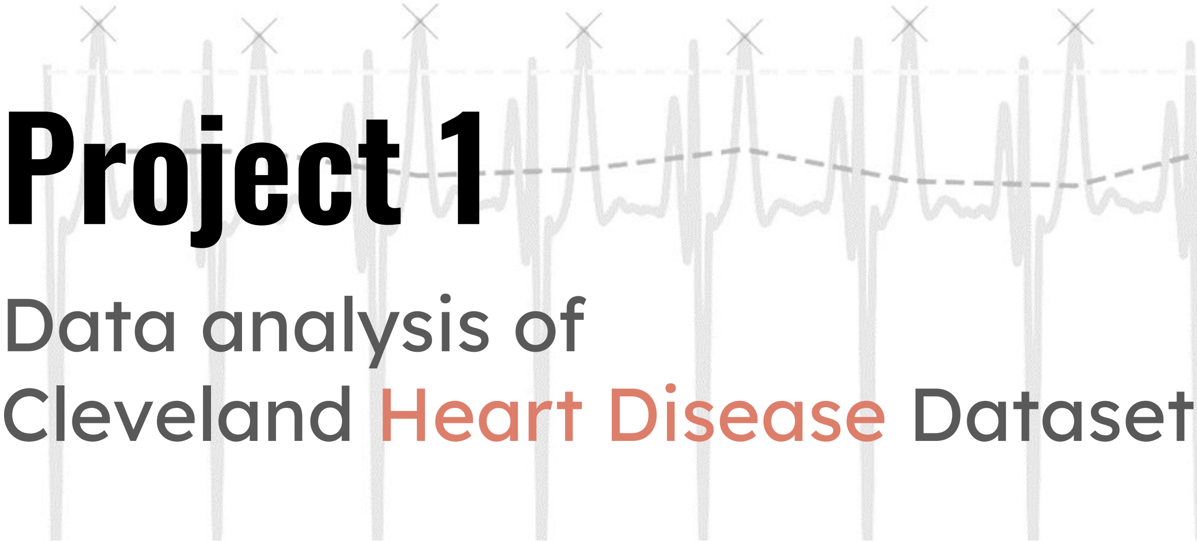
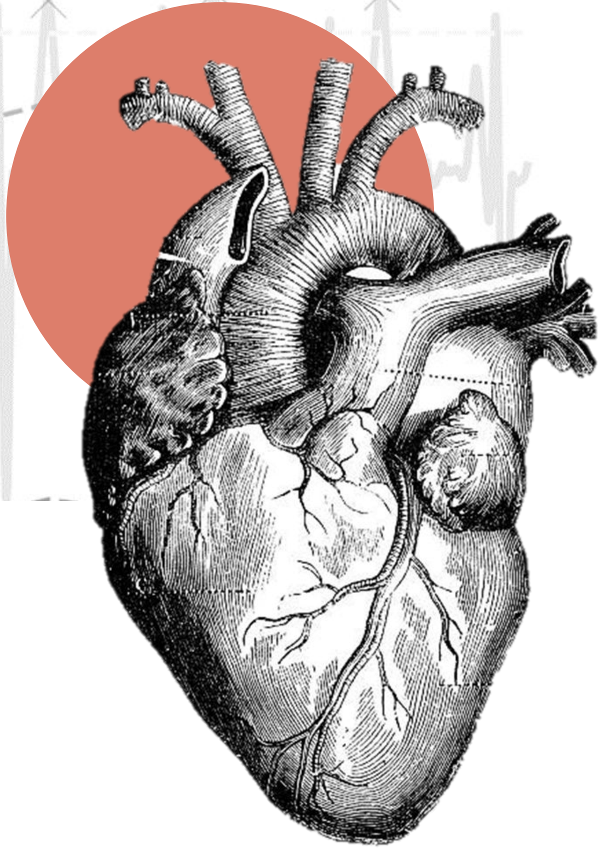


Project 1

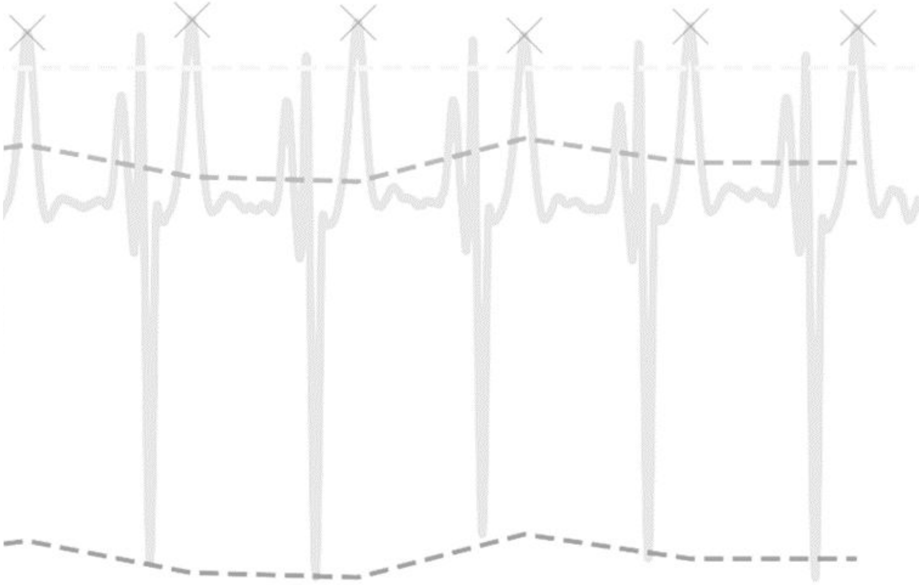
Data analysis of Cleveland Heart Disease Dataset



Group 11:
Nicolò Retis
Hang Mai Anh Vo (Emmy)
Rinoshan Parameswaran

Contents

1. Introduction
2. Exploratory Data Analysis (EDA)
3. Model Selection
4. Model Evaluation and Comparison
5. Key Findings and Conclusions

