ДЕРЖАВНИЙ УНІВЕРСИТЕТ «ЖИТОМИРСЬКА ПОЛІТЕХНІКА»

Факультет інформаційно-комп’ютерних технологій

Кафедра інженерії програмного забезпечення

Звіт

з лабораторної роботи № 1

з дисципліни

# «Системи штучного інтелекту»

Виконав студент групи:

ІПЗ-19-2

Федоренко Евеліна

Перевірив:

Пулеко Ігор Васильович

Житомир

2022

**Лабораторна робота № 1**

**ПОПЕРЕДНЯ ОБРОБКА ТА КОНТРОЛЬОВАНА КЛАСИФІКАЦІЯ**

**ДАНИХ**

***Мета:*** використовуючи спеціалізовані бібліотеки та мову програмування Python дослідити попередню обробку та класифікацію даних.

Github: <https://github.com/idontneedsleep/SII-LR-1>

ЗАВДАННЯ НА ЛАБОРАТОРНУ РОБОТУ ТА МЕТОДИЧНІ

РЕКОМЕНДАЦІЇ ДО ЙОГО ВИКОНАННЯ

**Завдання 2.1. Попередня обробка даних**

Як правило, при обробці ми маємо справу з великими обсягами необроблених вихідних даних. Алгоритми машинного навчання розраховані на те, що, перш ніж вони зможуть розпочати процес тренування, отримані дані будуть відформатовані певним чином. Щоб привести дані до форми, що прийнятна для алгоритмів машинного навчання, ми повинні попередньо підготувати їх і перетворити на потрібний формат.

2.1.1. Бінарізація

Цей процес застосовується в тих випадках, коли ми хочемо перетворити наші числові значення на булеві значення (0, 1). Скористаємося вбудованим методом для бінаризації вхідних даних, встановивши значення 2,1 як порогове.

2.1.2. Виключення середнього

Виключення середнього - методика попередньої обробки даних, що зазвичай використовується в машинному навчанні. Як правило, із векторів ознак (feature vectors) доцільно виключати середні значення, щоб кожна ознака (feature) центрувалася на нулі. Це робиться з метою, виключити з розгляду зміщення значень у векторах ознак.

2.1.3. Масштабування

У нашому векторі ознак кожне значення може змінюватись у деяких випадкових межах. Тому дуже важливо масштабувати ознаки, щоб вони були рівним ігровим полем для тренування алгоритму машинного навчання. Ми не хочемо, щоб будь-яка з ознак могла набувати штучно великого або малого значення лише через природу вимірів.

2.1.4. Нормалізація

Процес нормалізації полягає у зміні значень у векторі ознак таким чином, щоб для їх вимірювання можна було використовувати одну загальну шкалу. У машинному навчанні використовують різні форми нормалізації. У найбільш поширених з них, значення змінюються так, щоб їх сума дорівнювала 1. L1-нормалізація використовує метод найменших абсолютних відхилень (Least Absolute Deviations), що забезпечує рівність 1 суми абсолютних значень в кожному ряду. L2-нормалізація використовує метод найменших квадратів, що забезпечує рівність 1 суми квадратів значень. Взагалі, техніка L1-нормалізації вважається більш надійною по порівняно з L2-нормалізацією, оскільки вона менш чутлива до викидів.

Дуже часто дані містять викиди, і з цим нічого не вдієш. Ми хочемо використовувати безпечні методики, що дозволяють ігнорувати викиди у процесі обчислень. Якби ми вирішували завдання, в якому викиди грають важливу роль, то, ймовірно, найкращим вибором була б L2-нормалізація.

2.1.5. Кодування міток

Як правило, в процесі класифікації даних ми маємо справу з множиною міток (labels). Ними можуть бути слова, числа або інші об'єкти. Функції машинного навчання, що входять до бібліотеки sklearn, очікують, що мітки є числами. Тому, якщо мітки – це вже числа, ми можемо використовувати їх безпосередньо для того, щоб почати тренування. Однак, зазвичай, це не так. На практиці мітками служать слова, оскільки в такому вигляді вони краще всього сприймаються людиною. Ми позначаємо тренувальні дані словами, щоб полегшити відстеження відповідностей. Для перетворення слів у числа необхідно використовувати кодування. Під кодуванням міток (label encoding) мається на увазі процес перетворення словесних міток на числову форму. Завдяки цьому алгоритми можуть оперувати нашими даними.

Різниця між L1 та L2 в тому, що у L1-нормалізації використовується сума абсолютних значень вагових значень, а L2-нормалізації - сума квадратів вагових значень, через це L1-нормалізація іноді дає побічний ефект видалення непотрібних функцій (features), присвоюючи пов'язаним з ними ваги значення 0.0, але не працює без проблем з усіма формами навчання, при цьому L2-нормалізація працює з усіма формами навчання, але не забезпечує неявної селекції функцій.

Лістинг програми:

import numpy as np  
from sklearn import preprocessing  
  
input\_data = np.array([[5.1, -2.9, 3.3],  
 [-1.2, 7.8, -6.1],  
 [3.9, 0.4, 2.1],  
[7.3, -9.9, -4.5]])  
  
data\_binarized = preprocessing.Binarizer(threshold=2.1).transform(input\_data)  
print("\n Binarized data:\n", data\_binarized)  
  
print("\nBEFORE: ")  
print("Mean =", input\_data.mean(axis=0))  
print("Std deviation =", input\_data.std(axis=0))  
  
data\_scaled = preprocessing.scale(input\_data)  
print("\nAFTER: ")  
print("Mean =", data\_scaled.mean(axis=0))  
print("Std deviation =", data\_scaled.std(axis=0))  
  
data\_normalized\_l1 = preprocessing.normalize(input\_data, norm='l1')  
data\_normalized\_l2 = preprocessing.normalize(input\_data, norm='l2')  
print("\nl1 normalized data:\n", data\_normalized\_l1)  
print("\nl2 normalized data:\n", data\_normalized\_l2)  
  
input\_labels = ['red', 'blасk', 'red', 'green', 'blасk', 'yellow', 'white']  
  
encoder = preprocessing.LabelEncoder()  
encoder.fit(input\_labels)  
  
print("Label mapping:")  
for i, item in enumerate(encoder.classes\_):  
 print(item, '-->', i)  
  
test\_labels = ['green', 'red', 'blасk']  
encoded\_values = encoder.transform(test\_labels)  
print("\nLabels =", test\_labels)  
print("Encoded values =", list(encoded\_values))  
  
encoded\_values = [3, 0, 4, 1]  
decoded\_list = encoder.inverse\_transform(encoded\_values)  
print("\nEncoded values =", encoded\_values)  
print("Decoded labels =", list(decoded\_list))

