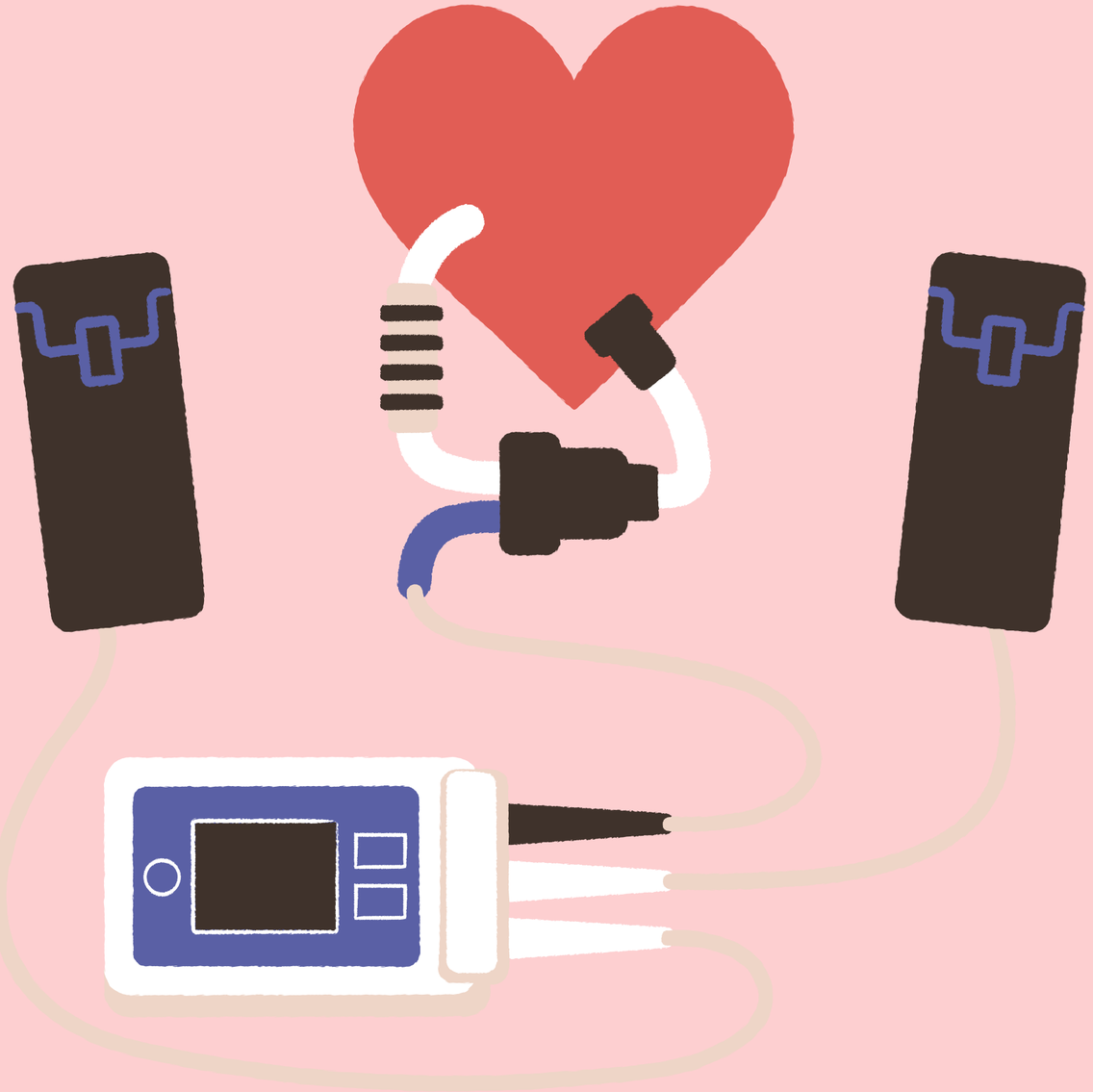
AI PREDICTING HEART FAILURE ANALYSIS



SC1015 PROJECT (FCMB - Team 8)
Veer Dosi, Sparsh Jain, Rowan Jaiswal

Problem Statement

Despite advancements in medical science, many individuals at risk of heart failure remain undiagnosed until the disease progresses to advanced stages. Traditional diagnostic methods can be invasive, expensive, and not universally accessible, creating a need for alternative early detection techniques.

Motivation

Total deaths in the United States from COVID-19 and other leading causes, 2020-2022

	Category	Total deaths (Jan.-Sept. 2022)	Total deaths (2021)	Total deaths (2020)
1	Heart disease	572,336	767,937	764,512
2	Cancer	454,176	604,358	599,607
3	COVID-19	234,434	475,059	343,566
4	Accidents	170,166	226,987	203,033
5	Stroke	123,215	162,769	159,248
6	Chronic respiratory	107,559	141,906	152,051
7	Alzheimer	87,866	119,442	134,271
8	Diabetes	74,716	103,197	101,355
9	Other respiratory	50,635	66,381	66,053
10	Renal failure	42,596	53,057	51,221

Notes: For 2022, the total death sum for each category is for January 1 - September 30, 2022, except deaths from accidents and suicides are from January - September 2021. Chronic respiratory is chronic lower respiratory disease.

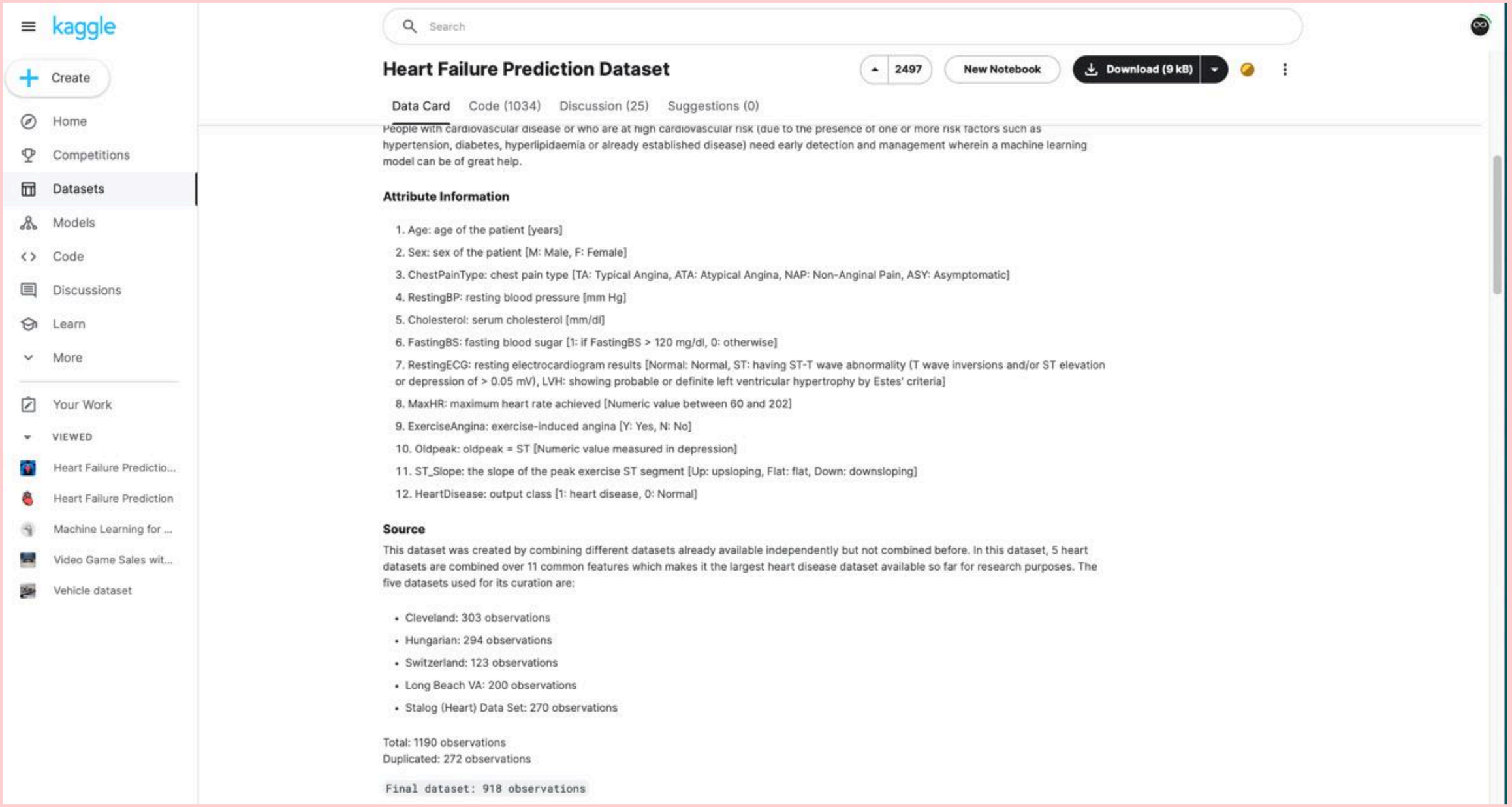
Source: KFF analysis of CDC mortality and KFF COVID-19 tracker data