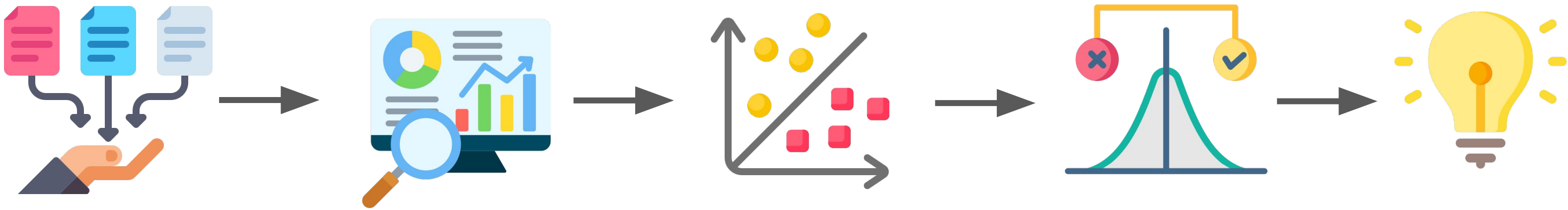


1. Introduction

Objective: From Cleveland Heart Disease Datasets

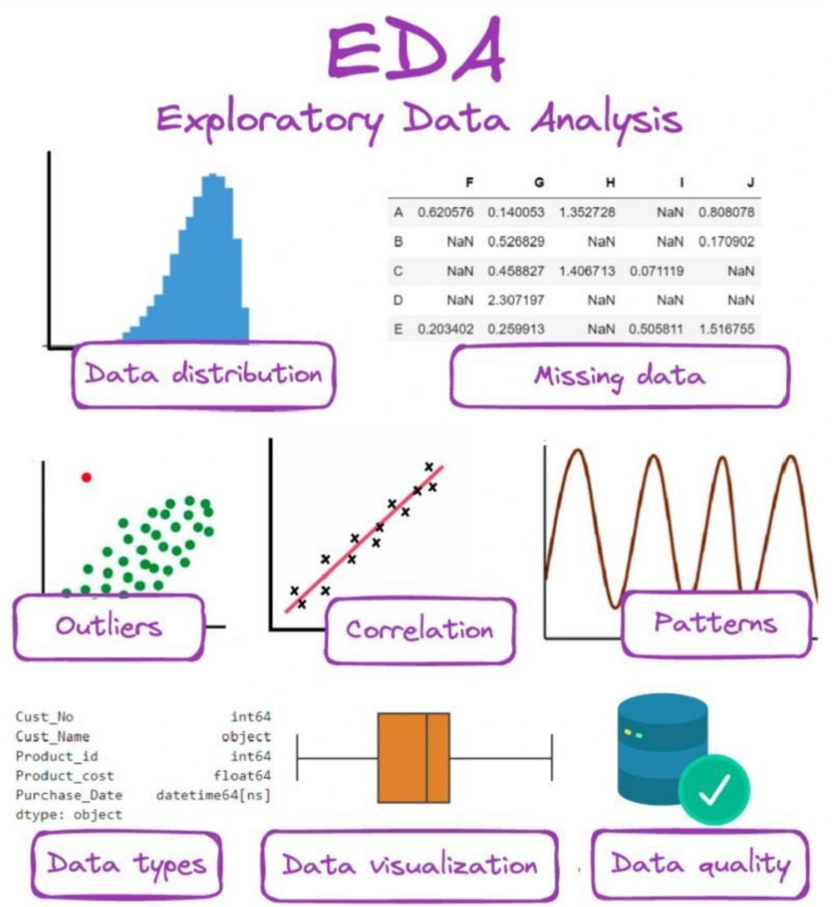
- To see the clinical pictures of heart disease patients
- To build prediction models to classify patients with and without heart diseases

Our work:



2. Exploratory Data Analysis (EDA)

Objective: To understand feature distributions and detect patterns associated with heart disease



	F	G	H	I	J
A	0.620576	0.140053	1.352728	NaN	0.808078
B	NaN	0.526829	NaN	NaN	0.170902
C	NaN	0.458827	1.406713	0.071119	NaN
D	NaN	2.307197	NaN	NaN	NaN
E	0.203402	0.259913	NaN	0.505811	1.516755

```
Cust_No                    int64
Cust_Name                 object
Product_id                 int64
Product_cost             float64
Purchase_Date     datetime64[ns]
dtype: object
```

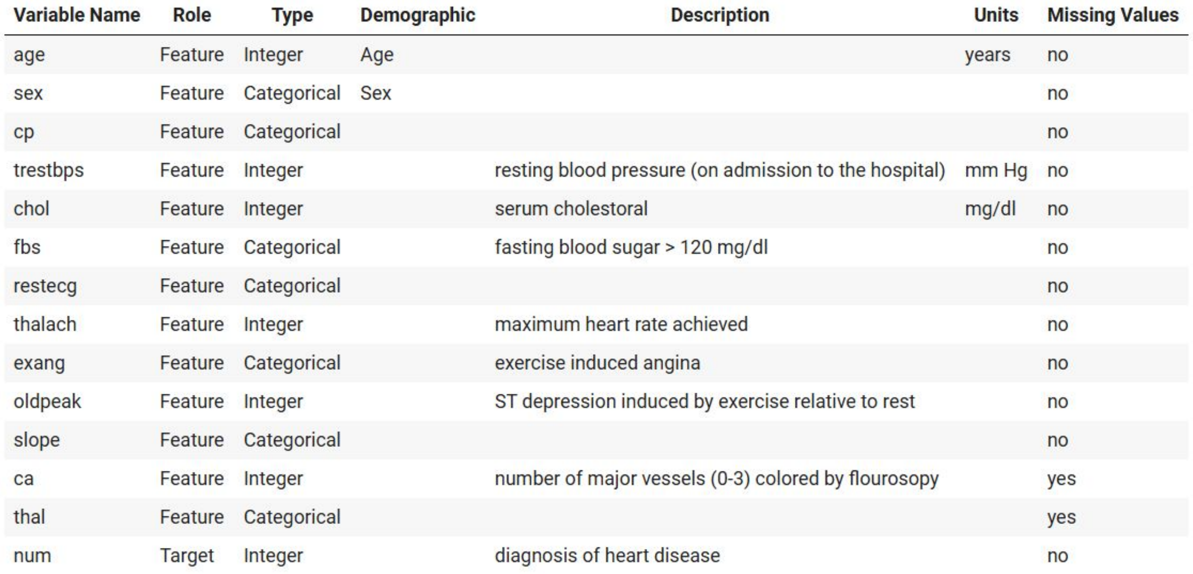
Source: https://www.markovml.com/blog/exploratory-data-analysis

What did we do?

- Handle missing data
- Correlation matrix
- Visualize important features in the dataset

Dataset overview

- 303 patient records. 6 records has missing data → remove.



