку шукає k-найближчі розмічені точки даних, після, використовуючи класи сусідів, вирішує як краще класифікувати нові дані.

Цей метод використовує дистанційну метрику, наприклад Евклидову:

$$\rho(x,x') = \sum_{i=1}^{n} (x_i - x_i')^2.$$

Вибір метрики залежить від типу даних, деякі можуть використовувати навіть дистанційну метрику на основі тренувальних даних. Цей метод потребує машинного навчання.

Алгоритм пошуку оптимальних параметрів реалізовано оптимальним значенням К, який знаходять по критерію змінного контролю з виключенням об’єктів по одному.

$$(k^*) = \operatorname{argmax}_k \; LOO(k;X^\ell), \text{ где } LOO(k;X^\ell) = \sum_{i=1}^{l} \left[y_i = a(x_i;X^l \setminus x_i;k)\right].$$

Алгоритм пошуку у разі нерівнозначних помилок

$$LOO(k;q;X^\ell) = \sum_{i=1}^{l} err(y_i, a(x_i;X^l/x_i;k;q)).$$

К-найближчий сусід – найпростіший алгоритм класифікації зображень. Для прогнозування нового екземпляра KNN обчислює евклідову відстань між новим екземпляром та усіма екземплярами у всьому навчальному наборі. Потім алгоритм шукає К найближчі (найбільш подібні) екземпляри і виводить клас із найвищою частотою (більшість голосів) як прогноз.

Запустили KNN з k від 1 до 20 та побудували точність для різних k (1:20):

```
> t(knn_acc)
      [,1]   [,2]   [,3]   [,4]   [,5]   [,6]
[1,] 1.0000 2.0000 3.0000 4.0000 5.0000 6.0000
[2,] 0.9734 0.9691 0.9751 0.9748 0.9751 0.9748
      [,7]   [,8]   [,9]   [,10]   [,11]
[1,] 7.0000 8.0000 9.0000 10.0000 11.0000
[2,] 0.9754 0.9754 0.9748  0.9738  0.9746
      [,12]  [,13]   [,14]   [,15]   [,16]
[1,] 12.0000 13.000 14.0000 15.0000 16.0000
[2,]  0.9734  0.973  0.9724  0.9718  0.9713
      [,17]  [,18]   [,19]  [,20]
[1,] 17.0000 18.000 19.0000 20.000
[2,]  0.9718  0.971  0.9709  0.971
```

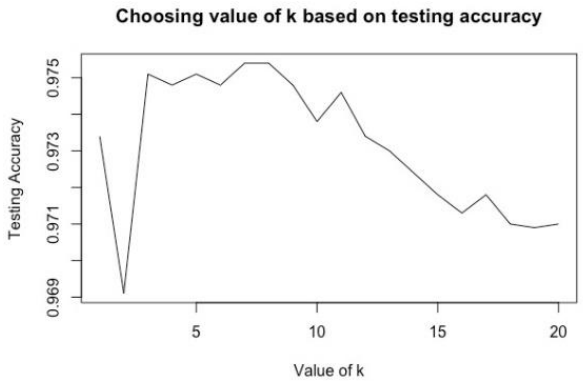


Рисунок 3.1 – Точність для різних k

Згідно графіку можна зрозуміти, що k = 8 дає найкращу точність.

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1    2    3    4    5    6    7    8    9
         0  971    0    8    0    1    3    2    0    4    3
         1    1 1131    0    1    4    0    5   17    0    4
         2    1    2 1002    1    0    0    0    4    3    2
         3    0    0    0  978    0    8    0    0   11    6
         4    0    0    1    1  953    1    3    4    5    7
         5    1    0    0   11    0  866    1    0    6    4
         6    5    1    2    0    5    8  947    0    2    1
         7    1    0   15    6    1    1    0  989    5    5
         8    0    0    4    5    1    3    0    0  933    4
         9    0    1    0    7   17    2    0   14    5  973

Overall Statistics

               Accuracy : 0.9743
                 95% CI : (0.971, 0.9773)
    No Information Rate : 0.1135
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9714

 Mcnemar's Test P-Value : NA

Statistics by Class:
```

Рисунок 3.2 – Матриця плутанини

Алгоритм вибрав 20 як початкове значення k з навчального набору квадратного кореня спостережень. Значення k = 8 з найкращою точністю вказує на те, що модель алгоритму машини KNN передбачить клас будь-якої вхідної цифри з 97,43% точністю, використовуючи 8 найближчих сусідів.

Як видно з матриці плутанини (рис.3.3), найбільше помилок при розрізнаванні було між цифрами 1 і 7 та 9 і 4, що зрозуміло оскільки при написанні арабським шрифотм, що використано в базі даних MNIST, ці цифри мають найбільшу схожість при написі, що підтверджується також відносно великою кількістю помилок між розпізнаванням цифри 3 і 8.

```
> table_prediction
          test_labels
prediction    0    1    2    3    4    5    6    7    8    9
         0  971    0   10    0    0    3    5    0    4    4
         1    1 1129    2    1    4    0    5   23    0    5
         2    1    2  996    2    0    0    0    4    3    3
         3    0    1    1  976    0   10    0    0   13    5
         4    0    0    1    1  946    1    2    2    5    9
         5    1    0    0    9    0  865    1    0    7    6
         6    5    2    3    0    7    9  945    0    2    1
         7    1    0   13    7    0    1    0  983    3    6
         8    0    0    6    9    2    0    0    0  930    1
         9    0    1    0    5   23    3    0   16    7  969
> accuracy = hit/sum(table_prediction)
> accuracy
[1] 0.9743
```

Рисунок 3.3 – Обчислення точності для k = 8

3.2 Random Forest

Другим алгоритмом, за яким будуть класифіковані дані MNIST, є випадковий ліс. Цей алгоритм базується на деревах рішень, які функціонують як набір вузлів та ребер. Точки даних перевіряються і на основі результату поділяються на підгрупи, де використовується інше відповідне запитання (= тест), і дані поділяються далі, поки всі кінцеві підгрупи не стануть «чистими»

- тобто всі вони потрапляють в одну і ту ж відповідь категорія останнього запитання, яке було задано. Однак цей алгоритм може бути легко переобладнаний, і він страждає від неточності. Випадковий ліс - одне з удосконалень цього методу, який було запроваджено. На відміну від дерев рішень, набір даних завантажується шляхом випадкового відбору зразків із вихідного набору даних, а потім на ньому запускається дерево рішень. Для кожного дерева створюється новий набір завантажених даних. Цей метод відомий як bagging або пакування в мішки. На додаток до пакування в мішки набору даних, випадковий ліс також використовує мішки для об'єктів, де випадкова вибірка об'єктів відбирається при кожному розділі дерева. Для алгоритму обчислюється помилка поза пакетом, оскільки не всі вибірки набору навчальних даних вибрані для сумки. Це відрізняється від помилки та точності тестування.

Як і в попередньому алгоритмі, основним методом оцінки буде точність, використовуючи функцію середнього значення (передбачені_мітки == істинні_мітки). Кількість ознак, які будуть розраховані на кожному кроці, м, буде визначено за допомогою аналогічного експерименту, як у попередньому розділі.

Для експерименту з випадковим лісом, спочатку кількість дерев було вибрано як 50, такий вибір пояснюється наявною потужністю ПК, на якому проводися експеримент, оскільки підвищення кількості дерев вимагає більше потужності та часу на проведення експериментів.

Потім запустили випадковий ліс з m від 1 до 10 і міру точності зберегли в матриці. В якості значень m було обрано від 1 до 10 черезтак зване «правило великого пальця», яке стверджує, що хорошим значенням m є квадратний корінь із числа об'єктів - в нашому випадку, в якості початкового значення використовувалося 50 компонент (зменшений набір даних), отже корінь квадратний з 50 становить близько 7,07. Матриця плутанини надана на рис.3.4. Як видно, точність значно нижча ніж у попередньому експерименті, більш того, поряд з плутаниною між 9 і 4, 1 і 7 з'явилися інші.