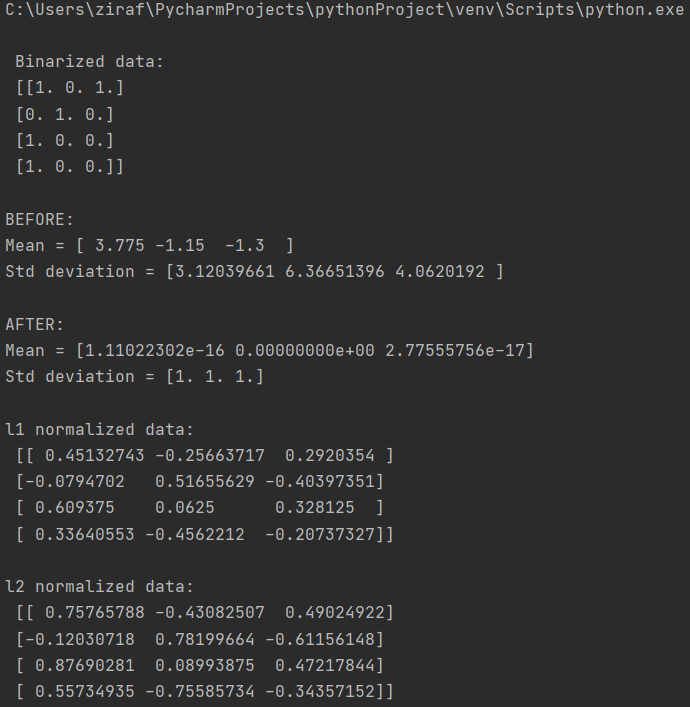


Рис. 1 Результат виконання програми

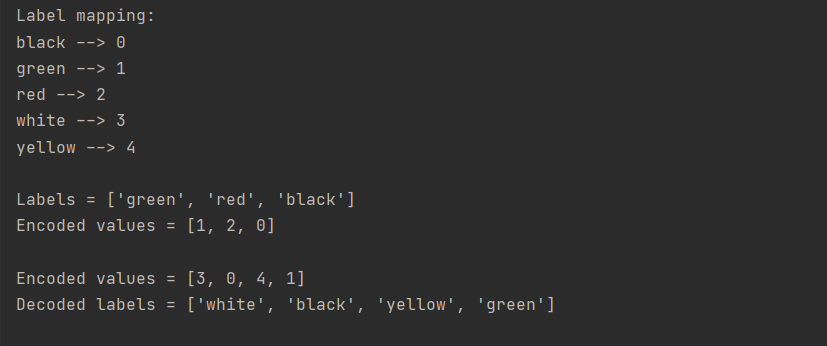


Рис. 2 Результат виконання програми

**Завдання 2.2. Попередня обробка нових даних**

У коді програми попереднього завдання поміняйте дані по рядках (значення змінної input\_data) на значення відповідно варіанту таблиці 1 та виконайте операції: Бінарізації, Виключення середнього, Масштабування, Нормалізації.

Варіант обирається відповідно номера за списком групи відповідно до

таблиці. Варіант 19

Значення змінної input\_data-4.1 -5.5 3.3 6.9 4.6 3.9 -4.2 3.8 2.3 3.9 3.4 -1.2

Поріг бінаризації 3.2

Лістинг програми:

import numpy as np  
from sklearn import preprocessing  
  
input\_data = np.array([[-4.1, -5.5, 3.3],  
 [6.9, 4.6, 3.9],  
 [-4.2, 3.8, 2.3],  
[3.9, 3.4, -1.2]])  
  
data\_binarized = preprocessing.Binarizer(threshold=3.2).transform(input\_data)  
print("\n Binarized data:\n", data\_binarized)  
  
print("\nBEFORE: ")  
print("Mean =", input\_data.mean(axis=0))  
print("Std deviation =", input\_data.std(axis=0))  
  
data\_scaled = preprocessing.scale(input\_data)  
print("\nAFTER: ")  
print("Mean =", data\_scaled.mean(axis=0))  
print("Std deviation =", data\_scaled.std(axis=0))  
  
data\_normalized\_l1 = preprocessing.normalize(input\_data, norm='l1')  
data\_normalized\_l2 = preprocessing.normalize(input\_data, norm='l2')  
print("\nl1 normalized data:\n", data\_normalized\_l1)  
print("\nl2 normalized data:\n", data\_normalized\_l2)

