МІНІСТЕРСТВО ОСВІТИ І НАУК УКРАЇНИ

ЗАХІДНОУКРАЇНСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

ФАКУЛЬТЕТ КОМП’ЮТЕРНИХ ІНФОРМАЦІНИХ ТЕХНОЛОГІЙ

Лабораторна робота №4

з дисципліни «Засоби аналізу даних на Python»

на тему:

Методи та алгоритми статистичної та попередньої обробки даних

Виконав:

Студент групи ІПЗ-22

Строгий Є. В.

Перевірив:

Тимчишин В. С.

Тернопіль 2024

**Мета роботи**: Надбання практичних навичок розв’язання задач регресії і класифікації, в лінійних моделях аналітичного аналізу даних за допомогою методів і засобів машинного навчання, реалізованих в бібліотеках Sklearn для ефективного розв’язання широкого кола задач в програмному середовищі Python.

Посилання на GitHub: <https://github.com/SheriffAltTab/PythonAnalysis-4>

**Завдання**

1. Написати програму для регресійного аналізу на основі даних засобами бібліотеки Sklearn.
2. Написати програму для класифікації об’єктів за характеристичними ознаками на основі даних засобами бібліотеки Sklearn і Pandas.

Для 1 завдання я обрав датасет Amazon Books Reviews

**Лістинг файлу Lab4.1:**In [66]:

from sklearn.datasets import fetch\_california\_housing  
  
# Завантаження даних  
data = fetch\_california\_housing()  
  
# Отримання характеристик та цільової змінної  
X = data.data  
y = data.target

In [67]:

from sklearn.model\_selection import train\_test\_split  
  
# Розділення даних на 80% для навчання та 20% для тестування  
X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

In [68]:

from sklearn.linear\_model import LinearRegression  
  
# Створення моделі лінійної регресії  
model = LinearRegression()  
  
# Навчання моделі на тренувальному наборі даних  
model.fit(X\_train, y\_train)

In [69]:

from sklearn.metrics import mean\_squared\_error  
  
# Прогнозування цін на будинки для тестового набору даних  
y\_pred = model.predict(X\_test)  
  
# Оцінка точності моделі за допомогою середньоквадратичної помилки (MSE)  
mse = mean\_squared\_error(y\_test, y\_pred)  
print("Середньоквадратична помилка:", mse)

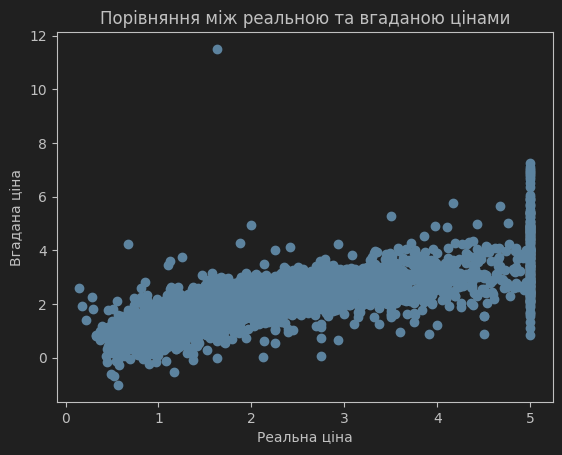
Out:

Середньоквадратична помилка: 0.5558915986952426

In [71]:

import matplotlib.pyplot as plt  
  
# Порівняння прогнозованих та реальних цін на будинки  
plt.scatter(y\_test, y\_pred)  
plt.xlabel("Реальна ціна")  
plt.ylabel("Прогнозована ціна")  
plt.title("Порівняння між реальною та прогнозованою цінами")  
plt.show()

Out:



**Лістинг файлу Lab4.2:**In [109]:

from sklearn.datasets import load\_breast\_cancer  
  
# Завантаження даних  
data = load\_breast\_cancer()  
  
# Отримання характеристик та цільової змінної  
X = data.data  
y = data.target

In [110]:

from sklearn.model\_selection import train\_test\_split  
  
# Розділення даних на 80% для навчання та 20% для тестування  
X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

In [111]:

from sklearn.neighbors import KNeighborsClassifier  
  
# Створення моделі K-найближчих сусідів (KNN)  
model = KNeighborsClassifier(n\_neighbors=5)  
  
# Навчання моделі на тренувальному наборі даних  
model.fit(X\_train, y\_train)

In [112]:

from sklearn.metrics import accuracy\_score  
  
# Прогнозування класів пухлини для тестового набору даних  
y\_pred = model.predict(X\_test)  
  
# Оцінка точності моделі за допомогою точності  
accuracy = accuracy\_score(y\_test, y\_pred)  
print("Точність моделі:", accuracy)

In [113]:

import matplotlib.pyplot as plt  
  
# Порівняння прогнозом та реальністю  
plt.scatter(y\_test, y\_pred)  
plt.xlabel("Реальність")  
plt.ylabel("Прогноз")  
plt.title("Порівняння між реальністю та прогнозом")  
plt.show()

Out:

