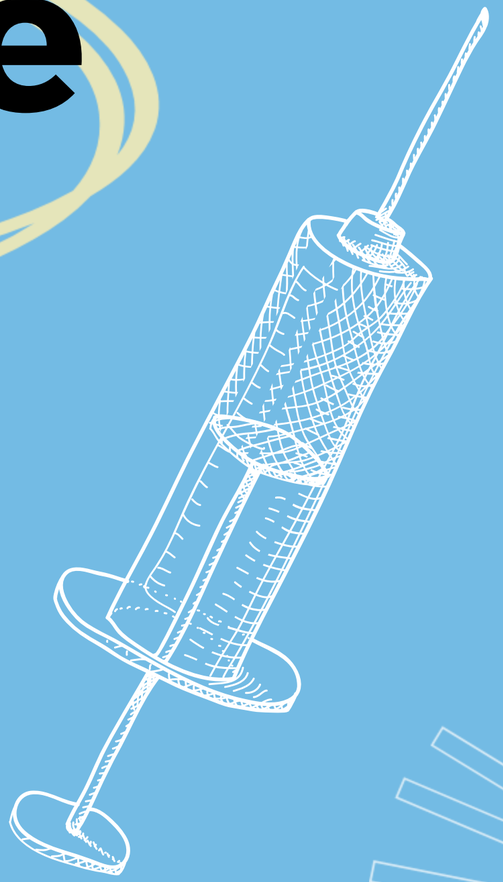
Heart Disease Prediction

Data Science in Healthcare
Group 06



PRESENTED BY

An Ni
Michaelsson Valtýr
Plavsic Filip
Roth Catalina

25 October 2024

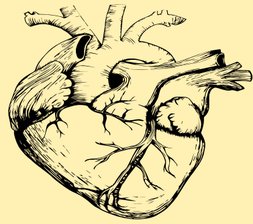
Motivation and Research Question

- Cardiovascular diseases are the leading cause of death worldwide, taking around 18 million lives annually.

- Common types include coronary heart disease, stroke and peripheral arterial disease.

- Significant financial burden due to direct and indirect costs.

Main question: Can CVD be predicted with an accuracy of 70% and recall ratio of 90% using machine learning models and logistic regression?

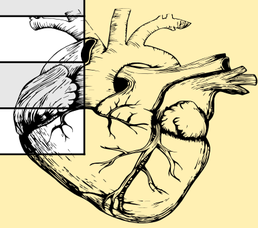


Variables Description

Dataset

Variable overview
Cleveland (1989)

Variable Name	Role	Type	Description	Missing Values
age	Feature	Integer	Age in years	No
caa	Feature	Integer	number of major vessels (0-3) colored by flourosopy	Yes
chol	Feature	Integer	Serum cholesterol in mg/dl	No
cp	Feature	Categorical	Chest pain type: 0 = typical angina; 1 = atypical angina; 2 = non-anginal pain; 3 = asymptomatic	No
exng	Feature	Categorical	Exercise induced angina	No
fbs	Feature	Categorical	Fasting blood sugar > 120 mg/dl 1 = true; 0 = false	No
output	Target	Integer	Diagnosis of heart disease	No
oldpeak	Feature	Integer	ST depression induced by exercise relative to rest	No
restecg	Feature	Categorical	Resting electrocardiographic results: 0 = normal 1 = having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria	No
sex	Feature	Categorical	Gender: 1 = male; 0 = female	No
slp	Feature	Categorical	The slope of the peak exercise ST segment: 0 = upsloping 1 = flat 2 = downsloping	No
thall	Feature	Categorical	Thalassemia	Yes
thalachh	Feature	Integer	Maximum heart rate achieved	No
trtbps	Feature	Integer	Resting blood pressure (on admission to the hospital) in mm Hg	No



Dataset and Preparation Process: A clinical dataset of 300 incidents with risk factors and label variable

Raw Dataset

- **Characteristics**:
  Attributes: 14
  Observations: 303

- **Target Variable (1):** Diagnosis of heart disease, binary, 57% positive diagnosis.

- **Control Factors (13)**:
  - Age, Sex, Resting Blood Pressure, Cholesterol, Max Heart Rate
  - High Fasting Glucose
  - Chest Pain Type, Exercise-Induced Angina, ST Depression, ST Segment Slope, Resting electrocardiographic
  - Major Vessels Stained, Thalassemia

Ready for supervised learning

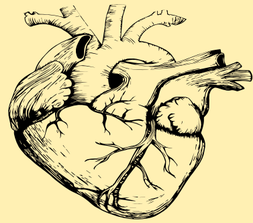
Preparation

- **Duplicate handling:** 1 duplicate, obvious double counting, remove.
- **Outlier investigation:** exclude the risk of measurement error by close investigation of the observation.
- **Verify data format:** keep the Numeric variables, ensure nominal and ordinal variables are labeled in numbers

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
74	43	0	2	122	213	0	1	165	0	0.2	1	0	2	1
19	69	0	3	140	239	0	1	151	0	1.8	2	2	2	1
49	53	0	0	138	234	0	0	160	0	0.0	2	0	2	1

