Business Data Understanding

Penyakit kardiovaskular (CVD) merupakan penyebab kematian nomor 1 secara global, merenggut sekitar 17,9 juta nyawa setiap tahun, yang menyumbang 31% dari semua kematian di seluruh dunia [1]. penyakit kardiovaskular adalah penyakit yang disebabkan gangguan fungsi jantung dan pembuluh darah. Ada banyak macam penyakit kardiovaskuler, salah satunya adalah penyakit jantung (Kemenkes RI, 2014). Empat dari lima kematian CVD kematian disebabkan oleh serangan jantung dan stroke, dan sepertiga dari kematian ini terjadi sebelum waktunya pada orang di bawah 70 tahun.

Gagal jantung atau heart failure (HF) adalah masalah kesehatan masyarakat global yang mempengaruhi jutaan orang. Orang dengan penyakit kardiovaskular atau yang berisiko kardiovaskular tinggi (karena adanya satu atau lebih faktor risiko seperti hipertensi, diabetes, hiperlipidemia) memerlukan deteksi dini dan manajemen di mana model machine learning dapat sangat membantu [2].

Diagnosis penyakit jantung melalui riwayat medis tradisional telah dianggap tidak dapat diandalkan dalam banyak aspek. Untuk mengklasifikasikan orang sehat dan orang dengan penyakit jantung, metode berbasis noninvasif seperti machine learning dapat diandalkan dan efisien. Dalam studi yang diusulkan, use case kali ini mengembangkan sistem diagnosis berbasis machine learning untuk prediksi penyakit jantung dengan menggunakan dataset penyakit jantung [3].

Sehingga berangkat dari latar belakang tersebut buatlah model machine learning untuk memprediksi apakah seseorang beresiko terhadap gagal jantung berdasarkan pada beberapa atribut. Logistic Regression dan random forest akan menjadi model dalam pembuatan machine learning memprediksi gagal jantung.

Data Understanding

A. Informasi Data Set
1. Deskripsi atribut

```
heart.columns
```

```
Index(['Age', 'Sex', 'ChestPainType', 'RestingBP', 'Cholesterol', 'FastingBS',
       'RestingECG', 'MaxHR', 'ExerciseAngina', 'Oldpeak', 'ST_Slope',
       'HeartDisease'],
      dtype='object')
```

Attributes	Description
Age	Usia pasien [Tahun]
Sex	Jenis kelamin pasien [M: Pria, F: Wanita]
ChestPainType	Jenis nyeri dada dimana ada empat jenis nilai berbeda [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic] yang ditentukan untuk atribut ini, setiap nilai menggambarkan tingkat nyeri dada.
RestingBP	Tekanan darah [mm Hg]

Cholesterol	Kolom ini menunjukkan kadar kolesterol yang pasien [mm/dl]
FastingBS	Atribut selanjutnya adalah menggambarkan kadar gula darah pada pasien [1: if FastingBS > 120 mg/dl, 0: otherwise]
RestingECG	Parameter ini menunjukkan hasil resting electrocardiogram [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]
MaxHR	Detak jantung maksimum tercapai [Numeric value between 60 and 202].
ExerciseAngina	Parameter ini digunakan untuk memahami tentang, apakah olahraga menginduksi angina atau tidak [Y: Yes, N: No].
Oldpeak	Atribut selanjutnya adalah mendefinisikan status depresi pasien.
ST_Slope	Kondisi pasien selama latihan puncak. Nilai ini didefinisikan menjadi tiga segmen [Up: upsloping, Flat: flat, Down: downsloping].
HeartDisease	output class [1: heart disease, 0: Normal]