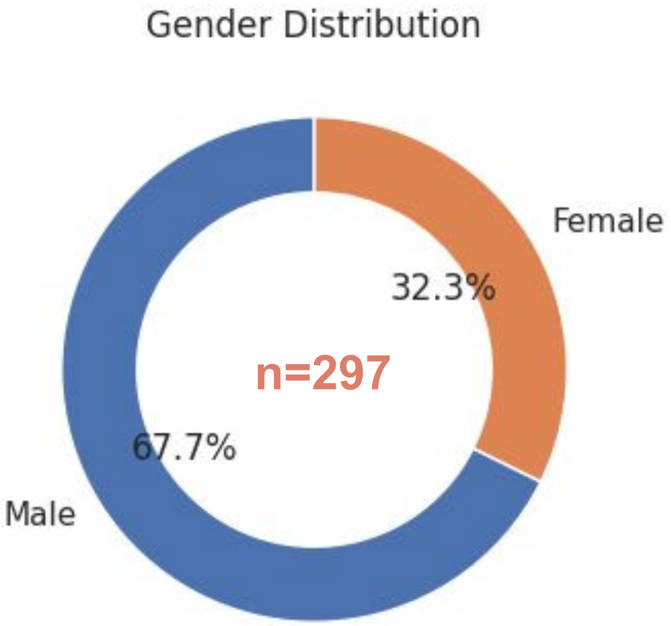
Variable Name	Role	Type	Demographic	Description	Units	Missing Values
age	Feature	Integer	Age		years	no
sex	Feature	Categorical	Sex			no
cp	Feature	Categorical				no
trestbps	Feature	Integer		resting blood pressure (on admission to the hospital)	mm Hg	no
chol	Feature	Integer		serum cholestoral	mg/dl	no
fbs	Feature	Categorical		fasting blood sugar > 120 mg/dl		no
restecg	Feature	Categorical				no
thalach	Feature	Integer		maximum heart rate achieved		no
exang	Feature	Categorical		exercise induced angina		no
oldpeak	Feature	Integer		ST depression induced by exercise relative to rest		no
slope	Feature	Categorical				no
ca	Feature	Integer		number of major vessels (0-3) colored by flourosopy		yes
thal	Feature	Categorical				yes
num	Target	Integer		diagnosis of heart disease		no

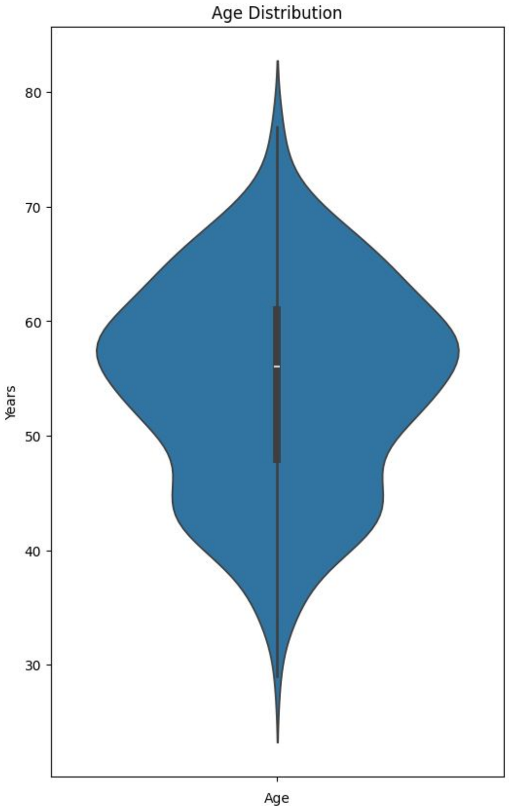
Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., ... & Froelicher, V. (1989). **International application of a new probability algorithm for the diagnosis of coronary artery disease**. *The American journal of cardiology*, *64*(5), 304-310.

Gender and age Distribution

- Age [29-77] years.

- Majority of males. This shown the same patterns with multiple epidemiological studies. [1]

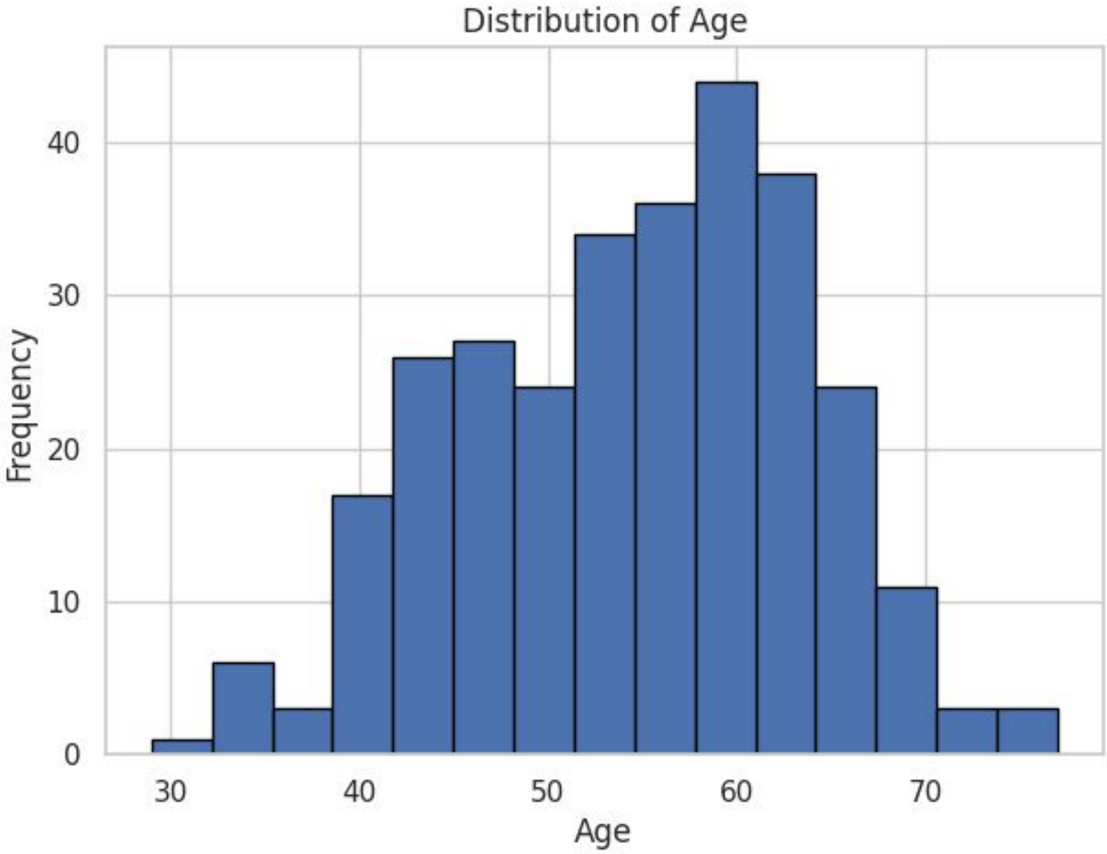




[1] Gao, Z., Chen, Z., Sun, A., & Deng, X. (2019). Gender differences in cardiovascular disease. *Medicine in Novel Technology and Devices*, *4*, 100025.