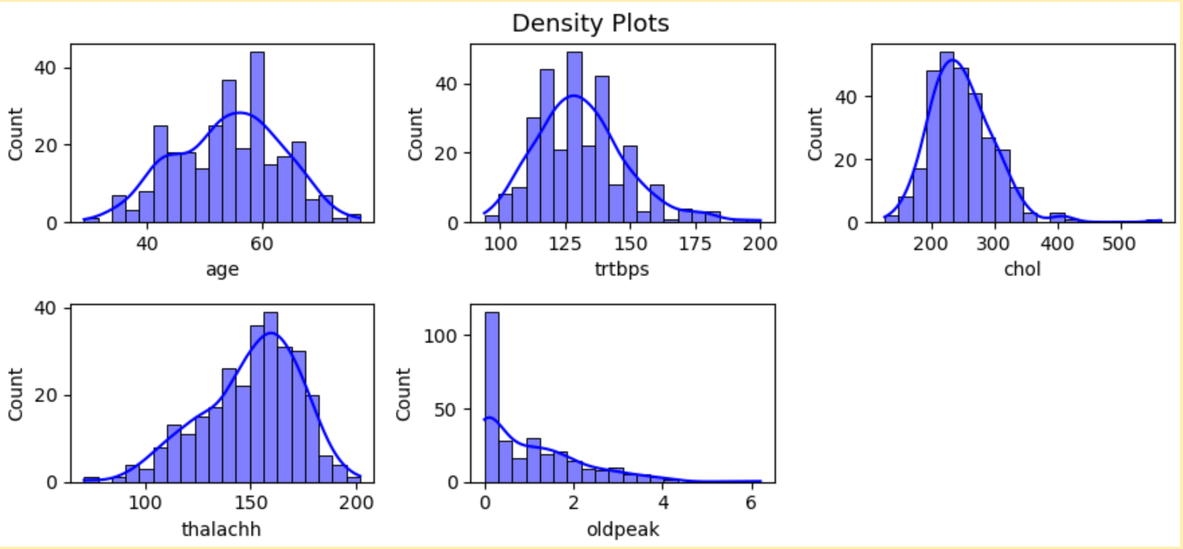
Sample of 3 out of 288 observations after transformation

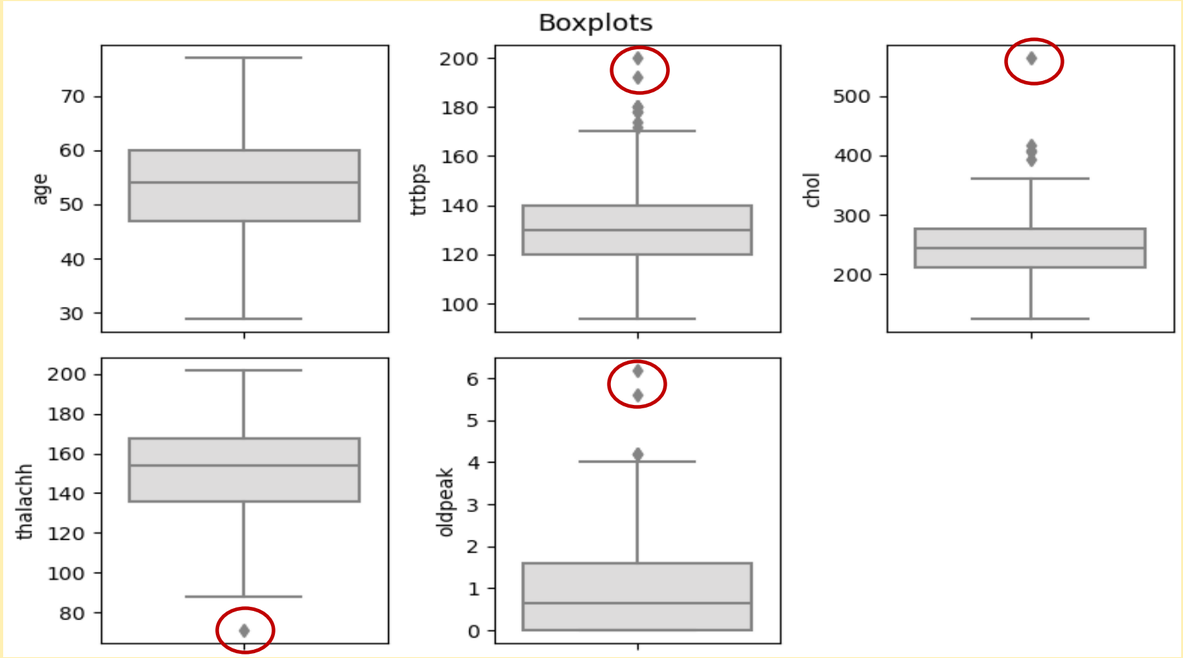
Source: from Cleveland Clinic (1989), a medical institute known for cardiology expertise.



