ЛАБОРАТОРНА РОБОТА № 1

ПОПЕРЕДНЯ ОБРОБКА ТА КОНТРОЛЬОВАНА КЛАСИФІКАЦІЯ ДАНИХ

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

Хід роботи:

Завдання 2.1: Попередня обробка даних

Бінарізація

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]
])

# Бінаризація даних
data_binarized = preprocessing.Binarizer(threshold=2.1).transform(input_data)
print("\n Binarized data:\n", data_binarized)
```

Результат виконання завдання

```
Binarized data:
[[1. 0. 1.]
[0. 1. 0.]
[1. 0. 0.]
[1. 0. 0.]]
```

Рис. 1.1 Результат виконання завдання

Виключення середнього

Лістинг програми:

```
# Виведення середнього значення та стандартного відхилення
print("\nBEFORE: ")
print("Mean =", input_data.mean(axis=0))
print("Std deviation =", input_data.std(axis=0))

# Виключення середнього
data_scaled = preprocessing.scale(input_data)
print("\nAFTER: ")
print("Mean =", data_scaled.mean(axis=0))
print("Std deviation =", data_scaled.std(axis=0))
```

Результат виконання завдання

```
BEFORE:
Mean = [ 3.775 -1.15  -1.3  ]
Std deviation = [3.12039661 6.36651396 4.0620192 ]

AFTER:
Mean = [1.11022302e-16 0.00000000e+00 2.77555756e-17]
Std deviation = [1. 1. 1.]
```

Рис. 1.2 Результат виконання завдання

Масштабування

Лістинг програми:

```
# Масштабування MinMax
data_scaler_minmax = preprocessing.MinMaxScaler(feature_range=(0, 1))
data_scaled_minmax = data_scaler_minmax.fit_transform(input_data)
print("\nMin max scaled data:\n", data_scaled_minmax)
```

Результат виконання завдання

```
Min max scaled data:
 [[0.74117647 0.39548023 1.        ]
 [0.         1.         0.        ]
 [0.6        0.5819209  0.87234043]
 [1.         0.         0.17021277]]
```

Рис. 1.3 Результат виконання завдання

Нормалізація

Лістинг програми:

```
# Нормалізація даних
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)
```

Результат виконання завдання

```
l1 normalized data:
 [[ 0.45132743 -0.25663717  0.2920354 ]
 [-0.0794702   0.51655629 -0.40397351]
 [ 0.609375    0.0625      0.328125  ]
 [ 0.33640553 -0.4562212  -0.20737327]]

l2 normalized data:
 [[ 0.75765788 -0.43082507  0.49024922]
 [-0.12030718  0.78199664 -0.61156148]
 [ 0.87690281  0.08993875  0.47217844]
 [ 0.55734935 -0.75585734 -0.34357152]]
```

Рис. 1.4 Результат виконання завдання

L1-нормалізація використовує метод найменших абсолютних відхилень, що забезпечує рівність 1 суми абсолютних значень в кожному ряду, в той час як, **L2-нормалізація –** рівність 1 суми квадратів значень в кожному ряду. Тому **L1-нормалізації** вважається більш надійною по порівняно з **L2-нормалізацією**, оскільки вона менш чутлива до викидів.

Кодування міток

Лістинг програми:

```python
import numpy as np
from sklearn import preprocessing

# Надання позначок вхідних даних
input_labels = ['red', 'Black', 'red', 'green', 'black', 'yellow', 'white']

# Створення кодувальника та встановлення відповідності між мітками та числа-ми
encoder = preprocessing.LabelEncoder()
encoder.fit(input_labels)

# Виведення відображення
print("\nLabel mapping:")
for i, item in enumerate(encoder.classes_):
    print(item, '-->', i)

# перетворення міток за допомогою кодувальника
test_labels = ['green', 'red', 'Black']
encoded_values = encoder.transform(test_labels)
print("\nLabels =", test_labels)
print("Encoded values =", list(encoded_values))

# Декодування набору чисел за допомогою декодера
encoded_values = [3, 0, 4, 1]
decoded_list = encoder.inverse_transform(encoded_values)
```