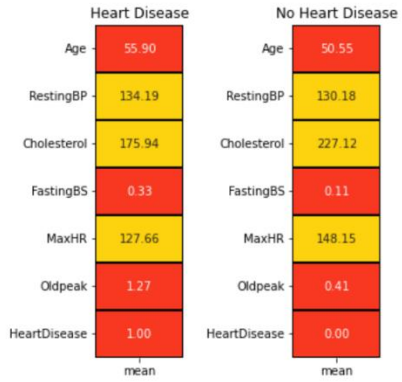
2. Melakukan pengecekan tipe data dan null values

```
heart.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 918 entries, 0 to 917
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   Age             918 non-null    int64  
 1   Sex             918 non-null    object 
 2   ChestPainType   918 non-null    object 
 3   RestingBP       918 non-null    int64  
 4   Cholesterol     918 non-null    int64  
 5   FastingBS       918 non-null    int64  
 6   RestingECG      918 non-null    object 
 7   MaxHR           918 non-null    int64  
 8   ExerciseAngina  918 non-null    object 
 9   Oldpeak         918 non-null    float64
 10  ST_Slope        918 non-null    object 
 11  HeartDisease    918 non-null    int64  
dtypes: float64(1), int64(6), object(5)
memory usage: 86.2+ KB
```

Semua kolom sudah memiliki tipe data yang sesuai dan tidak ada null values yang ada dalam data.

3. Describe data



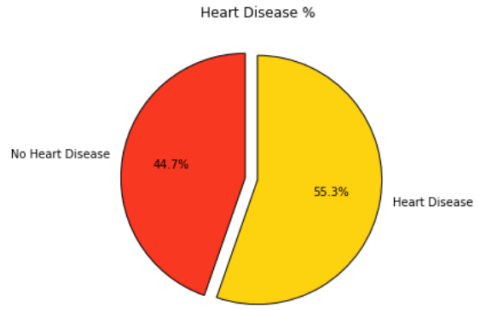
Nilai rata-rata dari semua fitur untuk kasus penyakit jantung dan penyakit non-jantung.

B. Analisis Data Eksploratif

1. Membagi fitur menjadi Numerik dan Kategori :

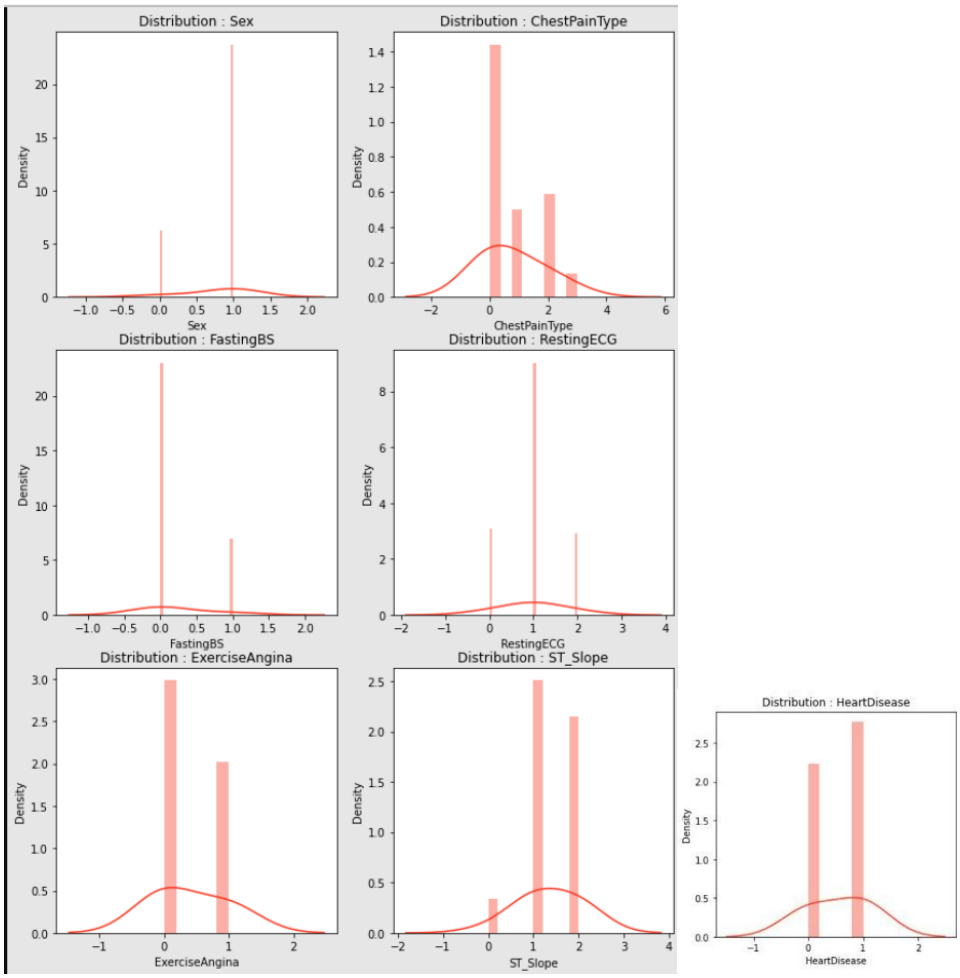
```
Categorical Features : Sex ChestPainType FastingBS RestingECG ExerciseAngina ST_Slope HeartDisease
Numerical Features : Age RestingBP Cholesterol MaxHR Oldpeak
```

