DESIGNING HEART DISEASE PREDICTION DASHBOARD

PORTFOLIO MACHINE LEARNING

– ERIKA BUDIARTI –

- Heart Disease Project -

An alarming number of patients are diagnosed with various cardiac symptoms. As data scientists, we aim to create a Machine Learning model to train our AI in classifying patients as either positive or negative for heart disease.

Problem Identification

Objective:

The goals to predict the likelihood of a patient being diagnosed with heart disease, a task involving a binary outcome:

Positive (+) = 1: Indicates that the patient has been diagnosed with heart disease.

Negative (-) = 0: Denotes that the patient has not been diagnosed with heart disease.

We will explore multiple Classification Models and determine which one yields the highest accuracy. Our analysis will involve scrutinizing patterns, analyzing trends, and identifying correlations present in the dataset. The ultimate objective is to pinpoint the crucial features contributing to the positive/negative diagnosis of heart disease.

Data Understanding (1/2)

Data obtained from the "Heart Disease dataset" by UCIML.
(https://archive.ics.uci.edu/dataset/45/heart+disease)
Let's look at the first 5 rows and the last 5 rows

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
1020	59	1	1	140	221	0	1	164	1	0.0	2	0	2	1
1021	60	1	0	125	258	0	0	141	1	2.8	1	1	3	0
1022	47	1	0	110	275	0	0	118	1	1.0	1	1	2	0
1023	50	0	0	110	254	0	0	159	0	0.0	2	0	2	1
1024	54	1	0	120	188	0	1	113	0	1.4	1	1	3	0

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1025 entries, 0 to 1024
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1025 non-null   int64
 1   sex       1025 non-null   int64
 2   cp        1025 non-null   int64
 3   trestbps  1025 non-null   int64
 4   chol      1025 non-null   int64
 5   fbs       1025 non-null   int64
 6   restecg   1025 non-null   int64
 7   thalach   1025 non-null   int64
 8   exang     1025 non-null   int64
 9   oldpeak   1025 non-null   float64
 10  slope     1025 non-null   int64
 11  ca        1025 non-null   int64
 12  thal      1025 non-null   int64
 13  target    1025 non-null   int64
dtypes: float64(1), int64(13)
memory usage: 112.2 KB
```

DQLab

Data Understanding (2/2)

The target (Y) – “Positive or Negative diagnosis of Heart Disease” is determined by 13 features (X):

1. **age** – (*Continuous*)
2. **sex** – (*Binary*) : 1= Male, 0= Female
3. **cp** (chest pain) – (*Ordinal* with 4 values) – 1: typical angina 2: atypical angina 3: non-anginal pain 4: asymptomatic
4. **trestbps** (resting blood pressure) – (*Continuous*)
5. **chol** (serum cholesterol) in mg/dl – (*Continuous*)
6. **fbs** (fasting blood sugar > 120 mg/dl – (*Binary*) : 1 = true, 0 = false
7. **restecg** (resting electrocardiography results) – (*Ordinal* with 3 values) – 0: normal 1: ST-T wave abnormalities 2: left ventricular hypertrophy
8. **thalach** (maximum heart rate achieved) – (*Continuous*)
9. **exang** (exercise induced angina (binary) – (*Binary*) : 1 = yes; 0 = no
10. **oldpeak** (ST depression induced by exercise relative to rest) – (*Continuous*)
11. **slope** (slope of the peak exercise ST segment – (*Ordinal* with 3 values) – 1: up sloping 2: flat 3: down sloping)
12. **ca** (number of major vessels) – (*Ordinal* with 4 values) colored by fluoroscopy – 0, 1, 2, 3
13. **thal** (result of thallium test) — (*Ordinal* with 3 values) – 1: normal 2: fixed defect 3: reversible defect

live Class

DQLab

The Tools Used

Google Colaboratory
Tools for data analysis, machine learning, deep learning that allows you to execute Python code.

Python
An interpreted, object-oriented, high-level programming language.

Visual Studio Code
Powerful source code editor runs on your desktop and on the web.

Anaconda
An open-source distribution used for data science, machine learning, deep learning, etc.

NumPy
NumPy is a general-purpose array-processing library.

pandas
Pandas is a powerful, flexible and easy to use data analysis and manipulation tool library.

matplotlib
Matplotlib is a comprehensive library for creating static and interactive visualizations.

seaborn
Seaborn is a library for data visualization and exploratory data analysis.

scikit learn
Scikit-Learn is an efficient library for predictive data analysis and machine learning.

Streamlit
Streamlit is a library to create and share beautiful, custom web apps for machine learning.

live Class

Data Quality (1/3)

Checking Categorical Data Information