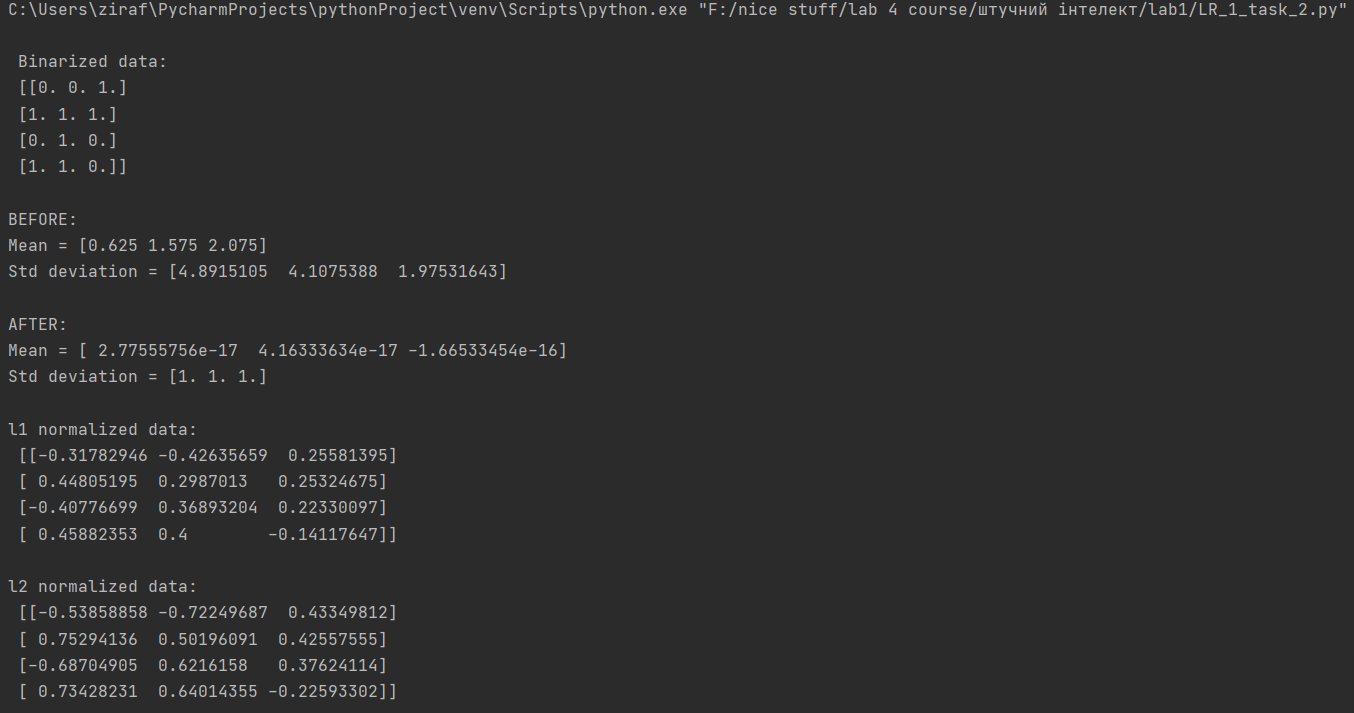


Рис. 3 Результат виконання програми

**Завдання 2.3. Класифікація логістичною регресією або логістичний**

**класифікатор**

Логістична регресія (logistic regression) - це методика, що використовується для пояснення відносин між вхідними та вихідними змінними. Вхідні змінні вважаються незалежними, вихідні – залежними. Залежна змінна може мати лише фіксований набір значень. Ці значення відповідають класам завдання класифікації. Метою є ідентифікація відносин між незалежними та залежними змінними за допомогою оцінки ймовірностей того, що та або інша залежна змінна відноситься до того чи іншого класу.

Лістинг програми:

import numpy as np  
from sklearn import linear\_model  
import matplotlib.pyplot as plt  
from utilities import visualize\_classifier  
  
X = np.array([[3.1, 7.2], [4, 6.7], [2.9, 8], [5.1, 4.5],  
 [6, 5], [5.6, 5], [3.3, 0.4],  
 [3.9, 0.9], [2.8, 1],  
 [0.5, 3.4], [1, 4], [0.6, 4.9]])  
y = np.array([0, 0, 0, 1, 1, 1, 2, 2, 2, 3, 3, 3])  
  
classifier = linear\_model.LogisticRegression(solver='liblinear', C=1)  
classifier.fit(X, y)  
  
visualize\_classifier(classifier, X, y)

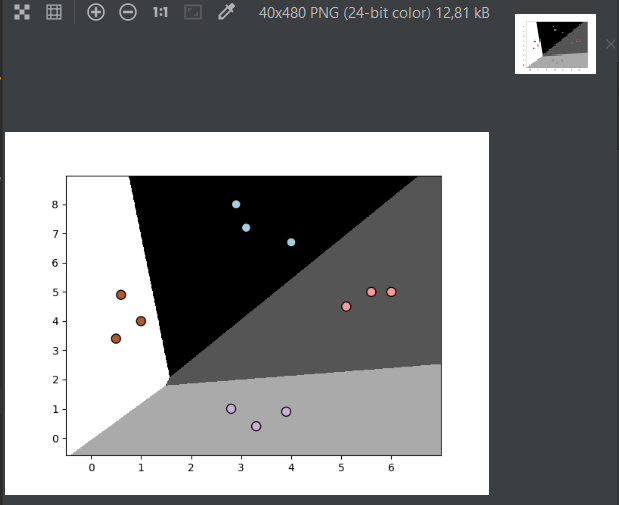


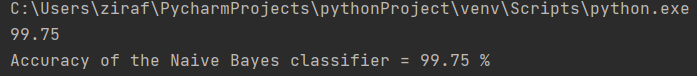
Рис. 4 Результат виконання програми

**Завдання 2.4. Класифікація наївним байєсовським класифікатором**

Наївний байєсовський :класифікатор (Na!ve Bayes classifier) - це простий класифікатор, заснований на використанні теореми Байєса, яка описує ймовірність події з урахуванням пов'язаних з нею умов. Такий класифікатор створюється за допомогою привласнення позначок класів екземплярам завдання. Останні представляються як векторів значень ознак. При цьому передбачається, що значення будь-якої заданої ознаки не залежить від значень інших ознак. Його припущення про незалежність ознак і становить наївну частину байєсовського класифікатора. Ми можемо оцінювати вплив будь-якої ознаки змінної класу незалежно від впливу інших ознак.

