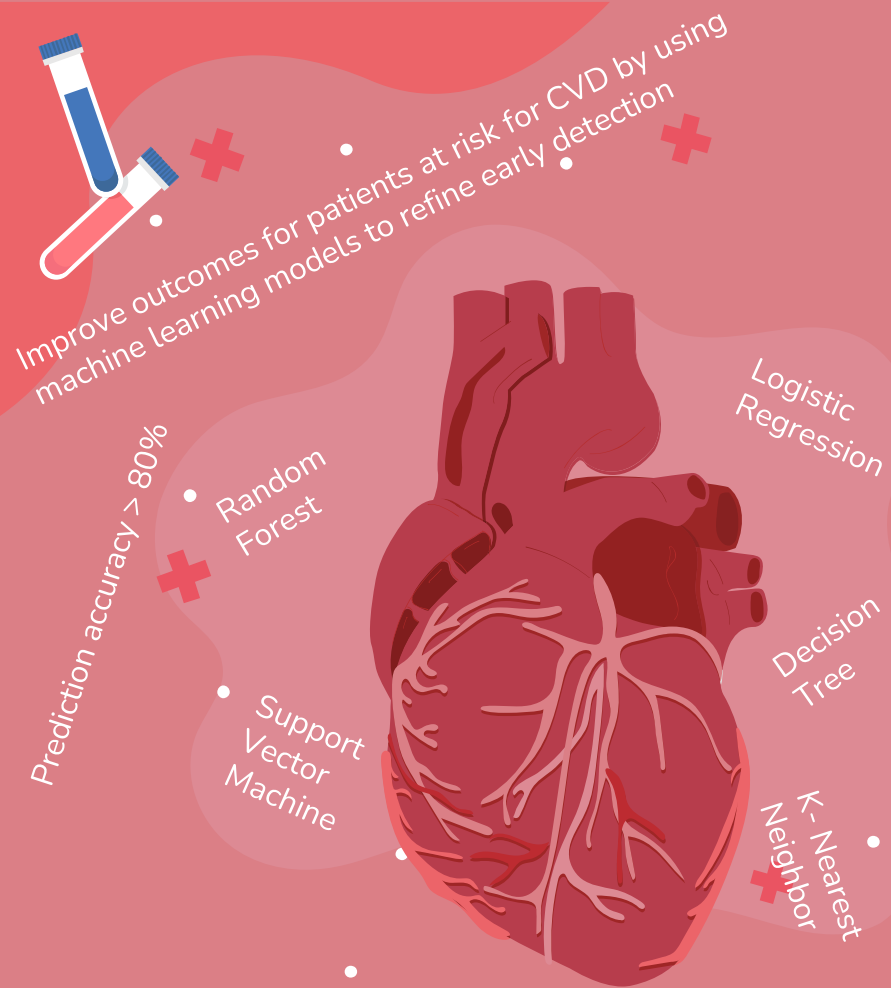


# Cardiovascular Disease Predictor

CS 6010 Project  
Amr Ibrahim & Vlad Yashaev

Fall 2022





# INTRODUCTION

Cardiovascular diseases (CVD) are leading cause of death globally

- 17.9 million lives per year
- Heart attacks and strokes account for 80% of CVDs
  - 1/3 of which occur in people under 70 years old

Identifying people with CVD and those who are at high cardiovascular risk can prevent premature deaths

