**VIETNAM NATIONAL UNIVERSITY – HCM**

**INTERNATIONAL UNIVERSITY**

A blue and white logo

Description automatically generated**DEPARTMENT OF COMPUTER SICENCE & ENGINEER**

SEMESTER 2 (2023-2024)

**IT159IU**

**Artificial Intelligence Project**

**Topic name: Heart failure classification**

*May 2024*

**Group 7:**

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **Full name - Student name** | **Student ID** | **Contribution (%)** |
| 1 | PHAM VU TUYET ANH | ITDSIU21073 | 35 |
| 2 | NGUYEN NHAT KHIEM | ITDSIU21091 | 35 |
| 3 | LIEU PHONG SON | ITITIU18200 | 30 |

**Table of Contents**

[I. INTRODUCTION (TUYET ANH – ITDSIU21073) 3](#_Toc167104323)

[1 Import needed libraries & dataset. 3](#_Toc167104324)

[2 Description 3](#_Toc167104325)

[3 Visualization 6](#_Toc167104326)

[II. Preprocessing (TUYET ANH – ITDSIU21073) 11](#_Toc167104327)

[1 Processing missing value 11](#_Toc167104328)

[**2** Encoding 11](#_Toc167104329)

[3 Normalization 12](#_Toc167104330)

[4 Check multicollinearity. 13](#_Toc167104331)

[III. APPLYING MODEL (NHAT KHIEM – ITDSIU21091): 14](#_Toc167104332)

[**1. Aim** 14](#_Toc167104333)

[**2. Methodology** 14](#_Toc167104334)

[IV. METRICS (NHAT KHIEM – ITDSIU21091) 16](#_Toc167104335)

[1. Compare 16](#_Toc167104336)

[2. Recommendation 16](#_Toc167104337)

[V. Exited model (PHONG SON – ITITIU18200) 17](#_Toc167104338)

[VI. References 20](#_Toc167104339)

# INTRODUCTION (TUYET ANH – ITDSIU21073)

## Import needed libraries & dataset.

It is necessary to import the needed libraries for

* Manipulation data:

import numpy as np

import pandas as pd

* Visualization:

import matplotlib.pyplot as plt

import seaborn as sns

* Applying models:

from sklearn.model\_selection import KFold, GridSearchCV,  StratifiedKFold, RepeatedStratifiedKFold, cross\_val\_score, train\_test\_split

from sklearn.naive\_bayes import GaussianNB

from sklearn.neural\_network import MLPClassifier

from sklearn.ensemble import RandomForestClassifier

from sklearn.neighbors import KNeighborsClassifier

from sklearn.linear\_model import LogisticRegression

from sklearn.svm import SVC

* Evaluating performance

from sklearn.metrics import accuracy\_score,f1\_score, recall\_score, classification\_report

from sklearn.metrics import roc\_auc\_score, confusion\_matrix

import time

import warnings

warnings.filterwarnings('ignore')

pd.options.display.float\_format = '{:.2f}'.format

The dataset can be imported to analyze by the command code with name “df” and path\_line(“heart.csv”):

df = pd.read\_csv("heart.csv")

## Description

Using “head” to see top five content of dataset:

A screenshot of a computer

Description automatically generateddf.head()

Dataset is divided into input data (used to predict and independent) and target output (will be predicted and dependent on input data).

* To predict the patient whether heart disease or normal which based on some information below:
* Age: age of patient [year]
* Sex: sex of the patient [M: Male, F: Female]
* Chest Pain Type: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
* RestingBP: resting blood pressure [mm Hg]
* Cholesterol: serum cholesterol [mm/dl]
* Fasting BS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
* Resting ECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
* Max HR: maximum heart rate achieved [Numeric value between 60 and 202]
* Exercise Angina: exercise-induced angina [Y: Yes, N: No]
* Oldpeak: oldpeak = ST [Numeric value measured in depression]
* ST\_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
* Target output: Heart disease [ 1: heart disease, 0: normal]
* In conclusion, it seems to be a task as a classification problem in supervised learning based on two main criteria. Firstly, the target output is categorical - 2 classes (heart disease / normal). Then, accessing to a labeled dataset for training.
* When the task is satisfied, supervised learning algorithms are recommended as well as Logistic Regression, Random Forests, SVM, KNN, Naive Bayes, Neural Networks.
* The shape of dataset:

print("The shape of dataset is: ",df.shape)