```
for i in categorical_col:
    print("Feature {} with {}".format(i, data[i].unique()))
    print()
```

Output:

```
Feature sex with ['Male' 'Female']

Feature cp with ['typical angina' 'atypical angina' 'non-anginal pain' 'asymtomatic']

Feature fbs with ['No' 'Yes']

Feature restecg with ['normal' 'probable or definite left ventricular hypertrophy'
 'ST-T Wave abnormal']

Feature exang with ['No' 'Yes']

Feature slope with ['upsloping' 'downsloping' 'flat']

Feature ca with ['Number of major vessels: 2' 'Number of major vessels: 0'
 'Number of major vessels: 1' 'Number of major vessels: 3' 4]

Feature thal with ['reversable defect' 'fixed defect' 'normal' 0]

Feature target with ['No disease' 'Disease']
```

Handling Variable:

- Feature 'ca' has 5 values ranging from 0 to 4, therefore the value “4” is substituted with “NaN” (as it shouldn't exist) - compare to the explanation on Data Understanding section
- Feature 'thal' has 4 values ranging from 0 to 3, therefore the value 0 is substituted with NaN (as it shouldn't exist) - compare to the explanation on Data Understanding section

Data Quality (2/3)

Checking Missing Value

```
data.isna().sum()
```

Output:

```
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
```

Result :
No Missing Value

Checking Duplicate Data

```
data.duplicated().sum()
```

Output:

```
723
```

Drop Duplicate Data

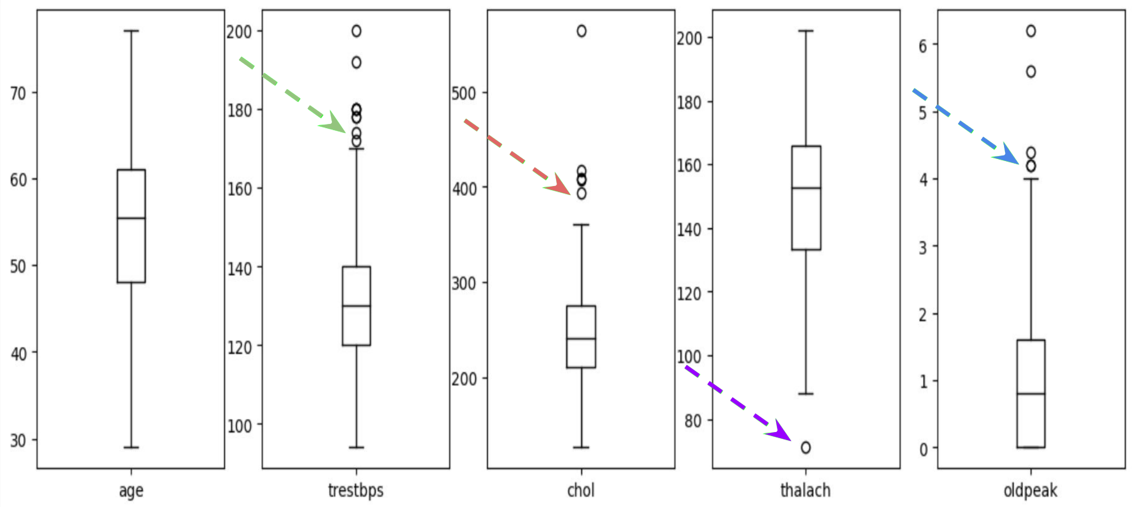
```
data.drop_duplicates(keep="first", inplace=True)
```

Data Quality (3/3)

Checking Outlier Using Box Plot