```
print("\nEncoded values =", encoded_values)
print("Decoded labels =", list(decoded_list))
```

Результат виконання завдання

```
Label mapping:
green --> 0
red --> 1
white --> 2
yellow --> 3
blаck --> 4
black --> 5

Labels = ['green', 'red', 'blаck']
Encoded values = [0, 1, 4]

Encoded values = [3, 0, 4, 1]
Decoded labels = ['yellow', 'green', 'blаck', 'red']

Process finished with exit code 0
```

Рис. 1.5 Результат виконання завдання

Завдання 2.2: Попередня обробка нових даних

Таблиця 1

№ варіанту	Значення змінної input_data												Поріг бінаризації
5	-1.3	3.9	4.5	-5.3	-4.2	-1.3	5.2	-6.5	-1.1	-5.2	2.6	-2.2	3.0

Лістинг програми:

```
import numpy as np
from sklearn import preprocessing

input_data = np.array([
    [-1.3, 3.9, 4.5],
    [-5.3, -4.2, -1.3],
    [5.2, -6.5, -1.1],
    [-5.2, 2.6, -2.2]
])

# Бінаризація даних
data_binarized = preprocessing.Binarizer(threshold=1.8).transform(input_data)
print("\n Binarized data:\n", data_binarized)

# Виведення середнього значення та стандартного відхилення
print("\nBEFORE: ")
print("Mean =", input_data.mean(axis=0))
print("Std deviation =", input_data.std(axis=0))

# Виключення середнього
data_scaled = preprocessing.scale(input_data)
```

```
print("\nAFTER: ")
print("Mean =", data_scaled.mean(axis=0))
print("Std deviation =", data_scaled.std(axis=0))

# Масштабування MinMax
data_scaler_minmax = preprocessing.MinMaxScaler(feature_range=(0, 1))
data_scaled_minmax = data_scaler_minmax.fit_transform(input_data)
print("\nMin max scaled data:\n", data_scaled_minmax)

# Нормалізація даних
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)
```

Результат виконання завдання

```
 Binarized data:
 [[0. 1. 1.]
 [0. 0. 0.]
 [1. 0. 0.]
 [0. 1. 0.]]

BEFORE:
Mean = [-1.65  -1.05  -0.025]
Std deviation = [4.27112397 4.40028408 2.64516068]

AFTER:
Mean = [-2.77555756e-17  5.55111512e-17  5.55111512e-17]
Std deviation = [1. 1. 1.]

Min max scaled data:
 [[0.38095238 1.         1.        ]
 [0.         0.22115385 0.13432836]
 [1.         0.         0.1641791 ]
 [0.00952381 0.875      0.        ]]

l1 normalized data:
 [[-0.13402062  0.40206186  0.46391753]
 [-0.49074074 -0.38888889 -0.12037037]
 [ 0.40625    -0.5078125  -0.0859375 ]
 [-0.52        0.26       -0.22      ]]

l2 normalized data:
 [[-0.21328678  0.63986035  0.7383004 ]
 [-0.76965323 -0.60991388 -0.18878287]
 [ 0.61931099 -0.77413873 -0.13100809]
 [-0.83653629  0.41826814 -0.3539192 ]]
```

Рис. 1.6 Результат виконання завдання

Завдання 2.3: Класифікація логістичною регресією або логістичний класифікатор

Лістинг програми:

```
import numpy as np
from sklearn import linear_model
import matplotlib.pyplot as plt
```

```
from utilities import visualize_classifier

# Визначення зразка вхідних даних
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])
# Створення логістичного класифікатора
classifier = linear_model.LogisticRegression(solver='liblinear', C=1)