```
Call:
 randomForest(x = train_final, y = train_labels, ntree = numTrees)
               Type of random forest: classification
                     Number of trees: 50
No. of variables tried at each split: 7

        OOB estimate of  error rate: 6.19%
Confusion matrix:
     0    1    2    3    4    5    6    7    8    9 class.error
0 5755    0   15   13   16   27   55    9   26    7  0.02836400
1    1 6602   43   22    7   10    8   13   27    9  0.02076535
2   51    8 5574   71   40   15   26   58   98   17  0.06445116
3   13    3  106 5610    7  122   21   51  145   53  0.08497798
4   12   20   48    8 5495    4   39   23   37  156  0.05939747
5   31    3   26  144   39 5015   56    8   60   39  0.07489393
6   31    7   21    9   15   72 5745    1   15    2  0.02923285
7    5   31   69   15   60   12    1 5914   20  138  0.05602554
8   23   44   80  177   40  149   39   33 5191   75  0.11280123
9   25    9   32   90  183   34    4  135   49 5388  0.09430156
```

Рисунок 3.4 – Матриця плутанини

Разом з тим, очікуваний рівень помилок (OOB estimate of error rate - оцінка поза мішком, - метод вимірювання помилки прогнозування випадкових лісів, підсилених дерев рішень та ін.) складає 6.19%. Графік точності відносно m показує, що найкраща точність тестування досягається при m = 7, що становить 93,81%. Ще однією особливістю випадкового лісового алгоритму, з яким можна експериментувати, є кількість дерев. Дивлячись на графік помилки, що вийшла з мішка, порівняно з кількістю дерев для m = 7, рівень помилок залишається досить постійним приблизно після 100 дерев, тому обрана кількість 300 дерев не повинна впливати на точність.

Випадковий ліс - це алгоритм класифікації, що складається з багатьох дерев рішень. Він використовує мішки та має випадковість при побудові кожного окремого дерева, щоб спробувати створити некорельований ліс з

дерев, прогнозування якого комітетом є більш точним, ніж будь-яке окреме дерево. Створимо випадкову вибірку індексів "numTrain"

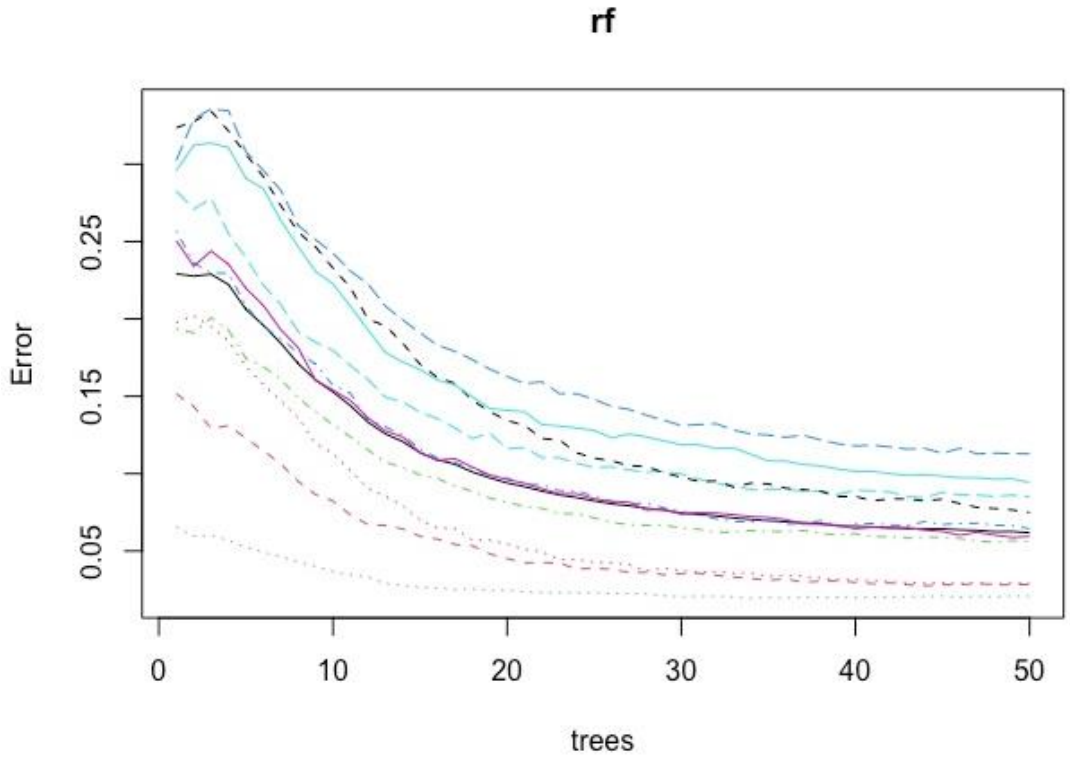


Рисунок 3.5 – Ділянка rf для різної кількості дерев

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1    2    3    4    5    6    7    8    9
         0  961    0   12    1    1    4    8    1    7    2
         1    0 1122    0    0    1    1    2    6    0    5
         2    2    4  967    6    3    5    3   15   10    6
         3    0    1   11  954    0   14    0    1   15   10
         4    0    0    6    0  935    3    3    8    8   22
         5    2    2    1   17    3  846    7    0   17    8
         6    9    3    6    2   10    7  933    0    4    2
         7    1    0   10    6    2    1    0  971    8   10
         8    3    2   18   16    5    7    2    3  896    9
         9    2    1    1    8   22    4    0   23    9  935

Overall Statistics

               Accuracy : 0.952
                 95% CI : (0.9476, 0.9561)
    No Information Rate : 0.1135
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9466

 Mcnemar's Test P-Value : NA

Statistics by Class:
```

Рисунок 3.6– Розрахунок прогнозу

Прогнози були подані та отримали оцінку 94,66%.

	Class: 0	Class: 1	Class: 2	Class: 3	Class: 4	Class: 5
Sensitivity	0.9806	0.9885	0.9370	0.9446	0.9521	0.9484
Specificity	0.9960	0.9983	0.9940	0.9942	0.9945	0.9937
Pos Pred Value	0.9639	0.9868	0.9471	0.9483	0.9492	0.9369
Neg Pred Value	0.9979	0.9985	0.9928	0.9938	0.9948	0.9949
Prevalence	0.0980	0.1135	0.1032	0.1010	0.0982	0.0892
Detection Rate	0.0961	0.1122	0.0967	0.0954	0.0935	0.0846
Detection Prevalence	0.0997	0.1137	0.1021	0.1006	0.0985	0.0903
Balanced Accuracy	0.9883	0.9934	0.9655	0.9694	0.9733	0.9711
	Class: 6	Class: 7	Class: 8	Class: 9		
Sensitivity	0.9739	0.9446	0.9199	0.9267		
Specificity	0.9952	0.9958	0.9928	0.9922		
Pos Pred Value	0.9559	0.9623	0.9324	0.9303		
Neg Pred Value	0.9972	0.9937	0.9914	0.9918		
Prevalence	0.0958	0.1028	0.0974	0.1009		
Detection Rate	0.0933	0.0971	0.0896	0.0935		
Detection Prevalence	0.0976	0.1009	0.0961	0.1005		
Balanced Accuracy	0.9846	0.9702	0.9564	0.9594		

Рисунок 3.7 – Порівняння класів

Експеримерт з випадковим лісом пройшов добре, був легким у навчанні та досить швидким.

Висновки до розділу 3

В розділі проведено дослідження алгоритмів класифікації зображень.

Під час тестування алгоритму методу найближчих сусідів було виявлено, що дана модель передбачить клас будь якої вхідної цифри з 97,43% точністю, використовуючи 8 найближчих сусідів.

Випадковий ліс, це другий алгоритм, за яким проводилися дослідження класифікації даних, за яким отримали оцінку 94,66%.

Саме випадковий ліс відпрацював швидше та був досить простим в навчанні.

4 ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ

В даному розділі було розглянуто загальні питання з охорони праці. Були проаналізовані умови праці, вимоги до приміщення та організації. Розглянуті заходи, які дозволяють забезпечити гігієну праці та виробничу санітарію. Також були розглянуті рекомендації щодо пожежної безпеки, електробезпеки, мікроклімату та освітлення.

4.1 Загальні питання з охорони праці

Згідно з законом “Про охорону праці” [1] охорона праці це – система правових, соціально-економічних, організаційно-технічних, санітарно-гігієнічних і лікувально-профілактичних заходів та засобів, спрямованих на збереження життя, здоров'я і працездатності людини у процесі трудової діяльності.

При роботі з обчислювальною технікою змінюються фізичні і хімічні фактори навколишнього середовища: виникає статична електрика, електромагнітне випромінювання, змінюється температура і вологість, рівень вміст кисню і озону в повітрі. Забезпечення цих умов покладається на власника або уповноважений ним орган (далі роботодавець). Умови праці на робочому місці, безпека технологічних процесів, машин, механізмів, устаткування та інших засобів виробництва, стан засобів колективного та індивідуального захисту, що використовуються працівником, а також санітарно-побутові умови повинні відповідати вимогам нормативних актів про охорону праці.

4.1.1 Правові та організаційні основи охорони праці

Основним організаційним напрямом у здійсненні управління в сфері

охорони праці є усвідомлення пріоритету безпеки праці і підвищення соціальної відповідальності держави, і особистої відповідальності працівників.

Державна політика в галузі охорони праці визначається відповідно до Конституції України Верховною Радою України і спрямована на створення належних, безпечних і здорових умов праці, запобігання нещасним випадкам та професійним захворюванням. Відповідно до статті 3 Закону України “Про охорону праці” [1] (далі – Закону) законодавство про охорону праці складається з Закону, Кодексу законів про працю України [2], Закону України “Про загальнообов'язкове державне соціальне страхування від нещасного випадку на виробництві та професійного захворювання, які спричинили втрату працездатності” [3] та прийнятих відповідно до них нормативно-правових актів, норм міжнародного договору (ратифіковані Конвенції і Рекомендації МОТ, директиви Європейської Ради).