```
data.plot(kind = 'box',
          subplots = True,
          layout = (2,7),
          sharex = False,
          sharey = False,
          figsize = (20, 10),
          color = 'k')
plt.show()
```

Output:

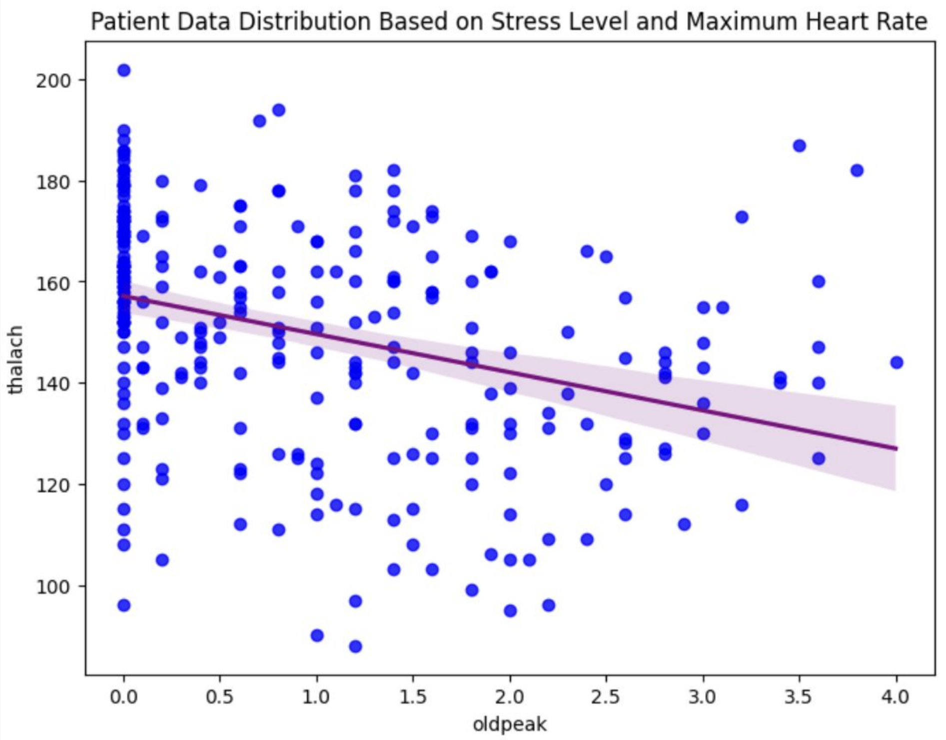


Data that has outliers can be replaced using:

- Maximum Value: Q3 + 1.5 IQR
- Minimum Value: Q1 - 1.5 IQR

Data Analysis (1/5)

Correlation between oldpeak and thalach



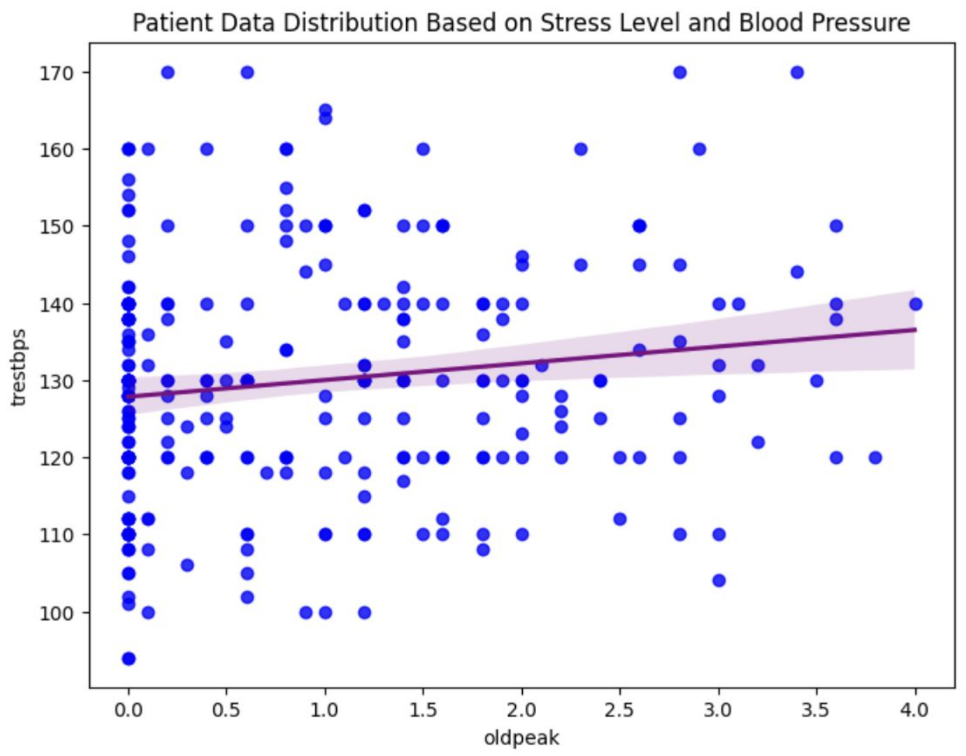
From the scatter plot, we can find the Negative Correlation between patient stress level (oldpeak) and maximum heart rate (thalach).

It means that as stress level increases, the maximum heart rate tends to decrease, and vice versa. This relationship can be explained by the physiological response of the body to stress. Stress, particularly psychological or emotional stress, triggers the release of stress hormones such as cortisol and adrenaline (epinephrine). These hormones prepare the body for a "fight or flight" response, which is an evolutionary adaptation to handle threats or challenges. However, chronic or prolonged stress can have negative effects on the body, including the cardiovascular system.