Exploratory Data Analysis (EDA)

Continuous Variables





- **Age:** peak between 50 – 60 years.
- **Trtbps (Resting Blood Pressure):** peak between 130 – 140 mm Hg.
- **Chol (Cholesterol):** skewed to the right with peak between 250 – 300 mg/dL.
- **Thalachh (Maximum Heart Rate):** skewed to the left with peak around 150 – 160 bpm.
- **Oldpeak (ST Depression)** skewed to the right. Most observations near zero.

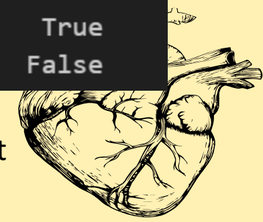
Note: ST Depression -- > Specific finding in electrocardiogram that can indicate a lack of oxygen.

Outlier Values

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	\
85	67	0	2	115	564	0	0	160	0	1.6	1	
204	62	0	0	160	164	0	0	145	0	6.2	0	
221	55	1	0	140	217	0	1	111	1	5.6	0	
223	56	0	0	200	288	1	0	133	1	4.0	0	
248	54	1	1	192	283	0	0	195	0	0.0	2	
272	67	1	0	120	237	0	1	71	0	1.0	1	

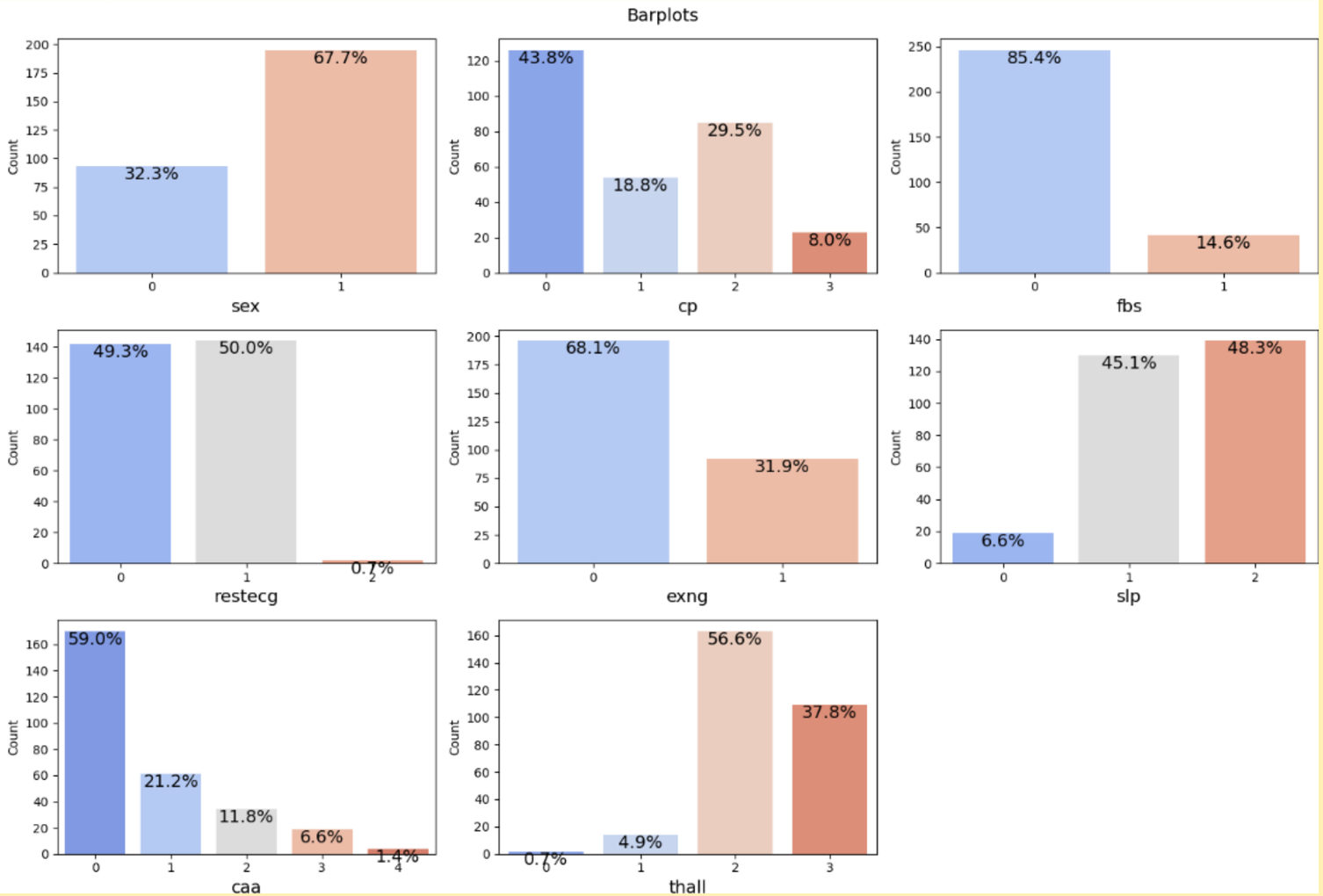
	caa	thall	output	high_chol	low_thalachh	high_oldpeak	high_trtbps
85	0	3	1	True	False	False	False
204	3	3	0	False	False	True	False
221	0	3	0	False	False	True	False
223	2	3	0	False	False	False	True
248	1	3	0	False	False	False	True
272	0	2	0	False	True	False	False

- There is no evidence that these outliers are caused by measurement error.
- Removing them could result in loss of meaningful insights that are important for the analysis and prediction.

