МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»

Факультет інформатики та обчислювальної техніки

Кафедра інформатики та програмної інженерії

Практикум № 5

з курсу «Аналіз даних в інформаційних системах»

на тему: «Регресійні моделі»

|  |  |
| --- | --- |
| Викладач:  Ліхоузова Т. А. | Виконав:  студент 2 курсу  групи ІП-11 ФІОТ  Головня О. Р. |

Київ-2023

**Основне завдання** Приклад виконання

Завантажити [дані про якість червоного вина](https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009)

1. Дослідити дані, підготувати їх для побудови регресійної моделі
2. Розділити дані на навчальну та тестову вибірки
3. Побудувати декілька регресійних моделей для прогнозу якості вина (12 - quality). Використати лінійну одномірну та багатомірну регресію та поліноміальну регресію обраного вами виду (3-5 моделей)
4. Використовуючи тестову вибірку, з'ясувати яка з моделей краща

Для виконання лабораторної роботи було використано Python.

Код програми:

import pandas as pd

import numpy as np

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LinearRegression

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.preprocessing import PolynomialFeatures

from sklearn.pipeline import make\_pipeline

from sklearn.metrics import r2\_score

from sklearn.metrics import mean\_squared\_error

# Зчитуємо CSV-файл

df = pd.read\_csv('winequality-red.csv',sep=',',encoding='utf-8')

#print(df.columns)

#print(df.head(5))

#print(df.isnull().sum())

sns.countplot(x='quality', data=df)

plt.show()

# підготовка даних для лінійної регресії

X = df.iloc[:, 11].values.reshape(-1, 1) # quality

y = df.iloc[:, 10].values.reshape(-1, 1) # alcohol

# розділення даних на навчальну та тестову вибірки

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=0)

# побудова та навчання моделі лінійної регресії

regressor = LinearRegression()

regressor.fit(X\_train, y\_train)

# прогнозування значень на тестовій вибірці

y\_pred = regressor.predict(X\_test)

# оцінка якості моделі

print('Linear Regression R-squared:', r2\_score(y\_test, y\_pred))

# підготовка даних для багатомірної регресії

X = df.iloc[:, :-1].values # всі ознаки, крім quality

y = df.iloc[:, 11].values.reshape(-1, 1) # quality

# розділення даних на навчальну та тестову вибірки

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=0)

# побудова та навчання моделі багатомірної регресії

regressor = LinearRegression()

regressor.fit(X\_train, y\_train)

# прогнозування значень на тестовій вибірці

y\_pred = regressor.predict(X\_test)

# оцінка якості моделі

print('Multivariate Linear Regression R-squared:', r2\_score(y\_test, y\_pred))

# підготовка даних для поліноміальної регресії

X = df.iloc[:, 10].values.reshape(-1, 1) # alcohol

y = df.iloc[:, 11].values.reshape(-1, 1) # quality

# розділення даних на навчальну та тестову вибірки

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=0)

# побудова та навчання моделі поліноміальної регресії

model = make\_pipeline(PolynomialFeatures(4), LinearRegression())

model.fit(X\_train, y\_train)

# прогнозування значень на тестовій вибірці

y\_pred = model.predict(X\_test)

# оцінка якості моделі

print('Polynomial Regression R-squared:', r2\_score(y\_test, y\_pred))

###########################################################################

#df1 = pd.read\_csv('Data4.csv',delimiter=";",encoding='ISO-8859-1',on\_bad\_lines='skip')

#df2 = pd.read\_csv('Data4t.csv',delimiter=";",encoding='ISO-8859-1',on\_bad\_lines='skip')

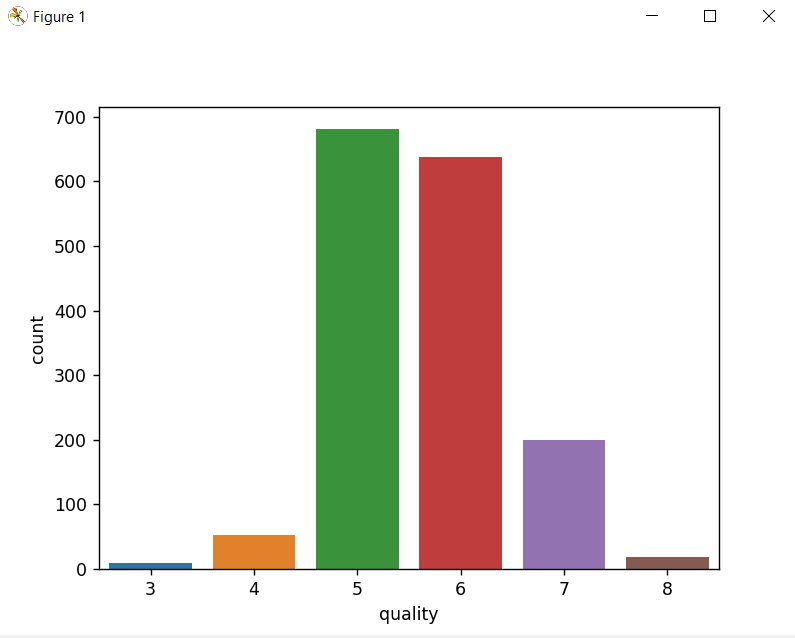
#print(df1.describe())

#print(df1.head(5))

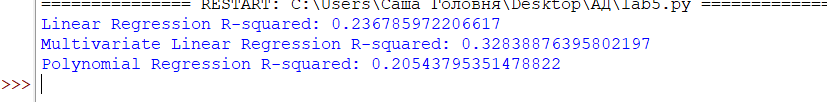
#corr\_matrix = df1.corr(numeric\_only=True)

#sns.heatmap(corr\_matrix, annot=True, cmap='coolwarm')

1. **Основне завдання:**



Розподіл по якості:

****Результати регресій:

(Чим більше число, тим краща модель, отже: багатомірна регресія краще)

**Висновок:**

Під час виконання лабораторної роботи я дослідив дані, підготував їх для побудови регресійної моделі, розділив дані на навчальну та тестову вибірки, побудував декілька регресійних моделей для прогнозу якості вина та з’ясував яка краща