2. Visualisasi perbandingan nilai variabel Target (HeartDisease)



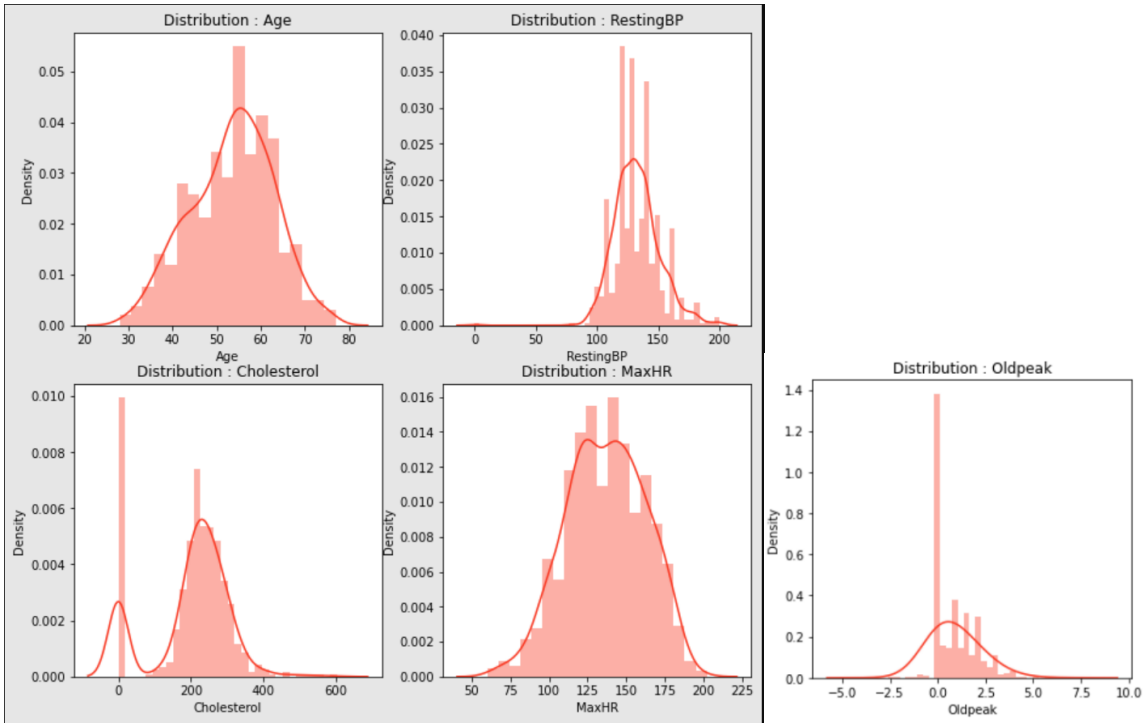

```
1    508
0    410
Name: HeartDisease, dtype: int64
```