Лістинг програми:

import numpy as np  
import matplotlib.pyplot as plt  
from sklearn.naive\_bayes import GaussianNB  
from sklearn.model\_selection import train\_test\_split, cross\_val\_score  
from utilities import visualize\_classifier  
  
input\_file = 'data\_multivar\_nb.txt'  
data = np.loadtxt(input\_file, delimiter=',')  
X, y = data[:, :-1], data[:, -1]  
  
classifier = GaussianNB()  
classifier.fit(X, y)  
y\_pred = classifier.predict(X)  
  
accuracy = 100.0 \* (y == y\_pred).sum() / X.shape[0]  
print(accuracy)  
print("Accuracy of the Naive Bayes classifier =", round(accuracy, 2), "%")  
visualize\_classifier(classifier, X, y)



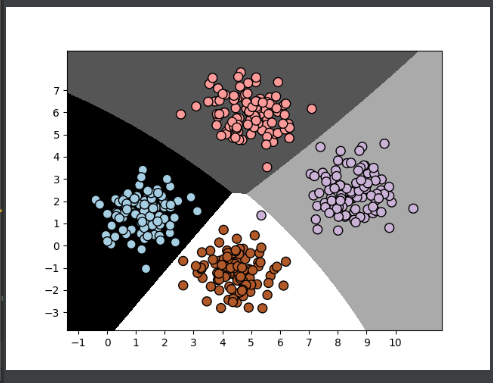


Рис. 5-6 Результат виконання програми

Лістинг програми:

import numpy as np  
import matplotlib.pyplot as plt  
from sklearn.naive\_bayes import GaussianNB  
from sklearn.model\_selection import train\_test\_split, cross\_val\_score  
from utilities import visualize\_classifier  
  
input\_file = 'data\_multivar\_nb.txt'  
data = np.loadtxt(input\_file, delimiter = ',')  
X, y = data[:, :-1], data[:, -1]  
X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size = 0.2, random\_state = 3)  
  
classifier = GaussianNB()  
classifier.fit(X\_train, y\_train)  
y\_test\_pred = classifier.predict(X\_test)  
  
accuracy = 100.0 \* (y\_test == y\_test\_pred).sum() / X\_test.shape[0]  
print(accuracy)  
print("Accuracy of the new Naive Bayes classifier =", round(accuracy, 2), "%")  
visualize\_classifier(classifier, X\_test, y\_test)  
  
num\_folds = 3  
accuracy\_values = cross\_val\_score(classifier, X, y, scoring = 'accuracy', cv = num\_folds)  
print("Accuracy: " + str(round(100 \* accuracy\_values.mean(), 2)) + "%")  
  
precision\_values = cross\_val\_score(classifier, X, y, scoring='precision\_weighted', cv=num\_folds)  
print("Precision: " + str(round(100 \* precision\_values.mean(), 2)) + "%")  
  
recall\_values = cross\_val\_score(classifier, X, y, scoring='recall\_weighted', cv=num\_folds)  
print("Recall: " + str(round(100 \* recall\_values.mean(), 2)) + "%")  
  
f1\_values = cross\_val\_score(classifier, X, y, scoring='f1\_weighted', cv=num\_folds)  
print("F1: " + str(round(100 \* f1\_values.mean(), 2)) + "%")