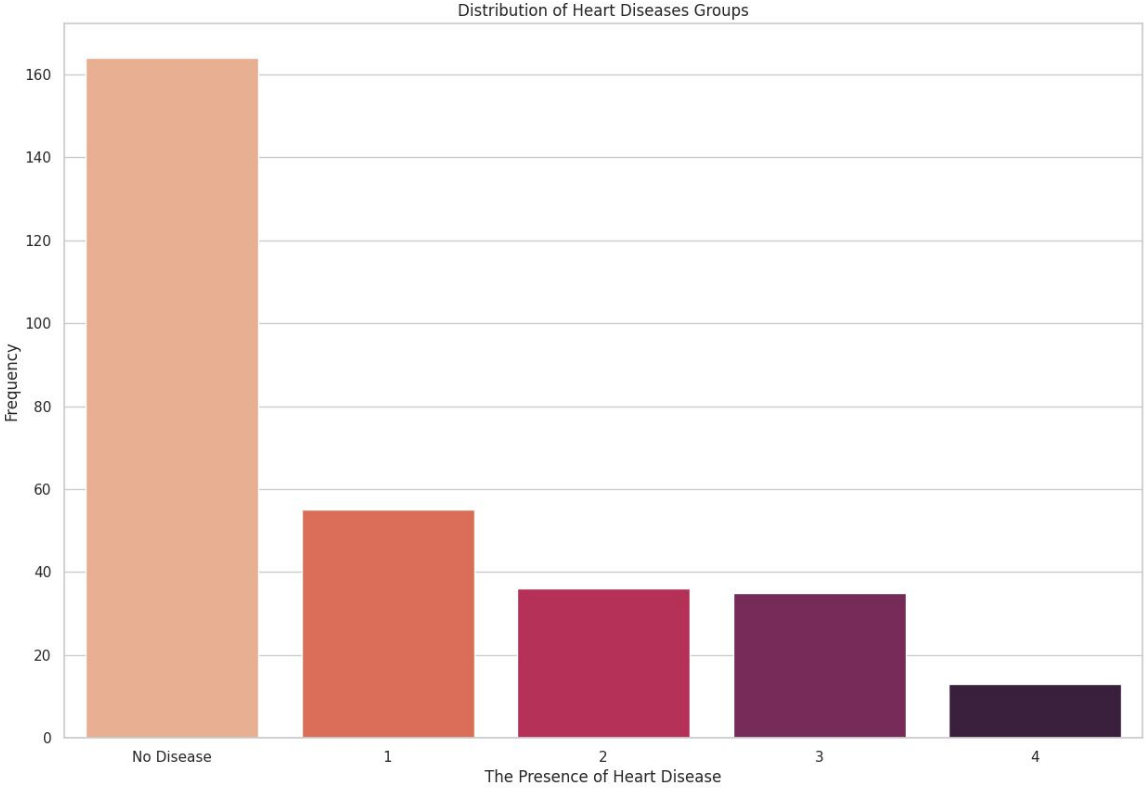
Age Countplot

- People under 40 and 70+ are under represented.
- Only 21 patients belonging to these groups.



Heart disease distribution

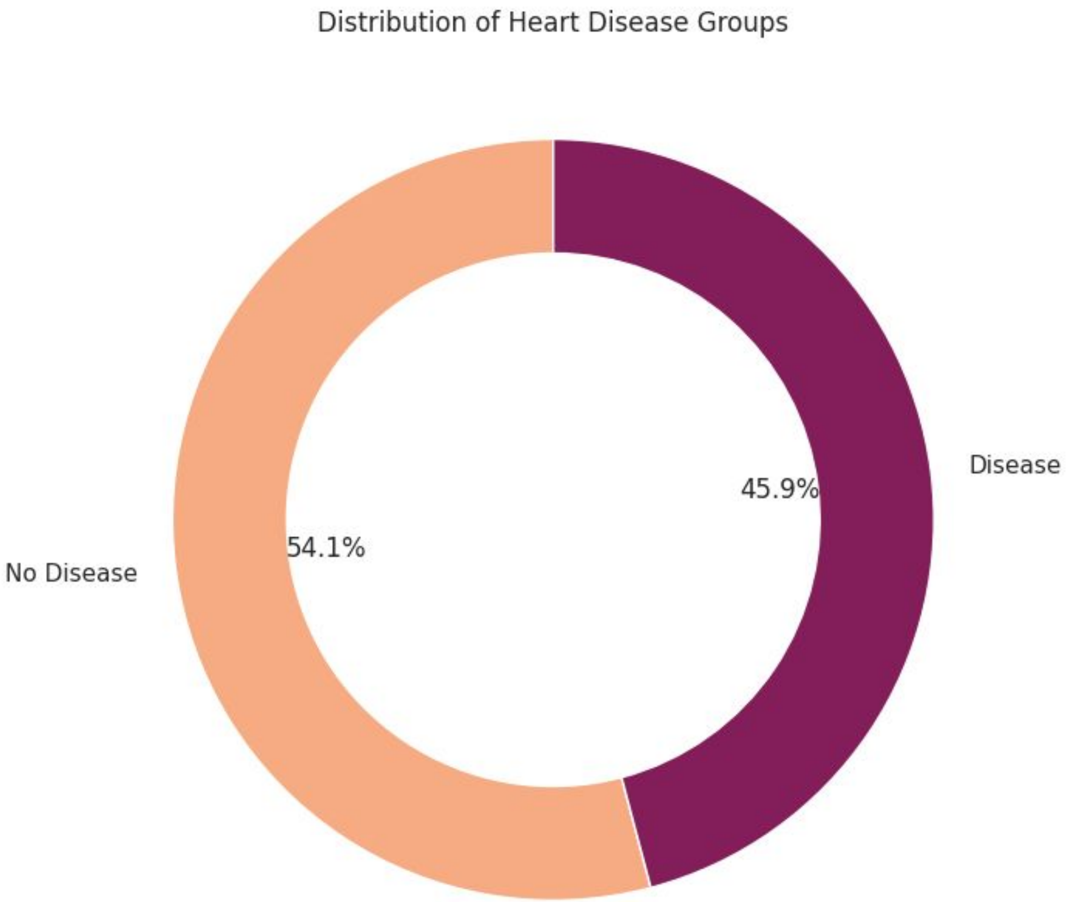
- The majority of the patients don’t have any heart disease.
- Number from 1 to 4 are used to rank the severe of heart disease.