На законодавчому рівні визначено такі пріоритетні напрямки з безпеки праці:

– кожен працівник несе безпосередню відповідальність за порушення зазначених Законом, нормами і правилами вимог;

– напрямки реалізації конституційного права громадян на їх життя і здоров’я в процесі трудової діяльності:

– пріоритет життя і здоров’я працівників по відношенню до результатів виробничої діяльності підприємства;

– повна відповідальність роботодавця за створення належних – безпечних і здорових умов праці;

– соціальний захист працівників, повне відшкодування збитків особам, які потерпіли від нещасних випадків на виробництві та професійних захворювань;

– комплексне розв’язання завдань охорони праці;

– підвищення рівня промислової безпеки шляхом забезпечення суцільного технічного контролю за станом виробництв, технологій та

продукції, а також сприяння підприємствам у створенні безпечних та нешкідливих умов праці;

– соціальний захист працівників, повне відшкодування збитків особам, які потерпіли від нещасних випадків на виробництві та професійних захворювань;

– використання економічних методів управління охороною праці, участь держави у фінансуванні заходів щодо охорони праці;

– використання світового досвіду організації роботи щодо поліпшення умов і підвищення безпеки праці на основі міжнародної співпраці.

Користувачі персональних комп'ютерів, для яких ця робота є головною, підлягають медичним оглядам: попереднім — під час влаштування на роботу і періодичним — протягом професійної діяльності раз на два роки.

Наявні трудові відносини між працівниками і роботодавцями в Україні за темою дипломного проекту регулюються Кодексом законів про працю (КЗпП) України, відповідно до якого права працюючої людини на охорону праці охороняються всебічно та норми охорони праці неухильно інтегровані до правил внутрішнього розпорядку організації/підприємства.

4.1.2 Організаційно-технічні заходи з безпеки праці

В організації/підприємстві проводиться навчання і перевірка знань з питань охорони праці відповідно до вимог Типового положення про порядок проведення навчання і перевірки знань з питань охорони праці, затвердженого наказом Держнаглядохоронпраці України від 26.01.2005 N 15, зареєстрованого в Міністерстві юстиції України 15.02.2005 за N 231/10511 [4].

Обов'язковими вимогами враховане наступне:

– ознайомлення з правилами безпеки праці, одержання відповідних

інструктажів засвідчується у журналі інструктажів.

– перед допуском до самостійної роботи кожен працівник має право на навчання з питань охорони праці і роботодавець зобов’язаний, і проводить таке навчання у вигляді двох інструктажів з питань охорони праці:

1) вступного, який проводять працівники служби охорони праці об’єкта господарювання з усіма працівниками, яких приймають на роботу незалежно від їхньої освіти та стажу роботи за програмою, в якій подають загальні питання охорони праці із врахуванням її особливостей на об’єкті господарювання;

2) первинного, який проводять керівники структурних підрозділів на місці праці з кожним працівником до початку їхньої роботи на цьому робочому місці.

Проходження працівником цих інструктажів з питань охорони праці підтверджується записами у відповідних журналах обліку інструктажів і скріплюється підписами осіб, які проводили інструктажі та осіб, які отримали інструктажі.

3) Повторний (не рідше одного разу в 6 місяців);

4) Позаплановий (при зміні правил охорони праці);

5) Поточний (проводять з працівниками перед виконанням робіт, на яких оформляється наряд-допуск)

– обов’язкові організаційні заходи перед початком, під час і після завершення роботи повинні включати перевірку (візуально) наявності і справності електрообладнання та його заземлення, а під час виконання роботи вимогу “не залишати без нагляду обладнання, яке працює”. Після закінчення роботи - вимагається прибирання робочого місця, відключення всіх електроприладів від електромережі.

4.2 Аналіз стану та умов праці

Для роботи над створенням система віддаленого моніторингу біофізичних параметрів людини достатньо однієї людини для якої надано робоче місце зі стаціонарним комп’ютером.

4.2.1 Вимоги до приміщення

Геометричні розміри приміщення зазначені в табл. 4.1.

Таблиця 4.1 – Розміри приміщення

Найменування	**Значення**
Довжина, м	5
Ширина, м	4
Висота, м	3
Площа, $м^2$	20
Об’єм, $м^3$	60

Згідно з ДСанПІН 3.3.2.007-98 “Державних санітарних правила і норми роботи з візуальними дисплейними терміналами електронно-обчислювальних машин” [5] розмір площі для одного робочого місця оператора персонального комп’ютера має бути не менше 6 кв. м, а об’єм — не менше 20 куб. м. Отже, дане приміщення цілком відповідає зазначеним нормам.

Також для дотримання визначеного рівня мікроклімату в будівлі встановлено систему опалення та кондиціонування.

Для забезпечення потрібного рівного освітленості кімната має вікно та систему загального рівномірного освітлення, що встановлена на стелі. Для

дотримання вимог пожежної безпеки встановлено порошковий вогнегасник та систему автоматичної пожежної сигналізації.

4.2.2 Вимоги до організації робочого місця

При порівнянні відповідності характеристик робочого місця к нормативним, основні вимоги до організації робочого місця за ДСанПіН 3.3.2.007-98 [5] і відповідними фактичними значеннями для робочого місця за ДСН 3.3.6.042-99 [6], констатуємо повну відповідність в таблиці 4.2.

Таблиця 4.2 - Характеристики робочого місця

Найменування параметра	Фактичне значення	Нормативне значення
Висота робочої поверхні, мм	700	680 ÷800
Висота простору для ніг, мм	680	не менше 600
Ширина простору для ніг, мм	550	не менше 500
Глибина простору для ніг, мм	700	не менше 650
Висота поверхні сидіння, мм	500	400 ÷500
Ширина сидіння, мм	400	не менше 400
Глибина сидіння, мм	400	не менше 400
Висота поверхні спинки, мм	600	не менше 300
Ширина опорної поверхні спинки, мм	400	не менше 380
Радіус кривини спинки в горизонтальній площині, мм	400	400
Відстань від очей до екрану дисплея, мм	800	700 ÷800

Приміщення кабінету знаходиться у вчасній одноповерховій будівлі і має об’єм 45 м3, площу — 18 м2.

Температура в приміщенні протягом року коливається у межах 18–24°С, відносна вологість — близько 50%. Система вентилювання приміщення — природна неорганізована, а опалення — централізоване.

Розміщення вікон забезпечує природне освітлення з коефіцієнтом природного освітлення не менше 1,5%, а загальне штучне освітлення, яке здійснюється за допомогою однієї люмінесцентної лампи, забезпечує рівень освітленості не менше 200 Лк. За ступенем пожежної безпеки приміщення належить до категорії В.

4.2.3 Навантаження та напруженість процесу праці

Під час виконання робіт використовують ПК та периферійні пристрої (лазерні та струменеві), що призводить до навантаження на окремі системи організму. Такі перекоси у напруженні різних систем організму, що трапляються під час роботи з ПК, зокрема, значна напруженість зорового аналізатора і довготривале малорухоме положення перед екраном, не тільки не зменшують загального напруження, а навпаки, призводять до його посилення і появи стресових реакцій.

Найбільшому ризику виникнення різноманітних порушень піддаються: органи зору, м’язово-скелетна система, нервово-психічна діяльність, репродуктивна функція у жінок.

Тобто наявні психофізіологічні небезпечні та шкідливі фактори:

а) фізичного перевантаження:

- статичного;
- динамічного;

б) нервово-психічного перевантаження:

- розумового перенапруження;
- монотонності праці;
- перенапруження аналізаторів;
- емоційних перевантажень.

Роботу за дипломним проектом визнано, таку, що займає 50% часу робочого дня та за восьмигодинної робочої зміни рекомендовано встановити додаткові регламентовані перерви тривалістю 15 хв через кожну годину роботи.

4.3 Виробнича санітарія

На підставі аналізу небезпечних та шкідливих факторів при виробництві (експлуатації), пожежної безпеки можуть бути надалі вирішені питання необхідності забезпечення працюючих достатньою кількістю освітлення, вентиляції повітря, організації заземлення, тощо.

4.3.1 Аналіз небезпечних та шкідливих факторів при виробництві (експлуатації) виробу

Роботу, пов'язану з ЕОП з ВДТ, у тому числі на тих, які мають робочі місця, обладнані ЕОМ з ВДТ і ПП, виконують із забезпеченням виконання НПАОП 0.00-7.15-18 [7] “Вимоги щодо безпеки та захисту здоров'я працівників під час роботи з екранними пристроями”, які встановлюють вимоги безпеки до обладнання робочих місць, до роботи із застосуванням ЕОМ з ВДТ і ПП. Переважно роботи за проектами виконують у кабінетах чи інших приміщеннях, де використовують різноманітне електрообладнання, зокрема персональні комп’ютери (ПК) та периферійні пристрої.

Основними робочими характеристиками персонального комп'ютера є наступні:

- робоча напруга U = +220В ± 5%;
- робочий струм I = 2А;
- споживана потужність P = 350 Вт.

Робоче місце має відповідати вимогам Державних санітарних правил і норм роботи з візуальними дисплейними терміналами електронно-

обчислювальних машин, затверджених постановою Головного державного санітарного лікаря України від 10.12.98 N 7 [5].