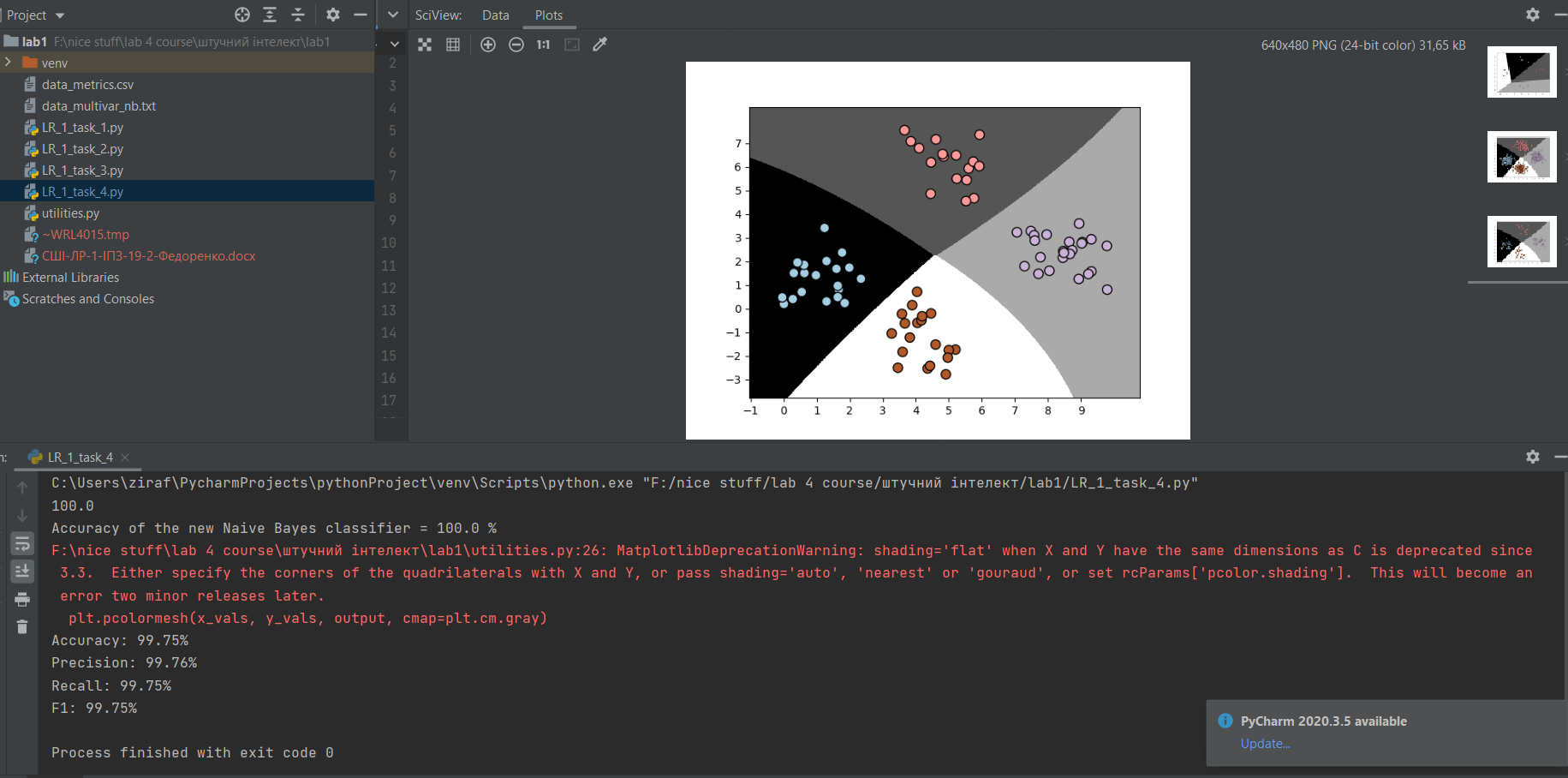


Рис. 7 Результат виконання програми

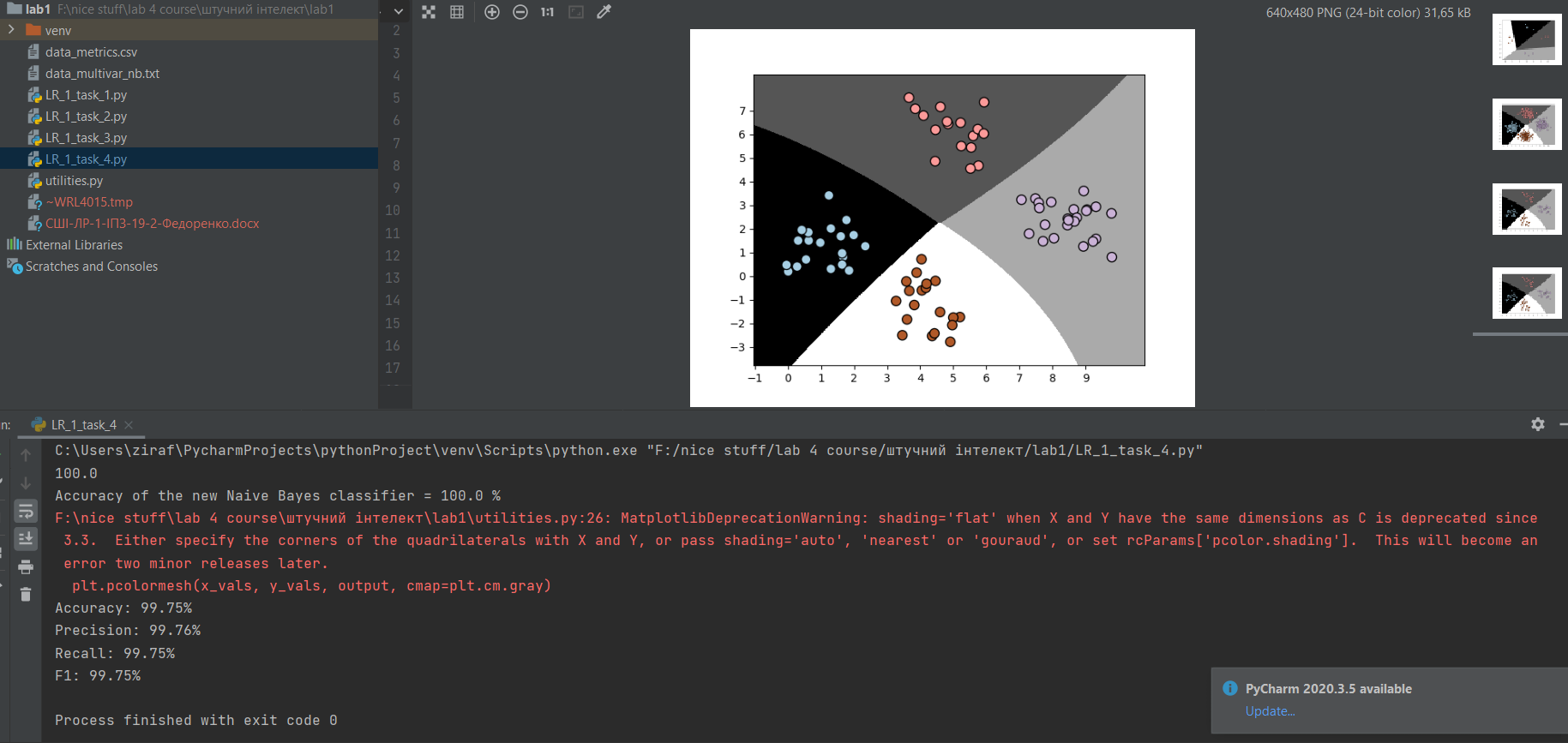


Рис. 8 Результат виконання програми (другий прогін)

Висновок: в першому результаті ми отримали accuracy 99.75%, після змін в другому запуску маємо 100%. Пояснюється це тим, що перший метод є не надійним, тому в другому виконуємо перехресну перевірку і розділяємо дані на тренувальні і тестувальні.

**Завдання 2.5. Вивчити метрики якості класифікації**

Лістинг програми:

import pandas as pd  
import numpy as np  
from sklearn.metrics import accuracy\_score, confusion\_matrix, recall\_score, precision\_score, f1\_score, roc\_curve, \  
 roc\_auc\_score  
import matplotlib.pyplot as plt  
  
df = pd.read\_csv('data\_metrics.csv')  
df.head()  
  
df['predicted\_RF'] = (df.model\_RF >= 0.5).astype('int')  
df['predicted\_LR'] = (df.model\_LR >= 0.5).astype('int')  
df.head()  
  
confusion\_matrix(df.actual\_label.values, df.predicted\_RF.values)  
  
  
def find\_TP(y\_true, y\_pred):  
 return sum((y\_true == 1) & (y\_pred == 1))  
  
  
def find\_FN(y\_true, y\_pred):  
 return sum((y\_true == 1) & (y\_pred == 0))  
  
  
def find\_FP(y\_true, y\_pred):  
 return sum((y\_true == 0) & (y\_pred == 1))  
  
  
def find\_TN(y\_true, y\_pred):  
 return sum((y\_true == 0) & (y\_pred == 0))  
  
  
print('TP:', find\_TP(df.actual\_label.values, df.predicted\_RF.values))  
print('FN:', find\_FN(df.actual\_label.values, df.predicted\_RF.values))  
print('FP:', find\_FP(df.actual\_label.values, df.predicted\_RF.values))  
print('TN:', find\_TN(df.actual\_label.values, df.predicted\_RF.values))  
  
  
def find\_conf\_matrix\_values(y\_true, y\_pred):  
 TP = find\_TP(y\_true, y\_pred)  
 FN = find\_FN(y\_true, y\_pred)  
 FP = find\_FP(y\_true, y\_pred)  
 TN = find\_TN(y\_true, y\_pred)  
 return TP, FN, FP, TN  
  