Grouping approach

Patients were splitted into 2 groups, based on whether they do or don’t have heart disease:

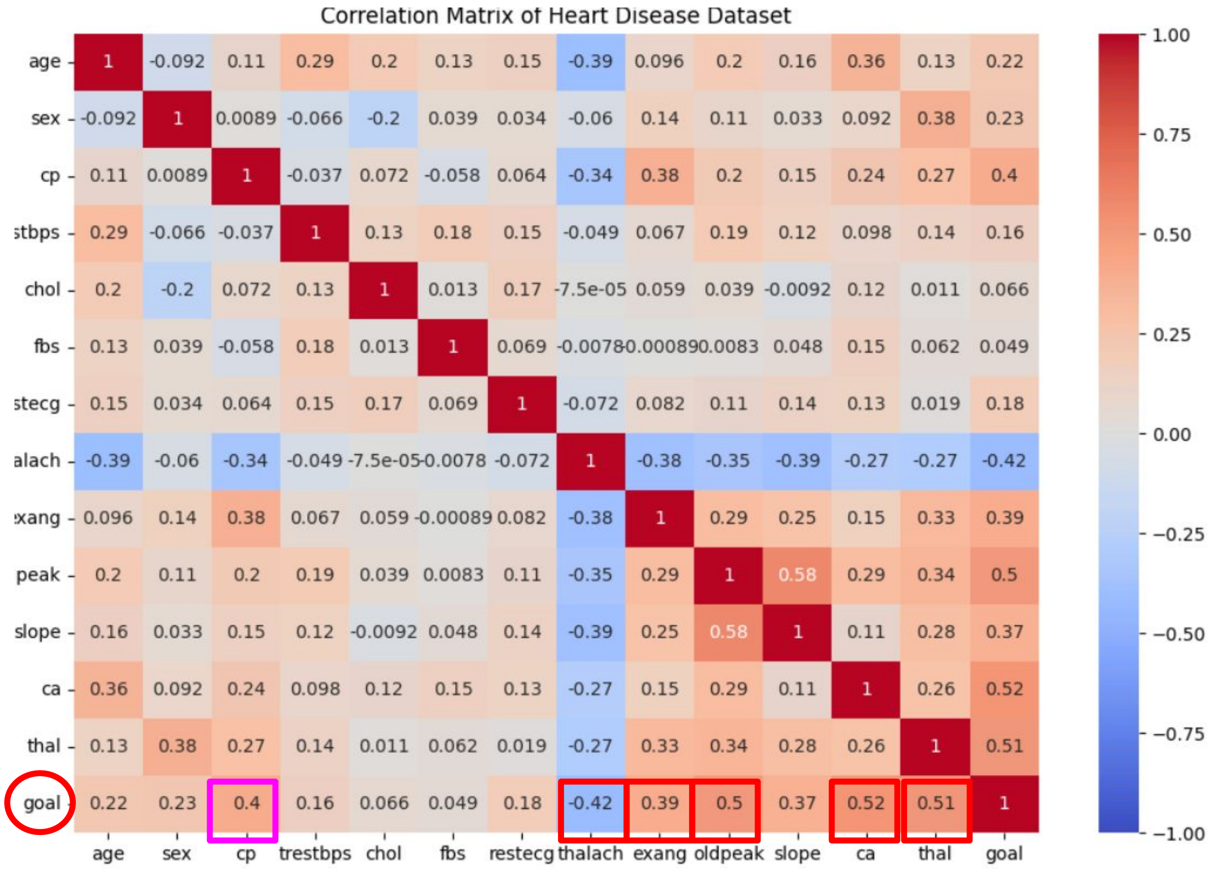
- More robust data inference
- Better understanding of the data
- Balancing the data



Correlation Matrix

Features with high corr.

- Chest Pain Type (`cp`)
- ST Depression (`oldpeak`)
- Exercise-Induced Angina (`exang`)
- Number of Blocked Vessels (`ca`)
- Thalassemia (`thal`)
- Maximum Heart Rate Achieved (`thalach`)