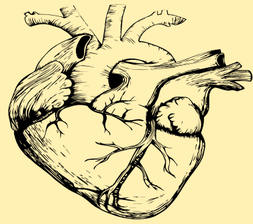


Exploratory Data Analysis (EDA)

Categorical variables

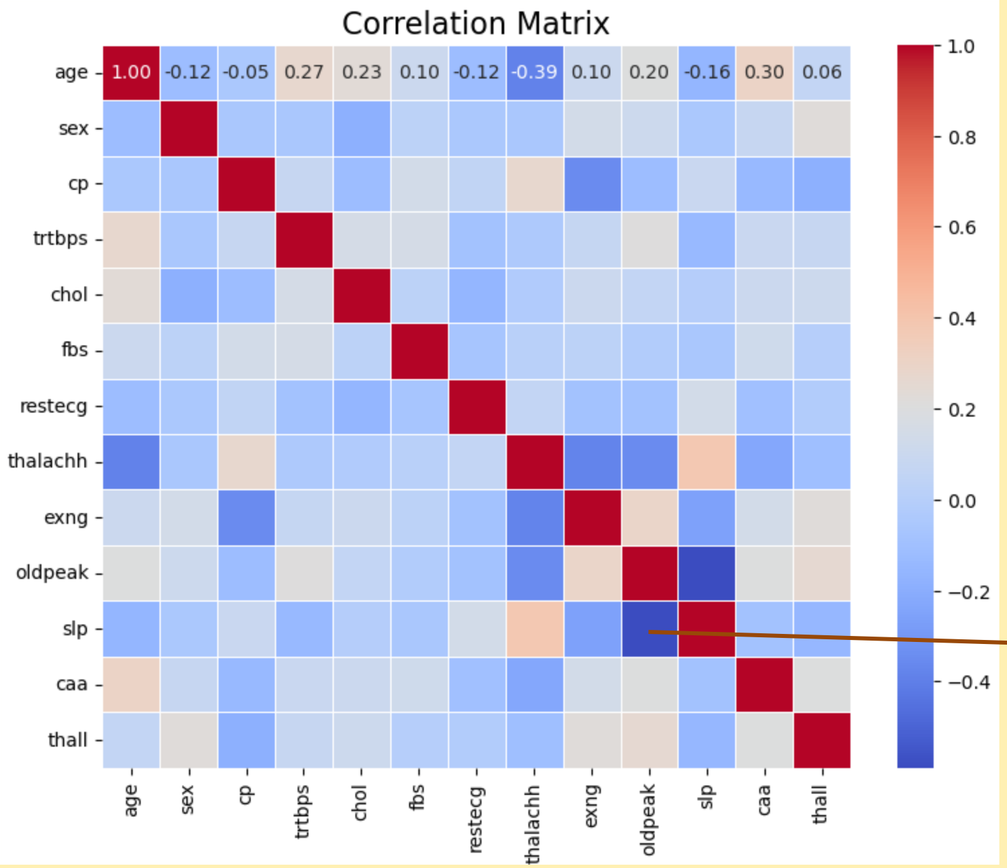


- **Sex:** 68% male and 32% female.
- **CP (Chest Pain Type):** There is a higher presence of type 0 (approximately 44%), which indicates typical angina.
- **Fbs (Fasting Blood Sugar):** 85.4% of the observations have a value below 120 mg/dl.
- **Restecg (Resting Electrocardiogram Results):** 49.3% of the individuals involved in the analysis have normal values, 50% have some abnormalities, and less than 1% may have possible ventricular hypertrophy.
- **Exng (Exercise-Induced Angina):** About 68% of individuals experience angina during exercise.
- **Slp (Slope of the Peak Exercise ST Segment):** Approximately 7% have upsloping, 45% have flat, and 48% have downsloping during exercise.
- **Caa (Number of Major Vessels Colored by Fluoroscopy):** 59% have an absence of vessels, while 41% have a presence in varying degrees.
- **Thall (Thallium Stress Test Results):** 4.9% have normal values, while approximately 95% have some defects (with 38% being reversible).