  
def fedorenko\_confusion\_matrix(y\_true, y\_pred):  
 TP, FN, FP, TN = find\_conf\_matrix\_values(y\_true, y\_pred)  
 return np.array([[TN, FP], [FN, TP]])  
  
  
fedorenko\_confusion\_matrix(df.actual\_label.values, df.predicted\_RF.values)  
  
assert np.array\_equal(fedorenko\_confusion\_matrix(df.actual\_label.values, df.predicted\_RF.values),  
 confusion\_matrix(df.actual\_label.values, df.predicted\_RF.values))  
  
  
def fedorenko\_accuracy\_score(y\_true, y\_pred):  
 TP, FN, FP, TN = find\_conf\_matrix\_values(y\_true, y\_pred)  
 return (TP + TN) / (TP + FP + TN + FN)  
  
  
assert fedorenko\_accuracy\_score(df.actual\_label.values, df.predicted\_RF.values) == accuracy\_score(df.actual\_label.values,  
 df.predicted\_RF.values), 'fedorenko\_accuracy\_score failed on RF'  
  
assert fedorenko\_accuracy\_score(df.actual\_label.values, df.predicted\_LR.values) == accuracy\_score(df.actual\_label.values,  
 df.predicted\_LR.values), 'fedorenko\_accuracy\_score failed on LR'  
  
print('Accuracy RF: %.3f' % (fedorenko\_accuracy\_score(df.actual\_label.values, df.predicted\_RF.values)))  
print('Accuracy LR: %.3f' % (fedorenko\_accuracy\_score(df.actual\_label.values, df.predicted\_LR.values)))  
  
accuracy\_score(df.actual\_label.values, df.predicted\_RF.values)  
recall\_score(df.actual\_label.values, df.predicted\_RF.values)  
  
  
def fedorenko\_recall\_score(y\_true, y\_pred):  
 TP, FN, FP, TN = find\_conf\_matrix\_values(y\_true, y\_pred)  
 return TP / (TP + FN)  
  
  
assert fedorenko\_recall\_score(df.actual\_label.values, df.predicted\_RF.values) == recall\_score(df.actual\_label.values,  
 df.predicted\_RF.values), 'fedorenko\_recall\_score failed on RF'  
  
assert fedorenko\_recall\_score(df.actual\_label.values, df.predicted\_LR.values) == recall\_score(df.actual\_label.values,  
 df.predicted\_LR.values), 'fedorenko\_recall\_score failed on LR'  
  
print('Recall RF: %.3f' % (fedorenko\_recall\_score(df.actual\_label.values, df.predicted\_RF.values)))  
print('Recall LR: %.3f' % (fedorenko\_recall\_score(df.actual\_label.values, df.predicted\_LR.values)))  
  
precision\_score(df.actual\_label.values, df.predicted\_RF.values)  
  
  
def fedorenko\_precision\_score(y\_true, y\_pred):  
 TP, FN, FP, TN = find\_conf\_matrix\_values(y\_true, y\_pred)  
 return TP / (TP + FP)  
  
  
assert fedorenko\_precision\_score(df.actual\_label.values, df.predicted\_RF.values) == precision\_score(  
 df.actual\_label.values, df.predicted\_RF.values), 'fedorenko\_precision\_score failed on RF'  
  
assert fedorenko\_precision\_score(df.actual\_label.values, df.predicted\_LR.values) == precision\_score(  
 df.actual\_label.values, df.predicted\_LR.values), 'fedorenko\_precision\_score failed on LR'  
  
print('Precision RF: %.3f' % (fedorenko\_precision\_score(df.actual\_label.values, df.predicted\_RF.values)))  
print('Precision LR: %.3f' % (fedorenko\_precision\_score(df.actual\_label.values, df.predicted\_LR.values)))  
  
f1\_score(df.actual\_label.values, df.predicted\_RF.values)  
  
  
def fedorenko\_f1\_score(y\_true, y\_pred):  
 TP, FN, FP, TN = find\_conf\_matrix\_values(y\_true, y\_pred)  
 recall = fedorenko\_recall\_score(y\_true, y\_pred)  
 precision = fedorenko\_precision\_score(y\_true, y\_pred)  
 return 2 \* (recall \* precision) / (recall + precision)  
  
  
assert fedorenko\_f1\_score(df.actual\_label.values, df.predicted\_RF.values) == f1\_score(df.actual\_label.values,  
 df.predicted\_RF.values), 'fedorenko\_f1\_score failed on RF'  
assert fedorenko\_f1\_score(df.actual\_label.values, df.predicted\_LR.values) == f1\_score(df.actual\_label.values,  
 df.predicted\_LR.values), 'fedorenko\_f1\_score failed on LR'  
  
print('F1 RF: %.3f' % (fedorenko\_f1\_score(df.actual\_label.values, df.predicted\_RF.values)))  
print('F1 LR: %.3f' % (fedorenko\_f1\_score(df.actual\_label.values, df.predicted\_LR.values)))  
  