DQLab

Data Analysis (2/5)

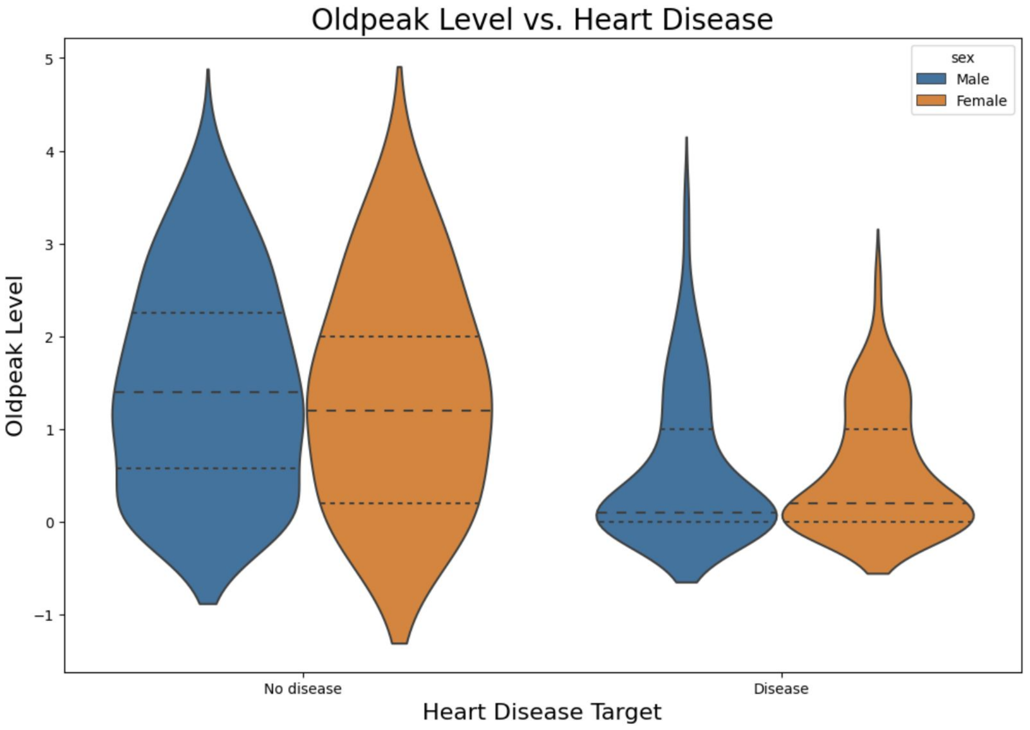
Correlation between oldpeak and trestbps



From the scatter plot, we can find the Positive Correlation between patient stress level (oldpeak) and blood pressure (trestbps). When a person experiences stress, whether it's due to emotional, psychological, or physical factors, their body undergoes a series of changes known as the "fight or flight" response. However, when this stress response is triggered frequently or for prolonged periods due to chronic stress, it can lead to consistent elevation of blood pressure. The increased heart rate and the narrowed blood vessels result in higher resistance against blood flow, which in turn leads to higher blood pressure levels.

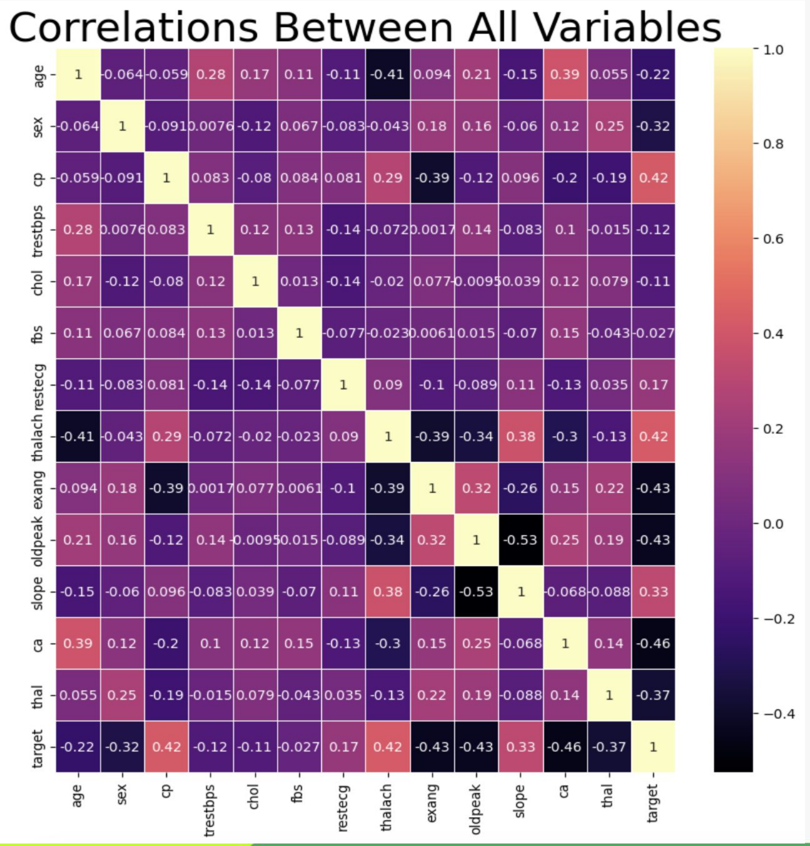
Data Analysis (3/5)

Correlation between oldpeak level (ST depression) and heart disease



We can see that the overall shape & distribution for negative & positive patients differ vastly. Positive patients exhibit a lower median for ST depression level & thus a great distribution of their data is between 0 & 2, while negative patients are between 1 & 3. In addition, we don't see many differences between male & female target outcomes.