Аналіз небезпечних та шкідливих виробничих факторів виконується у табличній формі (табл. 4.3).

Таблиця 4.3 – Аналіз небезпечних і шкідливих виробничих факторів

Небезпечні і шкідливі виробничі фактори	Джерела факторів (види робіт)	Кількісна оцінка	Нормативні документи
1	2	3	4
Фізичні			
- підвищений рівень напруги електричної мережі, замикання якої може відбутися через тіло людини	-//-	4	[8]
- недостатність природного світла	порушення умов праці (вимог до приміщень)	2	[10]
- недостатнє освітлення робочої зони	порушення гігієнічних параметрів виробничого середовища	3	[10]
Психофізіологічні:			
- нервово-психічна перевантаження (розумове, перенапруження аналізаторів-	- пошук інформації для постановки теми; - пошук та аналіз аналогів і літератури; - пошук наявних технологій,	4	[7] [5]

зорових)	моделювання та аналіз алгоритмів; - виконання роботи за темою диплома, тестування; - оформлення роботи		
Небезпечні і шкідливі виробничі фактори	Джерела факторів (види робіт)	Кількісна оцінка	Нормативні документи
1	2	3	4
Психофізіологічні:			
- фізичні (статичне – сидіння)	порушення умов праці (організації місця праці- сидіння користувача,) та організації робочого часу - безперервна робота)	2	[7] [5]

4.3.2 Рекомендації щодо пожежної безпеки

Пожежна безпека при застосуванні ПК забезпечується:

– системою запобігання пожежі,

– системою протипожежного захисту,

– організаційно-технічними заходами.

Згідно ДСТУ Б В.1.1-36:2016 [9] таке приміщення, площею 18 м2, відноситься до категорії “В” (пожежонебезпечної) та для протипожежного захисту в ньому проектом передбачено устаткування автоматичною пожежною сигналізацією із застосуванням датчиків-сповіщувачів РІД-1 (сповіщувач димовий ізоляційний) в кількості 1 шт., і застосуванням первинних засобів пожежогасіння.

Горючими матеріалами в приміщенні, де розташовані ПК, є:

– поліамід – матеріал корпусу мікросхем, горюча речовина, температура самозаймання 420 ˚С,

– полівінілхлорид – ізоляційний матеріал, горюча речовина, температура запалювання 335˚С, температура самозаймання 530 ˚С,

– склотекстоліт ДЦ – матеріал друкарських плат, важко горючий матеріал, показник горючості 1.74, не схильний до температурного самозаймання,

– пластикат кабельний №.489 – матеріал ізоляції кабелів, горючий матеріал, показник горючості більше 2.1,

– деревина – будівельний і обробний матеріал, з якого виготовлені меблі, горючий матеріал, показник горючості більше 2.1, температура запалювання 255 ℃, температура самозаймання 399 ˚С.

Простори усередині приміщень в межах, яких можуть утворюватися або знаходиться пожежонебезпечні речовини і матеріали відповідно до ДСТУ Б В.1.1-36:2016 [9] відносяться до пожежонебезпечної зони класу П-IIа. Це обумовлено тим, що в приміщенні знаходяться тверді горючі та важко займисті речовини та матеріали. Приміщенню, у якому розташоване робоче місце, присвоюється II ступень вогнестійкості.

Продуктами згорання, що виділяються на пожежі, є: окис вуглецю; сірчистий газ; окис азоту; синильна кислота; акромін; фосген; хлор і ін. При горінні пластмас, окрім звичних продуктів згорання, виділяються різні продукти термічного розкладання: хлорангідриді кислоти, формальдегіди, хлористий водень, фосген, синильна кислота, аміак, фенол, ацетон, стирол.

4.3.3 Електробезпека

Основним небезпечним фактором при роботі з ЕОМ є небезпека ураження людини електричним струмом, яка посилюється тим, що органи чуття людини не можуть на відстані виявити наявність електричної напруги на обладнанні.

Проходячи через тіло людини, електричний струм чинить на нього складний вплив, що є сукупністю термічної (нагрів тканин і біологічних середовищ), електролітичної (розкладання крові і плазми) і біологічної (роздратування і збудження нервових волокон та інших органів тканин організму) дій.

Тяжкість ураження людини електричним струмом залежить від цілого ряду чинників:

1) значення сили струму;

2) електричного опору тіла людини і тривалості протікання через нього струму;

3) типу і частоти струму;

4) індивідуальних властивостей людини і навколишнього середовища.

Приміщення для ЕОМ відноситься до приміщень без підвищеної небезпеки, тобто в приміщення, в яких відсутні умови, що створюють підвищену або особливу небезпеку. Небезпека ураження електричним струмом існує всюди, де використовуються електроустановки, тому приміщення без підвищеної небезпеки не можна назвати безпечними.

Електробезпека забезпечується:

1) відповідною конструкцією електроустановок;

2) застосуванням технічних способів і засобів захисту;

3) організаційними і технічними заходами.

Основними технічними способами і засобами захисту від ураження електричним струмом, що використовуються окремо або в поєднанні один з одним, є:

1) захисне заземлення;

2) занулення;

3) вирівнювання потенціалів;

4) мале напруга;

5) електричне поділ мереж;

6) захисне відключення;

7) ізоляція струмоведучих частин;

8) компенсація струмів замикання на землю;

9) захисні пристрої;

10) попереджувальна сигналізація, блокування, знаки безпеки;

11) ізолюючі захисні та запобіжні пристосування.

4.4 Гігієнічні вимоги до параметрів виробничого середовища

4.4.1 Мікроклімат

Мікроклімат робочих приміщень – це клімат внутрішнього середовища цих приміщень, що визначається діючої на організм людини з'єднанням температури, вологості, швидкості переміщення повітря. В даному приміщенні проводяться роботи, що виконуються сидячи і не потребують динамічного фізичного напруження, то для нього відповідає категорія робіт Іа. Отже оптимальні значення для температури, відносної вологості й рухливості повітря для зазначеного робочого місця відповідають ДСН 3.3.6.042-99 “Санітарні норми мікроклімату виробничих приміщень” [6] і наведені в таблиці 4.3.

Таблиця 4.3 – Норми мікроклімату робочої зони об’єкту

Період рок	Температура, ^{o}C	Відносна вологість,%	Швидкість вітру, м/с, не більше
Холодний	21 - 23	60 – 40	0.1
Теплий	22 - 24	60 - 40	0.2

4.4.2 Освітлення

Світло є природною умовою існування людини. Воно впливає на стан вищих психічних функцій і фізіологічні процеси в організмі. Хороше освітлення діє тонізуюче, створює гарний настрій, покращує протікання основних процесів вищої нервової діяльності.

Збільшення освітленості сприяє поліпшенню працездатності навіть в тих випадках, коли процес праці практично не залежить від зорового сприйняття. При поганому освітленні людина швидко втомлюється, працює менш продуктивно, виникає потенційна небезпека помилкових дій і нещасних випадків.

У проекті, що розробляється, передбачається використовувати суміщене освітлення. У світлий час доби використовуватиметься природне освітлення приміщення через віконні отвори, в решту часу використовуватиметься штучне освітлення.

Штучне освітлення в робочому приміщенні передбачається здійснювати з використанням люмінесцентних джерел світла у світильниках загального освітлення, оскільки люмінесцентні лампи мають високу потужність (80 Вт), тривалий термін служби (до 10000 годин), спектральним складом випромінюваного світла, близький до сонячного. При експлуатації ПК виконується зорова робота IV в розряду точності (середня точність). При цьому нормована освітленість на робочому місці (Ен) рівна 200 лк. Джерелом природного освітлення є сонячне світло. У приміщенні, де розташовані ЕОМ передбачається природне бічне освітлення, рівень якого відповідає ДБН В.2.5-28:2018 Природне і штучне освітлення [10]

Регулярно повинен проводитися контроль освітленості, який підтверджує, що рівень освітленості задовольняє [10]

Розрахунок освітлення.

Для виробничих та адміністративних приміщень світловий коефіцієнт приймається не менше -1/8, в побутових – 1/10:

$$S_b = \left(\frac{1}{5} \div \frac{1}{10}\right) \cdot S_n , \quad (4.1)$$

де Sb – площа віконних прорізів, м2;

Sn – площа підлоги, м2.

Sn = a × b = 4 × 5 = 20 м2 ,

S = 1/8 × 20 = 2.5м2 .

Приймаємо 1 вікно площею S = 2.5м2. Світильники загального освітлення розташовуються над робочими поверхнями в рівномірно-прямокутному порядку. Для організації освітлення в темний час доби передбачається обладнати приміщення, довжина якого складає 4м, ширина 4м, світильниками ЛПО2П, оснащеними лампами типу ЛБ (дві по 80Вт) з світловим потоком 5400лм кожна. Розрахунок штучного освітлення виробляється по коефіцієнтах використання світлового потоку, яким визначається потік, необхідний для створення заданої освітленості при загальному рівномірному освітленні.