Exploratory Data Analysis (EDA)

Correlation Between Variables

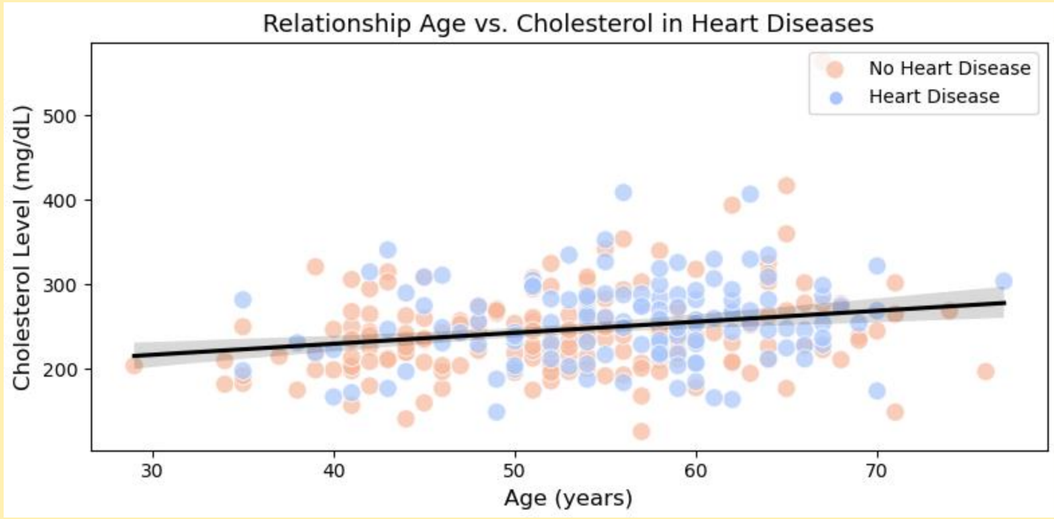


Corr = -0.59

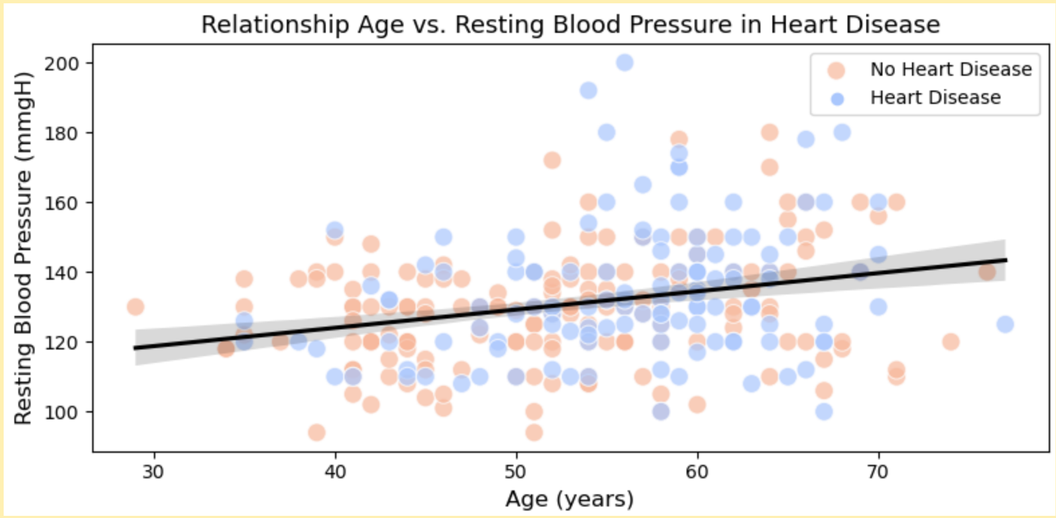
Interpretation: Pearson, standard correlation coefficient *shows that there is no variables with strong or very strong correlations, implying that we might not need to exclude variables due to collinearity.*

Rule of thumb: 00-.19 “very weak” .20-.39 “weak” .40-.59 “moderate” .60-.79 “strong” .80-1.0 “very strong” (https://www.statstutor.ac.uk/resources/uploaded/pearsons.pdf)

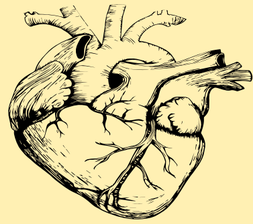
Relation between Cholesterol, Rest Blood Pressure and Age



Correlation Age and Cholesterol: 0.23



Correlation Resting Blood Pressure and Age: 0.27

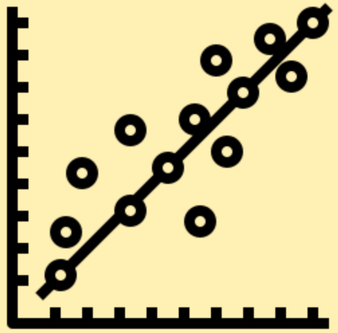


Machine Learnings Models

Classical Machine Learning Approaches

Typically based on statistical principles and are well-suited for structured data, such as tabular datasets

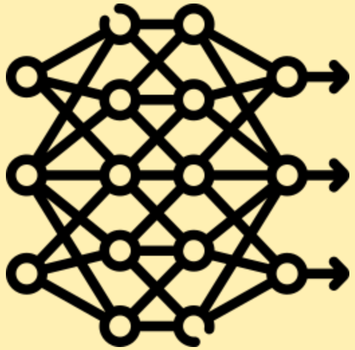
- Linear Regression
- Logistic Regression
- Decision Tree
- Random Forest
- k-Nearest Neighbor