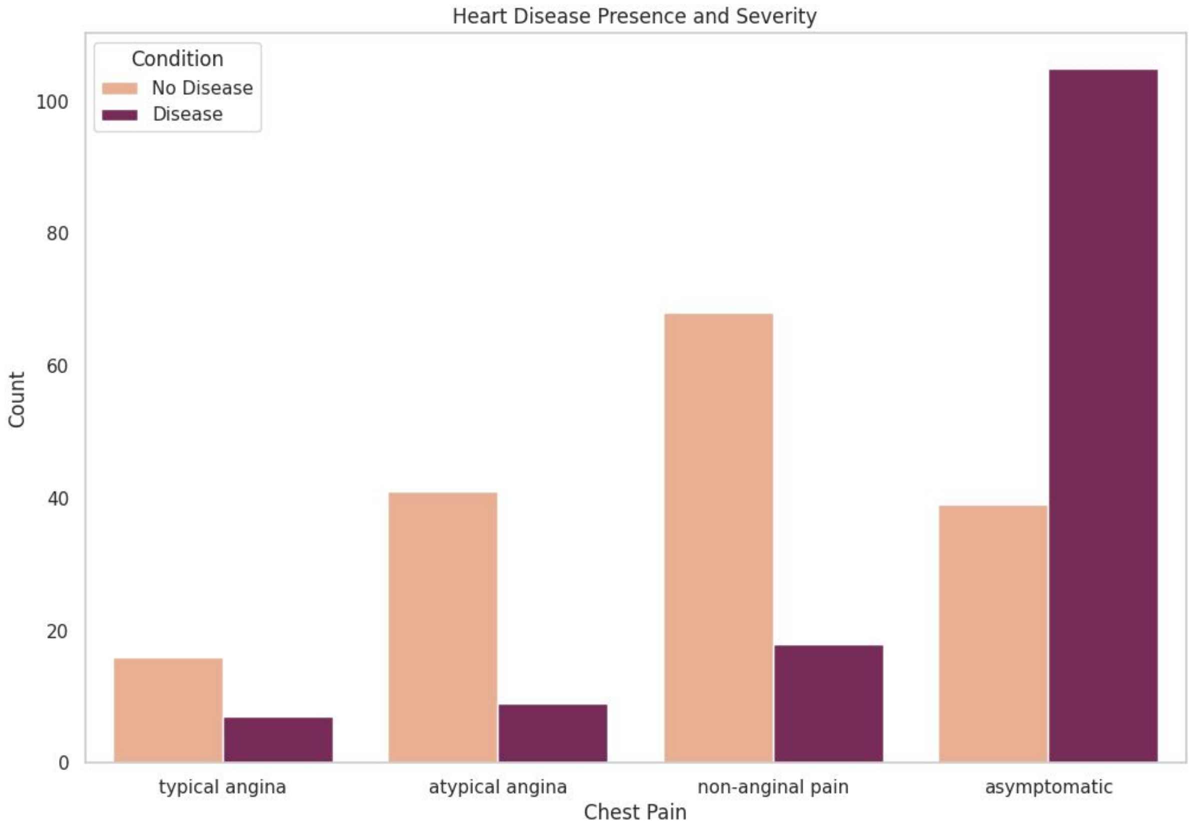


Correlation Matrix of Heart Disease Dataset

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	goal
age	1	-0.092	0.11	0.29	0.2	0.13	0.15	-0.39	0.096	0.2	0.16	0.36	0.13	0.22
sex	-0.092	1	0.0089	-0.066	-0.2	0.039	0.034	-0.06	0.14	0.11	0.033	0.092	0.38	0.23
cp	0.11	0.0089	1	-0.037	0.072	-0.058	0.064	-0.34	0.38	0.2	0.15	0.24	0.27	0.4
stbps	0.29	-0.066	-0.037	1	0.13	0.18	0.15	-0.049	0.067	0.19	0.12	0.098	0.14	0.16
chol	0.2	-0.2	0.072	0.13	1	0.013	0.17	-7.5e-05	0.059	0.039	-0.0092	0.12	0.011	0.066
fbs	0.13	0.039	-0.058	0.18	0.013	1	0.069	-0.0078	-0.00089	0.0083	0.048	0.15	0.062	0.049
stecg	0.15	0.034	0.064	0.15	0.17	0.069	1	-0.072	0.082	0.11	0.14	0.13	0.019	0.18
alach	-0.39	-0.06	-0.34	-0.049	-7.5e-05	-0.0078	-0.072	1	-0.38	-0.35	-0.39	-0.27	-0.27	-0.42
xang	0.096	0.14	0.38	0.067	0.059	-0.00089	0.082	-0.38	1	0.29	0.25	0.15	0.33	0.39
peak	0.2	0.11	0.2	0.19	0.039	0.0083	0.11	-0.35	0.29	1	0.58	0.29	0.34	0.5
slope	0.16	0.033	0.15	0.12	-0.0092	0.048	0.14	-0.39	0.25	0.58	1	0.11	0.28	0.37
ca	0.36	0.092	0.24	0.098	0.12	0.15	0.13	-0.27	0.15	0.29	0.11	1	0.26	0.52
thal	0.13	0.38	0.27	0.14	0.011	0.062	0.019	-0.27	0.33	0.34	0.28	0.26	1	0.51
goal	0.22	0.23	0.4	0.16	0.066	0.049	0.18	-0.42	0.39	0.5	0.37	0.52	0.51	1

Chest Pain Distribution

- Chest Pain correlates with disease.
- Patient with disease tend not to manifest chest pain.



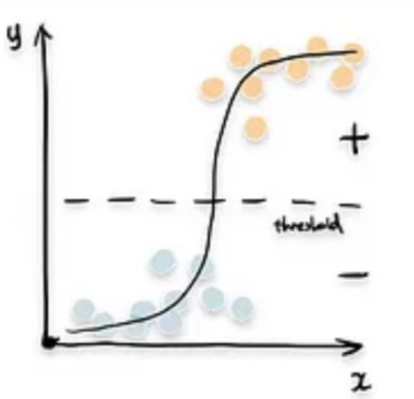
3. Model selection

Supervised machine learning models

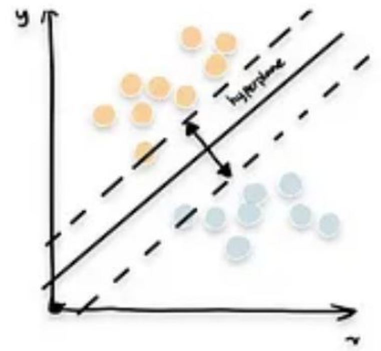
Objective: To find a good model to predict if a patient has heart diseases or not

Models

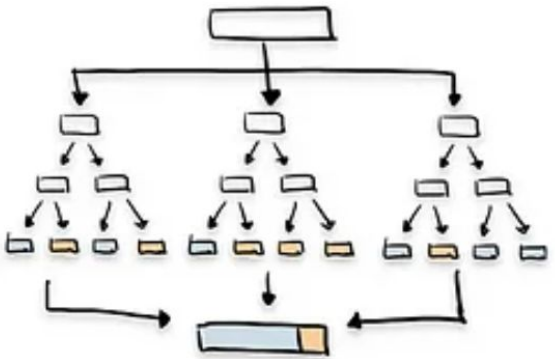
Logistic regression



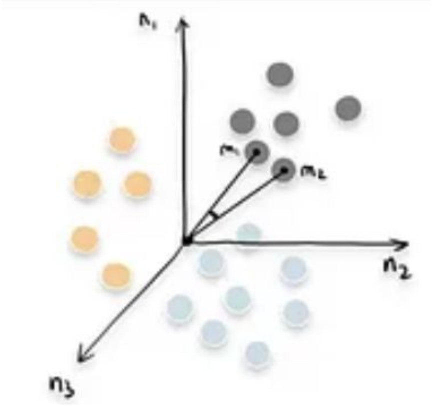
Support Vector Machine (SVM)



Random Forest



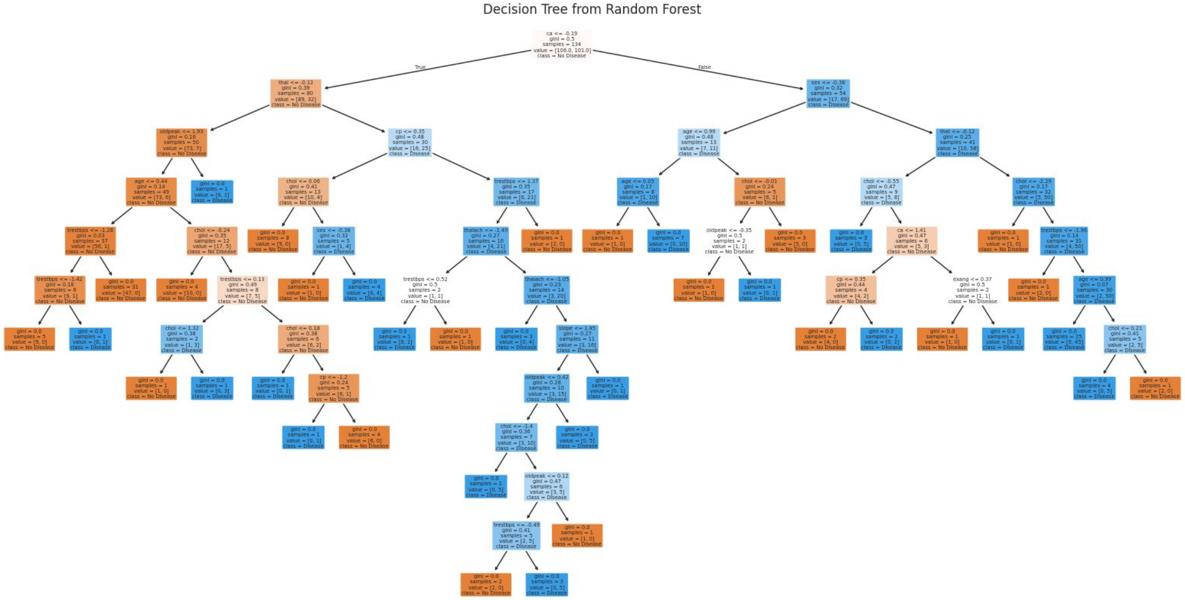
K Nearest Neighbor (KNN)



Images from: https://towardsdatascience.com/top-machine-learning-algorithms-for-classification-2197870ff501

