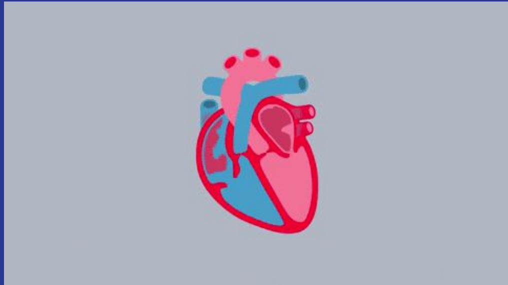
Heart Failure Prediction

IBM Supervised Machine Learning: Classification

Dam Minh Tien



For future reading, please go to my personal repository about IBM Machine Learning courses:

Github: damminhtien / machine-learning-ibm

Main objectives

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worlwide.

Heart failure is a common event caused by CVDs and this dataset contains 12 features that can be used to predict mortality by heart failure.

Most cardiovascular diseases can be prevented by addressing behavioural risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol using population-wide strategies.

People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

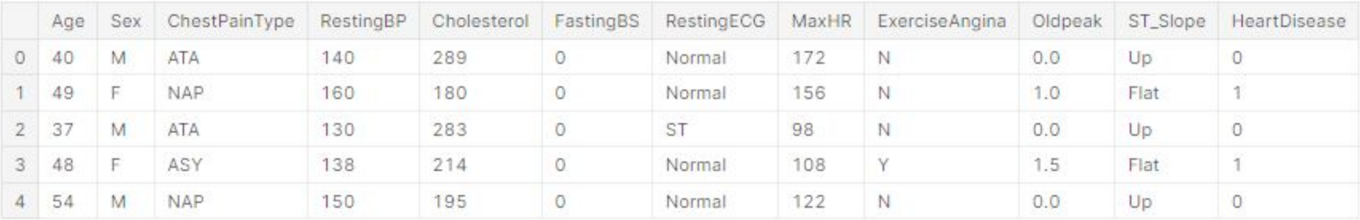
Dataset brief description

Source: [Kaggle | Heart failure clinical data]

Details: Dataset from Davide Chicco, Giuseppe Jurman: Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making 20, 16 (2020).

Dataset contains 12 clinical features por predicting death events.

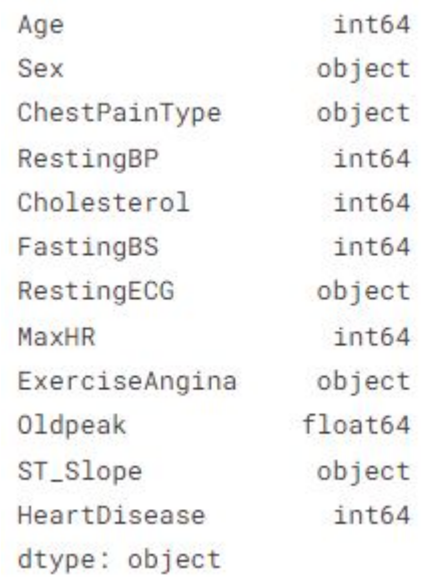
Top 5 rows:



	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0

Dataset brief description

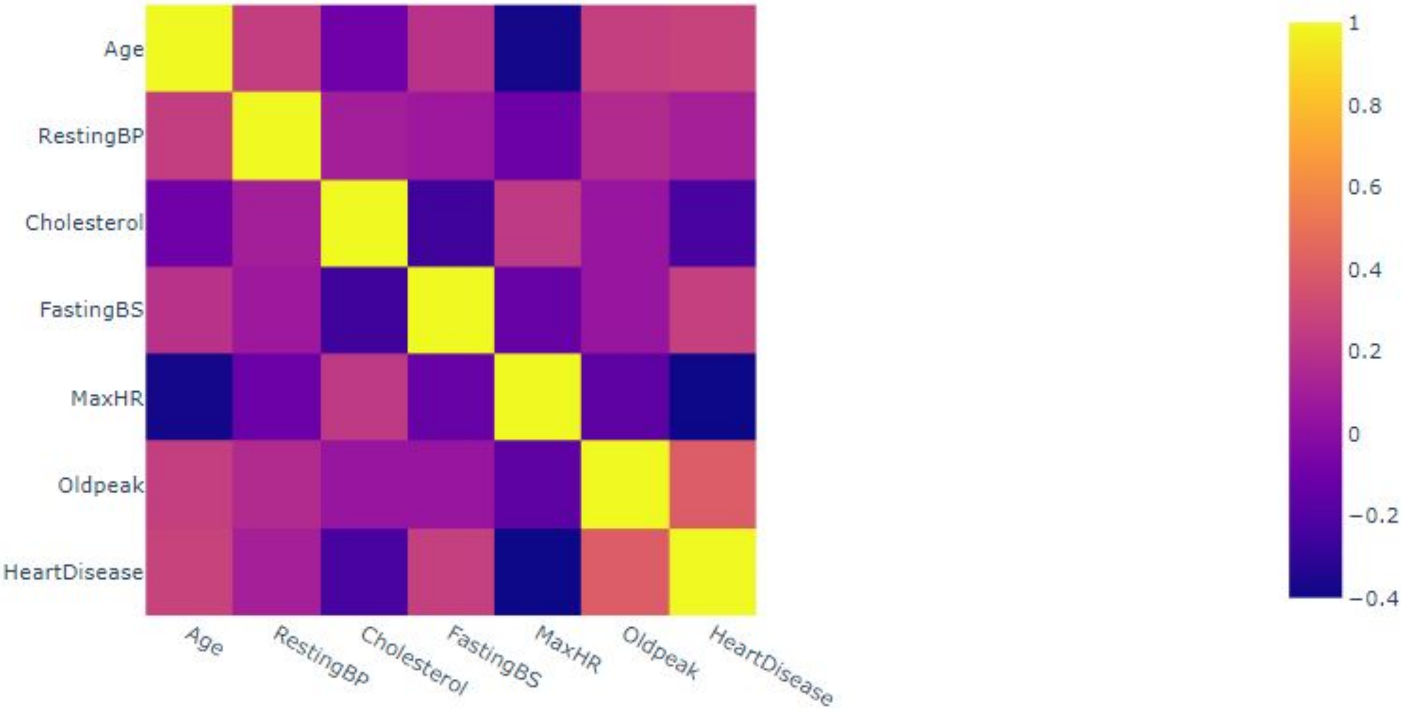
Column details



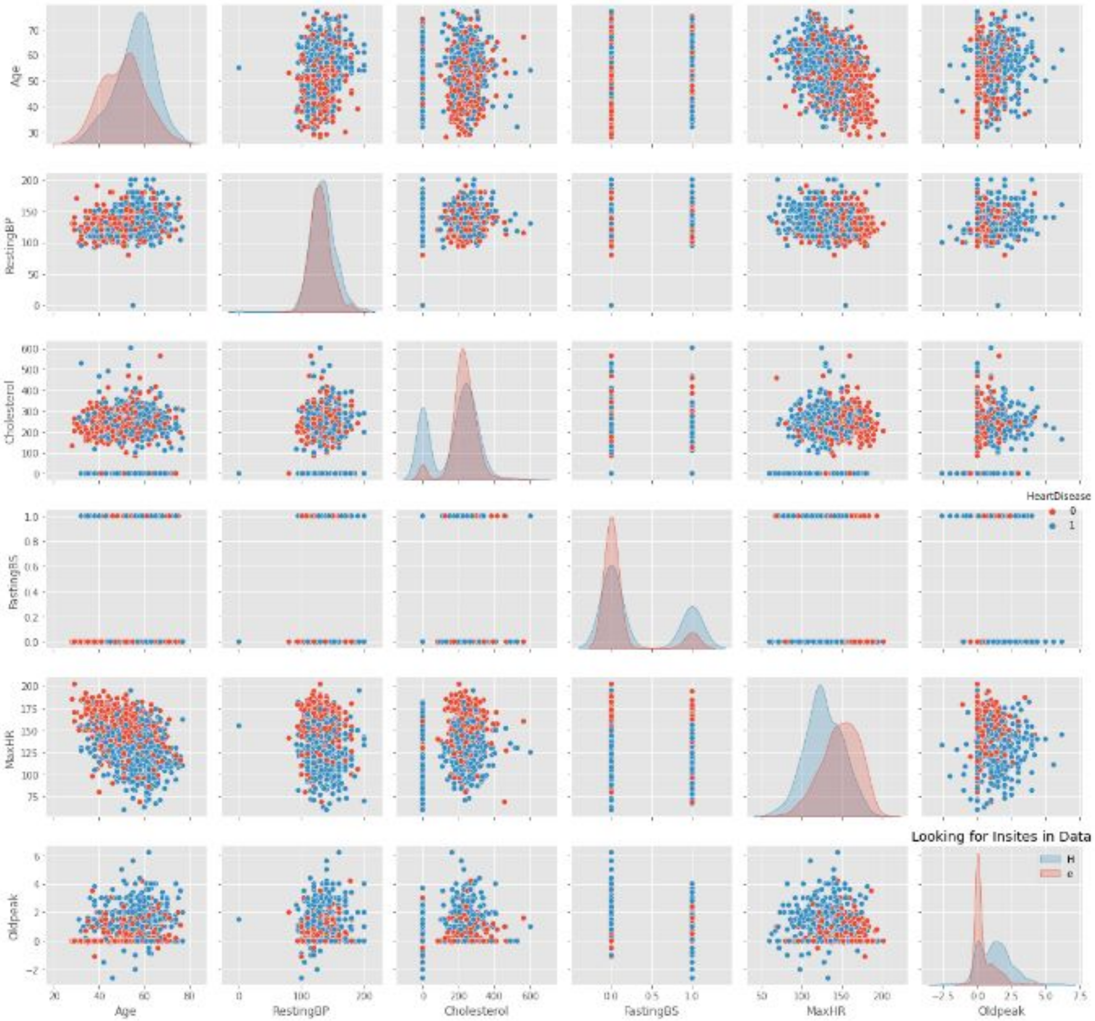
```
Age                 int64
Sex                object
ChestPainType      object
RestingBP           int64
Cholesterol         int64
FastingBS           int64
RestingECG         object
MaxHR               int64
ExerciseAngina     object
Oldpeak           float64
ST_Slope           object
HeartDisease        int64
dtype: object
```

1. Age: age of the patient [years]
2. Sex: sex of the patient [M: Male, F: Female]
3. ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. RestingBP: resting blood pressure [mm Hg]
5. Cholesterol: serum cholesterol [mm/dl]
6. FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7. RestingECG: 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]
8. MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
9. ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
10. Oldpeak: oldpeak = ST [Numeric value measured in depression]
11. ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12. HeartDisease: output class [1: heart disease, 0: Normal]