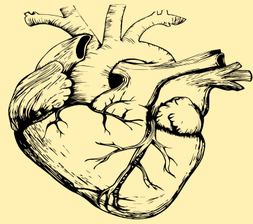


Newer Machine Learning Approaches

Often involve more sophisticated techniques that can handle complex and unstructured data

- Neural Network – multi layer perceptron

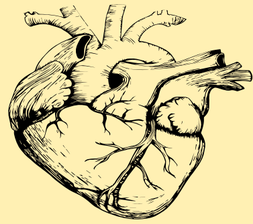




Models Comparison

Performance Metrics

	Logistic Regression	Decision Tree	Random Forest	Naive Bayes	K-Nearest Neighbor	Neural Network (Multi-layer)
Accuracy	**0.90**	0.62	0.79	**0.90**	0.76	0.86
Precision	**0.89**	0.65	0.78	**0.89**	0.75	0.84
Recall	**0.94**	0.76	0.88	**0.94**	0.88	**0.94**
F1 Score	**0.91**	0.70	0.83	**0.91**	0.81	0.89

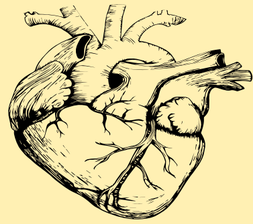


Conclusions

- Most models were able to predict CVD over 70% accuracy (all but Decision Tree).
- Three models had a recall score of 90% or higher with two more close at 88%. Decision tree was the poorest performer.
- Logistic regression and Naïve Bayes were the best performers

Shortcomings

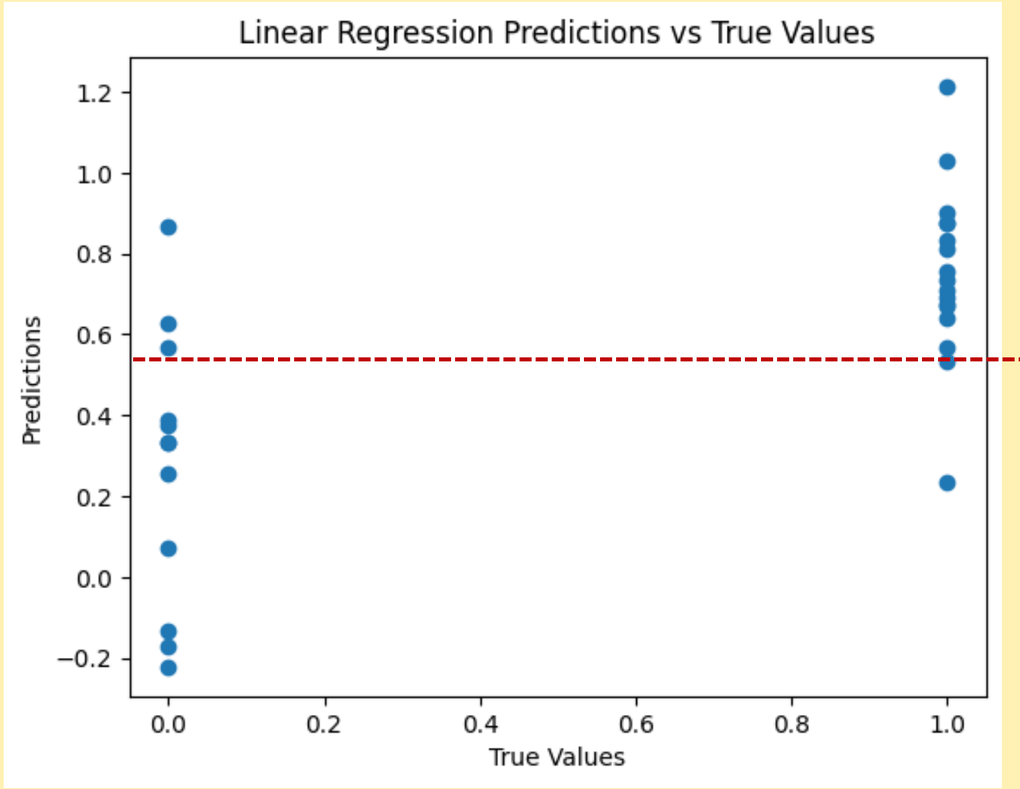
- Small dataset
- Narrow sample
- Target variable positive/negative split



Appendix: trial and error with linear regression

Linear Regression was also applied but True value – Prediction plot does not show optimal performance

Although not necessarily fitting the classification nature of our project, a linear regression was performed to explore the choice of model.

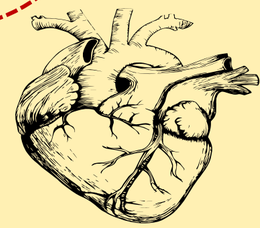


Performance Metrics

Metric	Value
Root Mean Squared Error	0.368746
R-squared	0.439443

Interpretation: With all variables used, R-squared 0.43 → linear regression model has moderate prediction power, but not very strong.

Interpretation: predictions are closer to the direction of true value, implying moderate prediction power. However, it also shows that linear regression has its limitation for this case, due to the binary nature of the target variable.