print('\nscores with threshold = 0.5')  
print('Accuracy RF: %.3f' % (fedorenko\_accuracy\_score(df.actual\_label.values, df.predicted\_RF.values)))  
print('Recall RF: %.3f' % (fedorenko\_recall\_score(df.actual\_label.values, df.predicted\_RF.values)))  
print('Precision RF: %.3f' % (fedorenko\_precision\_score(df.actual\_label.values, df.predicted\_RF.values)))  
print('F1 RF: %.3f' % (fedorenko\_f1\_score(df.actual\_label.values, df.predicted\_RF.values)))  
print('Accuracy LR: %.3f' % (fedorenko\_accuracy\_score(df.actual\_label.values, df.predicted\_LR.values)))  
print('Recall LR: %.3f' % (fedorenko\_recall\_score(df.actual\_label.values, df.predicted\_LR.values)))  
print('Precision LR: %.3f' % (fedorenko\_precision\_score(df.actual\_label.values, df.predicted\_LR.values)))  
print('F1 LR: %.3f' % (fedorenko\_f1\_score(df.actual\_label.values, df.predicted\_LR.values)))  
print('')  
print('scores with threshold = 0.25')  
print(  
 'Accuracy RF: %.3f' % (fedorenko\_accuracy\_score(df.actual\_label.values, (df.model\_RF >= 0.25).astype('int').values)))  
print('Recall RF: %.3f' % (fedorenko\_recall\_score(df.actual\_label.values, (df.model\_RF >= 0.25).astype('int').values)))  
print('Precision RF: %.3f' % (  
 fedorenko\_precision\_score(df.actual\_label.values, (df.model\_RF >= 0.25).astype('int').values)))  
print('F1 RF: %.3f' % (fedorenko\_f1\_score(df.actual\_label.values, (df.model\_RF >= 0.25).astype('int').values)))  
print(  
 'Accuracy LR: %.3f' % (fedorenko\_accuracy\_score(df.actual\_label.values, (df.model\_LR >= 0.25).astype('int').values)))  
print('Recall LR: %.3f' % (fedorenko\_recall\_score(df.actual\_label.values, (df.model\_LR >= 0.25).astype('int').values)))  
print('Precision LR: %.3f' % (  
 fedorenko\_precision\_score(df.actual\_label.values, (df.model\_LR >= 0.25).astype('int').values)))  
print('F1 LR: %.3f' % (fedorenko\_f1\_score(df.actual\_label.values, (df.model\_LR >= 0.25).astype('int').values)))  
  
fpr\_RF, tpr\_RF, thresholds\_RF = roc\_curve(df.actual\_label.values, df.model\_RF.values)  
fpr\_LR, tpr\_LR, thresholds\_LR = roc\_curve(df.actual\_label.values, df.model\_LR.values)  
  
auc\_RF = roc\_auc\_score(df.actual\_label.values, df.model\_RF.values)  
auc\_LR = roc\_auc\_score(df.actual\_label.values, df.model\_LR.values)  
print('AUC RF:%.3f' % auc\_RF)  
print('AUC LR:%.3f' % auc\_LR)  
  
plt.plot(fpr\_RF, tpr\_RF, 'r-', label='RF AUC: %.3f' % auc\_RF)  
plt.plot(fpr\_LR, tpr\_LR, 'b-', label='LR AUC: %.3f' % auc\_LR)  
plt.plot([0, 1], [0, 1], 'k-', label='random')  
plt.plot([0, 0, 1, 1], [0, 1, 1, 1], 'g-', label='perfect')  
plt.legend()  
plt.xlabel('False Positive Rate')  
plt.ylabel('True Positive Rate')  
plt.show()

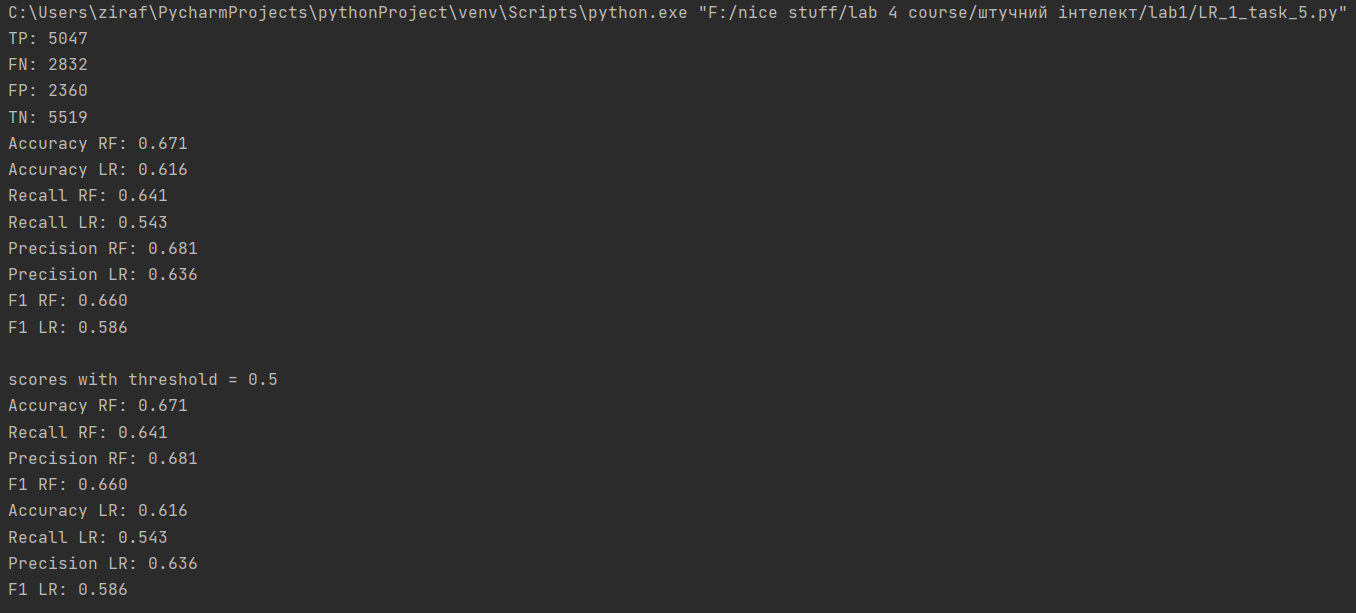


Рис. 9 Результат виконання програми



Рис. 10 Результат виконання програми

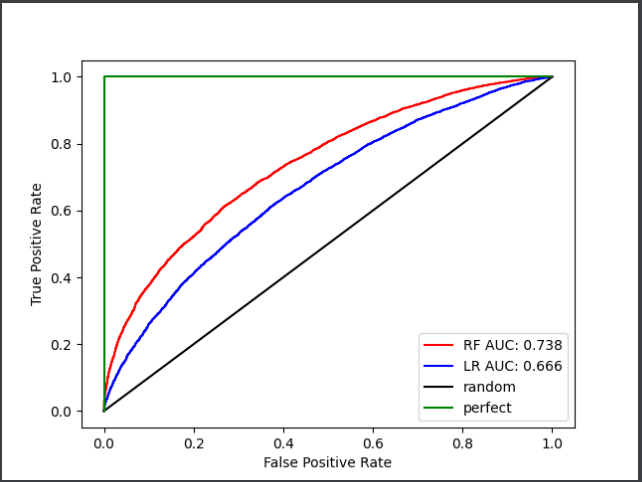


Рис. 11 Результат виконання програми

В даному завдані використовувались дві моделі – RF та LR. RF модель краще, так як є більш точною і повною, ніж LR. Це також можна побачити на графіку (рис. 11).