* Dataset records 12 information of 918 patient.
* Using “df.info()” to quickly get a summary of dataset

A screen shot of a computer

Description automatically generated

* This summary is useful for showing that there are no missing values in any columns; therefore, it is not necessary to preprocess for missing values. Additionally, there are three types of data: float, int, and categorical. This helps determine that the dataset needs to be scaled and encoded in the preprocessing step.
* It is easily to get a comprehensive statistical summary of dataset by using describe method.
* Numerical columns:

A screenshot of a graph

Description automatically generated

* The ranges of the numerical variables differ significantly, such as between FastingBS and Cholesterol. It seems to all numerical input variables need to be scaled which help to reduce biased results and improve the performance of machine learning models.
* A screenshot of a computer

  Description automatically generatedObject columns:
* Since these are not ordinal variables and do not have a ranking, one-hot encoding is a better choice than label encoding.

## Visualization

* First, it is necessary to visualizing the distribution of target output through pie chart by using command line:

# Overall about output of database

partition = *list*(df['HeartDisease'].value\_counts())

# returns a list (easier to work) containing frequency table of target output.

plt.pie( # Creating the pie chart

        partition, # Adding the value of pie chart

*labels* = ['Nomarl','Heart Disease'], # Adding the label

*autopct*='%1.1f%%', # Specifies the format for displaying the percentage values on the pie chart.

*startangle* = 90, #  It starts at 90 degrees (which is at the top).

*explode* = (0.1,0), # This parameter determines the fraction of the radius with which to offset each slice. Here, the first slice ('Normal') is exploded (offset) by 0.1 times the radius.

*shadow*= True, # Adding shadow effect

       )

plt.title('Heart Disease %'); # set title

A blue and orange pie chart

Description automatically generated

Figure I‑1 Distribution of output target

* Observing the total dataset, we note that 44.7% have heart disease, which is a substantial proportion. With many individuals affected by heart disease, there is a necessity to utilize machine learning to predict whether someone is at risk of developing the condition to intervene in a timely manner.
* Relationship between categorical features and output target