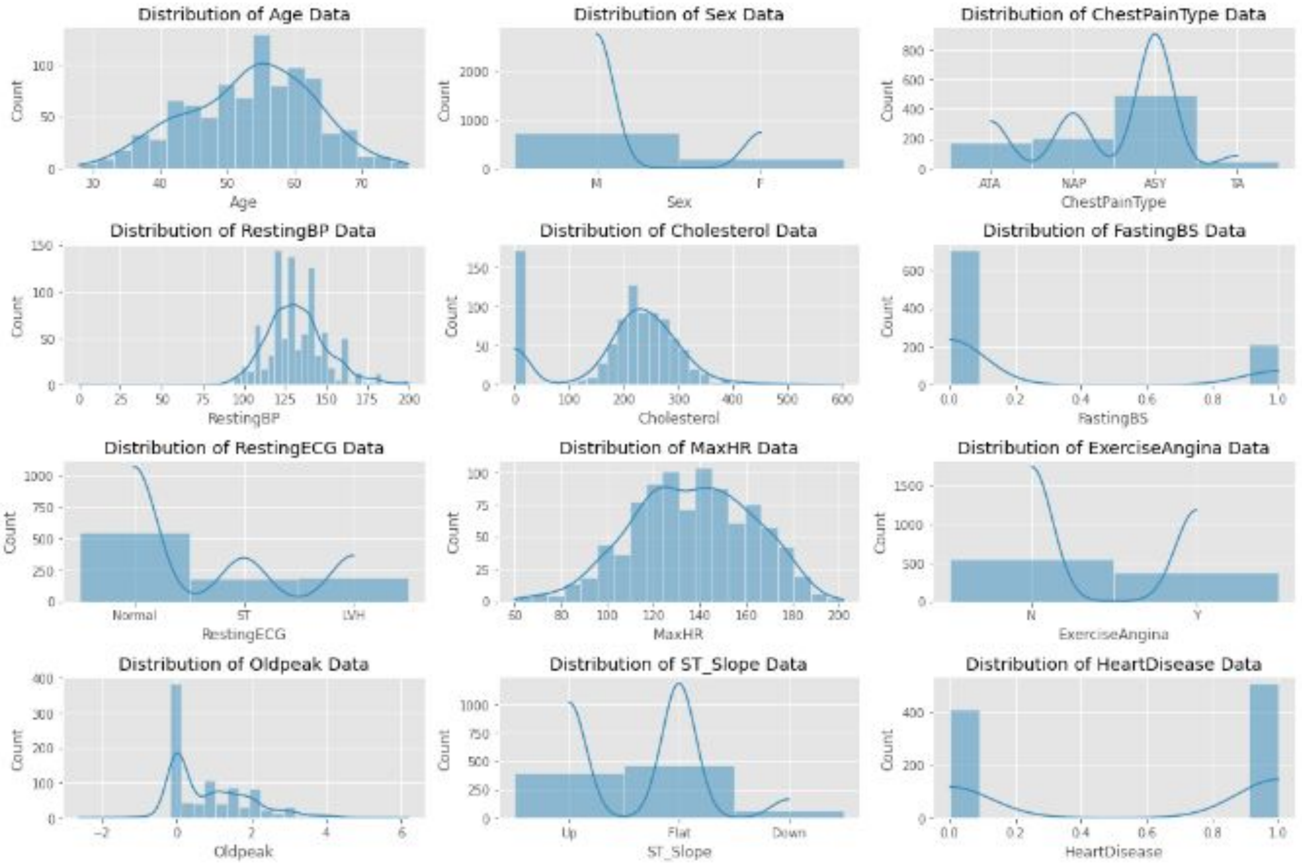
Data exploration - coefficient matrix



Data exploration - multiple pairwise bivariate distributions



Data exploration - columns’s values distributions



Distribution of Age Data

Distribution of Sex Data

Distribution of ChestPainType Data

Distribution of RestingBP Data

Distribution of Cholesterol Data

Distribution of FastingBS Data

Distribution of RestingECG Data

Distribution of MaxHR Data

Distribution of ExerciseAngina Data

Distribution of Oldpeak Data

Distribution of ST_Slope Data

Distribution of HeartDisease Data

Data preprocessing

We preprocess data through three steps:

- Handling null values
- Features Scaling
- Handling Categorical Variables

Data modeling

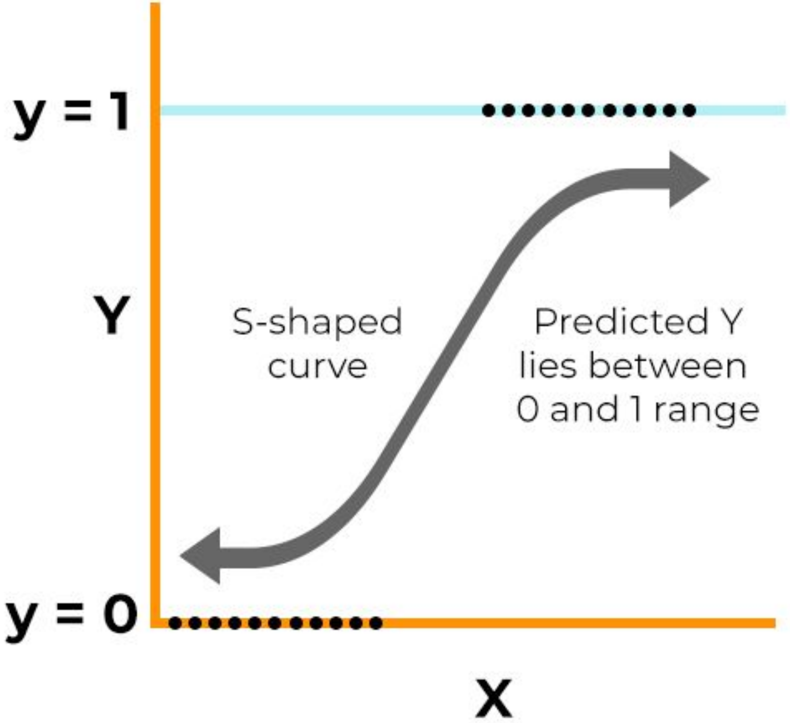
We decide to experience three difference models:

1. Logistic regression
2. Support vector machine
3. Random forest

Data modeling - logistic regression

Logistic regression is often used for classification and predictive analytics. Logistic regression estimates the probability of an event occurring, such as voted or didn’t vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1.

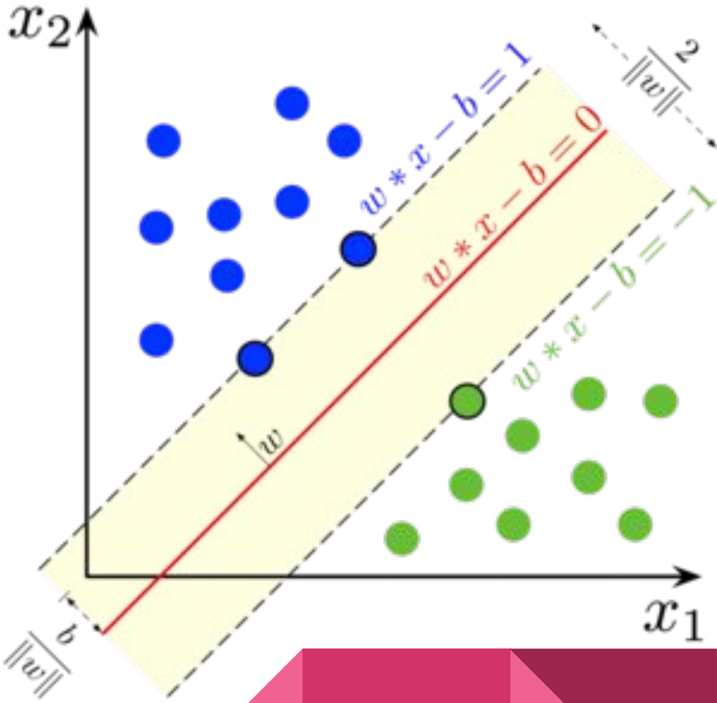
In logistic regression, a logit transformation is applied on the odds—that is, the probability of success divided by the probability of failure. This is also commonly known as the log odds, or the natural logarithm of odds.