# Тренування класифікатора
classifier.fit(X, y)
visualize_classifier(classifier, X, y)
```

Результат виконання завдання

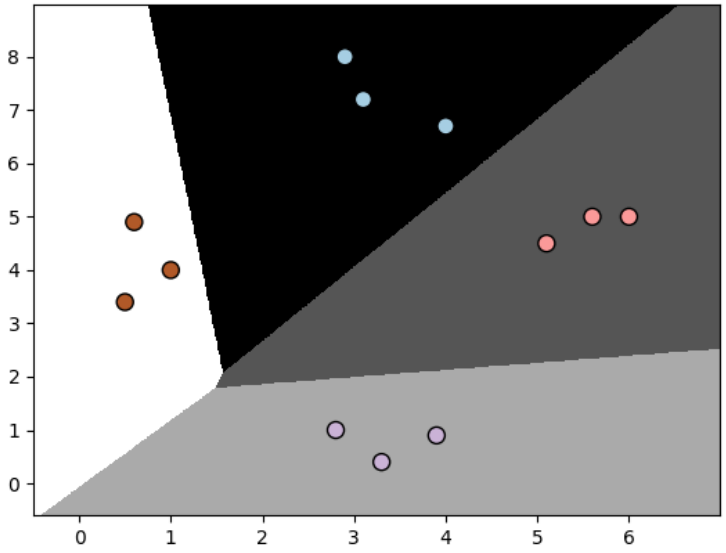

Рис. 1.7 Результат виконання завдання

Завдання 2.4: Класифікація наївним байєсовським класифікатором

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
from utilities import visualize_classifier

# Вхідний файл, який містить дані
```

```
input_file = 'data_multivar_nb.txt'

# Завантаження даних із вхідного файлу
data = np.loadtxt(input_file, delimiter=',')
X, y = data[:, :-1], data[:, -1]

# Створення наївного байєсовського класифікатора
classifier = GaussianNB()

# Тренування класифікатора
classifier.fit(X, y)

# Прогнозування значень для тренувальних даних
y_pred = classifier.predict(X)

# Обчислення якості класифікатора
accuracy = 100.0 * (y == y_pred).sum() / X.shape[0]
print("Accuracy of Naive Bayes classifier =", round(accuracy, 2), "%")

# Візуалізація результатів роботи класифікатора
visualize_classifier(classifier, X, y)
```

Результат виконання завдання

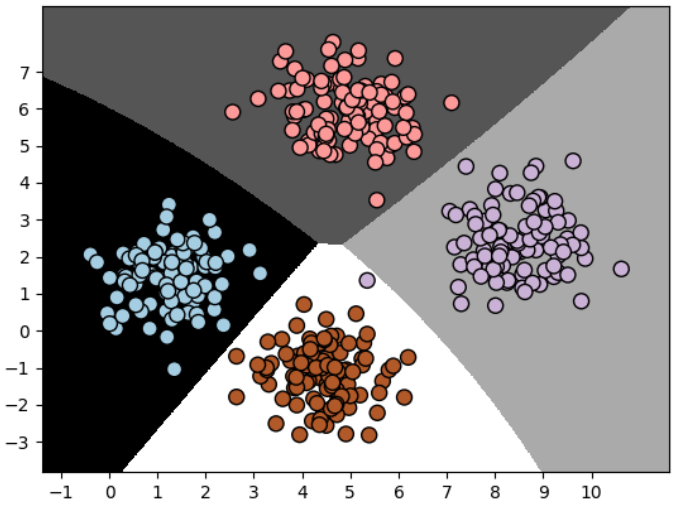

Рис. 1.8 Результат виконання завдання

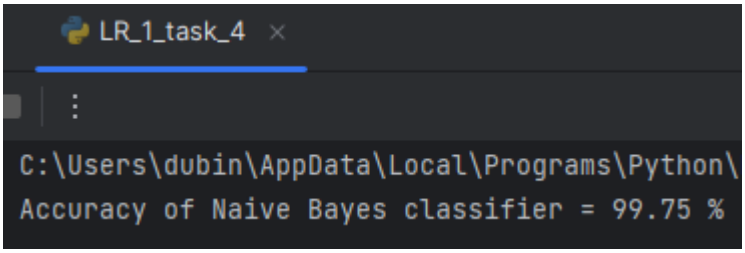

Рис. 1.9 Результат виконання завдання

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
```

```
# Розбивка даних на навчальний та тестовий набори
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=3)
classifier_new = GaussianNB()
classifier_new.fit(X_train, y_train)
y_test_pred = classifier_new.predict(X_test)