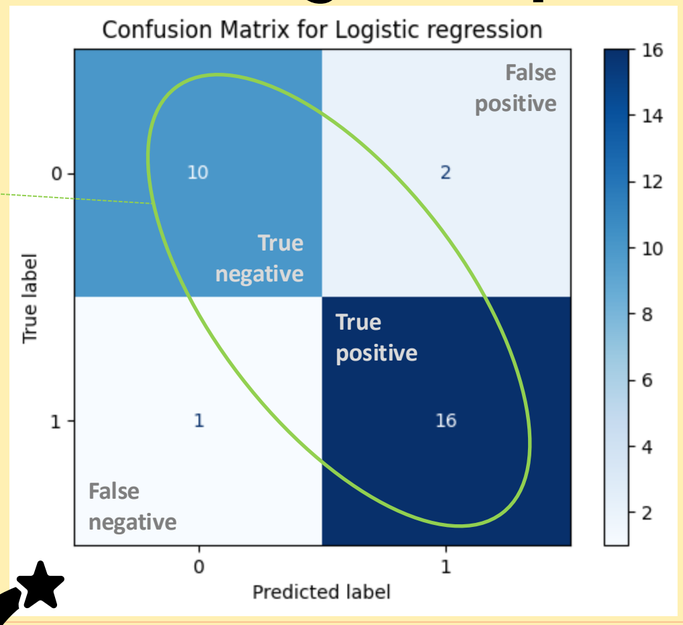
Appendix: Confusion Matrices

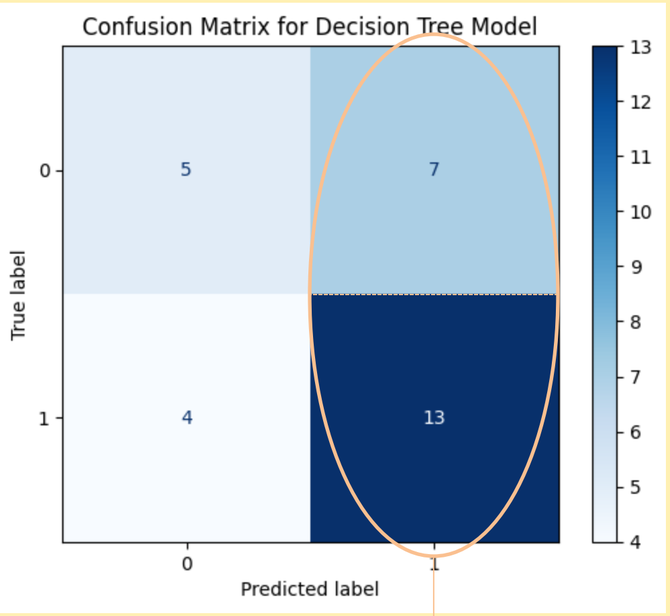
2 Focuses: Ensure general prediction → Accuracy; Reduce false negative → Recall

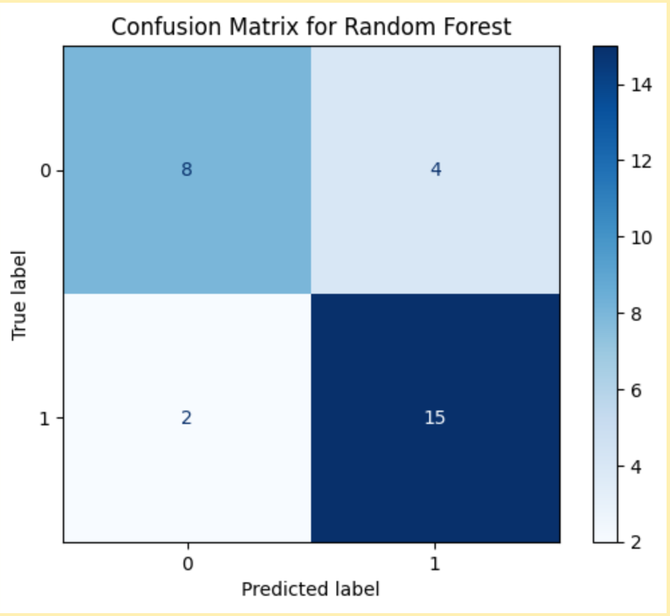
Accuracy
= diagonal cases / all cases
=(10+16)/(10+16+1+2)
= 0.89

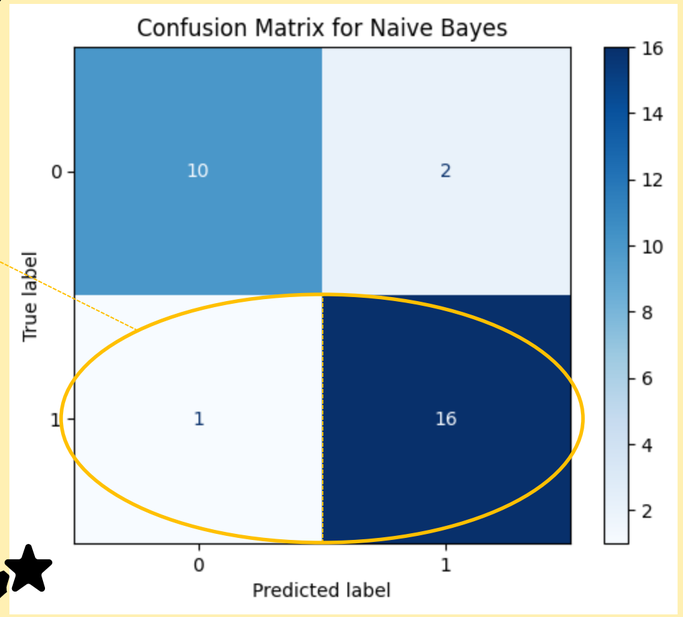
Precision
= true positive / predicted positive
= 13/(7+13) =0.65