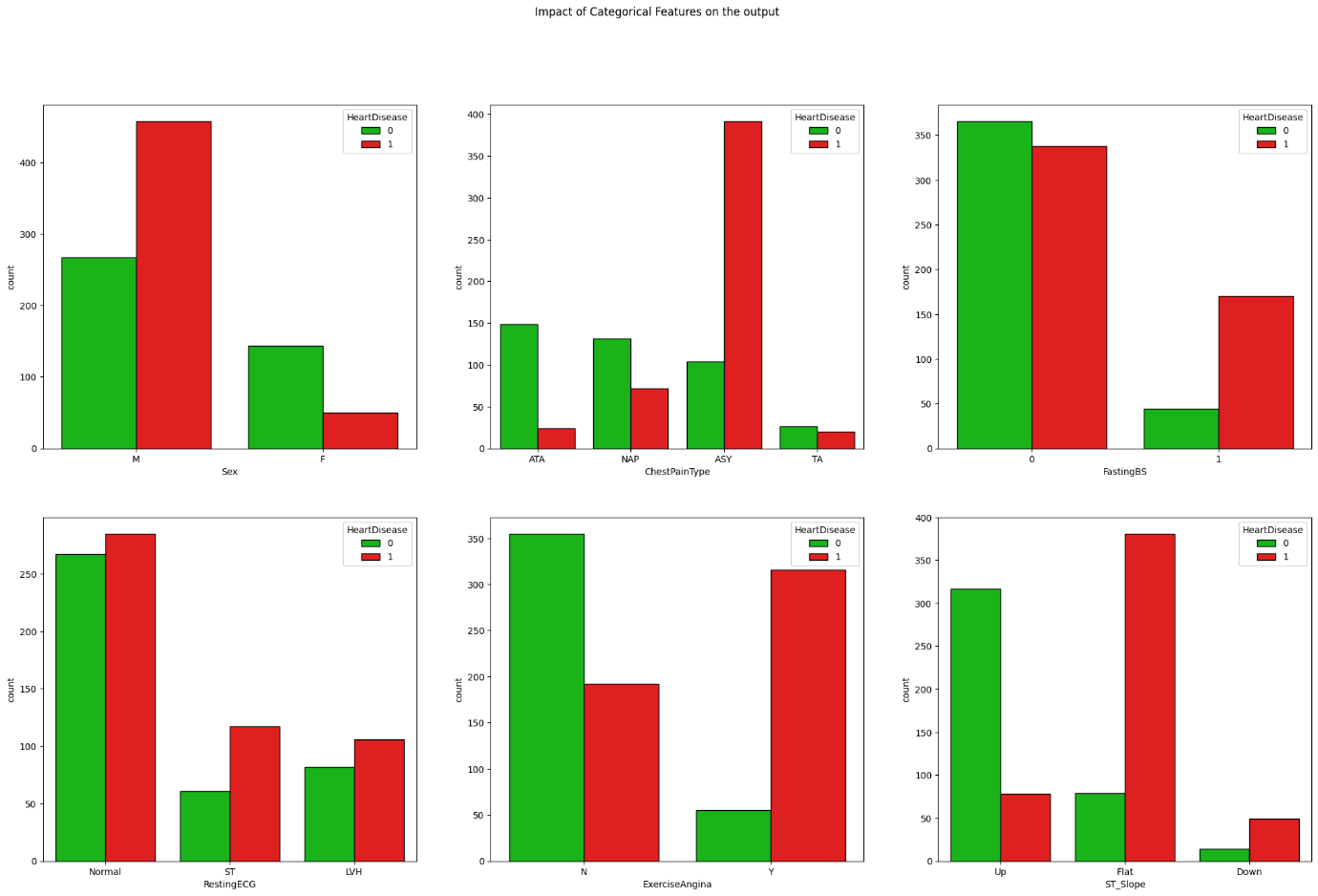


Figure I‑2 Impact of categorical features on output target

* Regarding gender, the rate of male suffering from heart failure is higher than that of female. The number of survey participants who were male was higher than that of female.
* Regarding chest pain type, the patient who has ASY is the highest ratio for heart failure
* Regarding FastingBS, the patien who has fasting blood sugar higher than 120 ml/dl, has higher proporttion.
* Regarding RestingECG, most people with heart failure have normal Resting ECG
* Regrading ExerciseAngina, most people with heart failure have exercise angina
* Regrading ST\_Slope, most people with heart failure have flat level
* Relationship between numerical features and heart disease.

A group of graphs showing different sizes of data

Description automatically generated with medium confidence

Figure I‑3 'Relationship between numerical features and heart disease

* Relationship between numerical features and no heart disease

A group of blue and white graphs

Description automatically generated

Figure I‑4Relationship between numerical features and no heart disease

* Most of person who age 55 - 60 years do not have high proportion of having heart disease
* Most of person who has 200-250mm/dl of cholesterol have high proportion of having heart disease
* Most of person who has 140-150 of MaxHR have high proportion of having heart disease
* Most of person who has 0 of Oldpeak do not have high proportion of having heart disease

# **Preprocessing (**TUYET ANH – ITDSIU21073**)**

## Processing missing value

A screenshot of a computer

Description automatically generated

Figure II‑1 Count the missing value

* The check for missing values with the command above reveals that there are no missing values in any column of the data, so the step can be skipped.

## Encoding

A screenshot of a computer

Description automatically generatedSince dealing with nominal variables, it is logical to use one-hot-encoding by command conde:

Figure II‑2 Encoding by one -hot encoding

This function from the pandas library is used to convert categorical variables into dummy/indicator variables. “drop\_first=True” means that it drops the first level of each categorical variable to avoid multicollinearity. For example, if there are two categories in a variable, only one dummy variable is created to represent them.