Розрахунок кількості світильників n визначається по формулі (4.2):

$$n = \frac{E \times S \times Z \times K}{F \times U \times M}, \quad (4.2)$$

Де E нормована освітленість робочої поверхні, визначається нормами – 300лк;

S – освітлювана площа, м2; S = 20 м2;

Z – поправочний коефіцієнт світильника (Z = 1.15 для ламп розжарювання та ДРЛ;

Z = 1,1 для люмінесцентних ламп) приймаємо рівним 1.1;

K – коефіцієнт запасу, що враховує зниження освітленості в процесі експлуатації – 1.5;

U – коефіцієнт використання, залежний від типу світильника, показника індексу приміщення і т.п. – 0.575;

M – число люмінесцентних ламп в світильнику – 2;

F – світловий потік лампи – 5400лм (для ЛБ-80). Підставивши числові значення у формулу (4.2), отримуємо:

$$n = \frac{300 \times 20 \times 1.1 \times 1.5}{5400 \times 0.575 \times 2} \approx 1.594 \quad (4.3)$$

Приймаємо освітлювальну установку, яка складається з 2-х світильників, які складаються з двох люмінесцентних ламп загальною потужністю 160Вт, напругою 220В.

4.4.3 Вентилювання

Здійснюється провітрювання приміщення, в залежності від погодних умов, тривалість повинна бути не менше 10 хв. Найкращий обмін повітря здійснюється при наскрізному провітрюванні.

4.4.4 Розрахунок захисного заземлення (забезпечення електробезпеки будівлі)

Загальний опір захисного заземлення визначається за формулою (4.4):

$$R_{ззп} = \frac{R_з \cdot R_n}{R_n \cdot n \cdot \eta_з + R_з \cdot \eta_n}, \quad (4.4)$$

Де Rз – опір заземлення, якими когут бать труби, опори, кути і т.п., Ом;

Rn – опір опори, яке з'єднує заземлювачі, Ом;

n – кількість заземлювачів;

ηз – коефіцієнт екранування заземлювача; приймається в межах 0,2 ÷ 0,9; ηз = 0,7

ηn – коефіцієнт екранування сполучної стійки; приймається в межах 0,1 ÷ 0,7; ηn = 0,5;

Опір заземлення визначається за формулою (4.5):

$$R_з = \frac{\rho}{2\pi \cdot l} \cdot \left(\ln\frac{2 \cdot l}{d} + \frac{1}{2}\ln\frac{4 \cdot t + l}{4 \cdot t - l}\right), \quad (4.5)$$

Де ρ – питомий опір ґрунту, залежить від типу ґрунту, Ом·м;

для піску – 400 ÷ 700 Ом·м; приймаємо ρ = 400 Ом·м;

l – довжина заземлювача, м; для труб - 2-3 м; l = 2 м;

d – діаметр заземлювача, м; для труб - 0,03-0,05 м; d = 0,03 м;

t – відстань від середини забитого в ґрунт заземлювача до рівня землі, м; t = 2 м.

Підставивши числові значення у формулу (4.5), отримуємо:

$$R_з = \frac{400}{2 \cdot 3{,}14 \cdot 2} \cdot \left(\ln\frac{2 \cdot 2}{0{,}03} + \frac{1}{2}\ln\frac{4 \cdot 2 + 2}{4 \cdot 2 - 2}\right) \approx 215\text{ , Ом} \quad (4.6)$$

Опір смуги, що з'єднує заземлювачі, визначається за формулою (4.7):

$$R_n = \frac{\rho}{2\pi \cdot L} \cdot \ln\frac{2 \cdot L^2}{b \cdot t^1}, \quad (4.7)$$

Де L – довжина смуги, що з'єднує заземлювачі (м) і приблизно дорівнює периметру будівлі: $П_{буд.}$ = 42·2 + 38·2 = 160 м; L = 160 м;

b – ширина смуги, м; b = 0,03 м;

t1 – глибина заземлення від рівня землі, м; t1 = 0,5 м.

Підставивши числові значення у формулу (4.7), отримуємо:

$$R_n = \frac{400}{2 \cdot 3{,}14 \cdot 160} \cdot \ln \frac{2 \cdot 160^2}{0{,}03 \cdot 0{,}5^1} = 5{,}99\text{ , Ом} \quad (4.8)$$

Кількість заземлювачів захисного заземлення визначається за формулою (4,9):

$$n = \frac{2 \cdot R_з}{4 \cdot \eta_з}, \quad (4.9)$$

Де 4 - допустимий загальний опір, Ом;

2 - коефіцієнт сезонності.

Визначаємо загальний опір захисного заземлення підставивши числові значення у формулу (4.4):

$$R_{ззп} = \frac{215 \cdot 5{,}99}{5{,}99 \cdot 154 \cdot 0{,}7 + 215 \cdot 0{,}5} = 1{,}7\text{ , Ом} \quad (4.10)$$

Висновок: дане захисне заземлення буде забезпечувати електробезпеку будівлі, так як виконується умова: Rззп < 4 Ом.

3) При виникненню пожеж при роботі на ПЕОМ від таких можливими джерел запалювання як:

– іскри і дуги коротких замикань;

– перегрів провідників, резисторів та інших радіодеталей ПЕОМ, від тривалої перевантаження та наявність перехідного опору;

– іскри при розмиканні і розмиканні ланцюгів;

– розряди статичної електрики;

– необережному поводженню з вогнем, а також вибухи газо-повітряних і паро-повітряних сумішей.

4.5 Екологія

Діяльність за темою магістерської роботи, а саме: оптимізація запитів до бази даних в про-цесі її виконання впливає на навколишнє природнє середовище і регламентується нормами дію-чого законодавства: Законом України “Про охорону навколишнього природного середовища” [11], Законом України “Про забезпечення санітарного та епідемічного благополуччя населення” [12], Законом України “Про відходи” [13].

В процесі діяльності з виконанням дипломного проектуванням виникають процеси поводження з відходами ІТ галузі. Нижче надано перелік відходів, що утворюються в процесі роботи:

- Відпрацьовані люмінесцентні лампи - I клас небезпеки.
- Змінні носії інформації - IV клас небезпеки.
- Відпрацьовані вогнегасники - IV клас небезпеки.
- Макулатура - IV клас небезпеки.

Висновок до розділу 4

У розділі “Охорона праці” виконаний аналіз потенційних небезпек при роботі із засобами обчислювальної техніки. Приведені рекомендації щодо організації робочого місця, електробезпеки та пожежної безпеки. Наведені розміри приміщення та наведено значення температури, вологості й рухливості повітря, необхідна кількість і потужність ламп та інші параметри, значення яких впливає на умови праці, рекомендації з охорони праці, техніки безпеки при роботі на комп’ютері.

Перелік джерел посилань до розділу 4

1.Закон України “Про охорону праці”. Вводиться в дію Постановою ВР № 2695-XII від 14.10.92, ВВР, 1992, № 49, ст.669. - Режим доступу: www. URL: https://zakon.rada.gov.ua/laws/show/2694-12

2.Кодекс законів про працю України. Затверджується Законом № 322-VIII від 10.12.71 ВВР, 1971. Режим доступу: www. URL: https://zakon.rada.gov.ua/laws/show/322-08

3.Закон України “Про загальнообов'язкове державне соціальне страхування від нещасного випадку на виробництві та професійного захворювання, які спричинили втрату працездатності”. Наказ від 21 грудня 2000 року N 2180-III. Режим доступу: www. URL: https://dnaop.com/html/2065/doc-zakon-ukrajini-pro-zagalynoobovjazkove-derzhavne-socialyne-strahuvannya-vid-neshhasnogo-vipadku-na-virobnictvi-ta-profesijnogo-z

4.Про затвердження Типового положення про порядок проведення навчання і перевірки знань з питань охорони праці НПАОП 0.00-4.12-05. Наказ від 26.01.2005 №15. Режим доступу: www. URL: https://zakon.rada.gov.ua/laws/show/z0231-05

5.Державні санітарні правила і норми роботи з візуальними дисплейними терміналами електронно-обчислювальних машин ДСанПІН 3.3.2.007-98. Затверджено Постановою Головного державного санітарного лікаря України 10 грудня 1998 р. N 7. Режим доступу: www. URL: https://zakon.rada.gov.ua/rada/show/v0007282-98

6.Санітарні норми мікроклімату виробничих приміщень ДСН 3.3.6.042-99. Постанова N 42 від 01.12.99. Режим доступу: www. URL: https://zakon.rada.gov.ua/rada/show/va042282-99

7.НПАОП 0.00-7.15-18 “Вимоги щодо безпеки та захисту здоров’я працівників під час роботи з екранними пристроями”. Зареєстровано в

Міністерстві юстиції України 25 квітня 2018 р. за № 508/31960. Режим доступу: www. URL: https://zakon.rada.gov.ua/laws/show/z0508-18

8.Електробезпека в будівлях і спорудах. Вимоги до захисних заходів від ураження електричним струмом. Наказ від 1 липня 2016 року N 204. Режим доступу: www. URL: http://epicentre.co.ua/dstu/doc28522.html