Data Analysis (4/5)



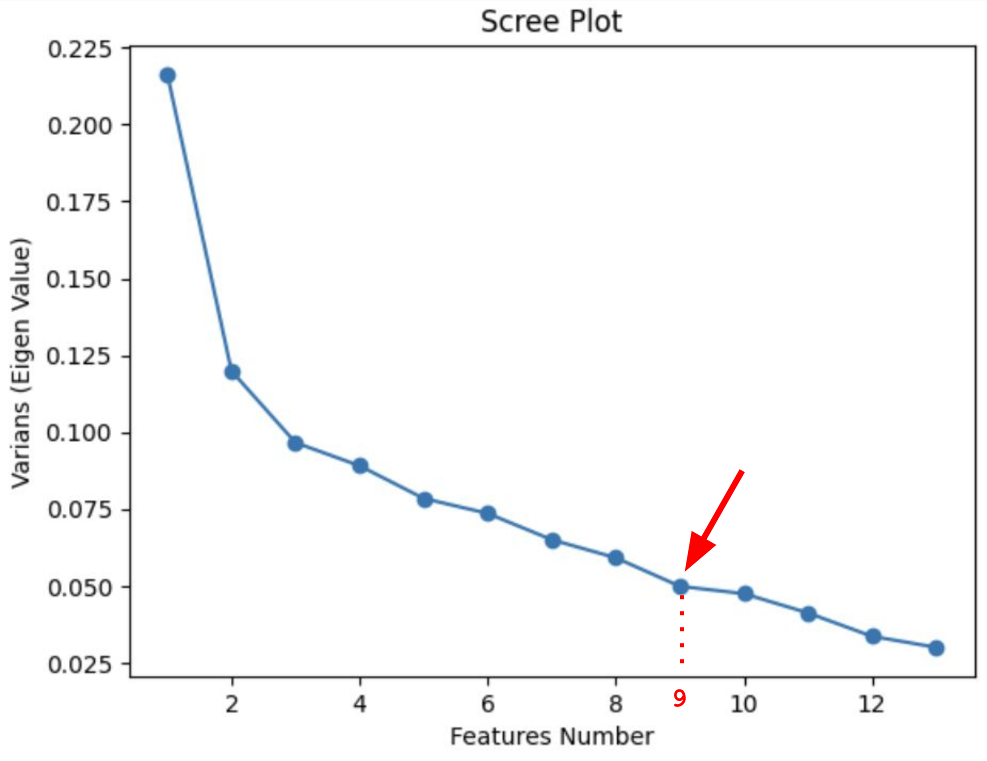
Looking for the positive correlation and negative correlation

----

The strongest positive correlation is between heart rate (thalach) and target. This makes sense because elevated heart rates are often associated with increased sympathetic nervous system activity, chronic stress, hypertension, inflammation, arrhythmias, and metabolic imbalances. These factors contribute to the development and progression of heart diseases.

The strongest negative correlation is between number of major blood vessel (ca) and target. This makes sense because as the number of major vessels with blockages increases, individuals tend to respond more aggressively to mitigate the risk of heart disease-related complications. This includes adopting healthier lifestyles, adhering to prescribed medications, etc.

Data Analysis (5/5)



Among the 13 features in this dataset, we have identified that 9 features show strong correlation, while 4 features show weak correlation. As the next step, we will proceed Data Modelling with 9 features.