3. Distribution of Categorical Features

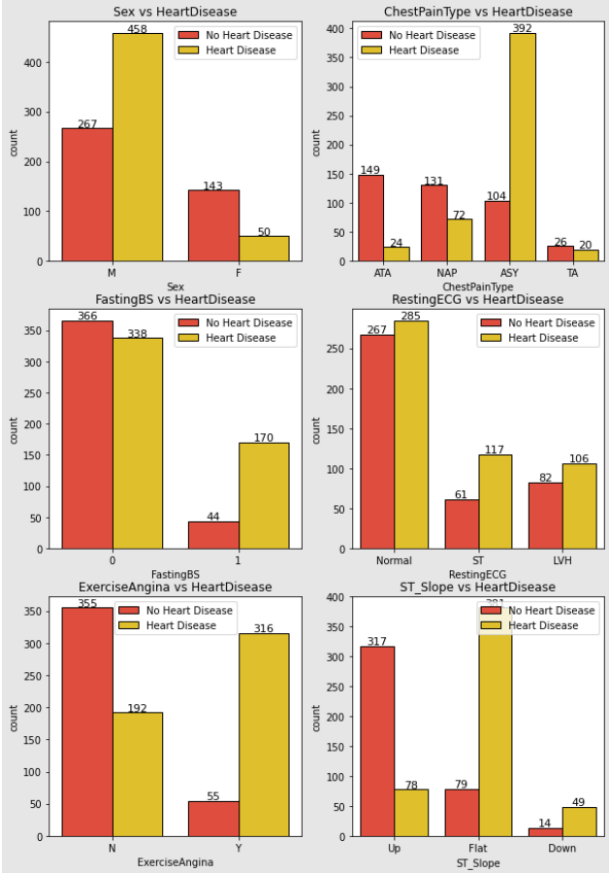


Semua fitur kategoris sudah dekat tentang Didistribusikan Secara Normal.

4. Distribusi Fitur Numerik



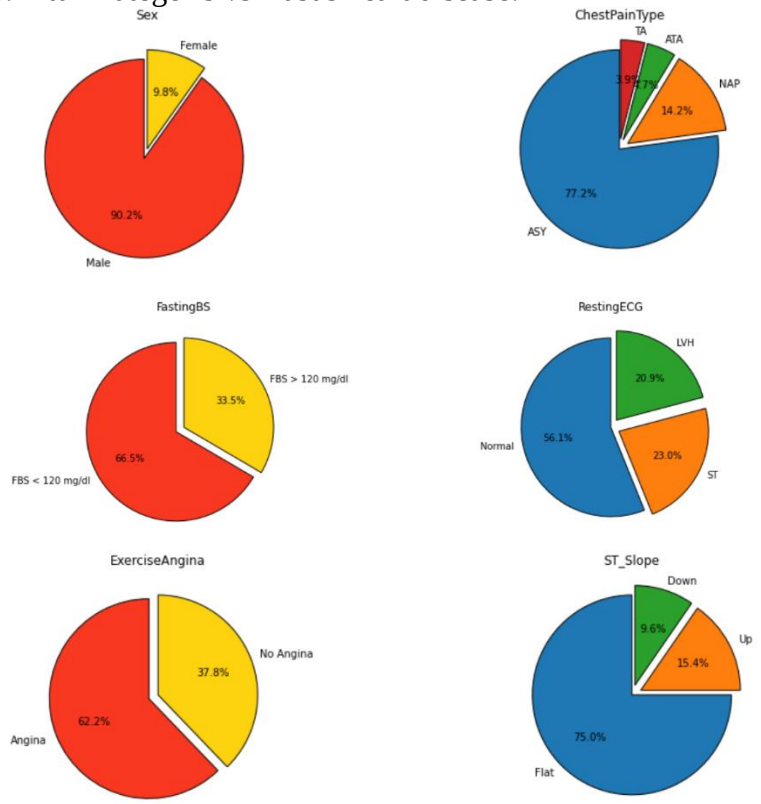
5. Fitur Kategoris vs Variabel Target (HeartDisease)



1. Jumlah pria memiliki lebih banyak pasien heart disease daripada no heart disease. Dalam kasus populasi Wanita, pasien heart disease lebih sedikit dari pasien no heart disease.
2. Jenis nyeri dada ASY menunjukkan kemungkinan besar penyakit jantung.