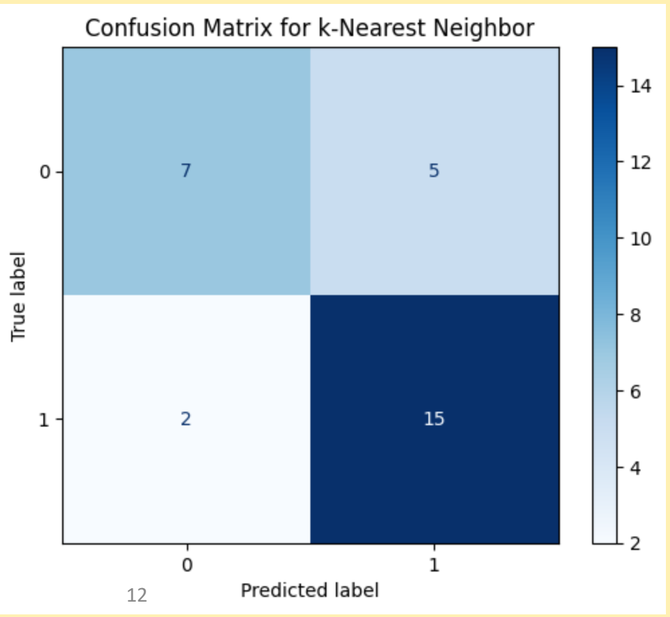
Recall
= predicted positive / actually positive
= 16/(1+16) =0.94

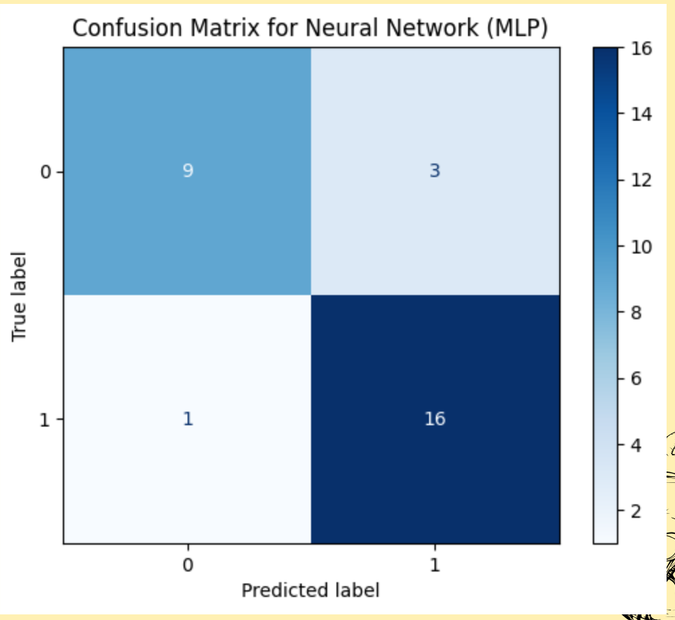












Appendix

References

Altman, D. G., & Bland, J. M. (1994). Diagnostic tests. 1: Sensitivity and specificity. *BMJ (Clinical research ed.)*, *308*(6943), 1552. https://doi.org/10.1136/bmj.308.6943.1552

Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1989). Heart Disease [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X

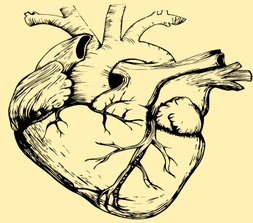
Lipid Panel. (2020, Dec. 4), Retrieved October 22, 2024, from https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/lipid-panel.

National Health Service. (n.d.). Cardiovascular disease. Retrieved September 29, 2024, from https://www.nhs.uk/conditions/cardiovascular-disease/

Sadhwani, P. (2020, June 8). A beginner's guide to collinearity: What it is and how it affects our regression model. Towards Data Science, Retrieved September 29, 2024, from https://towardsdatascience.com/a-beginners-guide-to-collinearity-what-it-is-and-how-it-affects-our-regression-model-d442b421ff95

World's Best Hospitals 2021. (2021)- Top 200 Global - Newsweek Rankings, Retrieved September 27, 2024, from https://www.newsweek.com/best-hospitals-2021

World Health Organization. (n.d.). Cardiovascular diseases (CVDs). Retrieved September 29, 2024, from https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1



thank you

Do you have any
Questions?