**Завдання 2.6. Розробіть програму класифікації даних в файлі data\_multivar\_nb.txt за допомогою машини опорних векторів (Support Vector Machine -** **SVМ). Розрахуйте показники якості класифікації. Порівняйте їх з показниками наївного байєсівського класифікатора. Зробіть висновки яку модель класифікації краще обрати і чому.**

Лістинг програми:

from sklearn.svm import SVC  
from sklearn.naive\_bayes import GaussianNB  
import pandas as pd  
import numpy as np  
from sklearn.model\_selection import train\_test\_split, cross\_val\_score  
import matplotlib.pyplot as plt  
from sklearn.metrics import roc\_curve  
from sklearn import preprocessing  
from utilities import visualize\_classifier  
  
input\_file = 'data\_multivar\_nb.txt'  
data = np.loadtxt(input\_file, delimiter=',')  
X, y = data[:, :-1], data[:, -1]  
  
X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=3)  
  
classifier\_svm = SVC()  
classifier\_svm.fit(X\_train, y\_train)  
  
classifier\_nb = GaussianNB()  
classifier\_nb.fit(X\_train, y\_train)  
  
y\_pred\_svm = classifier\_svm.predict(X\_test)  
y\_pred\_nb = classifier\_nb.predict(X\_test)  
  
num\_folds = 3  
  
print('Naive Bayes:')  
accuracy\_values = cross\_val\_score(classifier\_nb, X, y, scoring='accuracy', cv=num\_folds)  
print("Accuracy: " + str(round(100 \* accuracy\_values.mean(), 2))  
 + "%")  
  
precision\_values = cross\_val\_score(classifier\_nb, X, y, scoring='precision\_weighted', cv=num\_folds)  
print("Precision: " + str(round(100 \* precision\_values.mean(),  
 2)) + "%")  
  
recall\_values = cross\_val\_score(classifier\_nb, X, y, scoring='recall\_weighted', cv=num\_folds)  
print("Recall: " + str(round(100 \* recall\_values.mean(), 2)) +  
 "%")  
  
f1\_values = cross\_val\_score(classifier\_nb, X, y, scoring='f1\_weighted', cv=num\_folds)  
print("F1: " + str(round(100 \* f1\_values.mean(), 2)) + "%")  
  
print('\nSVM:')  
accuracy\_values = cross\_val\_score(classifier\_svm, X, y, scoring='accuracy', cv=num\_folds)  
print("Accuracy: " + str(round(100 \* accuracy\_values.mean(), 2))  
 + "%")  
  
precision\_values = cross\_val\_score(classifier\_svm, X, y, scoring='precision\_weighted', cv=num\_folds)  
print("Precision: " + str(round(100 \* precision\_values.mean(),  
 2)) + "%")  
  
recall\_values = cross\_val\_score(classifier\_svm, X, y, scoring='recall\_weighted', cv=num\_folds)  
print("Recall: " + str(round(100 \* recall\_values.mean(), 2)) +  
 "%")  
  
f1\_values = cross\_val\_score(classifier\_svm, X, y, scoring='f1\_weighted', cv=num\_folds)  
print("F1: " + str(round(100 \* f1\_values.mean(), 2)) + "%")  
  
visualize\_classifier(classifier\_nb, X\_test, y\_test)  
visualize\_classifier(classifier\_svm, X\_test, y\_test)

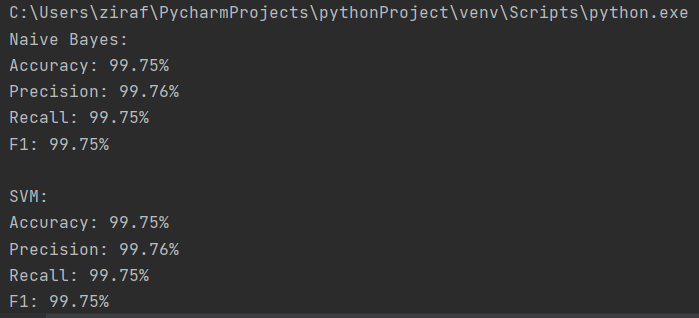


Рис. 12 Результат виконання програми

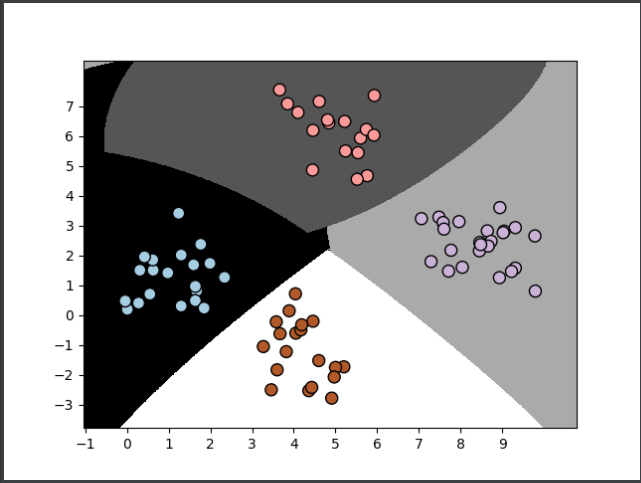


Рис. 13 Результат виконання програми

Висновок: в даному випадку ми не має великої різниці між показниками SVМ та показниками наївного байєсівського класифікатора, тому не модна визначити який з методів краще.

***Висновки:*** на даній лабораторній, ми використовуючи спеціалізовані бібліотеки та мову програмування Python, дослідити попередню обробку та класифікацію даних, розглянули бінаризацію, L1-нормалізацію та L2-нормалізацію, методи класифікації (наївний метод Баєса), різницю між моделями RF і LR.