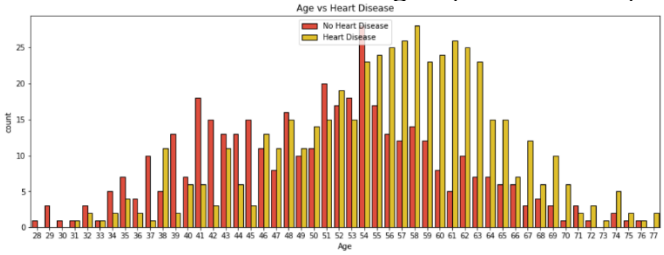
3. Jika dibandingkan antara no heart disease dan heart disease, Pasien yang di diagnosis dengan Fasting Blood Sugar dan no Fasting Blood Sugar memiliki pasien penyakit jantung yang cukup besar.
4. RestingECG tidak menampilkan dengan jelas kategori yang mempengaruhi potensi seseorang memiliki penyakit jantung karena semua nilainya menunjukan angka yang besar.
5. Dengan nilai ST_Slope, grafik flat menunjukkan kemungkinan yang sangat tinggi untuk didiagnosis dengan penyakit jantung. Grafik down juga menunjukkan output yang sama tetapi versi sedikit data.

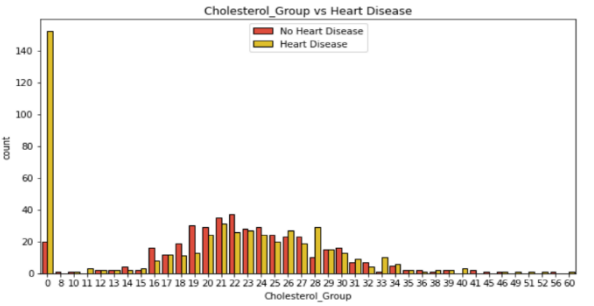
6. Fitur Kategoris vs Kasus heart disease:

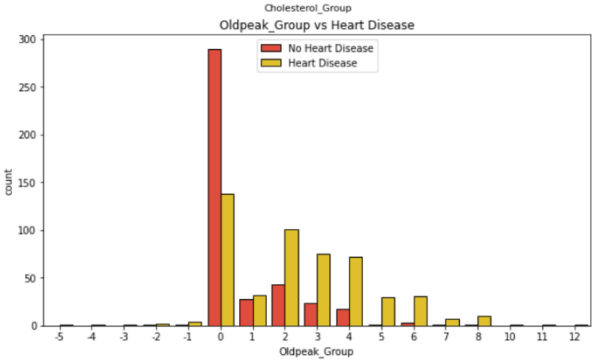


1. 90% pasien penyakit jantung adalah laki-laki.
2. Kolom chest pain type, tipe ASY 77% yang menyebabkan penyakit jantung.
3. Fasting Blood Sugar level < 120 mg/dl menunjukan kemungkinan besar penyakit jantung.
4. RestingECG, level normal menunjukan 56% kemungkinan penyakit jantung.Llebih besar dibandingkan level LVH dan ST.
5. Exercise Angina juga menunjuk ke arah penyakit jantung.
6. Data ST_Slope, level flat memperoleh 75% yang dapat membantu dalam mendeteksi masalah jantung.

7. Fitur numerik vs variabel target (HeartDisease) :







1. Kadar kolesterol antara 160 (16x10) - 340 (34x10) sangat rentan terhadap penyakit jantung.
2. Nilai oldpeak 0-4 menunjukkan probabilitas tinggi untuk didiagnosis dengan penyakit jantung.