On the other hand, the “ST\_Slope” is ordered variable which needs to be encoded by Label Encoding method as well as code below:

label\_encoder = LabelEncoder()

# Apply label encoding to the 'ST\_Slope' column

data['ST\_Slope'] = label\_encoder.fit\_transform(data['ST\_Slope'])

A screenshot of a computer

Description automatically generatedAfter encoding, the dataset covers all numerical features like:

Figure II‑3 Transform all data type to int

## Normalization

split\_index = int(data.shape[0] \* 0.9)

data1 = data.iloc[:split\_index, :]

data2 = data.iloc[split\_index:, :]

X = data1.drop(['HeartDisease'],*axis*=1)

y = data1['HeartDisease']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, *random\_state*=42, *test\_size*=0.3)

numerical\_features = ['Age', 'RestingBP','MaxHR']

robust\_features = ['Cholesterol', 'Oldpeak']

from sklearn.preprocessing import MinMaxScaler, StandardScaler

normalize= RobustScaler()

standard= StandardScaler()

X\_train[robust\_features]= normalize.fit\_transform(X\_train[robust\_features])

X\_train[numerical\_features]= standard.fit\_transform(X\_train[numerical\_features])

X\_test[robust\_features]= normalize.transform(X\_test[robust\_features])

X\_test[numerical\_features]= standard.transform(X\_test[numerical\_features])

## Check multicollinearity.

plt.figure(*figsize*=(15,10))

sns.heatmap(data1.corr(),*annot*=True,*cmap*='RdYlGn')*;*

A colorful squares with numbers

Description automatically generated

Figure II‑4 Result of multicollinearity

# APPLYING MODEL (NHAT KHIEM – ITDSIU21091):

## 1. Aim

The objective of this section is to train models for classifying heart failure patients into two categories: normal (0) and positive heart failure (1). Various machine learning models are used to achieve this classification and conducted a thorough evaluation to determine the best model based on accuracy and other performance metrics.

## 2. Methodology

1. **Grid Search for Model Selection:**

A grid search is performed to identify the most suitable machine learning models based on their accuracy scores on the training set.

We utilized GridSearchCV to perform an exhaustive search over specified parameter values for different machine learning algorithms along with kfold train-test split. The following models and their respective parameter grids were considered: KNeighborsClassifier, GaussianNB, SVC, MLPClassifier, LogisticRegression, RandomForestClassifier.

After performing GridSearch using sklearn, the output yeilds:

* KNeighborsClassifier: 86.5%, {'leaf\_size': 1, 'metric': 'manhattan', 'n\_neighbors': 9, 'weights': 'uniform'}
* GaussianNB: 84.26%, {'var\_smoothing': 0.03511191734215131}
* SVC: 85.64%, {'C': 1, 'gamma': 0.1, 'kernel': 'rbf'}
* MLPClassifier: 85.64%, {'activation': 'tanh', 'alpha': 0.0001, 'early\_stopping': False, 'hidden\_layer\_sizes': (50,100,50), 'learning\_rate': 'apdaptive', 'solver': 'sgd'}
* LogisticRegression: 83.91%, {'C': 5, 'penalty': 'l2', 'solver': 'liblinear'}
* RandomForestClassifier: 88.06%, {'criterion': 'gini', 'max\_depth': 5, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 50}

**Note:** Due to the utilization of n\_job = -1 parameter and some random nature of some algorithms, the output could vary after every run as result of parallel run of GridSearchCV.

1. **Model Implementation and Evaluation: Selected models are initialized, trained on the dataset.**

Each model was initialized with the optimal parameters determined from the grid search. The training process involved fitting the models on the training dataset, ensuring that they learned the underlying patterns and relationships between the features and the target variable (heart failure classification). The models were then prepared for subsequent evaluation to assess their performance on unseen test data.

The implementation process for each model is outlined below:

* KNeighborsClassifier: Initialized with the best parameters including leaf\_size, metric, n\_neighbors, and weights.
* GaussianNB: Configured with the optimal var\_smoothing parameter.
* SVC: Set up with the best values for C, gamma, and kernel.
* MLPClassifier: Defined with the optimal activation, alpha, early\_stopping, hidden\_layer\_sizes, learning\_rate, and solver.
* LogisticRegression: Implemented with the best C, penalty, and solver.
* RandomForestClassifier: Initialized with the best criterion, max\_depth, min\_samples\_leaf, min\_samples\_split, and n\_estimators.