# Обчислення якості класифікатора
accuracy = 100.0 * (y_test == y_test_pred).sum() / X_test.shape[0]
print("Accuracy of the new classifier =", round(accuracy, 2), "%")

# Візуалізація роботи класифікатора
visualize_classifier(classifier_new, 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)) + "%")
```

Результат виконання завдання

1 прогін

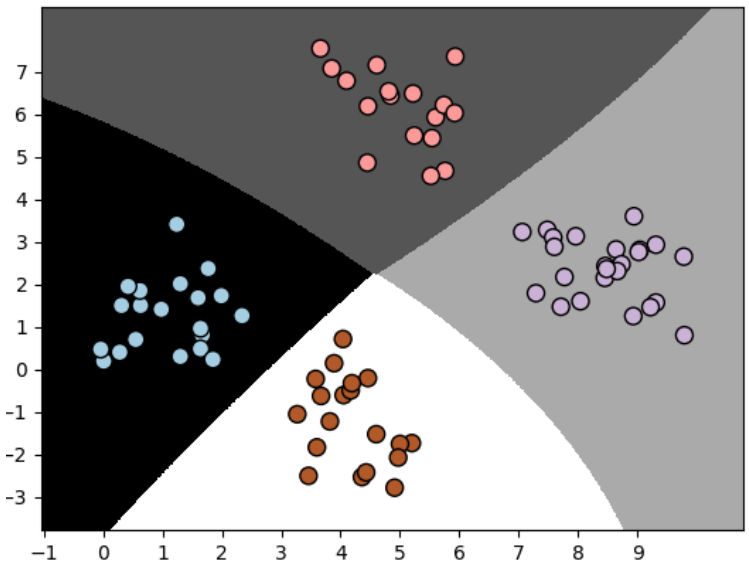

Рис. 1.10 Результат виконання завдання

```
Accuracy of the new classifier = 100.0 %
Accuracy: 99.75%
Precision: 99.76%
Recall: 99.75%
F1: 99.75%
```

Рис. 1.11 Результат виконання завдання

2 прогін

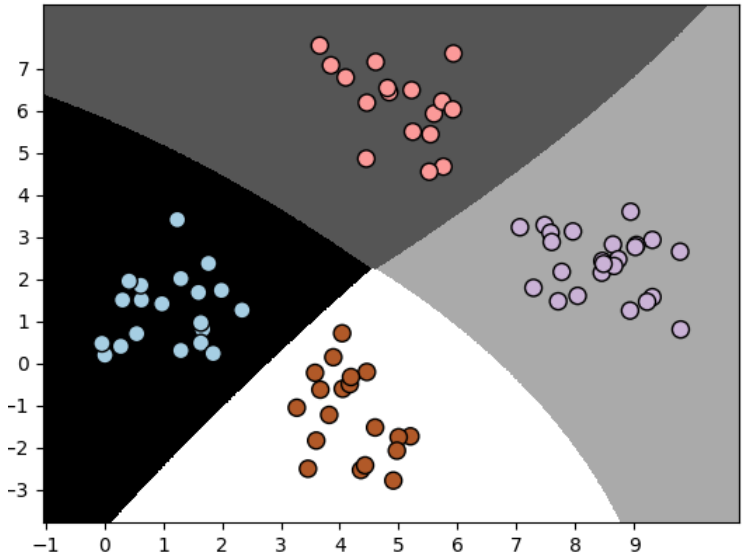

Рис. 1.12 Результат виконання завдання