Data Preparation

1. Resampling nilai agar sama rata dengan cara oversampling

```
1    410
0    410
Name: HeartDisease, dtype: int64
```

2. Lakukan One Hot Encoding

	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease	Sex_F	Sex_M	ChestPainType_ASY	ChestPainType_ATA	ChestPainType_NAP
0	40	140	289	0	172	0.0	0	0	1	0	1	0
1	49	160	180	0	156	1.0	1	1	0	0	0	1
2	37	130	283	0	98	0.0	0	0	1	0	1	0
3	48	138	214	0	108	1.5	1	1	0	1	0	0
4	54	150	195	0	122	0.0	0	0	1	0	0	1

ChestPainType_NAP	ChestPainType_TA	RestingECG_LVH	RestingECG_Normal	RestingECG_ST	ExerciseAngina_N	ExerciseAngina_Y	ST_Slope_Down	ST_Slope
0	0	0	1	0	1	0	0	
1	0	0	1	0	1	0	0	
0	0	0	0	1	1	0	0	
0	0	0	1	0	0	1	0	
1	0	0	1	0	1	0	0	

ExerciseAngina_Y	ST_Slope_Down	ST_Slope_Flat	ST_Slope_Up
0	0	0	1
0	0	1	0
0	0	0	1
1	0	1	0
0	0	0	1

Modeling

1. Bagikan data yang jadi bahan prediksi dengan target dengan menggunakan sumbu X dan y

```
X = heart.loc[:, heart.columns != 'HeartDisease']
y = heart["HeartDisease"]
```

2. Split sumbu X and y menjadi data train dan data test

```
# split sumbu X and y menjadi data train dan data test
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.4,
                                                    random_state=1)
```

3. Lakukan modelling

a. Logistic Regression :

```
from sklearn.linear_model import LogisticRegression

#Logistic Regression menggunakan Limited-memory Broyden-Fletcher-Goldfarb-Shanno Solver
lr = LogisticRegression(solver='lbfgs', max_iter=1000) #max_iter = maksimal iterasi
lr = lr.fit(X_train, y_train)
```

B. Random Forest

```
from sklearn.ensemble import RandomForestClassifier

classifier_rf = RandomForestClassifier(random_state=42, n_jobs=-1, max_depth=5,
                                       n_estimators=100, oob_score=True)

classifier_rf.fit(X_train, y_train)
```

```
RandomForestClassifier(max_depth=5, n_jobs=-1, oob_score=True, random_state=42)
```

Evaluation

```
# Logistic Regression
y_lr = lr.predict(X_test)

# Random Forest Awal
y_rf_before = classifier_rf.predict(X_test)
```

#Evaluasi Menggunakan Confusion Matrix

```
Logistic Regression :
 [[133  20]
 [ 29 186]]
Random Forest Awal :
 [[128  25]
 [ 18 197]]
```

#Evaluasi Menggunakan AUC

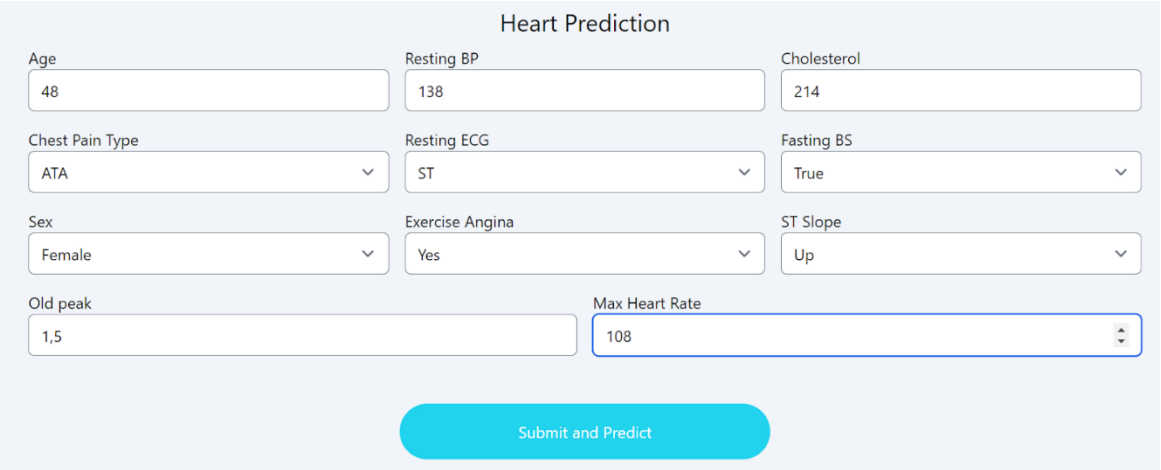
```
Logistic Regression :  0.8671986624107006
Random Forest Awal :  0.8764401884784923
```