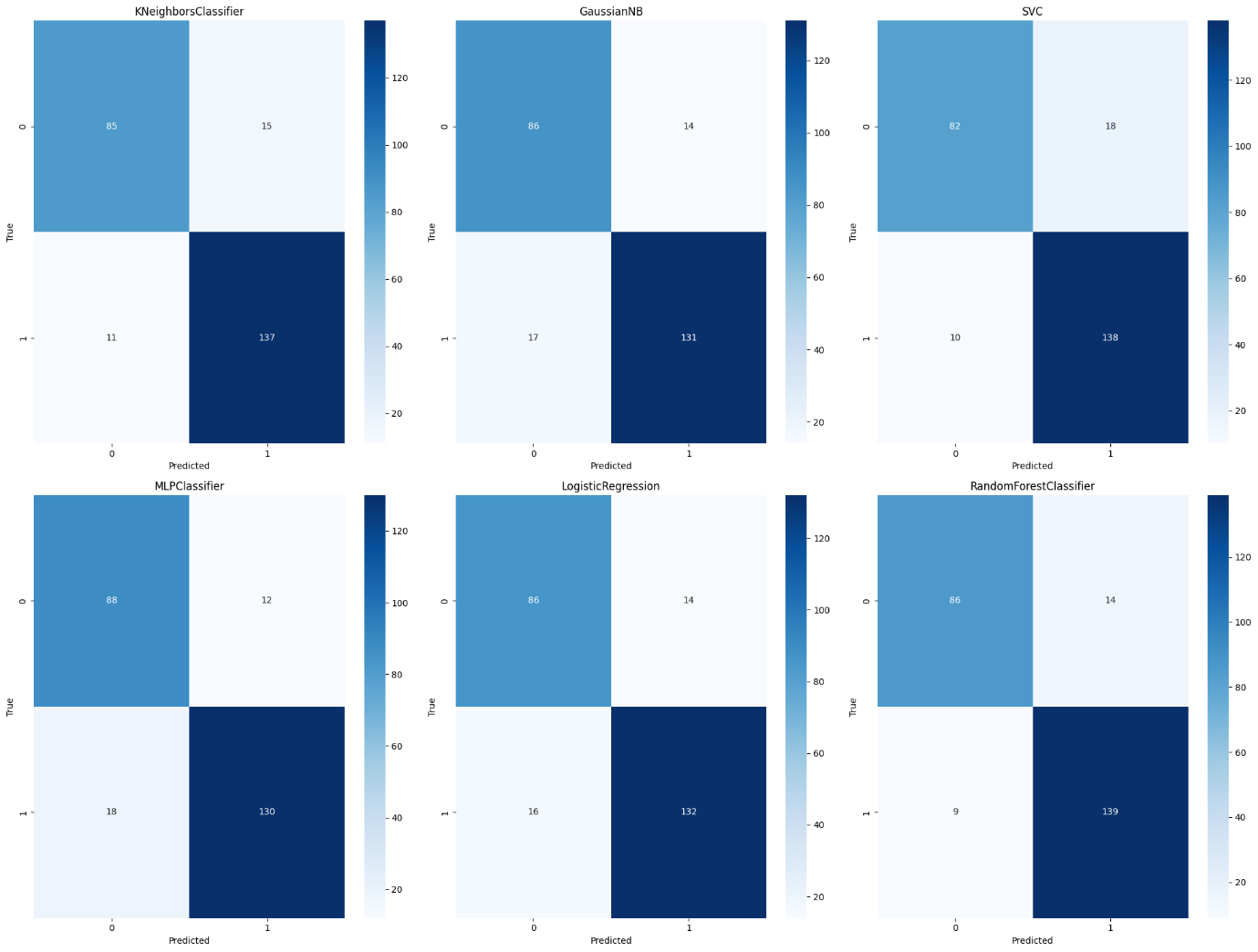
****After training outcome returns this:

Figure III-1 Confusion matrix of different models

# METRICS (NHAT KHIEM – ITDSIU21091)

## **A screenshot of a computer Description automatically generated**1. Compare

Figure IV – 1 comparison of different models.

Overall, all models performed well, each demonstrating strengths in various evaluation metrics. The RandomForestClassifier achieved the highest performance across several key metrics, with an accuracy of nearly 89%, a recall of 93%, an F1 score of 91%, and a ROC AUC score of 87%. These metrics indicate that the RandomForestClassifier not only correctly identified the majority of positive cases but also maintained a high balance between precision and recall. Additionally, it showed robust generalization capability as evidenced by its cross-validation score of 92% while the amount time taken for both training and testing is tiny.