```
Accuracy of the new classifier = 100.0 %
Accuracy: 99.75%
Precision: 99.76%
Recall: 99.75%
F1: 99.75%
```

Рис. 1.13 Результат виконання завдання

Отримані результати після двох прогонів ідентичні, бо тренування відбувалися на однакових початкових значеннях.

Завдання 2.5: Вивчити метрики якості класифікації

Лістинг програми:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.metrics import recall_score
from sklearn.metrics import precision_score
from sklearn.metrics import f1_score
from sklearn.metrics import roc_curve
from sklearn.metrics import roc_auc_score

df = pd.read_csv('data_metrics.csv')
df.head()

thresh = 0.5
df['predicted_RF'] = (df.model_RF >= 0.5).astype('int')
df['predicted_LR'] = (df.model_LR >= 0.5).astype('int')
df.head()
```

```
print(confusion_matrix(df.actual_label.values, df.predicted_RF.values))
print(confusion_matrix(df.actual_label.values, df.predicted_LR.values))


def find_TP(y_true, y_pred):
    # counts the number of true positives (y_true = 1, y_pred = 1)
    return sum((y_true == 1) & (y_pred == 1))


def find_FN(y_true, y_pred):
    # counts the number of false negatives (y_true = 1, y_pred = 0)
    return sum((y_true == 1) & (y_pred == 0))


def find_FP(y_true, y_pred):
    # counts the number of false positives (y_true = 0, y_pred = 1)
    return sum((y_true == 0) & (y_pred == 1))


def find_TN(y_true, y_pred):
    # counts the number of true negatives (y_true = 0, y_pred = 0)
    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):
    # calculate TP, FN, FP, TN
    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 my_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]])


print(my_confusion_matrix(df.actual_label.values, df.predicted_RF.values))
print(my_confusion_matrix(df.actual_label.values, df.predicted_LR.values))
assert np.array_equal(my_confusion_matrix(df.actual_label.values,
df.predicted_RF.values),
                      confusion_matrix(df.actual_label.values,
df.predicted_RF.values)), \
    'my_confusion_matrix() is not correct for RF'
assert np.array_equal(my_confusion_matrix(df.actual_label.values,
df.predicted_LR.values),
                      confusion_matrix(df.actual_label.values,
df.predicted_LR.values)), \
    'my_confusion_matrix() is not correct for LR'

print(accuracy_score(df.actual_label.values, df.predicted_RF.values))
print(accuracy_score(df.actual_label.values, df.predicted_LR.values))

def my_accuracy_score(y_true, y_pred):
    # calculates the fraction of samples
```

```
    TP, FN, FP, TN = find_conf_matrix_values(y_true, y_pred)
    return (TP + TN) / (TP + TN + FP + FN)


assert my_accuracy_score(df.actual_label.values, df.predicted_RF.values) ==
accuracy_score(
    df.actual_label.values,
    df.predicted_RF.values), 'my_accuracy_score failed on assert
my_accuracy_score(df.actual_label.values, ' \
                         'df.predicted_LR.values) ==
accuracy_score(df.actual_label.values, df.predicted_LR.values),' \
                         'my_accuracy_score failed on LR'
print('Accuracy RF: % .3f' % (my_accuracy_score(df.actual_label.values,
df.predicted_RF.values)))
print('Accuracy LR: % .3f' % (my_accuracy_score(df.actual_label.values,
df.predicted_LR.values)))

print(recall_score(df.actual_label.values, df.predicted_RF.values))
print(recall_score(df.actual_label.values, df.predicted_LR.values))


def my_recall_score(y_true, y_pred):
    # calculates the fraction of positive samples predicted correctly
    TP, FN, FP, TN = find_conf_matrix_values(y_true, y_pred)
    return TP / (TP + FN)


assert my_recall_score(df.actual_label.values, df.predicted_RF.values) ==
recall score(df.actual label.values,

df.predicted_RF.values), \
    'my_recall_score failed on RF'
assert my_recall_score(df.actual_label.values, df.predicted_LR.values) ==
recall_score(df.actual_label.values,

df.predicted_LR.values), \
    'my_recall_score failed on LR'
print('Recall RF: %.3f' % (my_recall_score(df.actual_label.values,
df.predicted_RF.values)))
print('Recall LR: %.3f' % (my_recall_score(df.actual_label.values,
df.predicted_LR.values)))

print(precision_score(df.actual_label.values, df.predicted_RF.values))
print(precision_score(df.actual_label.values, df.predicted_LR.values))


def my_precision_score(y_true, y_pred):
    # calculates the fraction of predicted positives samples that are actu-ally
positive
    TP, FN, FP, TN = find_conf_matrix_values(y_true, y_pred)
    return TP / (TP + FP)


assert my_precision_score(df.actual_label.values, df.predicted_RF.values) ==
precision_score(
    df.actual_label.values, df.predicted_RF.values), 'my_precision_score failed on
RF'
assert my_precision_score(df.actual_label.values, df.predicted_LR.values) ==
precision_score(
    df.actual_label.values, df.predicted_LR.values), 'my_precision_score failed on
LR'
print('Precision RF: %.3f' % (my_precision_score(df.actual_label.values,
df.predicted_RF.values)))
```

```
print('Precision LR: %.3f' % (my_precision_score(df.actual_label.values,
df.predicted_LR.values)))

print(f1_score(df.actual_label.values, df.predicted_RF.values))
print(f1_score(df.actual_label.values, df.predicted_LR.values))


def my_f1_score(y_true, y_pred):
    # calculates the F1 score
    recall = my_recall_score(y_true, y_pred)
    precision = my_precision_score(y_true, y_pred)
    return (2 * (precision * recall)) / (precision + recall)


assert my_f1_score(df.actual_label.values, df.predicted_RF.values) ==
f1_score(df.actual_label.values,

df.predicted_RF.values), 'my_f1_score failed on RF'
assert my_f1_score(df.actual_label.values, df.predicted_LR.values) ==
f1_score(df.actual_label.values,

df.predicted_LR.values), 'my_f1_score failed on LR'
print('F1 RF: %.3f' % (my_f1_score(df.actual_label.values,
df.predicted_RF.values)))
print('F1 LR: %.3f' % (my_f1_score(df.actual_label.values,
df.predicted_LR.values)))

print('scores with threshold = 0.5')
print('Accuracy RF: %.3f' % (my accuracy score(df.actual label.values,
df.predicted_RF.values)))
print('Recall RF: %.3f' % (my_recall_score(df.actual_label.values,
df.predicted_RF.values)))
print('Precision RF: %.3f' % (my_precision_score(df.actual_label.values,
df.predicted_RF.values)))
print('F1 RF: %.3f' % (my_f1_score(df.actual_label.values,
df.predicted_RF.values)))
print('')
print('scores with threshold = 0.25')
print('Accuracy RF: %.3f' % (
    my_accuracy_score(df.actual_label.values, (df.model_RF >=
0.25).astype('int').values)))
print('Recall RF: %.3f' % (my_recall_score(df.actual_label.values, (df.model_RF >=
0.25).astype('int').values)))
print('Precision RF: %.3f' % (
    my_precision_score(df.actual_label.values, (df.model_RF >=
0.25).astype('int').values)))
print('F1 RF: %.3f' % (my_f1_score(df.actual_label.values, (df.model_RF >=
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)

plt.plot(fpr_LR, tpr_LR, 'b-', label='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()

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()
```