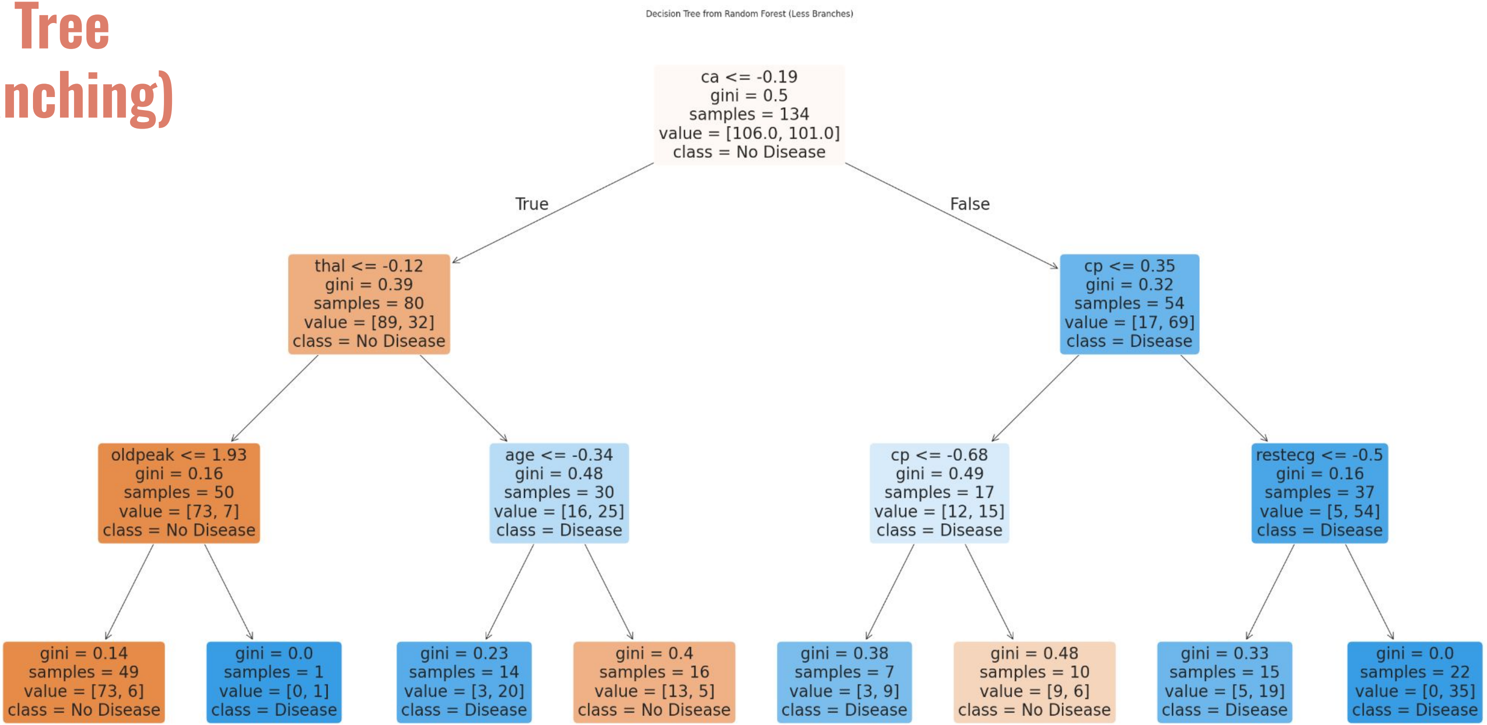
Decision Tree

- Can be interpreted by physician.
- Although reducing the branching may be necessary for the understanding and to limit overfitting.



Decision Tree from Random Forest

Decision Tree (less branching)

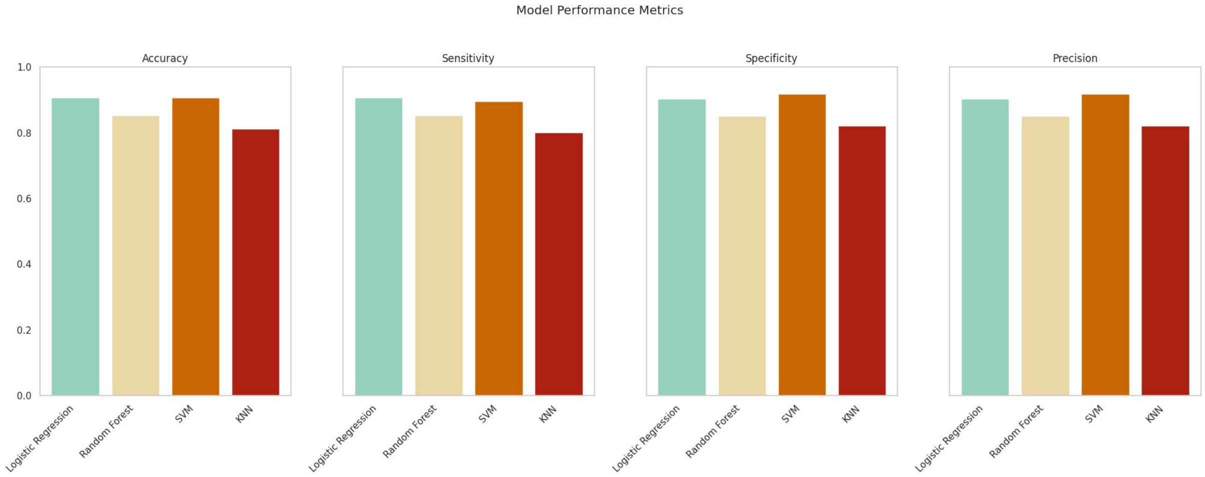


4. Model evaluation

Objective: To understand the performance of each selected models with the provided dataset

Model comparison (model metrics)