Data modeling - support vector machine

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

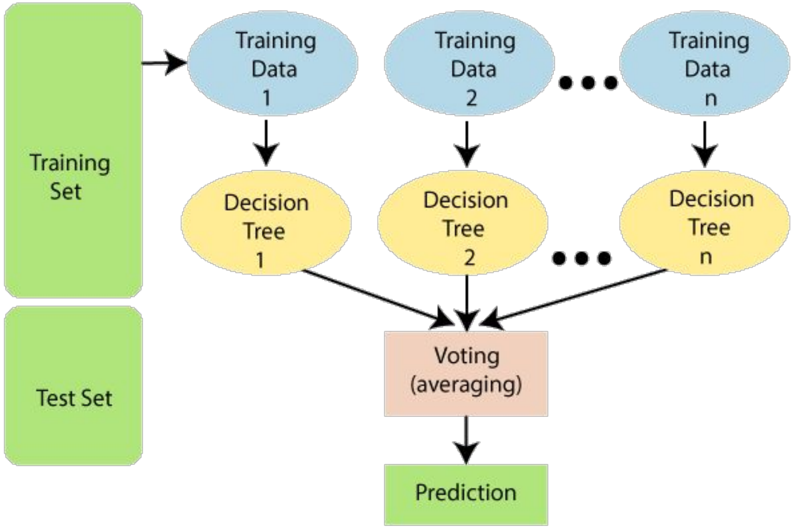
SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane.



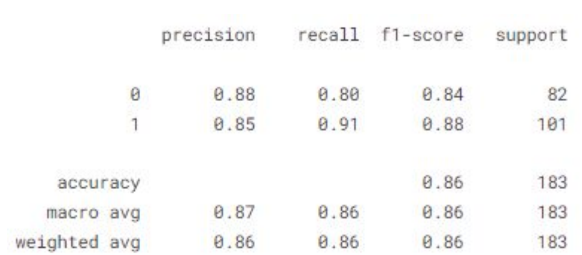
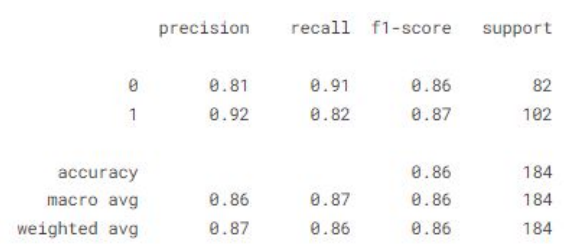
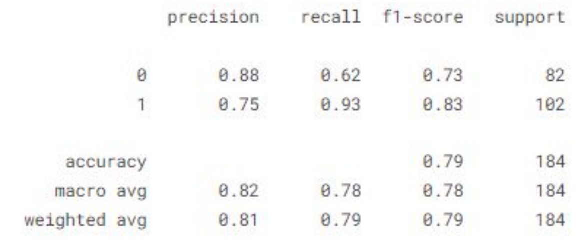
Data modeling - random forest

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.



Evaluation - confusion matrix

Logistic regression	Support vector machine	Random forest
		

```
              precision    recall  f1-score   support

           0       0.88      0.80      0.84        82
           1       0.85      0.91      0.88       101

    accuracy                           0.86       183
   macro avg       0.87      0.86      0.86       183
weighted avg       0.86      0.86      0.86       183
```

```
              precision    recall  f1-score   support

           0       0.81      0.91      0.86        82
           1       0.92      0.82      0.87       102

    accuracy                           0.86       184
   macro avg       0.86      0.87      0.86       184
weighted avg       0.87      0.86      0.86       184
```

```
              precision    recall  f1-score   support

           0       0.88      0.62      0.73        82
           1       0.75      0.93      0.83       102

    accuracy                           0.79       184
   macro avg       0.82      0.78      0.78       184
weighted avg       0.81      0.79      0.79       184
```

Key Findings and Insights

Support vector machine and logistic regression provide highest accuracy (0.86).

Random forest give lower accuracy when comparing with two above (0.79).

In production we tent to use the simpler model that is logistic regression.

Future work

There are several paths we can try to enhance the performance:

- Try neural network and ensemble models
- Better data preprocessing
- Try data normalise and regularization

Thanks for your reading!