The GaussianNB model also showed strong performance, particularly in terms of training efficiency. It achieved an accuracy of 88%, a recall of 89%, an F1 score of 89%, and a ROC AUC score of 87%. Despite these metrics being slightly lower than those of the RandomForestClassifier, GaussianNB had the shortest training time at less than 1 second. This makes it an ideal option for scenarios where computational resources or time are limited.

The remaining models, including KNeighborsClassifier, SVC, MLPClassifier, and LogisticRegression, also performed well, with accuracies ranging from 87% to 89% and consistent cross-validation scores around 92%. Each model has its unique advantages, such as the simplicity and interpretability of LogisticRegression or the flexibility of the MLPClassifier with its ability to capture complex patterns through neural networks.

**Note:** Training and testing in this part also make use of parallel computation, therefore there would be non-deterministic behaviours of the results.

## 2. Recommendation

Based on the evaluation metrics and training time, the RandomForestClassifier is recommended for heart failure classification. It achieved the highest accuracy, recall, F1 score, and ROC AUC score. Although it takes longer to train than some other models, its performance makes it suitable for this classification task.

For scenarios where training time is critical, GaussianNB can be considered as it offers a good balance of performance and efficiency, with the second-highest accuracy and the shortest training time.

# Exited model (PHONG SON – ITITIU18200)

1. Aiming

The purpose of this part is about training and testing the new dataset that we have split the old dataset into the new dataset by taking 10% of its data. From part IV, we have chosen out the model RandomForestClassifier which gives the highest accuracy value into the new dataset this time.

A screen shot of a black screen

Description automatically generated

Figure V-1 Table showing the result of the six diferent models

2. Application

Now we apply the RandomForestClassifier model to test our new dataset by running this code to separate the second data.

A computer screen shot of a code

Description automatically generated

Then this part of the code will show the results

# Apply the chosen model to predict a new data: RandomForest

prediction = rf.predict(X\_new)

# Accuracy Score on exited data

accuracy\_test = accuracy\_score(y\_new,prediction)

print('Accuracy score on test dataset : ', accuracy\_test)

And finally, this part of the code will classification report and the confusion matrix

# Print classification report

print("Classification Report:")

print(classification\_report(y\_new, prediction))

# Print confusion matrix

conf\_matrix = confusion\_matrix(y\_new, prediction)

plt.figure(*figsize*=(8, 6))

sns.heatmap(conf\_matrix, *annot*=True, *fmt*="d", *cmap*="Blues", *cbar*=False)

plt.xlabel("Predicted Labels")

plt.ylabel("True Labels")

plt.title("Confusion Matrix")

plt.show()

And here is the results of the test after applying the model and all of the remaining parts



A screenshot of a computer screen

Description automatically generatedFigure V-2 Accuracy score tested on the second dataset.

Figure V-3 Full evaluation tested on the second dataset.

A blue squares with white text

Description automatically generated

3. Conclusion

Although the result is not good compared to the old dataset but with the accuracy of 0.75, it is still useful for applying this method to any dataset. And without being said, the table still has the low value compared to the first dataset in the part IV so we would like to say that every model will show the different values and not of all them have the same accuracy and same predict. In another way that we can use ***Voting*** technique to allow us to combine multiple machine learning modules and use a majority vote or a weighted vote to make predictions. It’s a way to ensemble different models for potentially better performance. By this technique, the final result will be clearer than the traditional one.

# References

* tanmay111999. (2022). Heart Failure Prediction:CV Score(90%+): 5 Models. Retrieved from https://www.kaggle.com/code/tanmay111999/heart-failure-prediction-cv-score-90-5-models
* Fedesoriano. (2021). Heart Failure Prediction Dataset. Retrieved from https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction
* Géron, A. (2022). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (3rd ed.). [O’Reilly Media](https://raw.githubusercontent.com/data-science-projects-and-resources/Data-Science-EBooks/main/Machine%20Learning/Hands-on-Machine-Learning.pdf)
* learn. (n.d.). Retrieved from <https://scikit-learn.org/stable/>
* **GitHub Link**: <https://github.com/phamvutuyetanh/AI_Project.git>