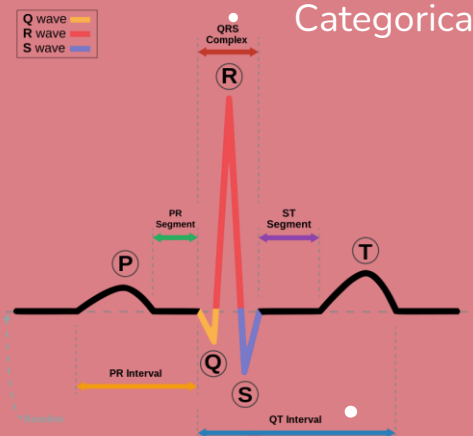




# Data Description

Data was obtained from kaggle Heart Failure Prediction Dataset

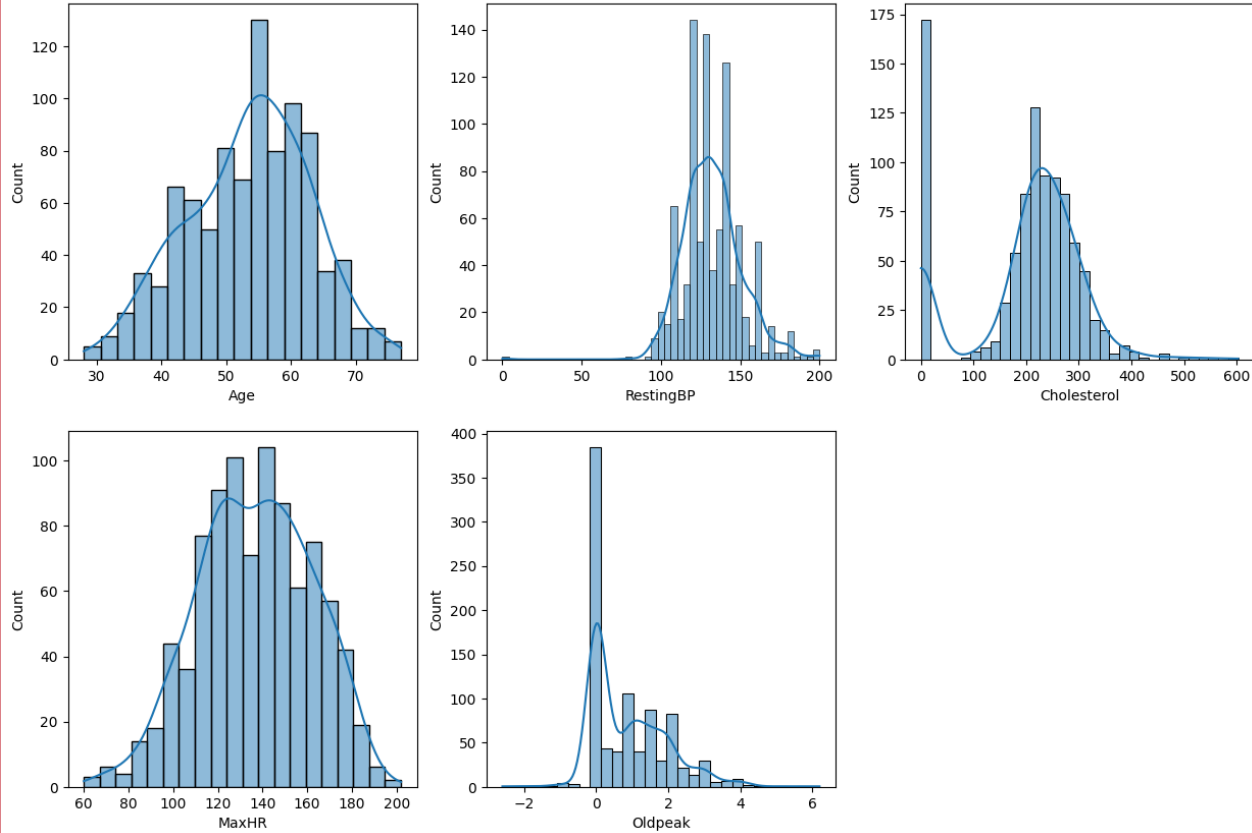
- CSV file with 12 features and 918 observations
- Features included are
  - Numerical (int64 & float64)
  - Categorical (object)