Peterson-KFF Health System Tracker

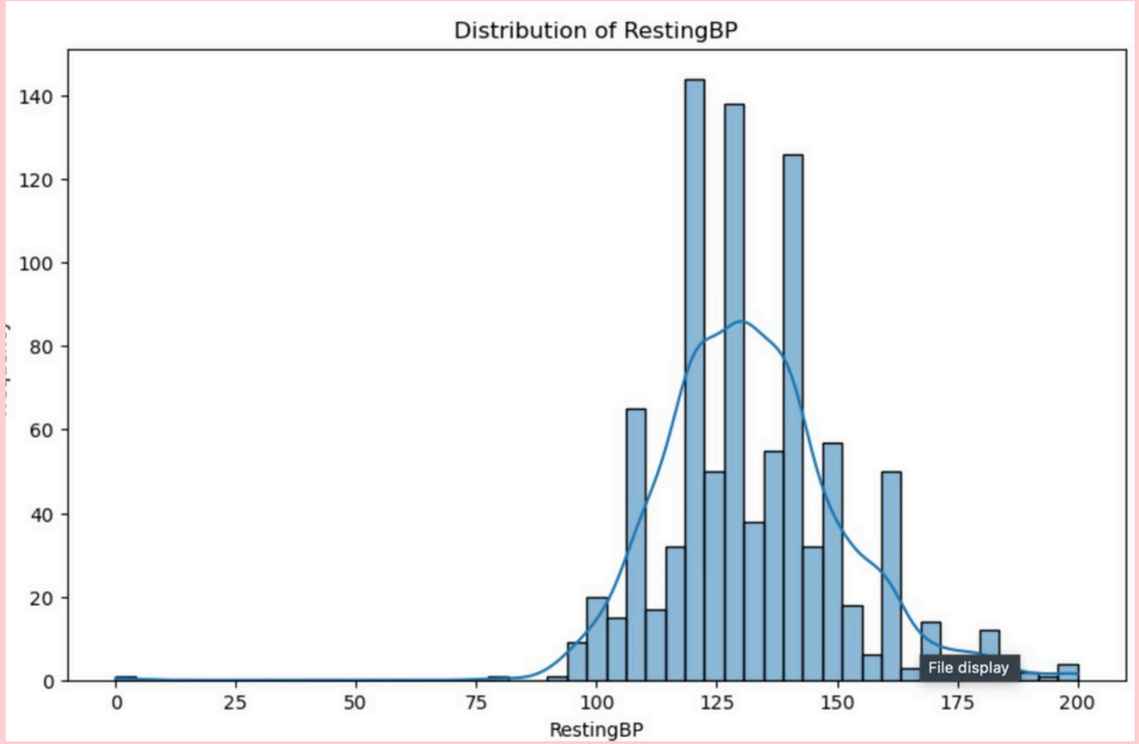
“A few years ago, a beloved family member passed away due to misdiagnosed heart condition. It motivated me to develop a tool that could assist healthcare providers in making more accurate diagnoses of heart disease, potentially saving lives.”

Heart disease remains the **leading cause of death** globally, claiming millions of lives each year. Misdiagnosing heart disease can delay critical treatment and ultimately increase the risk of fatal outcomes."

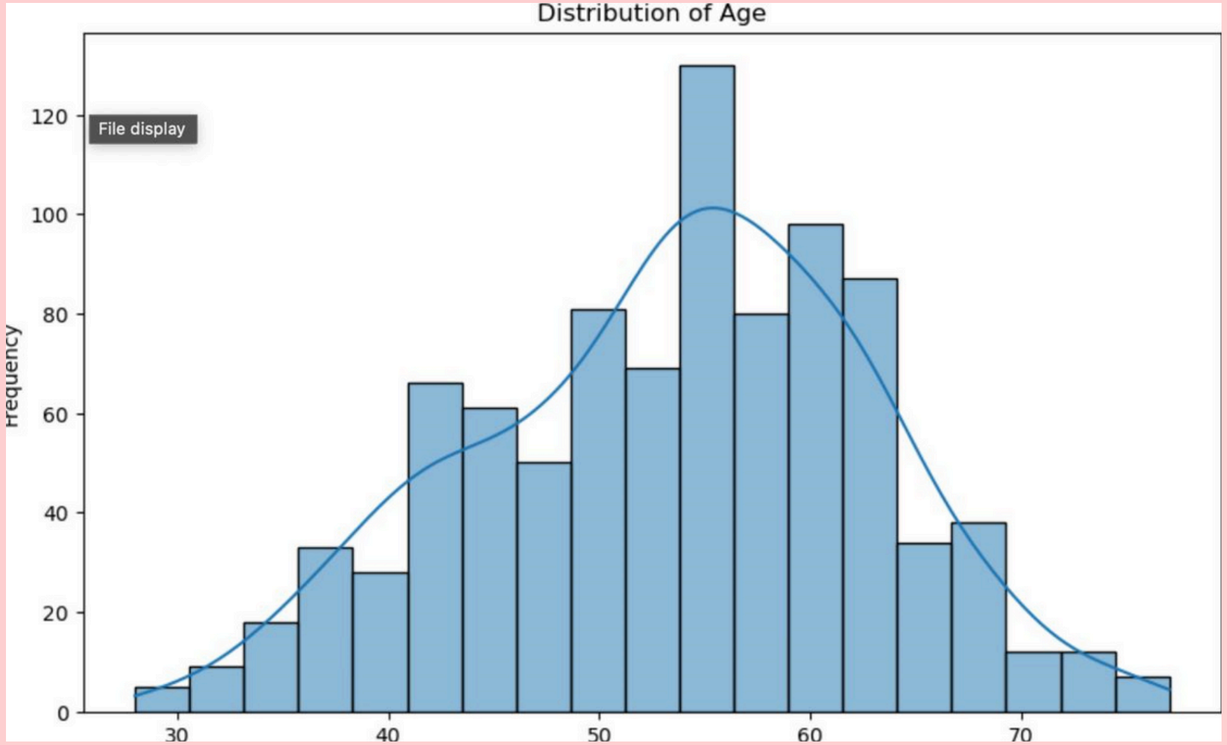
OUR DATASET


kaggle
Search
Create
Home
Competitions
Datasets
Models
Code
Discussions
Learn
More
Your Work
VIEWED
Heart Failure Predictio...
Heart Failure Prediction
Machine Learning for ...
Video Game Sales wit...
Vehicle dataset
Heart Failure Prediction Dataset
2497
New Notebook
Download (9 kB)
Data Card
Code (1034)
Discussion (25)
Suggestions (0)
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.
Attribute Information
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]
Source
This dataset was created by combining different datasets already available independently but not combined before. In this dataset, 5 heart datasets are combined over 11 common features which makes it the largest heart disease dataset available so far for research purposes. The five datasets used for its curation are:
• Cleveland: 303 observations
• Hungarian: 294 observations
• Switzerland: 123 observations
• Long Beach VA: 200 observations
• Stalog (Heart) Data Set: 270 observations
Total: 1190 observations
Duplicated: 272 observations
Final dataset: 918 observations

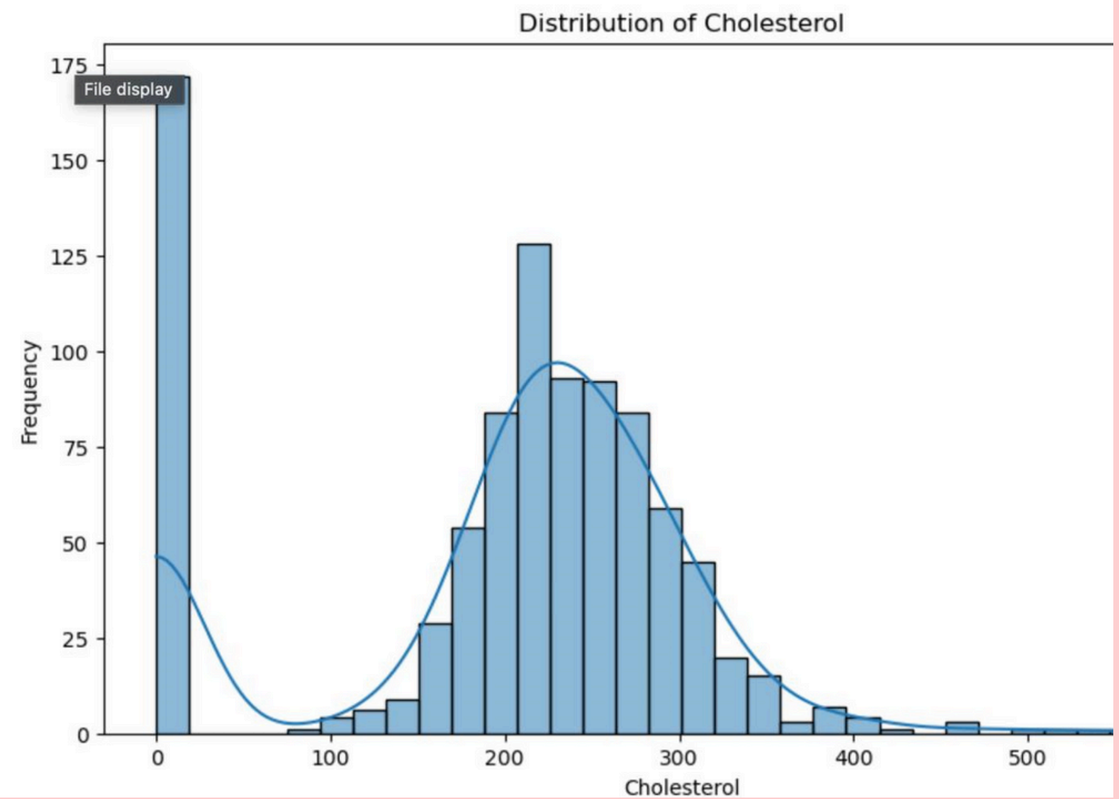
EXPLORATORY DATA ANALYSIS



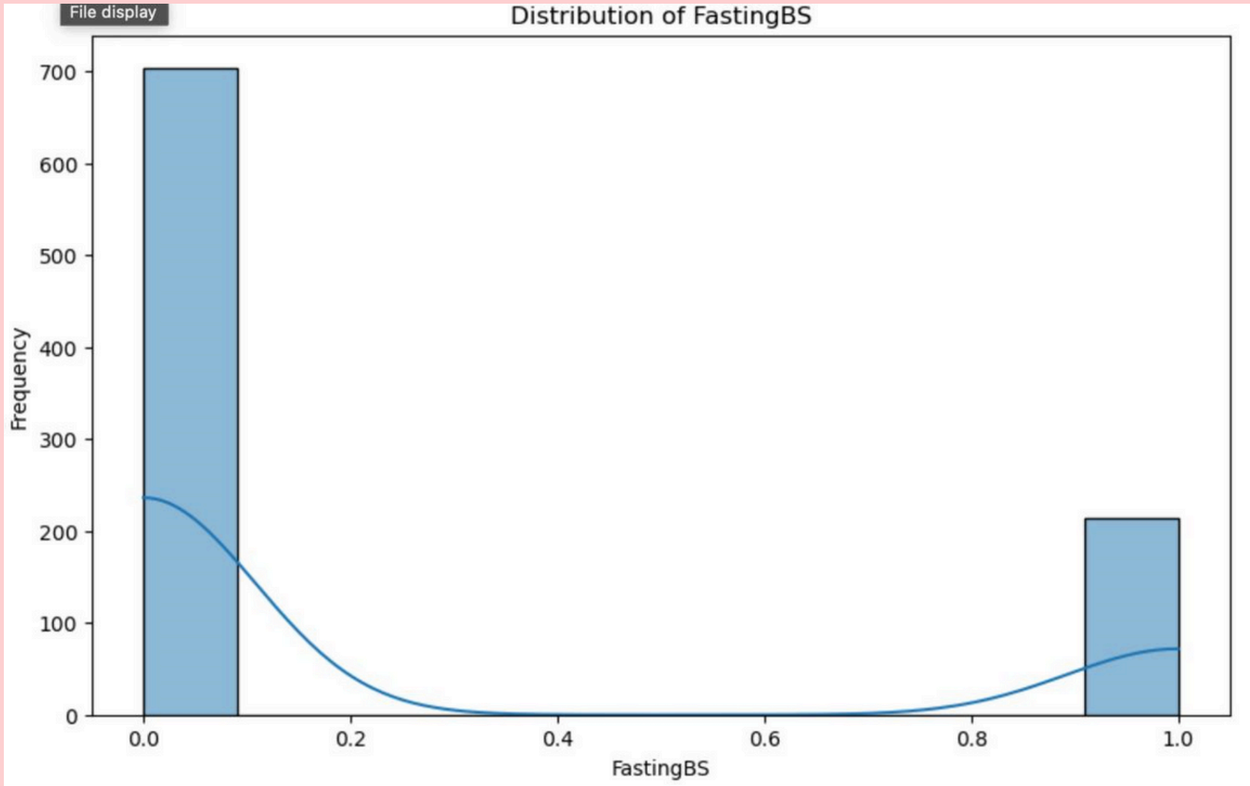
The presence of subjects with **higher values** of resting BP suggests a segment **at risk**



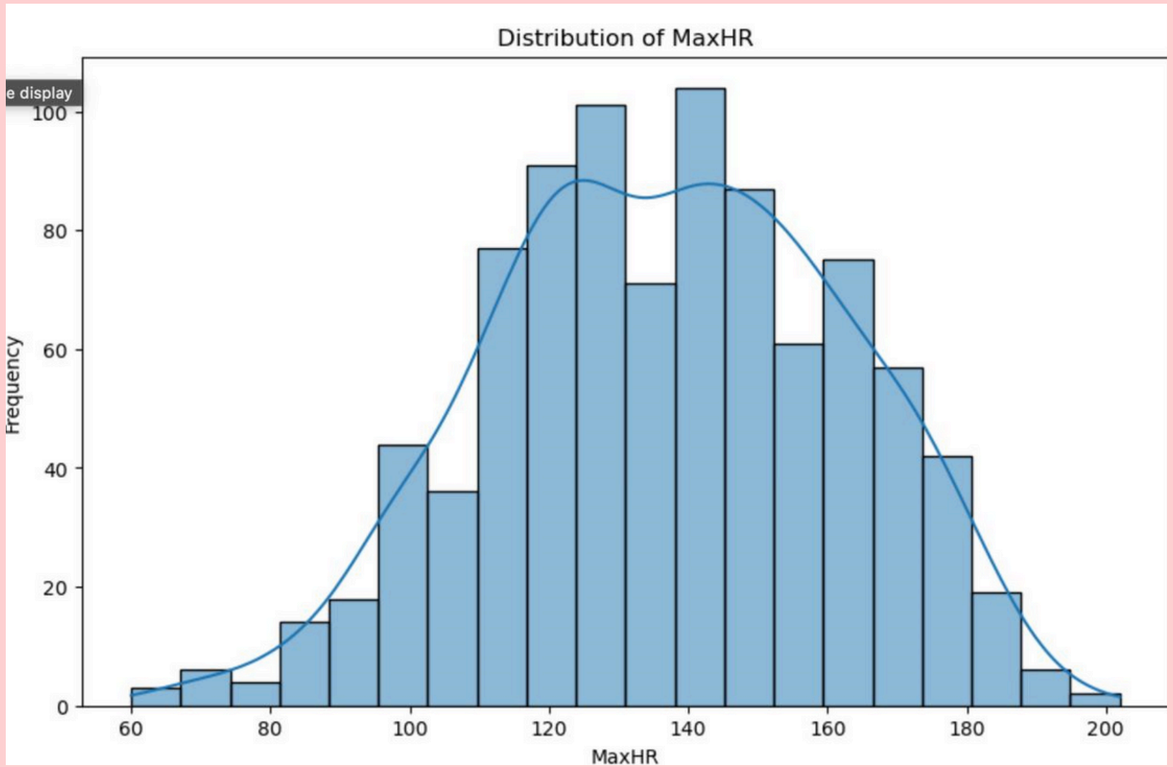
Most subjects fall between **50 and 60** years old



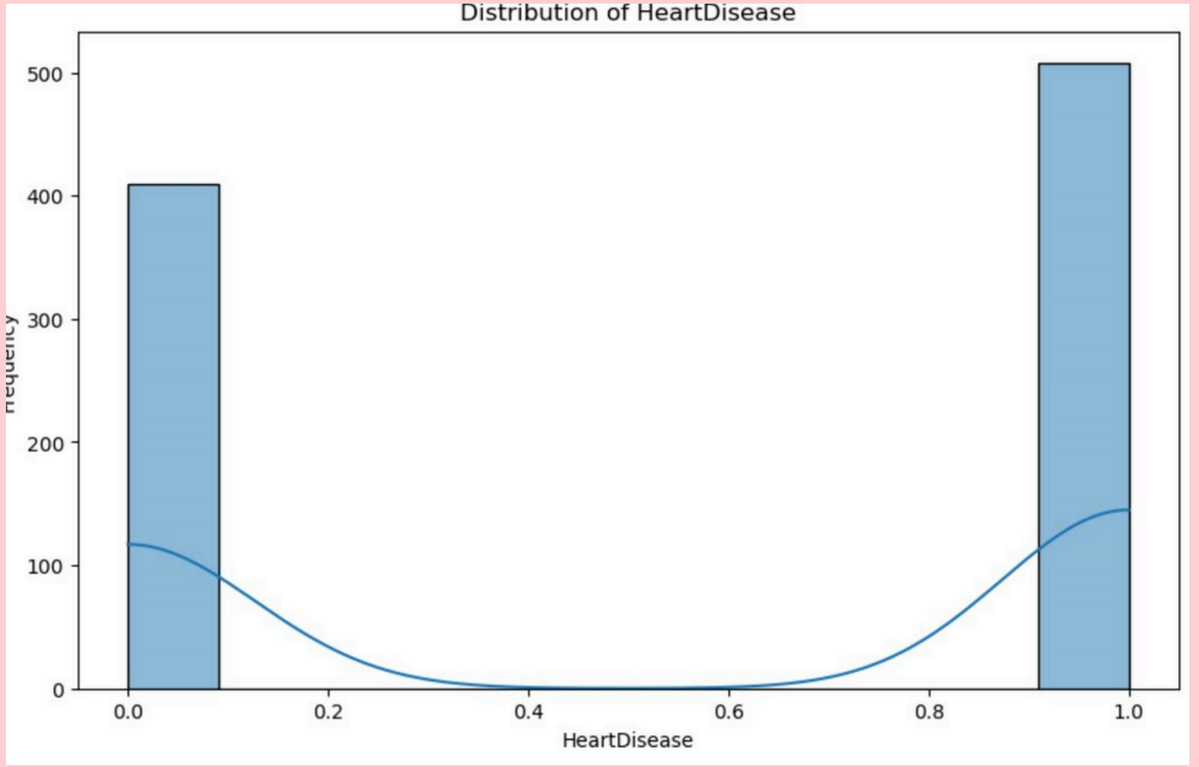
Its **negative correlation** with 'MaxHR' suggests that individuals with **higher cholesterol** might have reduced cardiac performance



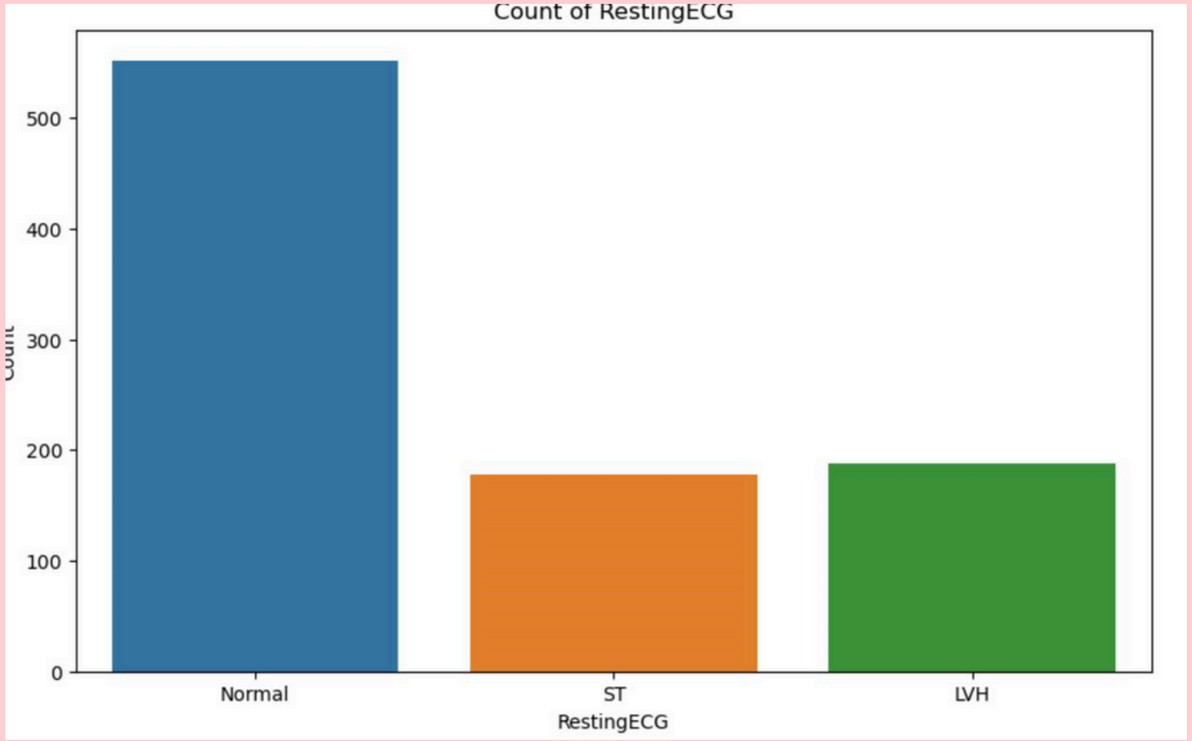
Most individuals have a fasting blood sugar below the diabetes threshold. A **substantial** number **exceed it**, highlighting a subgroup **at elevated risk**



A higher maximum heart rate during exercise is generally healthier



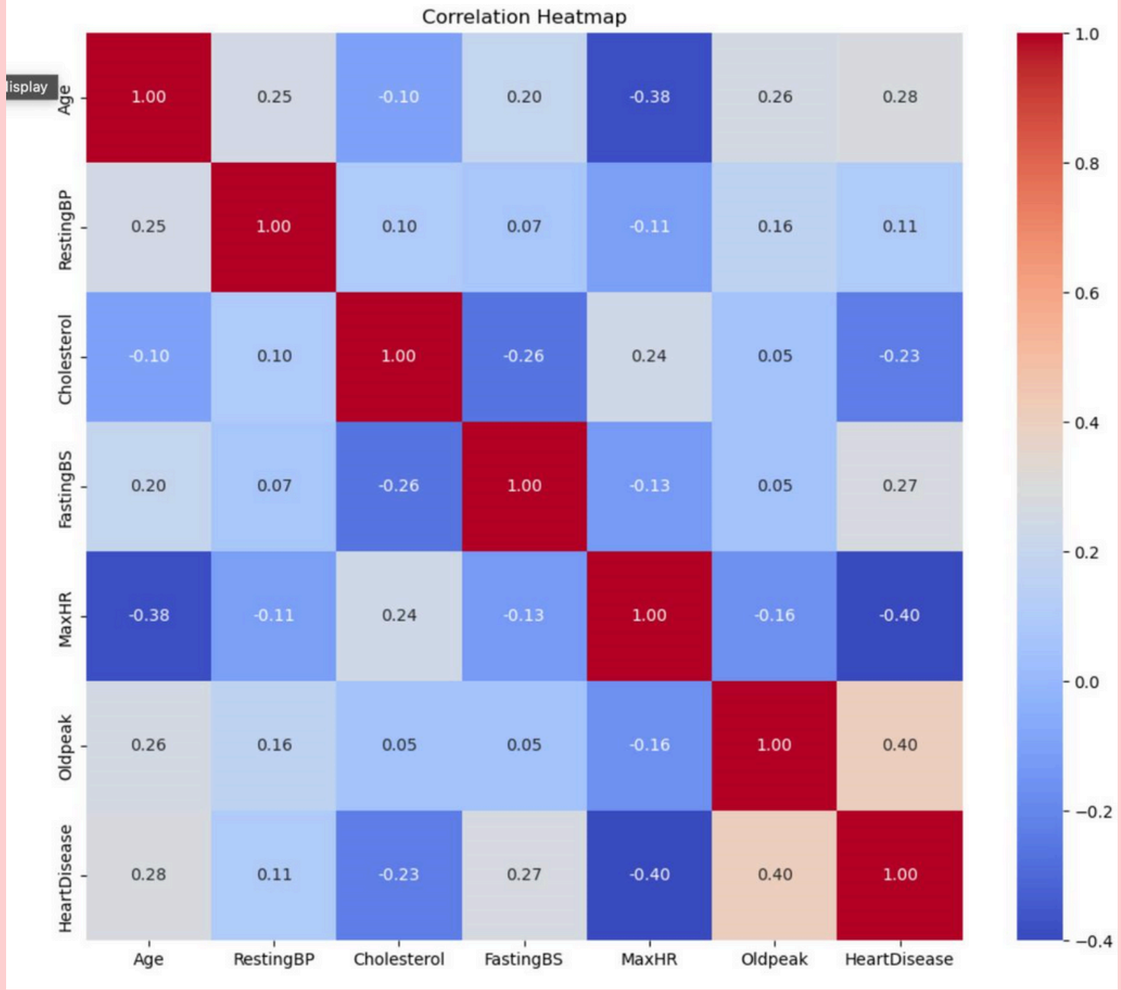
Even split in the dataset -> suggests dataset has a balanced representation of cases with and without heart disease which is **ideal**.



'Normal' category-> Most frequent

'ST' -> Abnormalities

'LVH' -> The least common



We can observe that **'MaxHR'** is negatively correlated with **'HeartDisease'**, suggesting that higher maximum heart rates are associated with lower heart disease risk.

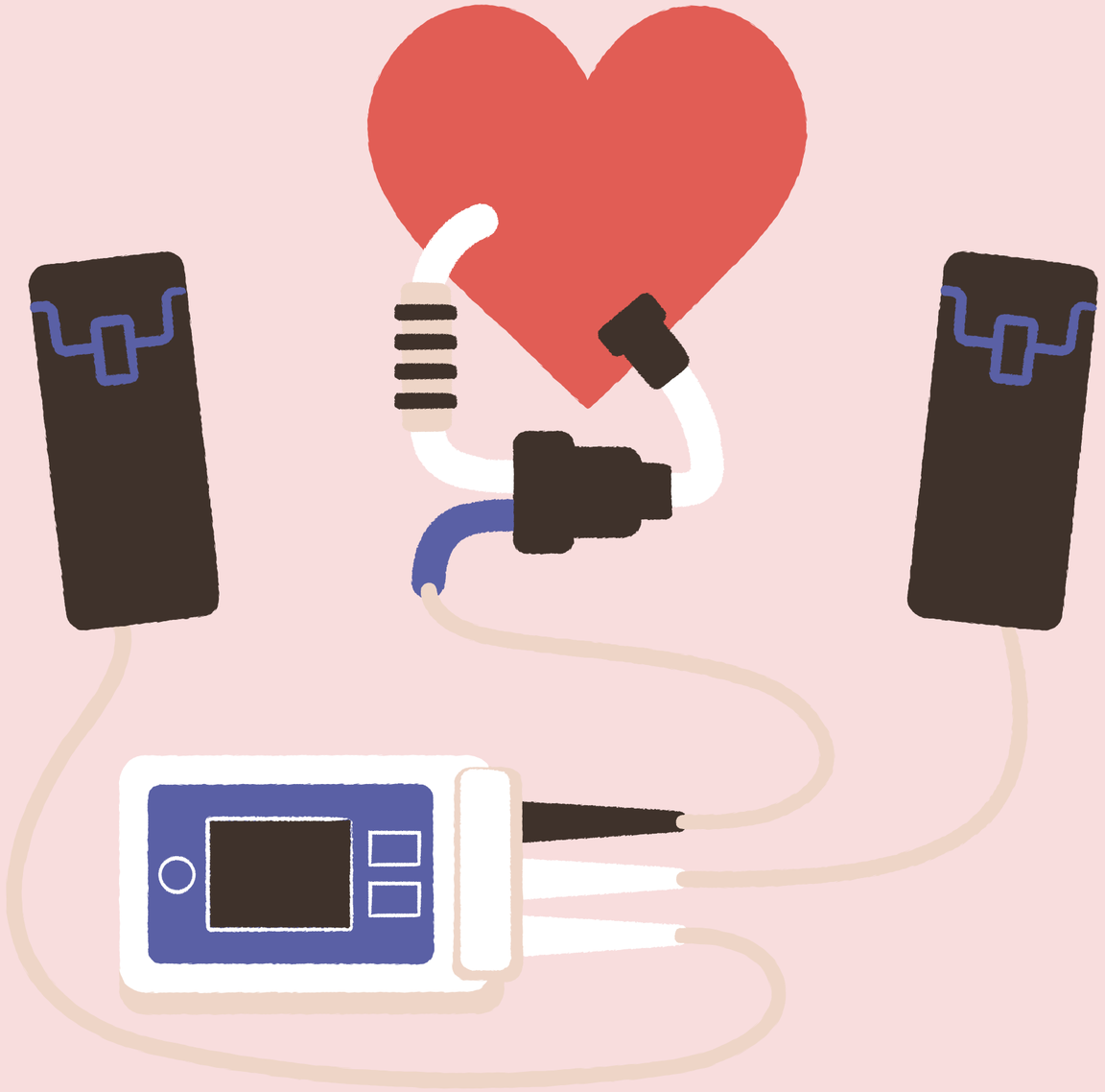
RADAR CHART

- **Cholesterol Levels:** It's notable that Group 1 have a larger area under the curve for cholesterol, suggesting higher average cholesterol levels in this group compared to Group 0.

- **Resting Blood Pressure :** Average resting blood pressure is higher in the group with heart disease. This is indicated by the more extended radial line for Group 1, pointing to a potential link between elevated blood pressure and the risk of heart disease.

- **Maximum Heart Rate:** In contrast, the maximum heart rate shows a reverse trend. Group 0, without heart disease, has a higher average maximum heart rate than Group 1. This could suggest that those without heart disease have better cardiac function, as evidenced by a higher achievable heart rate.

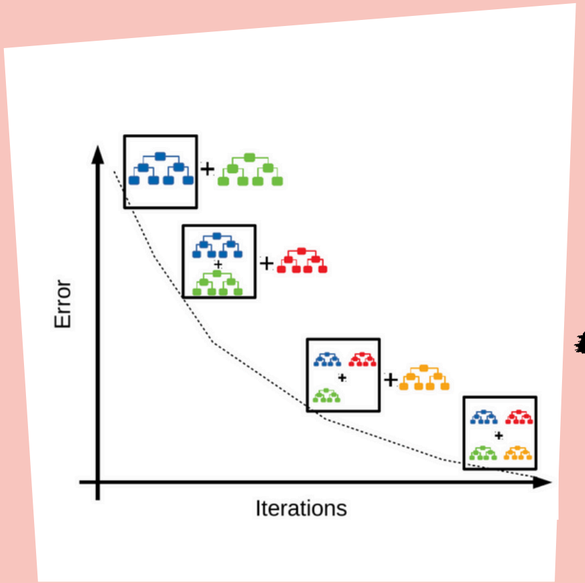
CORE ANALYSIS



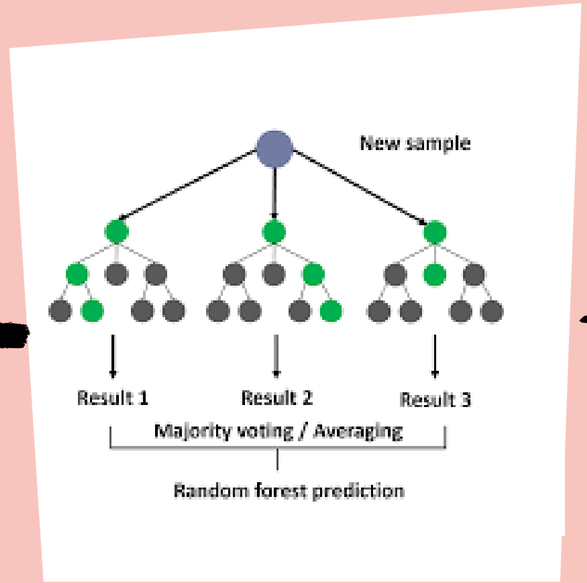
PRE-PROCESSING

```
df.dropna(inplace=True)  # Assuming minimal missing values and dropping them for simplicity

# Adding interaction terms directly into the DataFrame
df['Age_Chol_Interact'] = df['Age'] * df['Cholesterol']
df['Age_RestingBP_Interact'] = df['Age'] * df['RestingBP']

# Polynomial Features for Age and MaxHR
poly = PolynomialFeatures(degree=2, include_bias=False)
poly_features = poly.fit_transform(df[['Age', 'MaxHR']])
poly_feature_names = poly.get_feature_names_out(['Age', 'MaxHR'])

# Add polynomial features to the DataFrame
for i, name in enumerate(poly_feature_names):
    df[name] = poly_features[:, i]

# Update the feature lists
categorical_features = ['Sex', 'ChestPainType', 'FastingBS', 'RestingECG', 'ExerciseAngina', 'ST_Slope']
continuous_features = ['Age', 'RestingBP', 'Cholesterol', 'MaxHR', 'Oldpeak', 'Age_Chol_Interact', 'Age_Restir

# Prepare the dataset
X = df.drop('HeartDisease', axis=1)
y = df['HeartDisease']

# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Preprocessing pipeline for numerical and categorical features
preprocessor = ColumnTransformer(
    transformers=[
        ('num', StandardScaler(), continuous_features),
        ('cat', OneHotEncoder(), categorical_features)
    ])
```

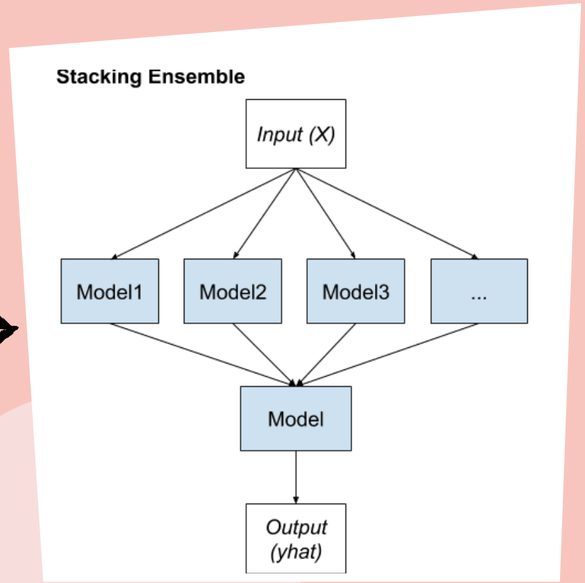
ML MODELS USED



GRADIENT BOOSTING



RANDOM FOREST



STACKING

HOW WE USED ML TO TACKLE OUR PROBLEM

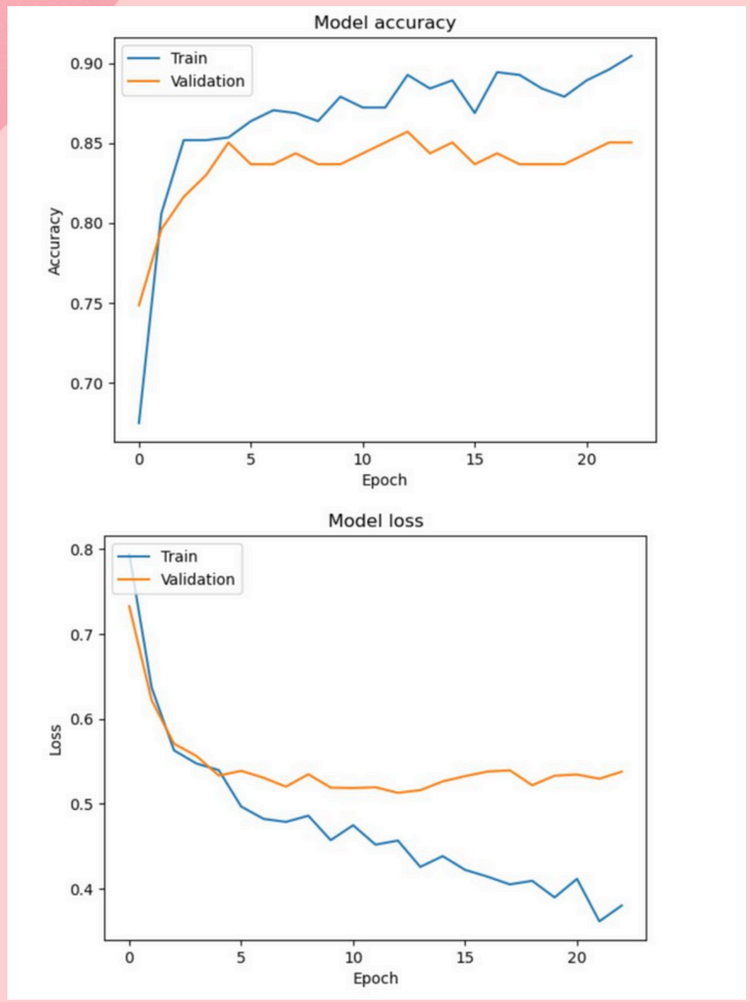
Our approach harnesses these techniques together.

- **Random Forest** -> Provides a diverse set of classifiers
- **Gradient Boosting** -> Offers refinement through focusing on mistakes.
- **Stacking** -> Takes these predictions as input
- Logistic Regression serves as the meta-model to finalize the prediction.

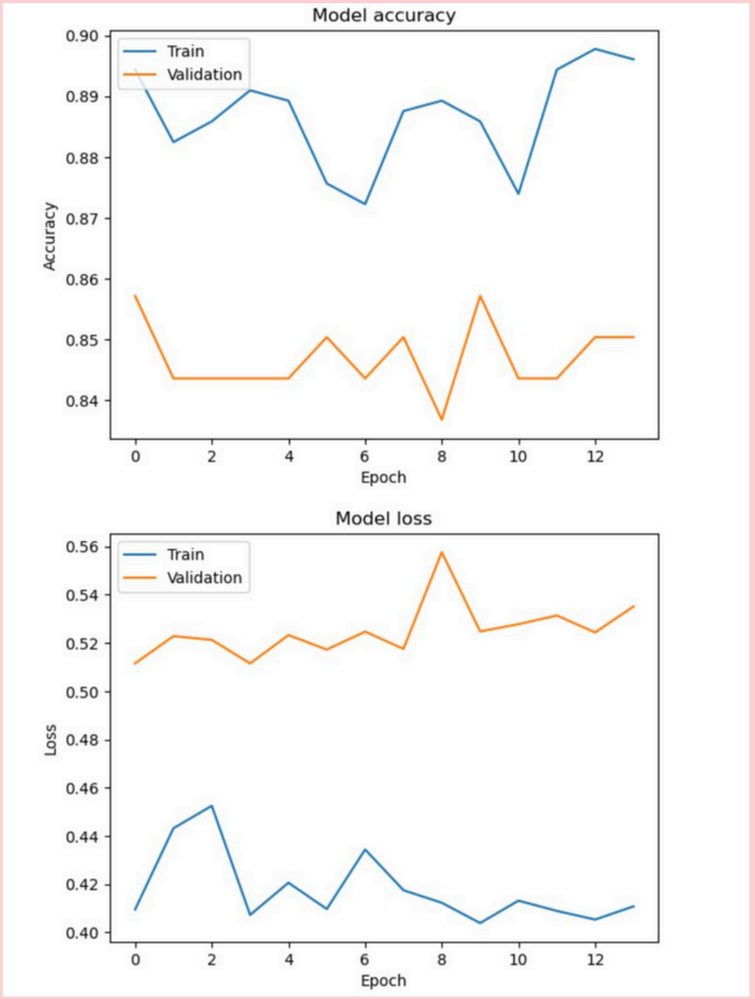
This creates a model that learns from itself, improving as it encounters new data

We've developed a system that's not only accurate but also adaptive

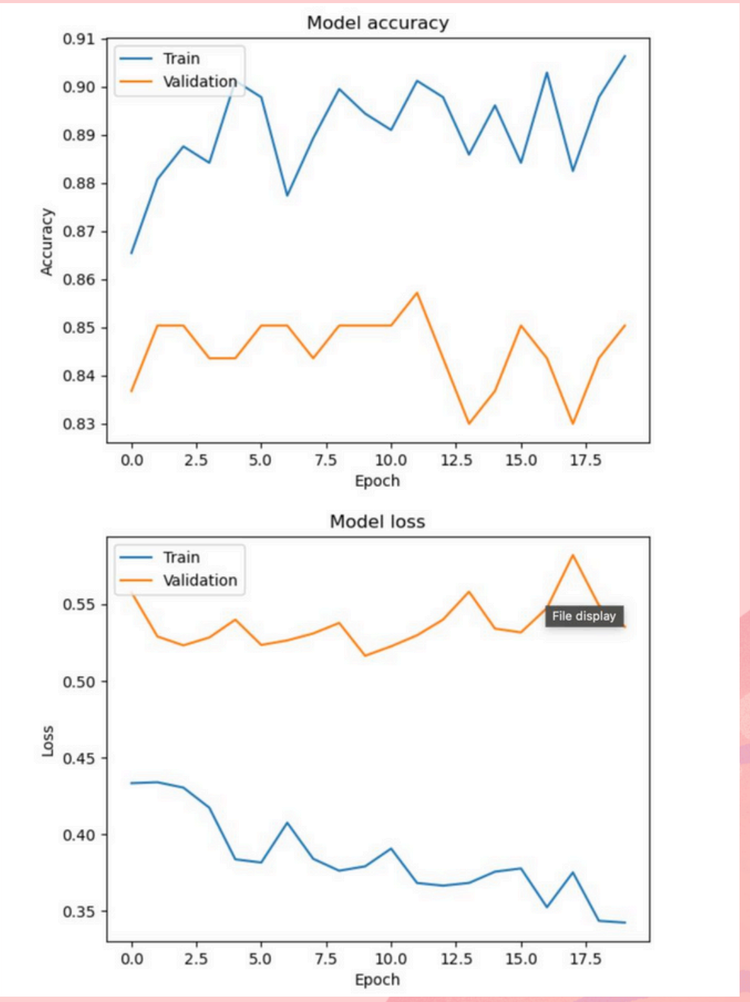
INDIVIDUAL NEURAL NETWORK OPTIMIZERS



ADAM

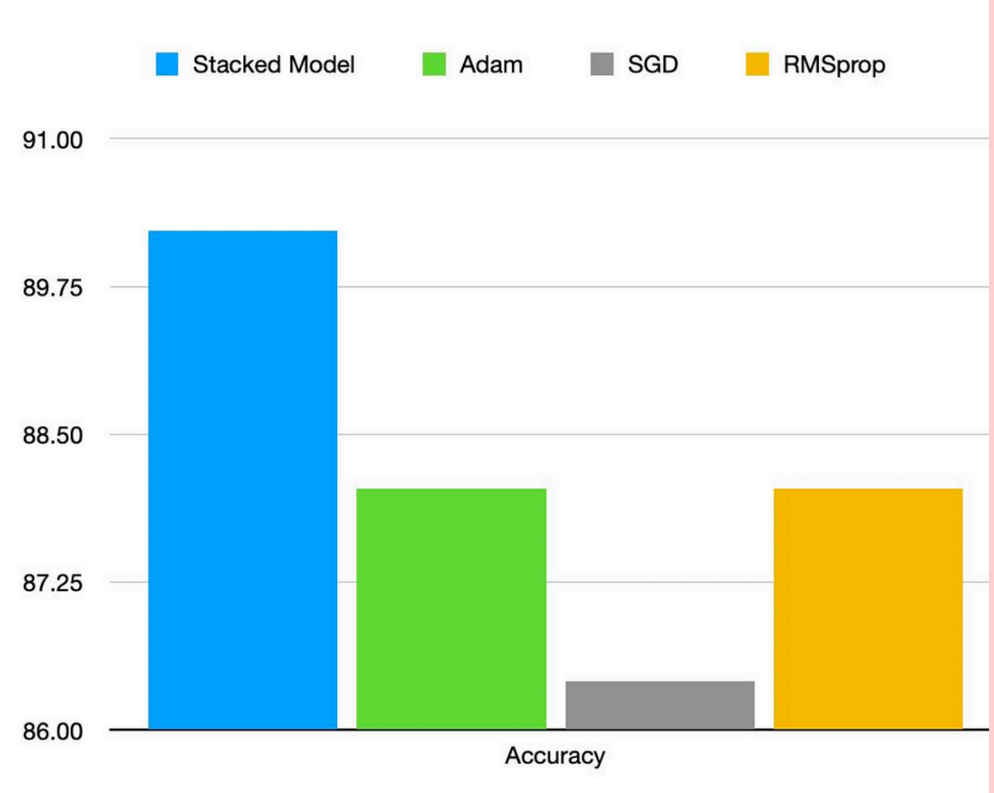


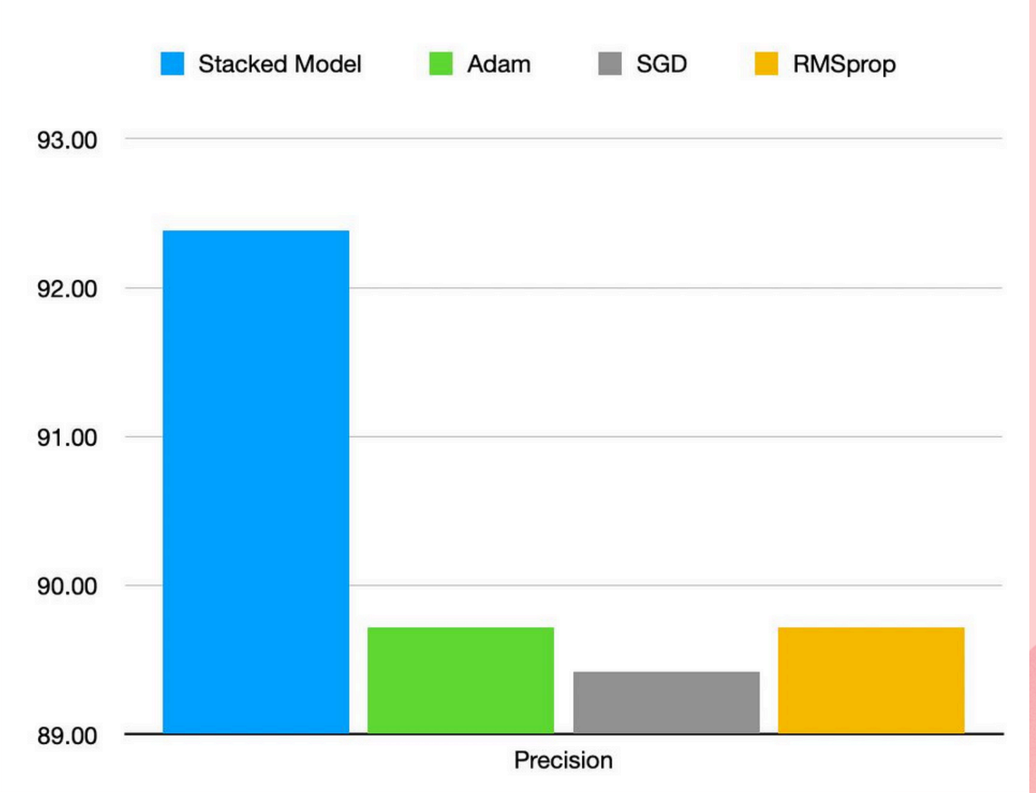
SGD



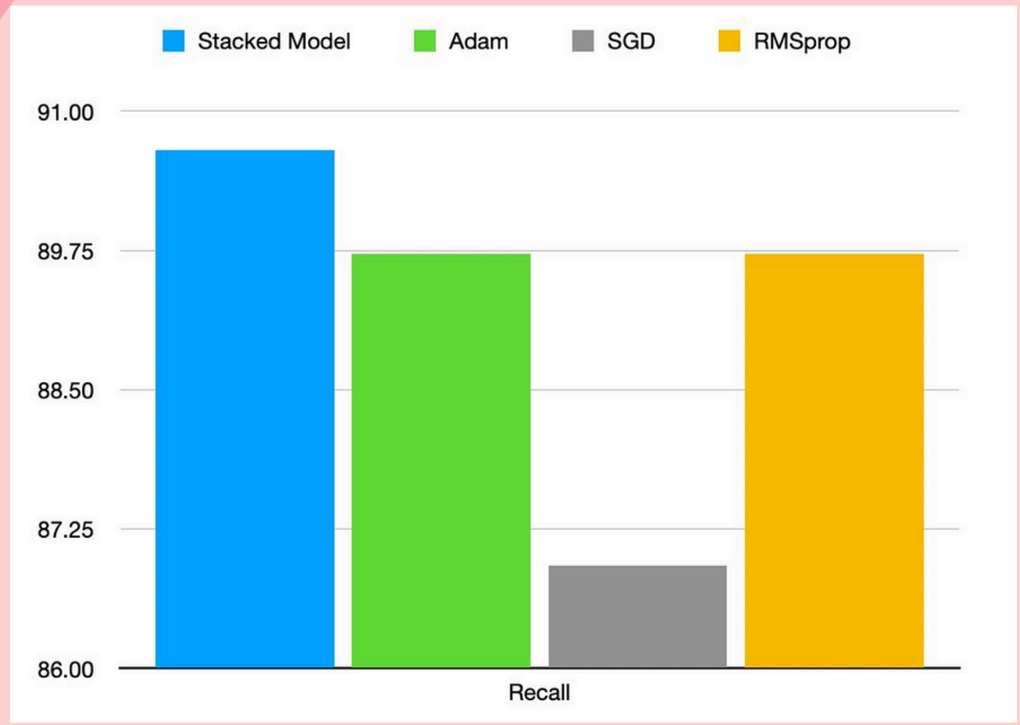
RMSprop

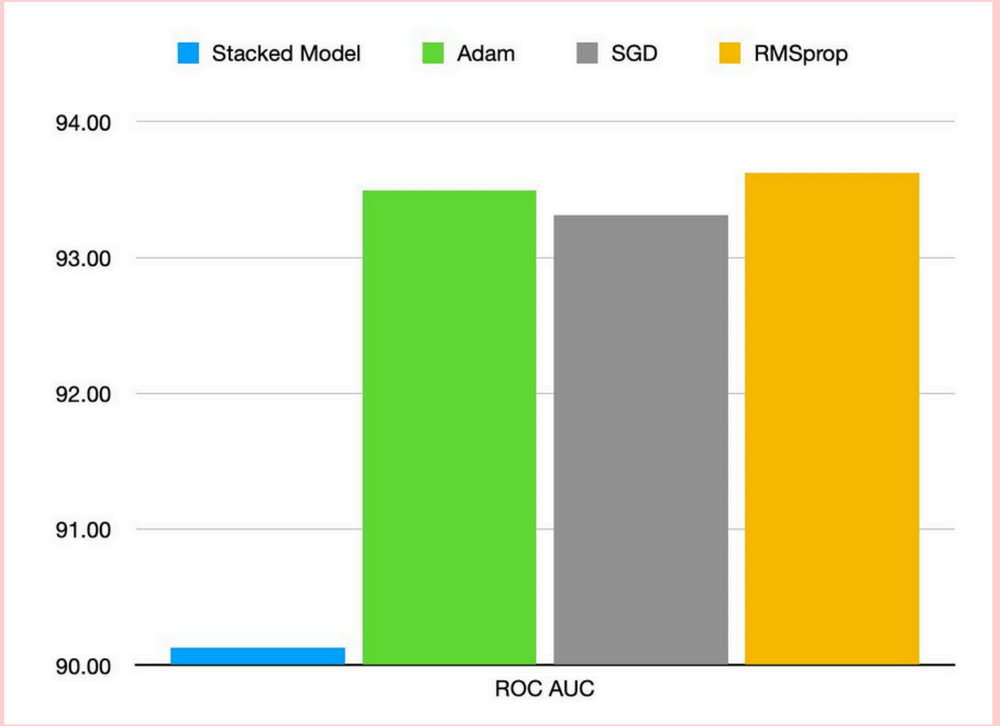
DATA DRIVEN INSIGHTS





DATA DRIVEN INSIGHTS





This showcases the power of ensemble methods, which combine the predictive capacity of multiple learners to achieve greater accuracy than any single learner alone.

```python
user_input_1 = {
    "Age": 20,
    "Sex": "F",
    "ChestPainType": "ATA",
    "RestingBP": 110,
    "Cholesterol": 270,
    "FastingBS": 1,
    "RestingECG": "LVH",
    "MaxHR": 150,
    "ExerciseAngina": "N",
    "Oldpeak": 1.4,
    "ST_Slope": "Up"
}

classification_1, probability_1 = predict_heart_failure(user_input_1, stack_pipeline)

# Output the model response
print(f"Prediction (Class): {'Heart Disease' if classification_1 == 1 else 'No Heart Disease'}")
print(f"Prediction (Probability of Heart Disease): {probability_1:.4f}")
```

Python

```
Prediction (Class): No Heart Disease
Prediction (Probability of Heart Disease): 0.1214
```

```python
user_input_2 = {
    "Age": 50,
    "Sex": "M",
    "ChestPainType": "NAP",
    "RestingBP": 130,
    "Cholesterol": 250,
    "FastingBS": 0,
    "RestingECG": "Normal",
    "MaxHR": 155,
    "ExerciseAngina": "N",
    "Oldpeak": 1.4,
    "ST_Slope": "Flat"
}

classification_2, probability_2 = predict_heart_failure(user_input_2, stack_pipeline)

# Output the model response
print(f"Prediction (Class): {'Heart Disease' if classification_2 == 1 else 'No Heart Disease'}")
print(f"Prediction (Probability of Heart Disease): {probability_2:.4f}")
```

Python

```
Prediction (Class): Heart Disease
Prediction (Probability of Heart Disease): 0.5513
```

FINAL OUTCOME

- While neural networks with advanced optimizers like **RMSprop** offer **high discrimination capabilities**, the **stacked model excels in overall accuracy** and balance of metrics.
- It demonstrates the **strength of ensemble techniques** in making nuanced predictions, particularly in complex domains like medical diagnosis.
- Our project confirms that sophisticated machine learning approaches, especially those combining multiple models, can significantly enhance predictive analytics in healthcare.

These findings open the door to future applications where such models could be deployed to save lives by predicting heart disease with high accuracy and reliability.

Thank you!