9.ДСТУ Б В.1.1-36:2016 «Визначення категорій приміщень, будинків та зовнішніх установок за вибухопожежною та пожежною небезпекою». Наказ від 15.06.2016 №158. Режим доступу: www. URL: https://zakon.rada.gov.ua/rada/show/v0158858-16

10. ДБН В.2.5-28:2018 «Природне і штучне освітлення». Режим доступу: www. URL: http://www.minregion.gov.ua/wp-content/uploads/2018/12/V2528-1.pdf

11. Закон України “Про охорону навколишнього природного середовища”. Вводиться в дію Постановою ВР № 1268-XII від 26.06.91, ВВР, 1991, № 41, ст.547. Режим доступу: www. URL: https://zakon.rada.gov.ua/laws/show/1264-12

12. Закони України “Про охорону навколишнього природного середовища”. Вводиться в дію Постановою ВР № 4005-XII від 24.02.94, ВВР, 1994, № 27, ст.219. Режим доступу: www. URL: https://zakon.rada.gov.ua/laws/show/4004-12

13. Закон України “Про відходи”. Відомості Верховної Ради України (ВВР), 1998, № 36-37, ст.242. Режим доступу: www. URL: https://zakon.rada.gov.ua/laws/show/187/98-вр

ВИСНОВКИ ДО РОБОТИ

Мої результати показали, що Knn перевершив дерева прийняття рішень, відповідно до відомих плюсів і мінусів двох розглянутих алгоритмів.

KNN:

+ Легко зрозуміти

+ Швидкий на тренуванні

+ підвищена точність

- За умови прокляття розмірності, якщо не застосовується PCA

- Багато обчислювальної потужності, необхідної для прогнозування результатів

Дерева рішень:

+ Легко пояснити

+ Не вимагає нормалізації даних

+ Може приймати як категоріальні, так і безперервні предикторські змінні

- Набагато нижчий рівень точності

Порівняно з результатами на веб-сторінці набору даних MNIST, вони цілком пристойні. Похибка для k-nn нижча, ніж помилки, перераховані для запусків без попередньої обробки, і хоча результати з коефіцієнтом помилок до 0,52% були досягнуті за допомогою k-nn, досягнута тут кількість помилок залишається досить хорошою. Що стосується випадкового лісу, то на веб-сторінці MNIST його немає. Однак є приклади посилених пнів, що є ще одним методом, заснованим на деревах рішень з використанням 1-рівневих дерев та методом посилення. Оскільки посилення є покращенням порівняно із випадковим лісом, очікується, що досягнуті показники помилок становитимуть покращення на 4,22%. За винятком простих підсилених пнів, це було так, з найнижчим рівнем помилок, досягнутим 0,87%. Цікаво, що це вище, ніж найнижчий зафіксований коефіцієнт помилок k-nn.

Як вже згадувалося раніше в розділах 3 та 4, існують інші способи порівняння цих методів, окрім простого вимірювання точності та рівня помилок. Оцінка AUC була проілюстрована протягом всієї роботи як ще один метод з'ясування того, наскільки гарними були прогнози алгоритмів. Однак не лише питання доброго прогнозування робить метод бажаним. Як обговорювалося в розділі 4, можна врахувати ефективність алгоритму, тобто наскільки ефективно використовується доступна обчислювальна потужність і наскільки трудомістким є використання алгоритму. Хоча класифікаційна помилка є властивістю, яку слід враховувати, ефективність алгоритму також може бути врахована. Якщо два методи показують однакові показники помилок, і один з них значно більш трудомісткий, ніж інший, швидший метод буде кращим.

Ця робота показала два можливі підходи до класифікаційних проблем та обговорила різні методи оцінки придатності алгоритмів. Хоча результати були багатообіцяючими з точки зору точності, існувало кілька обмежень. Дані були лише попередньо оброблені з точки зору масштабування та центрування. На веб-сайті набору даних MNIST, здається, більше попередньої обробки, як правило, призводило до кращого рівня помилок. Крім того, особливо у випадку алгоритму k-найближчих сусідів, були обмеження в обчислювальній потужності та часі, оскільки досліджувались лише k's від 2 до 10, і можливо, що кращий результат точності був би досягнутий при більш високому значенні k. Тоді порівняння між двома методами базувалось виключно на вимірі точності та на одному тестовому запуску на одному наборі даних тестування. Для кращого вивчення точності та продуктивності алгоритмів було б бажано провести більше тестів на більшій кількості наборів даних, щоб побачити, як вони будуть працювати на практиці.

На закінчення, на основі представленого порівняння методу k-найближчих сусідів та методу випадкового лісу, алгоритм k-nn з найбільш підходящим параметром мав нижчий рівень помилок, ніж алгоритм

випадкового лісу з найбільш підходящим параметром. У вакуумі цієї роботи метод k-nn показав кращі результати щодо точності, однак, щоб вважати ці результати остаточними, потрібно буде провести подальші тести.

ДОДАТОК А ПРЕЗЕНТАЦІЯ

МАГІСТЕРСЬКА РОБОТА

На тему: «Аналіз ефективності методів машинного навчання для систем розпізнавання письмових символів»

Виконав студент гр. КІ 19 дм
Доброжан З.Т.
Науковий керівник роботи
Скарга-Бандурова І.С.

Мета роботи:

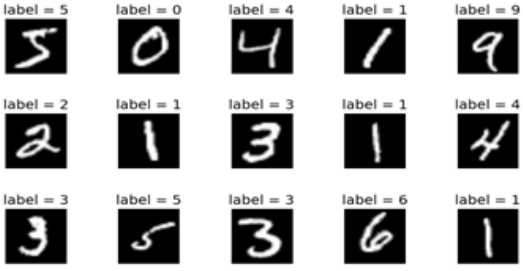
- є оцінка алгоритмів машинного навчання, що використовуються для класифікації записаних цифр із набору даних MNIST. Використаними алгоритмами є k-найближчі сусіди та випадковий ліс, а методом оцінки є точність. Перед запуском алгоритмів дані зменшуються за допомогою аналізу головних компонентів.

Машинне навчання

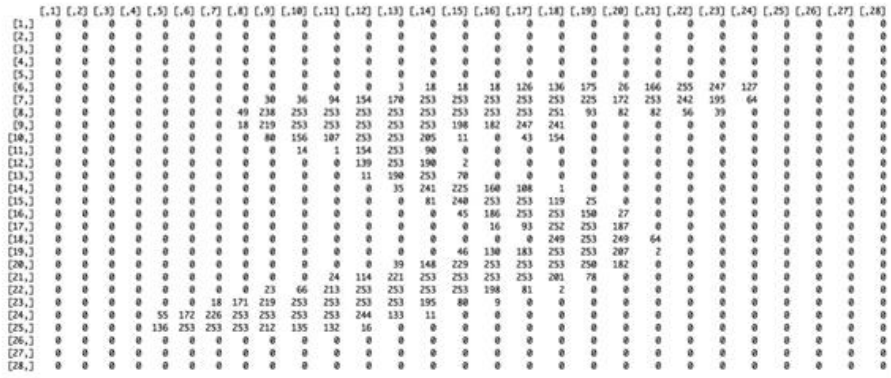
- це підгалузь штучного інтелекту в галузі інформатики, яка часто застосовує статистичні прийоми для надання комп'ютерам здатності «навчатися» з даних, без того, щоби бути програмованими явно.
- Одною з основних задач машинного навчання є **класифікація.** Вона передбачає собою визначення категорій та розподілу об'єктів згідно із заданими ознаками.

Набір даних Multi-NIST

- Це база даних, яка містить зразки рукописних варіантів написання чисел від 0 до 9, для калібрування і зіставлення методів розпізнавання зображень за допомогою машинного навчання.
- Набір даних представляє собою 60000 зображень для навчання і 10000 для тестування.



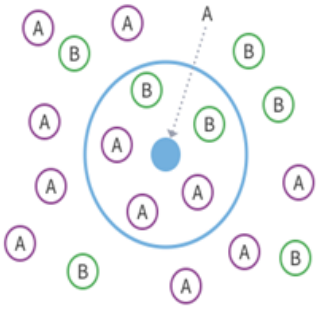
Кожна цифра набору представлена цілочисельною матрицею розміром 28*28, значення кожної комірки матриці можуть приймати значення від 0 до 255



Приклад цифри «5»

Метод найближчих сусідів

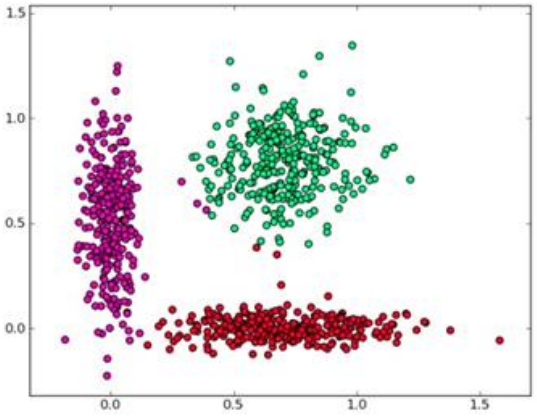
- метричний класифікатор, який базується на оцінюванні подібності об'єктів. Оцінюваний об'єкт відноситься до того класу, якому належать сусідні до нього об'єкти в навчальній виборці.