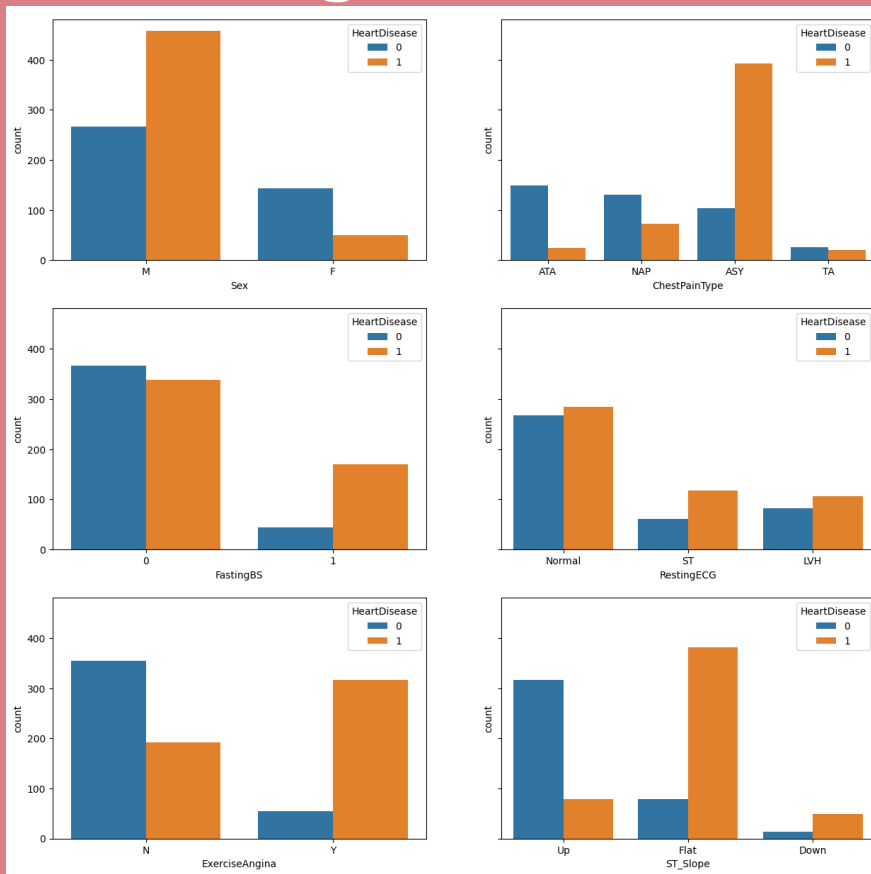
#	Column	Non-Null Count	Dtype
0	Age	918 non-null	int64
1	Sex	918 non-null	object
2	ChestPainType	918 non-null	object
3	RestingBP	918 non-null	int64
4	Cholesterol	918 non-null	int64
5	FastingBS	918 non-null	int64
6	RestingECG	918 non-null	object
7	MaxHR	918 non-null	int64
8	ExerciseAngina	918 non-null	object
9	Oldpeak	918 non-null	float64
10	ST_Slope	918 non-null	object
11	HeartDisease	918 non-null	int64



# Numerical Data



# Categorical Data



# Data Cleaning & Preprocessing



1

2

3

4

## NA Values

There were no  
NA values

## Transformations

All categorical  
variables were  
transformed into  
numerical

## Outliers

Rescale data to  
reduce effect of  
outliers

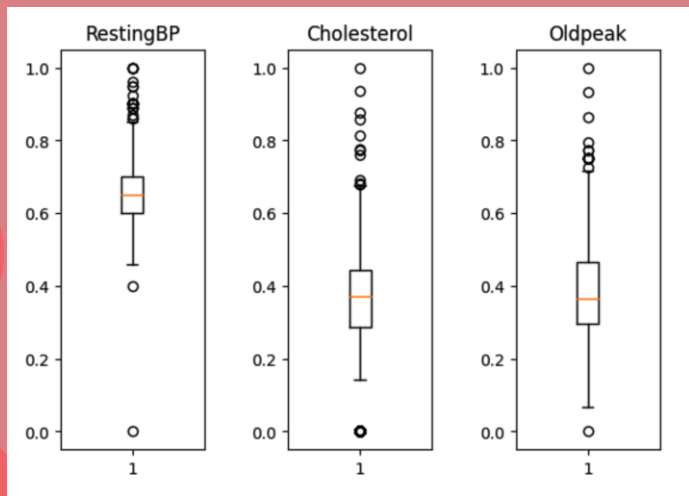
## Features

Evaluate features  
for possible  
dimensionality  
reduction



## .Outliers & Feature Analysis

# Outliers

[illegible]



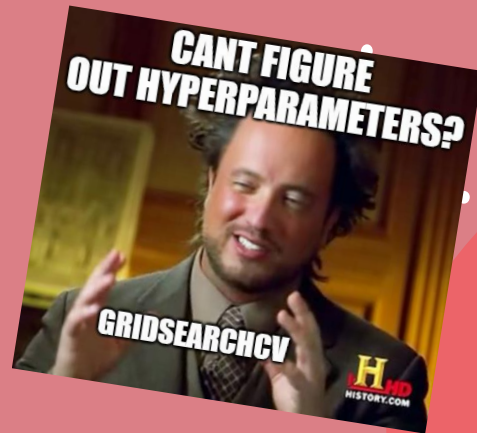
# Modeling and Evaluation

## 1. Parameter Tunning (GridSearchCV)

- LogisticRegression(C=100.0)
- SVC(C=0.1, gamma=0.1)
- RandomForestClassifier(max\_depth=40, n\_estimators=50)
- KNeighborsClassifier(n\_neighbors=11, weights='distance')
- DecisionTreeClassifier(max\_depth=5, min\_samples\_leaf=5)






