Heart Disease Prediction

using Analytics

By Shisir Gurung, Trupal Prajapati, Vaibhavi Shastri, Rithi Veronica

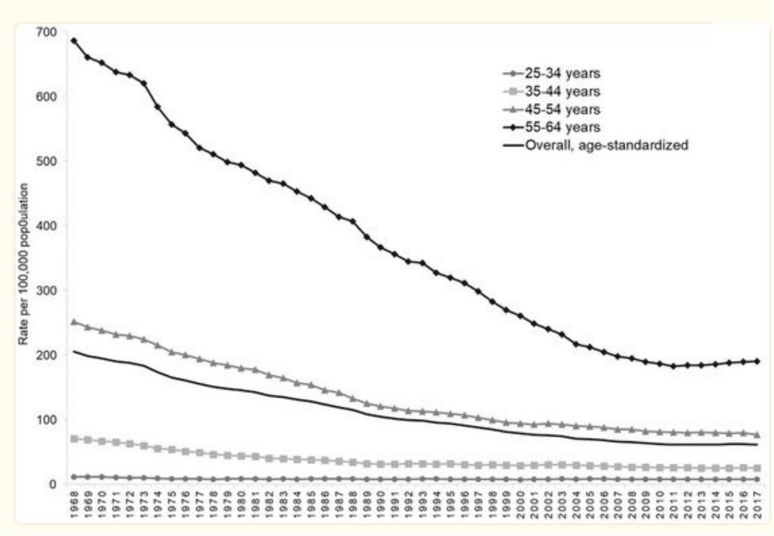
Background

1

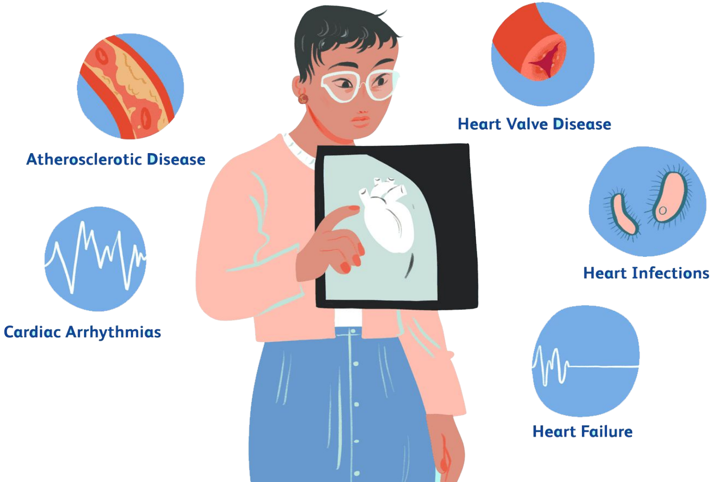
Death every 33 seconds

695,000

1 in every 5 Deaths



$240B



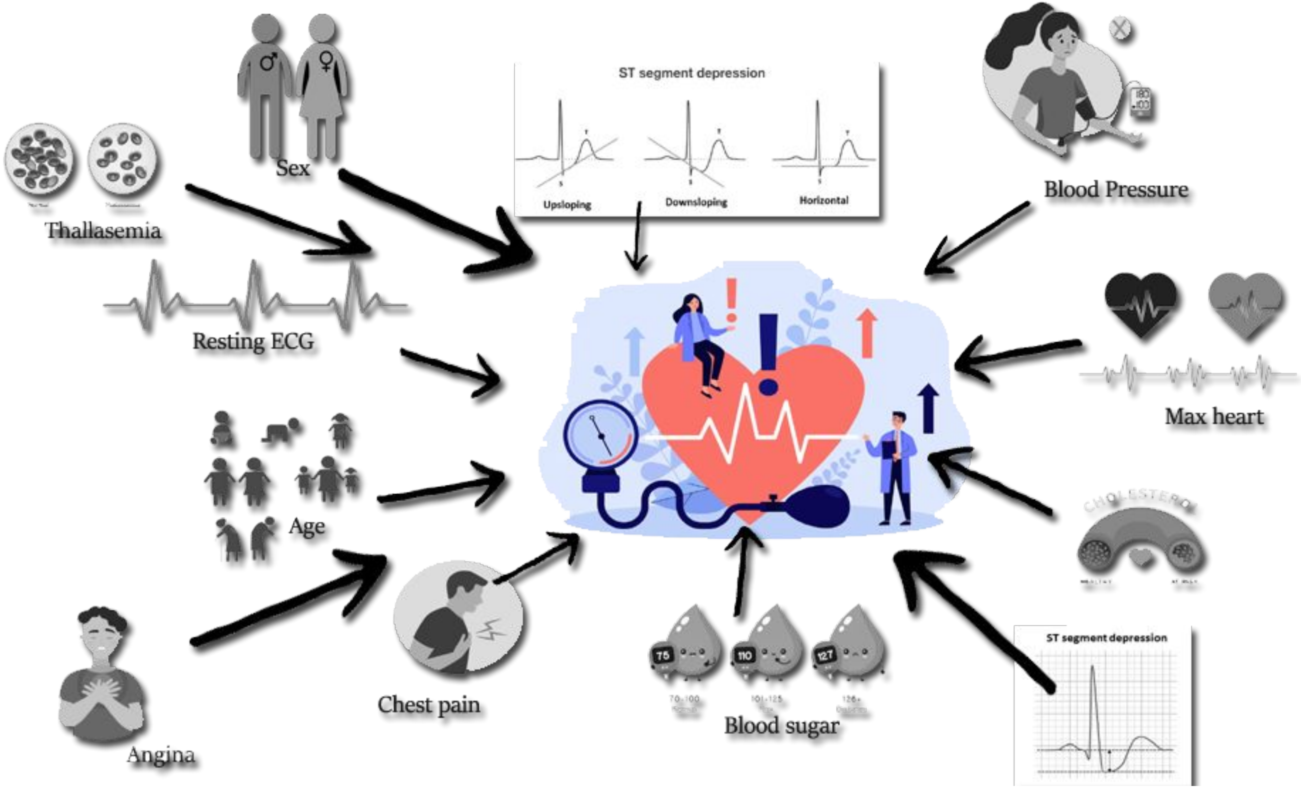
3rd

Globally

Source; Dattani et. al (2023), CDC(2024), Ritchie et. al (2020)

The Question

Can We Predict Heart Disease Based on Biomarkers?

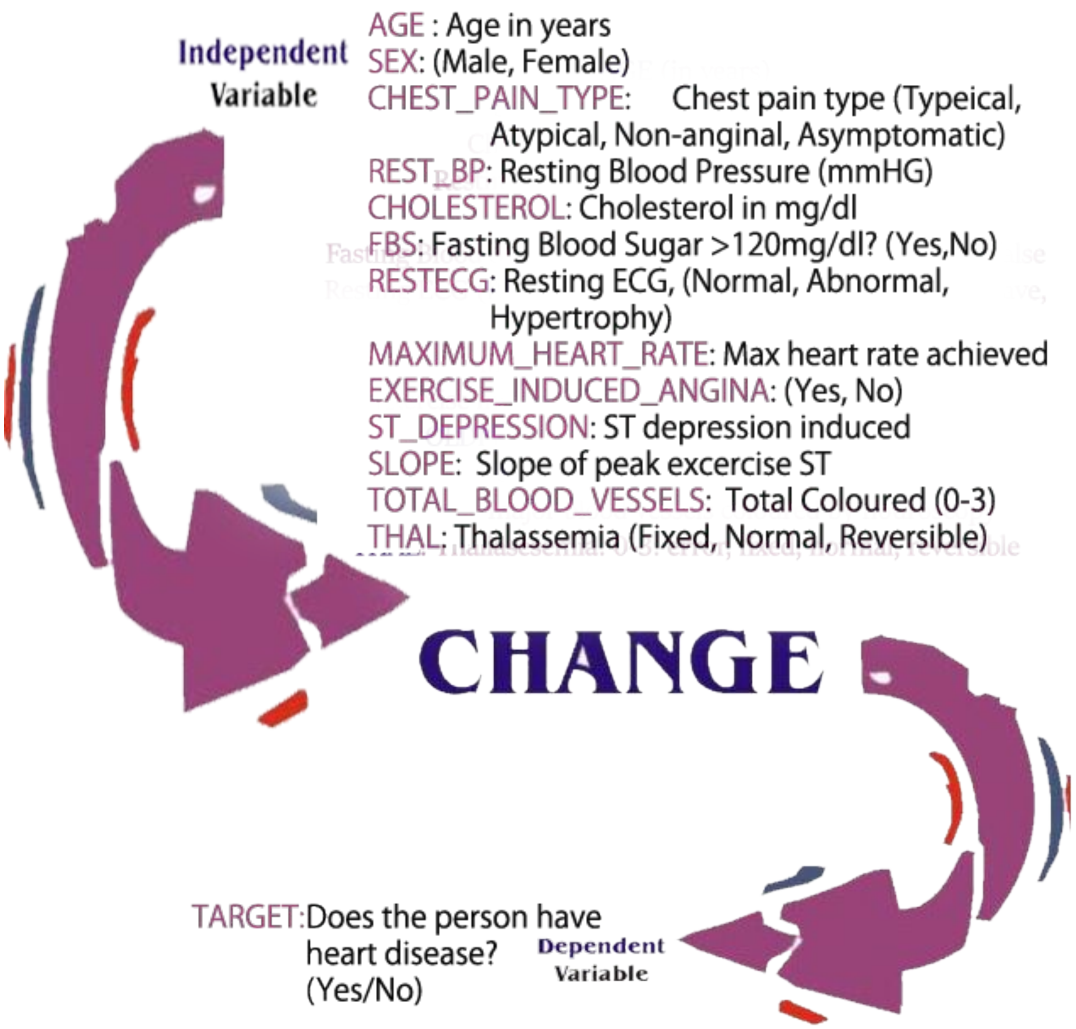


Data Description

Data source:
https://www.kaggle.com/code/desalegngeb/heart-disease-predictions

13
biomarkers

4242
Data points



Data Exploration

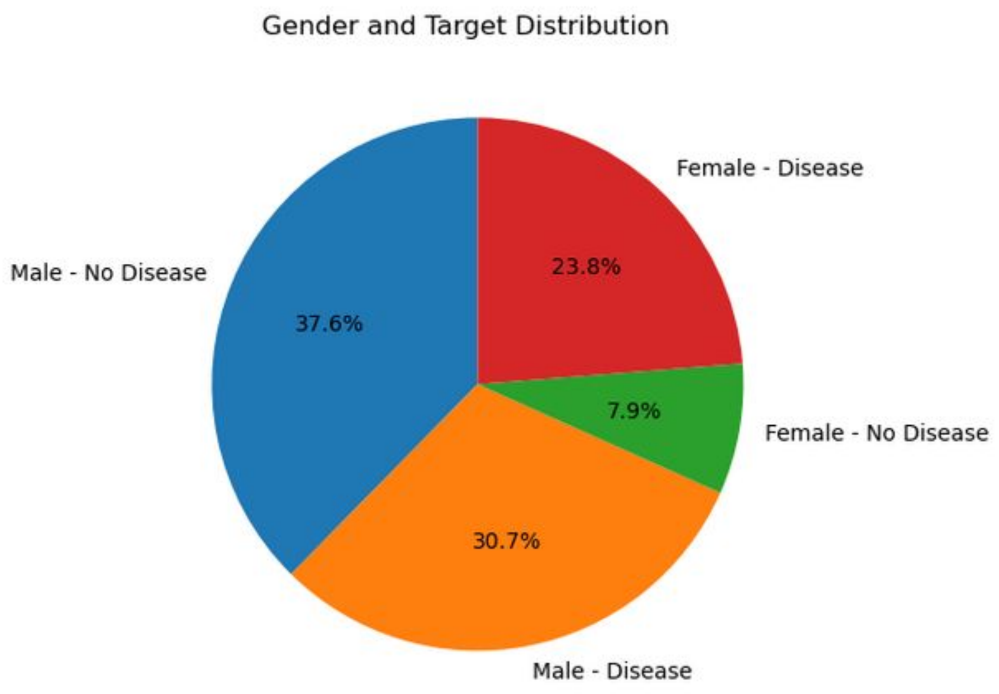
Disease : No Disease

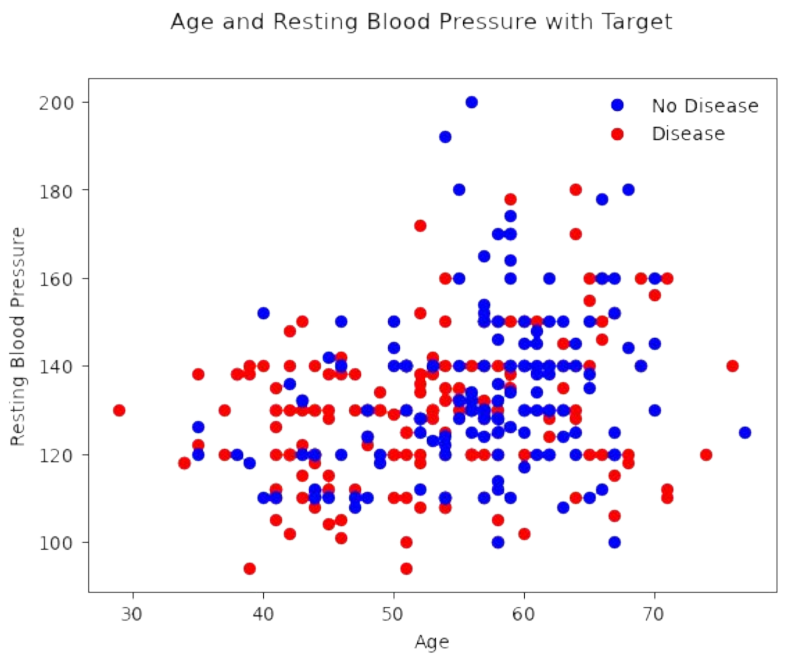
54.1% : 45.9%

Numerical variables summary (post renaming)

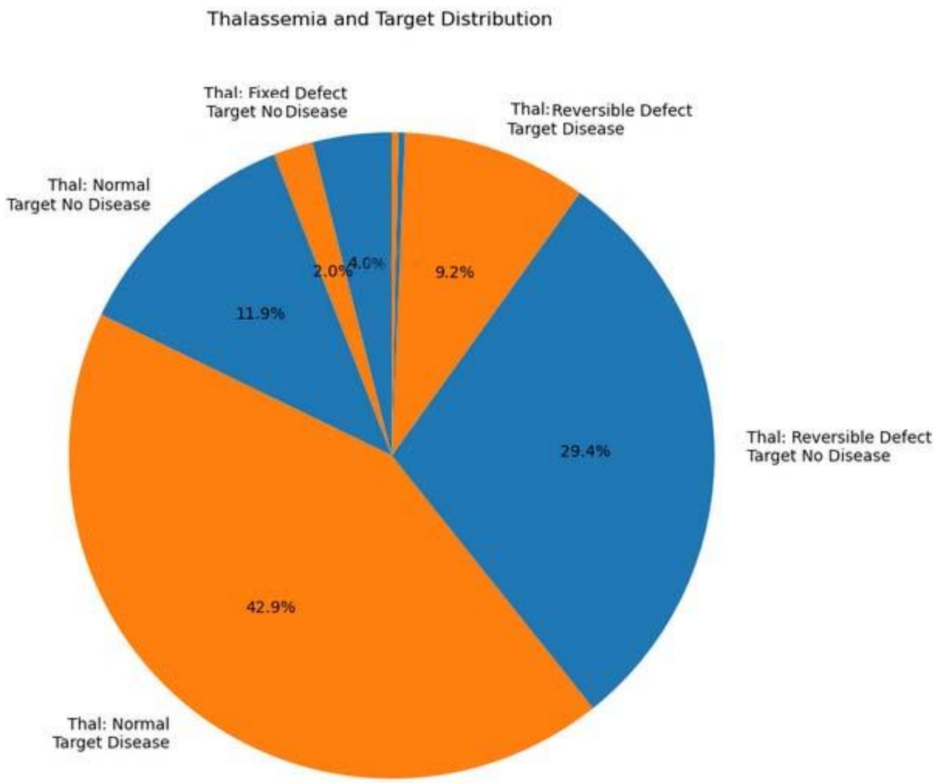
	age	rest_bp	cholesterol	maximum_heart_rate	ST_depression	total_blood_vessels
count	303.00	303.00	303.00	303.00	303.00	303.00
mean	54.37	131.62	246.26	149.65	1.04	0.73
std	9.08	17.54	51.83	22.91	1.16	1.02
min	29.00	94.00	126.00	71.00	0.00	0.00
25%	47.50	120.00	211.00	133.50	0.00	0.00
50%	55.00	130.00	240.00	153.00	0.80	0.00
75%	61.00	140.00	274.50	166.00	1.60	1.00
max	77.00	200.00	564.00	202.00	6.20	4.00

Data Exploration





Data Exploration



Summary of approaches

Multi-classification model:

- Support Vector Machines
  - Imbalanced Data
  - Future prospect not good with computational expense and scalability issues
- Random Forest
  - Finding the right hyperparameter tuning difficult
  - Computationally intensive, effect on future prospects
  - Interpretability issues
- Gradient Boosting
  - Parameter tuning, scalability issues wrt computational requirements
- Logistic Regression
  - Perfect fit for binary output
  - Less computation required

Review of Approaches

Support Vector Machines:

88%

Accuracy (kernel, c)
(linear, 0.5)

Random Forest:

85.2%

(estimators 100, 9 feats)

Gradient Boosting:

86.88%

learning_rate: 0.01
estimators : 50
max_depth : 5

Review of Approaches

Feature Selections

Support Vector Machines

1. Sex
2. Thal
3. Chest_pain_type
4. Total_blood_vessels
5. Exercise_Induced_Angina

Random Forest

1. Thal
2. Total_blood_vessels
3. Chest_pain_type
4. Maximum_heart_rate
5. ST_depression

Gradient Boosting

1. Total_blood_vessels
2. Chest_pain_type
3. Thal
4. ST_depression
5. Age

Summary of the Final Approach

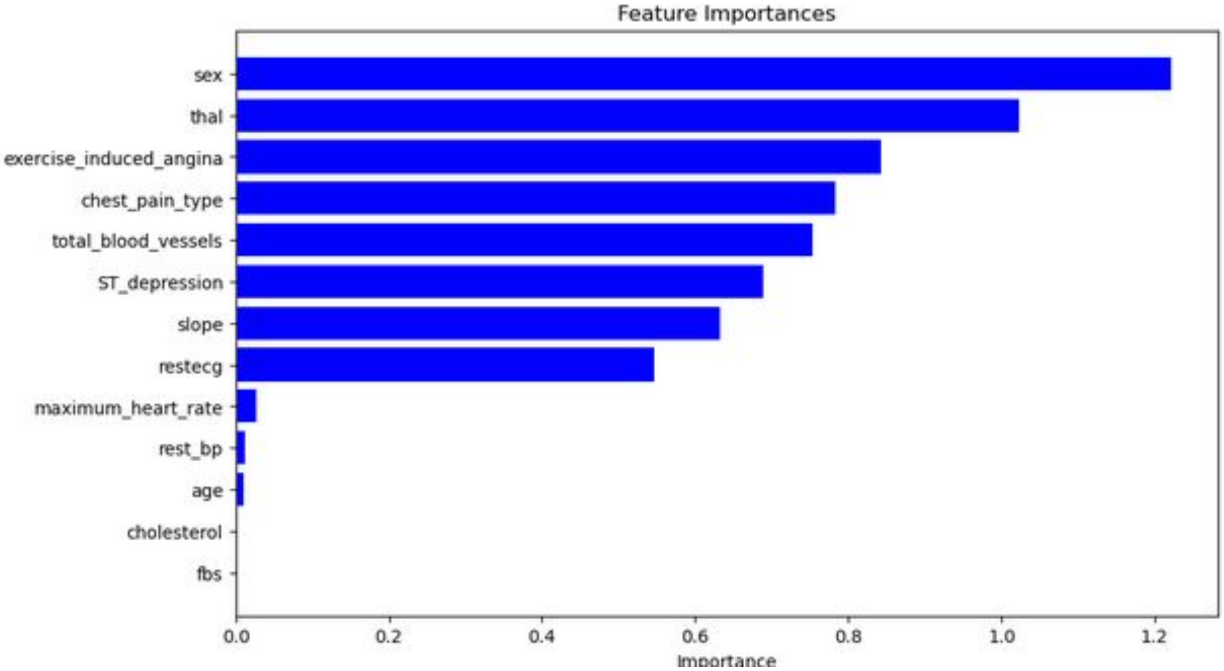
Logistic Regression

- Simplicity: Interpretability
- Binary Classification: Target variable binary
- Linearity Assumption: Biomarkers mostly linear
- Less Prone to Overfitting
- Low computation and memory requirements

88.5%

Accuracy

Review of the Final Approach



Logistic regression equation

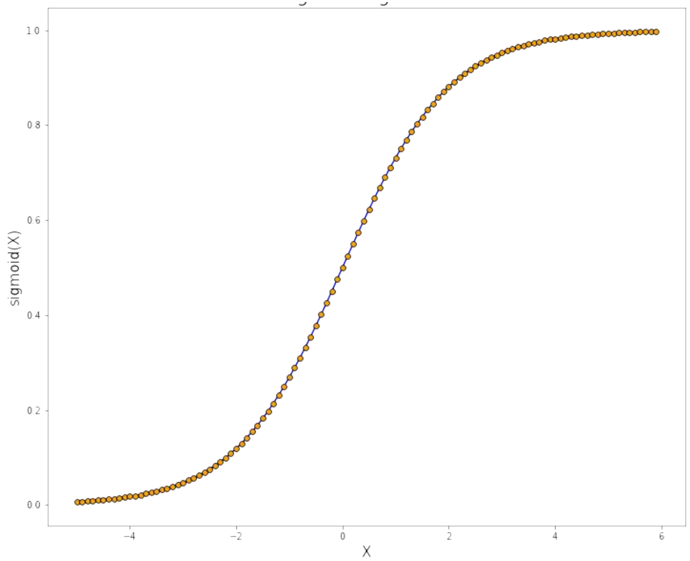
y = (0.010 * *age*) + (-1.221 * *sex*) + (0.783 * *chest_pain_type*) + (-0.012 * *rest_bp*) + (-0.002 * *cholesterol*) + (-0.001 * *fbs*) + (0.546 * *restecg*) + (0.028 * *maximum_heart_rate*) + (-0.843 * *exercise_induced_angina*) + (-0.689 * *ST_depression*) + (0.633 * *slope*) + (-0.753 * *total_blood_vessels*) + (-1.023 * *thal*) + (0.030)

The Final Approach: Logistic Regression

1. Easy to implement

2. Easy to Interpret

3. Economical usage of Computation and Time



Conclusions

Out of SVM, Random Forest,Gradient Booting, Logistic Regression, LR performed best.

Prominent Features: Thal (Thalassemia), Chest_pain_type, Total_blood_vessels

Finally Logistic Regression performed the best with regards to speed, efficiency, accuracy and interpretability.

Implications

1. Fast: Doctors and provide diagnostics/prediction fast

2. Easy to train on less data, so adapts well to new data/scenario

3. Good with limited resources: Applicable in more hospitals (medium to large)

References

CDC (2024) *Heart Disease Facts*. Available from: https://www.cdc.gov/heartdisease/facts.htm

Dattani, S., Samborska, V., Ritchie, H., Roser, M (2023) *Cardiovascular Disease*. Available from: https://ourworldindata.org/cardiovascular-diseases

Ritchey MD, Wall HK, George MG, Wright JS (2020) *US trends in premature heart disease mortality over the past 50 years*. Trends Cardiovasc Med. 2020. 30(6):364-374