Метод k-найближчих сусідів для підвищення своєї надійності, відносить об'єкт до того класу, до якого належать більшість його сусідів з навчальної вибірки. Для виключення ситуацій, коли однакова кількість сусідів належать різним класам, беруть непарне число сусідів.

Кластерний аналіз

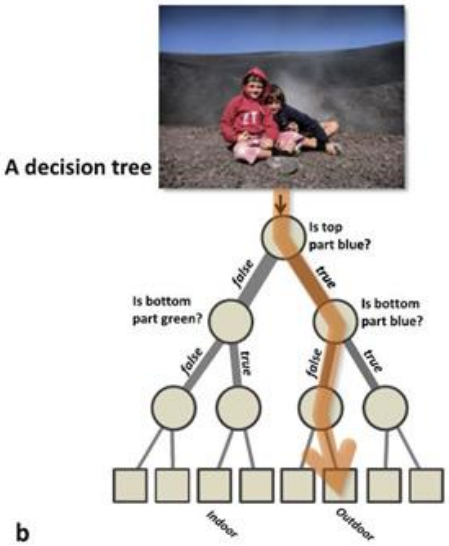
- Основна мета кластерного аналізу – знаходження груп схожих об'єктів у вибірці. Такі групи звуться кластерами. При істотній або абсолютній схожості об'єктів в кластері, вони повинні істотно відрізнятися від об'єктів в інших кластерах.



Дерево ухвалення рішень

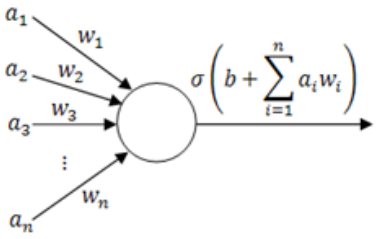
- Ухвалення рішення - це процес раціонального або ірраціонального вибору альтернатив, що має на меті досягнення усвідомлюваного результату. Один з методів автоматичного аналізу даних є дерево рішень.
- Дерево ухвалення рішень містить такі елементи як «листок» та «гілки». Гілки поєднують між собою листки та мають атрибути, відповіді, які розрізняють випадки. В листках записано значення цільової функції. Щоб класифікувати новий випадок, треба поступово, відповідно до атрибутів, спуститися до листа і видати відповідне до нього значення.
- Використовувані в Data Mining дерева рішень, бувають двох основних типів: регресивний аналіз дерева та аналіз дерева класифікації.

Приклад генерації рішення



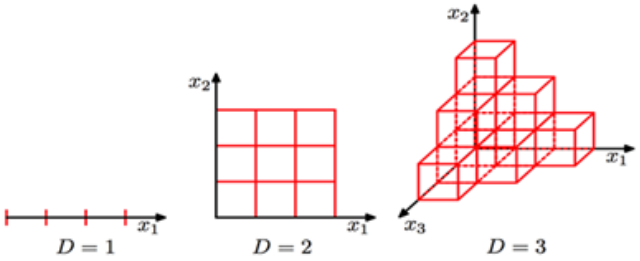
Штучні нейронні мережі

- системою взаємодіючих і з'єднаних між собою штучних нейронів, які отримують вхід, змінюють свій внутрішній стан, відповідно до цього входу і виробляють вихід. Вихід формується відповідно до входу та збудження, тобто зміни внутрішнього стану.



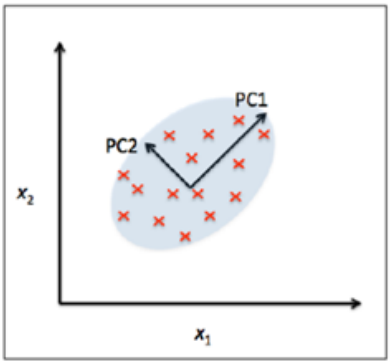
Задача зменшення розмірності

- це перетворення даних із високомірного простору в низьковимірний простір, так що другий варіант зберігає деякі основні властивості вихідних даних. Такі перетворення часто використовують в тих галузях, де використовують велику кількість даних або спостережень, таких як розпізнавання мови, обробка сигналів, біоінформатика та нейроінформатика.



Метод головних компонент

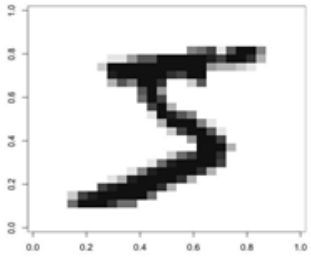
- це основна лінійна техніка зменшення розмірності, який здійснює лінійне відображення даних в менш вимірний простір таким чином, що максимізується дисперсія даних у маловимірному представленні. Будується матриця коваріації або кореляції даних, і обчислюються власні вектори цієї матриці.



Реалізація методу головних компонент

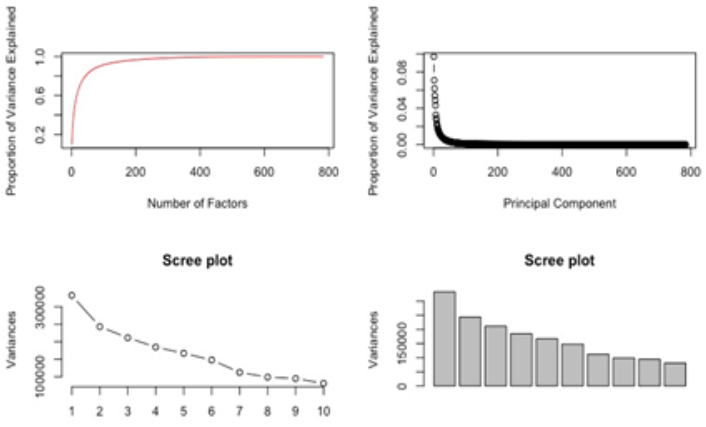
- Підготовка набору даних

Початковий вигляд першої цифри з набору даних



Щоб перевести всі значення числових стовпців у стандартний масштаб в наборі даних Було використано нормалізацію.

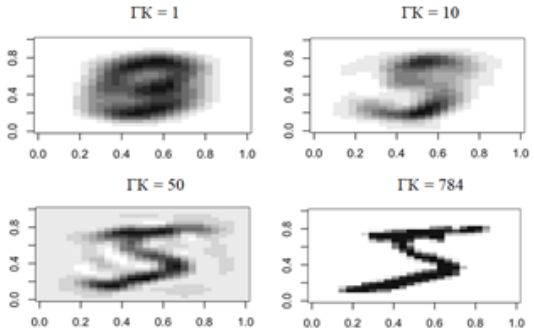
Графіки даних



Характеристика перших 5 компонент

	PC1	PC2	PC3	PC4	PC5
Standard deviation	576.82291	493.23822	459.89930	429.85624	408.56680
Proportion of Variance	0.09705	0.07096	0.06169	0.05389	0.04869
Cumulative Proportion	0.09705	0.16801	0.22970	0.28359	0.33228

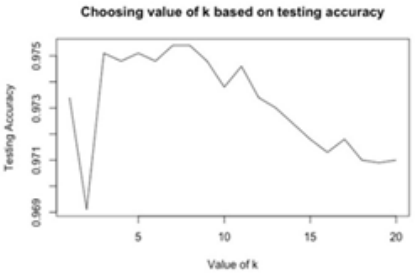
Реконструкція цифри 5



Використання методу k-найближчих сусідів

Запустили KNN з k від 1 до 20 та побудували точність для різних k (1:20):

```
> t(knn_acc)
      [,1]   [,2]   [,3]   [,4]   [,5]   [,6]
[1,] 1.0000 2.0000 3.0000 4.0000 5.0000 6.0000
[2,] 0.9734 0.9691 0.9751 0.9748 0.9751 0.9748
      [,7]   [,8]   [,9]   [,10]   [,11]
[1,] 7.0000 8.0000 9.0000 10.0000 11.0000
[2,] 0.9754 0.9754 0.9748  0.9738  0.9746
      [,12]  [,13]   [,14]   [,15]   [,16]
[1,] 12.0000 13.000 14.0000 15.0000 16.0000
[2,]  0.9734  0.973  0.9724  0.9718  0.9713
      [,17]  [,18]   [,19]  [,20]
[1,] 17.0000 18.000 19.0000 20.000
[2,]  0.9718  0.971  0.9709  0.971
```



Згідно графіку можна зрозуміти, що k = 8 дає найкращу точність

Обчислення точності для k = 8

```
> table_prediction
          test_labels
prediction    0    1    2    3    4    5    6    7    8    9
         0  971    0   10    0    0    3    5    0    4    4
         1    1 1129    2    1    4    0    5   23    0    5
         2    1    2  996    2    0    0    0    4    3    3
         3    0    1    1  976    0   10    0    0   13    5
         4    0    0    1    1  946    1    2    2    5    9
         5    1    0    0    9    0  865    1    0    7    6
         6    5    2    3    0    7    9  945    0    2    1
         7    1    0   13    7    0    1    0  983    3    6
         8    0    0    6    9    2    0    0    0  930    1
         9    0    1    0    5   23    3    0   16    7  969
> accuracy = hit/sum(table_prediction)
> accuracy
[1] 0.9743
```

Значення k = 8 з найкращою точністю вказує на те, що модель алгоритму машини KNN передбачить клас будь-якої вхідної цифри з **97,43%** точністю, використовуючи **8** найближчих сусідів.