Результат виконання завдання

```
[[5519 2360]
 [2832 5047]]
[[5425 2454]
 [3600 4279]]
TP: 5047
FN: 2832
FP: 2360
TN: 5519
[[5519 2360]
 [2832 5047]]
[[5425 2454]
 [3600 4279]]
0.6705165630156111
0.6158141896179719
Accuracy RF:  0.671
Accuracy LR:  0.616
0.6405635232897576
0.5430892245208783
Recall RF: 0.641
Recall LR: 0.543
0.681382476036182
0.6355265112134264
Precision RF: 0.681
Precision LR: 0.636
0.660342797330891
0.5856830002737475
F1 RF: 0.660
F1 LR: 0.586
scores with threshold = 0.5
Accuracy RF: 0.671
Recall RF: 0.641
Precision RF: 0.681
F1 RF: 0.660

scores with threshold = 0.25
Accuracy RF: 0.502
Recall RF: 1.000
Precision RF: 0.501
F1 RF: 0.668
```

Рис. 1.14 Результат виконання завдання

Акуратність для більшого порогу є кращою за акуратність для меншого. Але чутливість навпаки для більшого порогу є меншою. Точність для порогу 0.5 виявилася більшою за точність для 0.25. Оцінка f1 є майже ідентичною для обох порогів, а так як цей показник є одним з точніших для визначення пріоритетної моделі, то можемо зробити висновок, що обидва пороги мають право життя.