## 2. Model Evaluation (Talk about your modelAverages function)

- Function modelAverages(model, modelName, xTrain, yTrain, xTest, yTest)
  - Runs each model 50 times and returns averages for model accuracy, recall, precision, and F1 values





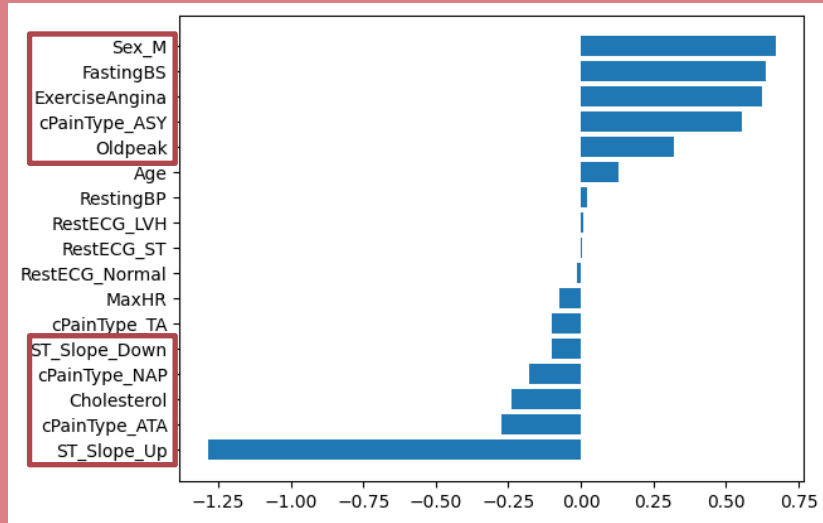
# Evaluation



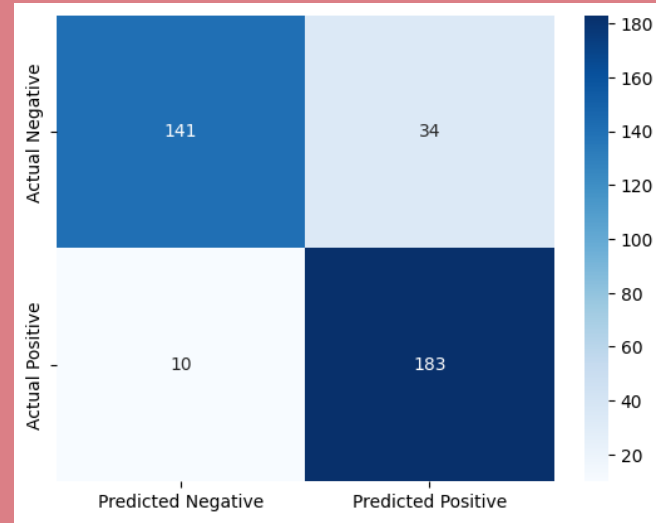
Model	Accuracy	Recall	Precision	F1 Score
Logistic Regression	88.32	92.23	86.41	89.22
SVM	88.04	94.82	84.33	89.27
Random Forest	87.20	92.19	84.73	88.31
KNN	85.87	89.64	84.39	86.93
Decision Tree	85.87	92.34	82.72	87.25

# SVM

## Feature Importance



## Confusion Matrix





# Discussion

- Results may vary for each iteration of model fitting and target predicting.
- Scaling data with outliers provides higher predictive accuracy.
- With enough training data and further parameter tuning, machine learning can be useful tool in refining early detection of CVD.
- Linear Regression performed the best on scaled data without outliers. However, for scaled data with outliers (i.e., more observations for the training data), SVM performed best.

## What we learned:

- Available resources exist to help figure out ideal parameter setting to increase accuracy metrics.
- It is better to test multiple models as different models may perform better for different dataset.
- It is important to clean up the data, then try different versions of the data for model training purposes
  - playing around with the number of features to retain.
  - Running the models on training data where outliers and/or NA values are handled differently.





# Conclusion

- Cardiovascular diseases (CVD) are the leading cause of death globally and early detection of CVD patients can prevent premature deaths.
- With a relatively small training dataset, machine learning models can be trained to help in the early detection of cardiovascular diseases in patients.

# Future Direction

- Test models on new test data.
- Evaluate accuracy of the models with less available features.
- More statistical analysis on variations using averages vs not.
- Review if the models performed better with outliers due to more observations or due to unique combinations of values in features.



# THANKS!

**CREDITS:** This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**