Випадковий ліс

- Другим алгоритмом, за яким будуть класифіковані дані MNIST, є випадковий ліс. Цей алгоритм базується на деревах рішень, які функціонують як набір вузлів та ребер.

Прогнози були подані та отримали оцінку **94,66%**

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1    2    3    4    5    6    7    8    9
         0  961    0   12    1    1    4    8    1    7    2
         1    0 1122    0    0    1    1    2    6    0    5
         2    2    4  967    6    3    5    3   15   10    6
         3    0    1   11  954    0   14    0    1   15   10
         4    0    0    6    0  935    3    3    8    8   22
         5    2    2    1   17    3  846    7    0   17    8
         6    9    3    6    2   10    7  933    0    4    2
         7    1    0   10    6    2    1    0  971    8   10
         8    3    2   18   16    5    7    2    3  896    9
         9    2    1    1    8   22    4    0   23    9  935

Overall Statistics

               Accuracy : 0.952
                 95% CI : (0.9476, 0.9561)
    No Information Rate : 0.1135
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9466

 Mcnemar's Test P-Value : NA

Statistics by Class:
```

Випадковий ліс

- Другим алгоритмом, за яким будуть класифіковані дані MNIST, є випадковий ліс. Цей алгоритм базується на деревах рішень, які функціонують як набір вузлів та ребер.

Прогнози були подані та отримали оцінку **94,66%**

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1    2    3    4    5    6    7    8    9
         0  961    0   12    1    1    4    8    1    7    2
         1    0 1122    0    0    1    1    2    6    0    5
         2    2    4  967    6    3    5    3   15   10    6
         3    0    1   11  954    0   14    0    1   15   10
         4    0    0    6    0  935    3    3    8    8   22
         5    2    2    1   17    3  846    7    0   17    8
         6    9    3    6    2   10    7  933    0    4    2
         7    1    0   10    6    2    1    0  971    8   10
         8    3    2   18   16    5    7    2    3  896    9
         9    2    1    1    8   22    4    0   23    9  935

Overall Statistics

               Accuracy : 0.952
                 95% CI : (0.9476, 0.9561)
    No Information Rate : 0.1135
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9466

 Mcnemar's Test P-Value : NA

Statistics by Class:
```

ДОДАТОК Б ЛІСТИНГ

```
# Load dataset
     load_mnist <- function() {
       load_image_file <- function(filename) {
         ret = list()
         f = file(filename,'rb')
         readBin(f,'integer',n=1,size=4,endian='big')
         ret$n = readBin(f,'integer',n=1,size=4,endian='big')
         nrow = readBin(f,'integer',n=1,size=4,endian='big')
         ncol = readBin(f,'integer',n=1,size=4,endian='big')
         x                                                    =
     readBin(f,'integer',n=ret$n*nrow*ncol,size=1,signed=F)
         ret$x = matrix(x, ncol=nrow*ncol, byrow=T)
         close(f)
         ret
       }
       load_label_file <- function(filename) {
         f = file(filename,'rb')
         readBin(f,'integer',n=1,size=4,endian='big')
         n = readBin(f,'integer',n=1,size=4,endian='big')
         y = readBin(f,'integer',n=n,size=1,signed=F)
         close(f)
         y
       }
     ---------------------------------------------


library(dslabs)
mnist <- read_mnist()
train_images <- mnist$train$images
test_images <- mnist$test$images
train_labels <- mnist$train$labels
test_labels <- mnist$test$labels


# Inspect contents
#summary(train_images)
summary(train_labels)

# How the picture looks like
train_images[1,]
show_digit <- function(arr784, col=gray(12:1/12), ...) {
  image(matrix(arr784, nrow=28)[,28:1], col=col, ...)
}
show_digit(train_images[1,])


# Normalize the numeric data
train_normalized   <-   as.data.frame(scale(train_images,scale   =
FALSE,center = TRUE))
```

```
test_normalized   <-   as.data.frame(scale(test_images,scale   =
FALSE,center = TRUE))

# Calculate covariance matrix
#digits_covMatrix <- cov(train_normalized)


# PCA
set.seed(132)
# pca_train <- prcomp(digits_covMatrix)
pca_train <- prcomp(train_normalized)
#pca_train[["x"]]

# The percent of the variances in data
variance_explained                                              <-
as.data.frame(pca_train$sdev^2/sum(pca_train$sdev^2))
variance_explained              <-              cbind(c(1:784),
cumsum(variance_explained))
colnames(variance_explained) <- c("NmbrPCs","CumVar")
# Look at 50th variance
variance_explained[50, ]
# head(variance_explained,20)

#By Heuristics that the cumulative percentage of the explained
variance is more than 80%. We choose 50 PCs that with 50 PCs
contain 82.5% of the variances in data.



# Plot the data.
par(mfrow=c(2,2))
plot (variance_explained$NmbrPCs, variance_explained$CumVar,
      xlab = "Number of Factors", ylab = "Proportion of Variance
Explained",
      type = "l", col = "red")
plot  (pca_train$sdev^2/sum(pca_train$sdev^2),  xlab  =  "Principal
Component",
     ylab = "Proportion of Variance Explained", type = "b" )
plot(pca_train, type = "l", main = "Scree plot")
plot(pca_train, type = "barplot", main = "Scree plot")


# summary(pca_train)
summary(pca_train)$importance[,1:5]


#  Do  the  actual  dimension  reduction  (matrix  of  784  columns
should be converted to matrix of 50 columns)
pca_rot <- pca_train$rotation[,1:50]
train_final <- as.matrix(train_images)%*%(pca_rot)
test_final <- as.matrix(test_images)%*%(pca_rot)
test_final <- data.frame(test_final)
train_final <- data.frame(train_final)
```

```
head(train_final)


# Visualize the different classes in two dimensions
library(ggplot2)
ggplot(train_final,      aes(PC1,PC2,color=train_labels))      +
geom_point() +
  labs(x= "PC1",y= "PC2") + ggtitle("PCA")


# Inspect the reconstruction of the original data from these two
components
mean.train = colMeans(train_images)
PCA <- c(1,10,50,784)
par(mfrow=c(2,2))
i <- 1
while(i <= length(PCA))
{
  nComp = PCA[i]
  train.pca.re        =        pca_train$x[,1:nComp]        %*%
t(pca_train$rotation[,1:nComp])
  train.pca.re    =    scale(train.pca.re,    center    =    -
mean.train,scale=F)
  show_digit(train.pca.re[1,])
  i<-i+1
}

# So PCA = 50 is acceptable to see the images of the digits.
```

K-nn

```
# Run KNN with k from 1 to 20
library(class)
set.seed(132)
knn_acc <- matrix(nrow=20,ncol=2)
i=1
for (k in 1:20){
  prediction = knn(train_final,test_final,train_labels,k=k)
  knn_acc[i,] <- c(k,mean(prediction==test_labels))
  i = i+1
}
t(knn_acc)


# Plot the accuracy for different k (1:20)
par(mfrow=c(1,1))
plot(knn_acc[,1],knn_acc[,2],type = "l",
     xlab = "Value of k",
     ylab = "Testing Accuracy",
     main = "Choosing value of k based on testing accuracy")
```

```
# From the plot, choose the k = 8, which gives the best
accuracy.
best_prediction <- knn(train_final,test_final,train_labels,k=8)
test_labels_knn <- factor(test_labels)

# Confusion Matrix
library(caret)
confusionMatrix(best_prediction, test_labels_knn)


# Calculate the accuracy for k = 8
hit = 0
for (i in 1:10){
  hit = hit + table(best_prediction,test_labels)[i,i]
}
table_prediction <- table(prediction,test_labels)
table_prediction
accuracy = hit/sum(table_prediction)
accuracy
```

Random Forest

```
# Random Forest
library(randomForest)
library(readr)
library(lattice)
set.seed(132)
numTrain <- 40000
numTrees <- 50

# Generate a random sample of "numTrain" indexes
rows <- sample(1:nrow(train_final), numTrain)
train_labels <- factor(train_labels)
rf <- randomForest(train_final, train_labels, ntree=numTrees)
rf
plot(rf)

# Make prediction
pred <- predict(rf, test_final)
test_labels_rf <- factor(test_labels)

# Confusion Matrix
confusionMatrix(pred, test_labels_rf)
```

Boosting

```
# ### Boosting
# install.packages("C50")
# library(C50)
```

```
#
# # build model
# tree = C5.0(test_labels ~ ., data = mnist, trials=10)
#
# #make predictions
# table(predict(tree,newdata = mnist), train_labels)
```