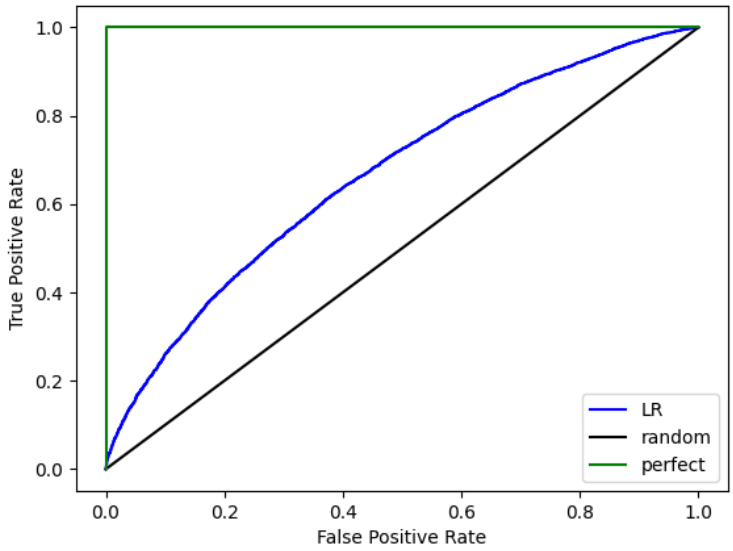

Рис. 1.15 Крива ROC для обох моделей

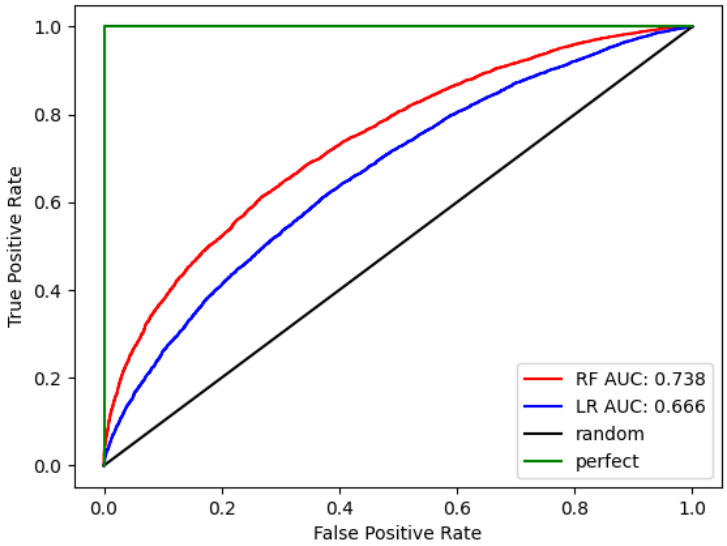

Рис. 1.16 Крива ROC для обох моделей (з урахуванням площ під кривою)

Площа під кривою для моделі RF (AUC = 0,738) краще, ніж LR (AUC = 0,666). Отже, згідно вищевказаної метрики робимо висновок, що модель RF краще.

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

Лістинг програми:

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.svm import SVC
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score
from utilities import visualize_classifier

# Вхідний файл, який містить дані
input_file = 'data_multivar_nb.txt'

# Завантаження даних із вхідного файлу
data = np.loadtxt(input_file, delimiter=',')
X, y = data[:, :-1], data[:, -1]

# Створення класифікатора SVM
classifier = SVC()

# Тренування класифікатора
classifier.fit(X, y)

# Прогнозування значень для тренувальних даних
y_pred = classifier.predict(X)

# Обчислення якості класифікатора
accuracy = 100.0 * (y == y_pred).sum() / X.shape[0]
print("Accuracy of Support Vector Machine classifier =", round(accuracy, 2), "%")