#Evaluasi Menggunakan Classification Report

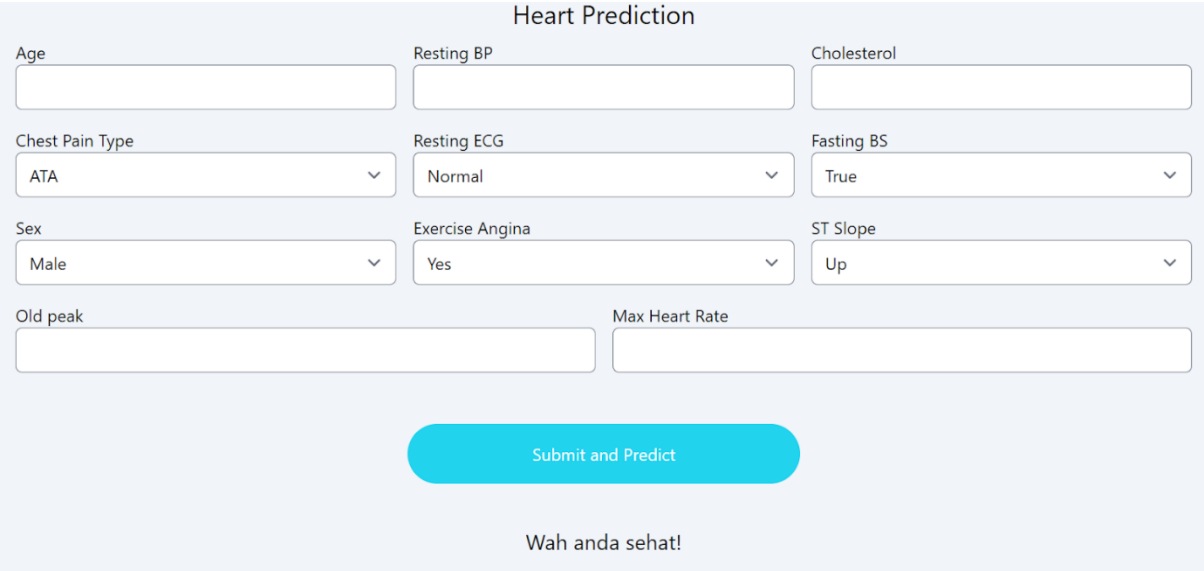
```
Logistic Regression :

              precision    recall  f1-score   support

           0       0.82      0.87      0.84       153
           1       0.90      0.87      0.88       215

    accuracy                           0.87       368
   macro avg       0.86      0.87      0.86       368
weighted avg       0.87      0.87      0.87       368


Random Forest Awal :

              precision    recall  f1-score   support

           0       0.88      0.84      0.86       153
           1       0.89      0.92      0.90       215

    accuracy                           0.88       368
   macro avg       0.88      0.88      0.88       368
weighted avg       0.88      0.88      0.88       368
```

Conclusion
Setelah melakukan serangkaian metode evaluasi, bisa disimpulkan bahwa model dengan skor terbaik adalah model yang dilakukan oleh **Random Forest Awal.**

Deployment
(Heart Disease Detector (edelindell.pythonanywhere.com)



Pengisian data



Setelah di klik submit and predict akan keluar hasilnya

Referensi

[1] Alotaibi, F. S. (2019). Implementation of machine learning model to predict heart failure disease. *International Journal of Advanced Computer Science and Applications*, *10*(6).

[2] Awan, S. E., Bennamoun, M., Sohel, F., Sanfilippo, F. M., & Dwivedi, G. (2019). Machine learning-based prediction of heart failure readmission or death: implications of choosing the right model and the right metrics. *ESC heart failure*, *6*(2), 428-435.

[3] Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., & Sun, R. (2018). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. *Mobile Information Systems*, *2018*.