DQLab

Data Modelling (1/9)

Model 1 : "Logistic Regression"

The Classification Report of Logistic Regression Classifier

	precision	recall	f1-score	support
0	0.73	0.83	0.78	23
1	0.87	0.79	0.83	34
accuracy			0.81	57
macro avg	0.80	0.81	0.80	57
weighted avg	0.81	0.81	0.81	57

Data Modelling (2/9)

Model 2 : "Random Forest"

The Classification Report of Random Forest Classifier

	precision	recall	f1-score	support
0	0.82	0.78	0.80	23
1	0.86	0.88	0.87	34
accuracy			0.84	57
macro avg	0.84	0.83	0.83	57
weighted avg	0.84	0.84	0.84	57

DQLab

Data Modelling (3/9)

Model 3 : “Decision Tree”

The Classification Report of Decision Tree Classifier

	precision	recall	f1-score	support
0	0.75	0.65	0.70	23
1	0.78	0.85	0.82	34
accuracy			0.77	57
macro avg	0.77	0.75	0.76	57
weighted avg	0.77	0.77	0.77	57

live Class

Data Modelling (4/9)

Model 4 : “K-Nearest Neighbors”

The Classification Report of K_Nearest Neighbors Classifier

	precision	recall	f1-score	support
0	0.77	0.74	0.76	23
1	0.83	0.85	0.84	34
accuracy			0.81	57
macro avg	0.80	0.80	0.80	57
weighted avg	0.81	0.81	0.81	57