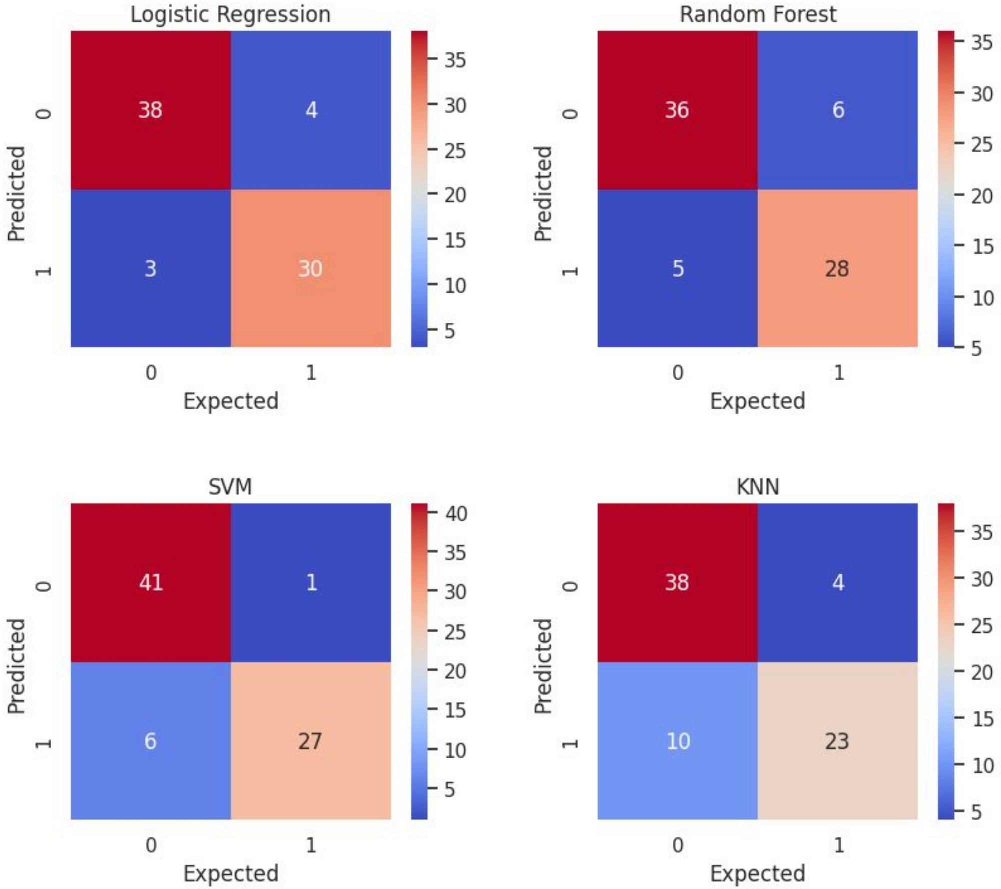
- Logistic regression and SVM perform similarly.
- Random Forest metrics are slightly better than KNN classifier.



Model comparison (confusion matrix)

- Logistic regression and SVM perform similarly.

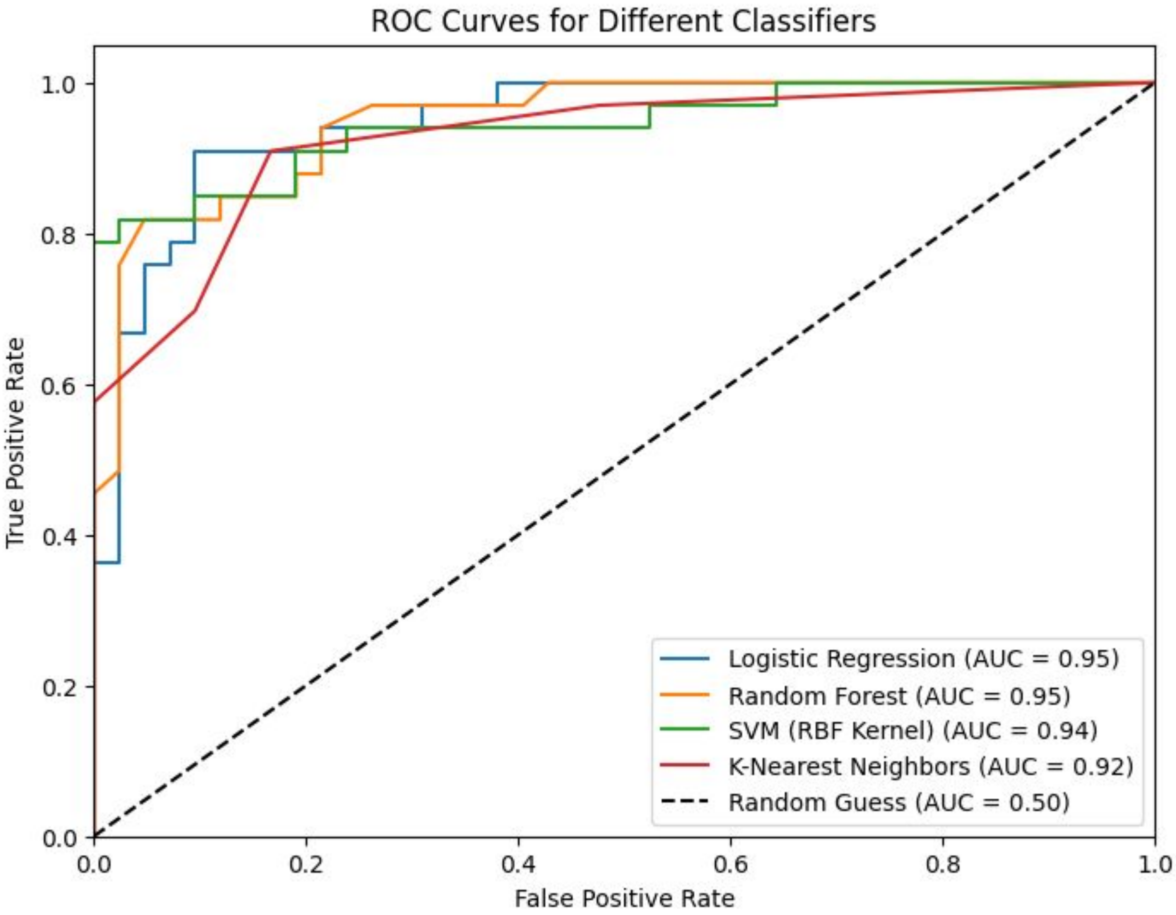
- Random Forest metrics are slightly better than KNN classifier.



Model comparison (ROC)

- Logistic regression and SVM perform similarly.

- Random Forest metrics are slightly better than KNN classifier.



5. Insights and Conclusion

Predictive Features: Blocked vessels (`ca`), ST depression (`oldpeak`), maximum heart rate (`thalach`) are strong indicators of heart disease (corr. value >= 0.4)

Chest pain type (`cp`), thalassemia (`thal`) is not even it has high corr. value because the number not reflect the severity, but just for categorizing.

Binary Classification: Converting to a binary target improves performance metrics, focusing on diagnosing disease presence rather than categorizing severity.

Model Evaluation: Random Forest, Logistic Regression, and SVM provide reliable predictions.

High AUC scores from Random Forest and SVM suggesting these models may well-suited for real-world clinical applications.

Clinical Application: These insights could aid in early identification of at-risk patients, helping healthcare providers prioritize testing and intervention.

Thank you

- This is the end -

Chest Pain

Patients divided in:

- 1: Typical angina
- 2: Atypical Anginal Pain.
- 3: Non Anginal Pain.
- 4: Asymptomatic.

What is **ANGINA?**

Angina is a type of chest pain caused by reduced blood flow to the heart. Angina is a symptom of coronary artery disease. Angina is often described as squeezing, pressure, heaviness, tightness or pain in the chest. It may feel like a heavy weight lying on the chest.

#WordWednesday