# # Візуалізація результатів роботи класифікатора
visualize_classifier(classifier, X, y)

# Розбивка даних на навчальний та тестовий набори
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=3)
classifier_new = SVC()
classifier_new.fit(X_train, y_train)
y_test_pred = classifier_new.predict(X_test)

# Обчислення якості класифікатора
accuracy = 100.0 * (y_test == y_test_pred).sum() / X_test.shape[0]
print("Accuracy of the new classifier =", round(accuracy, 2), "%")

# Візуалізація роботи класифікатора
visualize_classifier(classifier_new, 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)) + "%")
```

Класифікація Support Vector Machine - SVM

Результат виконання завдання

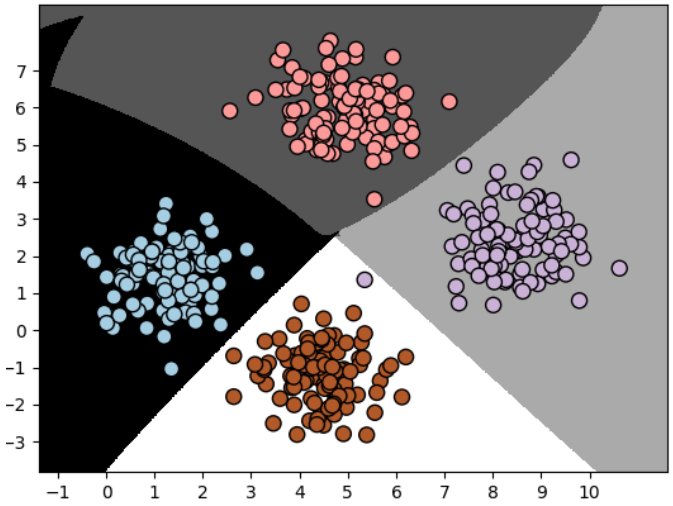

Рис. 1.17 Результат виконання завдання

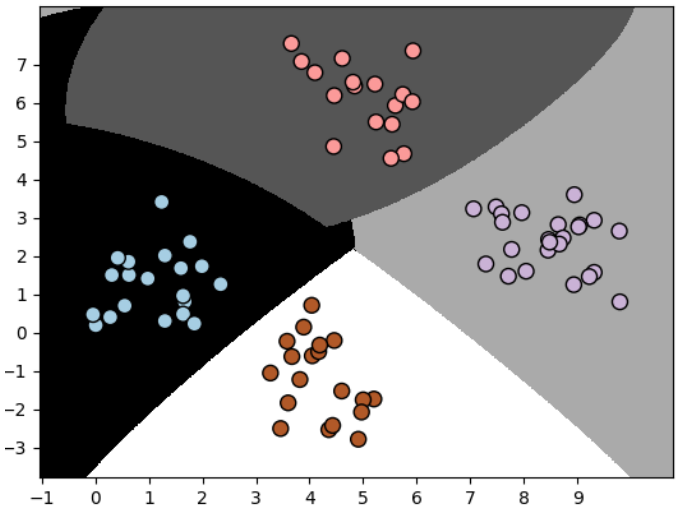

Рис. 1.18 Результат виконання завдання

```
Accuracy of Support Vector Machine classifier = 99.75 %
Accuracy of the new classifier = 100.0 %
Accuracy: 99.75%
Precision: 99.76%
Recall: 99.75%
F1: 99.75%
```

Рис. 1.19 Результат виконання завдання

Показники отримані з показниками обох класифікаторів ідентичні. Тому визначити який класифікатор краще неможливо на даному прикладі. Але зважаючи на те, що наївний байєсівський класифікатор визначає кожну ознаку як незалежну, важко отримати повну картину. Через це доцільніше використовувати класифікатор методу опорних векторів, а, також, він є найпопулярнішим методом класичної класифікації.

Посилання на GitHub: https://github.com/BogdanStelmah/Basics-of-AI_labs

Висновок: На даній лабораторній роботі ми використовуючи спеціалізовані бібліотеки та мову програмування Python дослідили попередню обробку та класифікацію даних.