Data Modelling (5/9)

Model 5 : “Support Vector Machine”

The Classification Report of Support Vector Machine Classifier

	precision	recall	f1-score	support
0	0.76	0.83	0.79	23
1	0.88	0.82	0.85	34
accuracy			0.82	57
macro avg	0.82	0.82	0.82	57
weighted avg	0.83	0.82	0.83	57

Data Modelling (6/9)

Model 6 : “Naives Bayes”

The Classification Report of Naives Bayes Classifier

	precision	recall	f1-score	support
0	0.70	0.83	0.76	23
1	0.87	0.76	0.81	34
accuracy			0.79	57
macro avg	0.79	0.80	0.79	57
weighted avg	0.80	0.79	0.79	57

DQLab

Data Modelling (7/9)

Model 7 : "XG Boost"

The Classification Report of XG Boost Classifier

	precision	recall	f1-score	support
0	0.75	0.78	0.77	23
1	0.85	0.82	0.84	34
accuracy			0.81	57
macro avg	0.80	0.80	0.80	57
weighted avg	0.81	0.81	0.81	57

live Class

DQLab

Data Modelling (8/9)

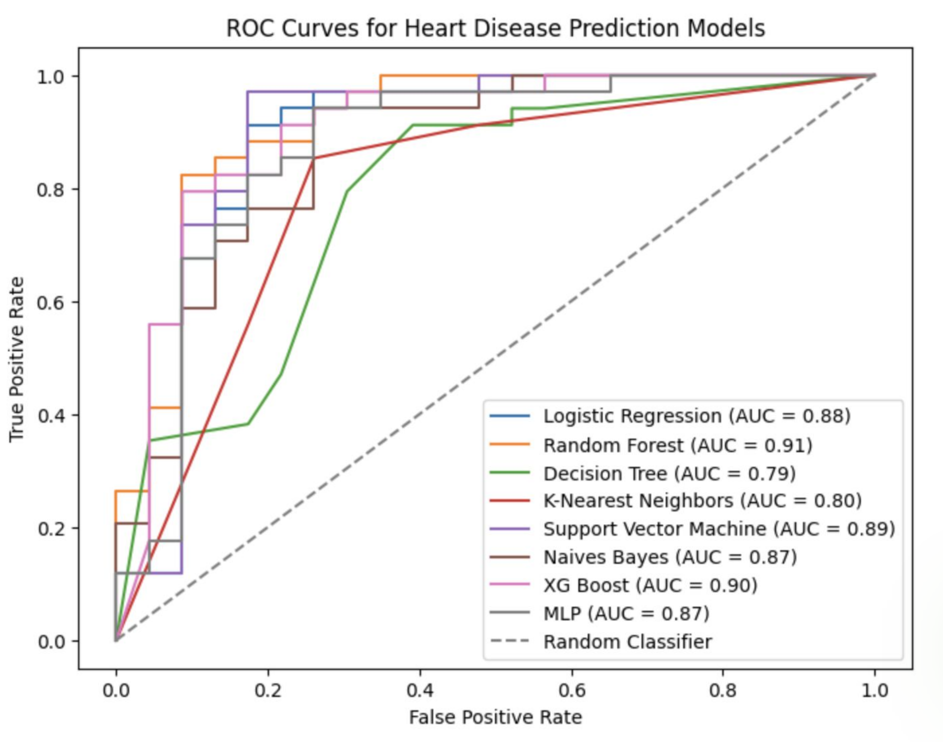
Model 8 : “MLP Classifier”

The Classification Report of MLP Classifier

	precision	recall	f1-score	support
0	0.70	0.83	0.76	23
1	0.87	0.76	0.81	34
accuracy			0.79	57
macro avg	0.79	0.80	0.79	57
weighted avg	0.80	0.79	0.79	57

Data Modelling (9/9)

AUC - ROC Analysis




```
AUC-ROC for Logistic Regression: 0.8849104859335037
AUC-ROC for Random Forest: 0.9143222506393862
AUC-ROC for Decision Tree: 0.7851662404092071
AUC-ROC for K-Nearest Neighbors: 0.8005115089514067
AUC-ROC for Support Vector Machine: 0.8938618925831202
AUC-ROC for Naives Bayes: 0.870843989769821
AUC-ROC for XGBoost: 0.9028132992327366
AUC-ROC for MLP: 0.8721227621483376
```

The best performance of this Machine Learning Model is "***the Random Forest***", indicated by the highest accuracy value in the classification report, which is 84%, and the largest AUC-ROC score compared to the other 7 models, which is 91%.

DQLab

Interactive Dashboard

liveClass

Deploy
Machine Learning Model

Streamlit

User Input Features:

Upload your input CSV file

Drag and drop file here

Limit 200MB per file • CSV

Browse files

Manual Input

Chest pain type

2

1 4

Type of Chest pain : Atypical angina

Maximum heart rate achieved

80

71 202

Slope of the peak exercise ST segment

1

0 2

ST depression induced

1.00

0.00 6.20

Exercise induced angina

1

0 1

Number of major vessels

1

0 3

Result of thallium test

1

1 3

Sex

Female

Age

30

29 77

Click Here To Predict

Welcome to ERIKA's Machine Learning Dashboard

This app predicts the Heart Disease.

Data obtained from the Heart Disease dataset by UCIML.

man-heart-attack.jpg

woman-heart-attack.jpg

	cp	thalach	slope	oldpeak	exang	ca	thal	sex	age
0	2	80	1	1	1	1	1	0	30

Prediction:

Prediction of this app is Yes Heart Disease

Conclusion

1. Among the 13 distinct attributes that were under scrutiny, our analysis identified the leading 9 features that played a pivotal role in distinguishing between positive and negative diagnoses. These differentiating features encompassed chest pain type (referred to as 'cp'), maximum heart rate achieved during exercise (referred to as 'thalach'), count of major blood vessels (referred to as 'ca'), the extent of ST depression induced by exercise in comparison to the resting state (referred to as 'oldpeak'), slope of the peak exercise ST segment (referred to as 'slope'), exercise induced angina (referred to as 'exang'), the result of thallium test (referred to as 'thal'), sex and age.

2. Our machine learning algorithm has now reached a proficient stage where it can accurately categorize patients afflicted with Heart Disease. This breakthrough enables us to provide precise diagnoses and facilitate the delivery of timely intervention and care that is essential for patient recovery. The ability to identify these crucial indicators at an early stage holds the potential to avert the escalation of symptoms and the emergence of more severe conditions in the future.

3. In the realm of accuracy metrics, our implementation of the Random Forest algorithm showcased an impressive performance, achieving a notable accuracy rate of 84%. While a threshold of 70% accuracy is generally considered commendable, it's essential to exercise caution, as excessively high accuracy values could indicate a phenomenon known as overfitting, where the model becomes too tailored to the training data. Therefore, an accuracy range of 70% to 80% is regarded as the optimal balance that indicates reliable model performance.

ERIKA BUDIARTI

LinkedIn :

linkedin.com/in/erika-budiarti

Assistant Finance Director
PT. ANDRO PERKASA MULIA
2023 - Present

Thanks!