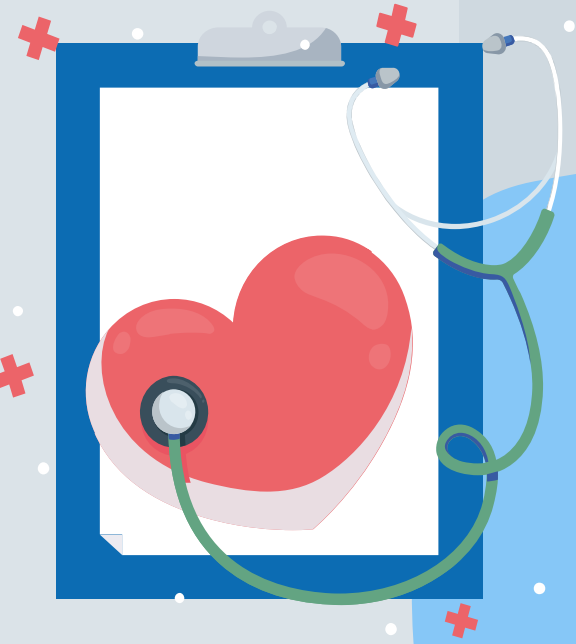


The Evaluation of relationship between Health Metrics on Cardiovasular Disease

Presented By – Annie Udhani
Graduate Student NJCU

INTRODUCTION

- Heart is a vital organ of the body which gives life to us and every cell.
- It supplies oxygen and essential nutrients and removes carbon-di-oxide and other toxins.
- Cardiovascular diseases (CVD) – a group of disorders affecting the heart or blood vessels due to build-up of fatty deposits resulting in a heart failure/death.
- Cardiovascular diseases increasingly becoming the major factor in human mortality

FACTS

In USA alone, 659k people dies every year at a rate of one person every 36 seconds

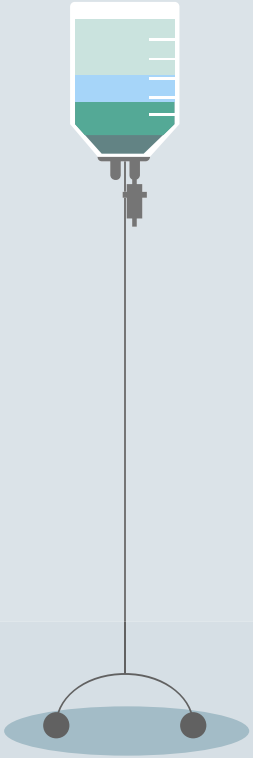
Costs around $363B every year in financial losses

17.9 million deaths every year as per WHO reports.

Cause of 32% of worldwide deaths.

Mortality rate of 85% worldwide.

SYMPTOMS



01
CHEST PAIN

02
DIZZINESS

03
SHORTNESS OF BREATH

04
FATIGUE

05
SWOLLEN KNEE

06
CHEST DISCOMFORT

KEY REASONS

- High Blood Pressure
- High Cholesterol level
- Hypertension
- Obesity
- Human Lifestyle

PROBLEM STATEMENT

- Healthcare professionals doing their best to save lives from CVD.
- As a Machine Learning engineers, how can we play our part against these deadly heart diseases.
- Formally, how can we use inexpensive and non-invasive method of diagnosis to detect and diagnose cardiovascular disease early in the cycle.

LiTERATURE REVIEW

- A lot of work done in this field applying simple and complex models like classification to neural network
- Zhang et al. applied SVM in 2017 to classify clinical data leaving result to the interpretation of others
- Guidi et al. in 2014 presented a Clinical Decision Support system evaluating Heart Failure severity using neural network, SVM and random forest
- Srinivas et al studied the likelihood of a coal mine worker getting CVDs

DATA COLLECTION

01 The dataset used here is a subset of Public health dataset.

02 The original dataset has 76 attributes including target/dependent variable.

03 We will be using 12 most used features by all the published researches.

04 This dataset was downloaded from Kaggle for the analysis.

DATA ATTRIBUTES

S. No.	Feature Name	Feature Description	Type
1	**Age**	age of the patient [years]	Independent
2	**Sex**	Sex of the patient	Independent
3	**Chest Pain Type**	chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]	Independent
4	**RestingBP**	resting blood pressure [mm Hg]	Independent
5	**Cholestrol**	Serum Cholestrol [mm/dl]	Independent
6	**FastingBS**	fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]	Independent
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]	Independent
8	**MaxHR**	maximum heart rate achieved [Numeric value between 60 and 202]	Independent
9	**ExerciseAngina**	exercise-induced angina [Y: Yes, N: No]	Independent
10	**Oldpeak**	oldpeak = ST [Numeric value measured in depression]	Independent
11	**ST_Slope**	the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, down: down sloping]	Independent

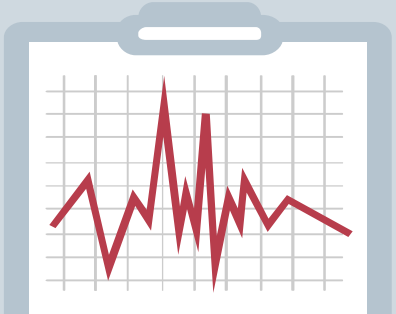
TARGET VARIABLE

- HeartDisease is target variable
- Signifying whether a person in records is healthy or has CVD
- output class [1: heart disease, 0: Normal]

Exploratory Data Analysis

A LOOK AT THE DATA

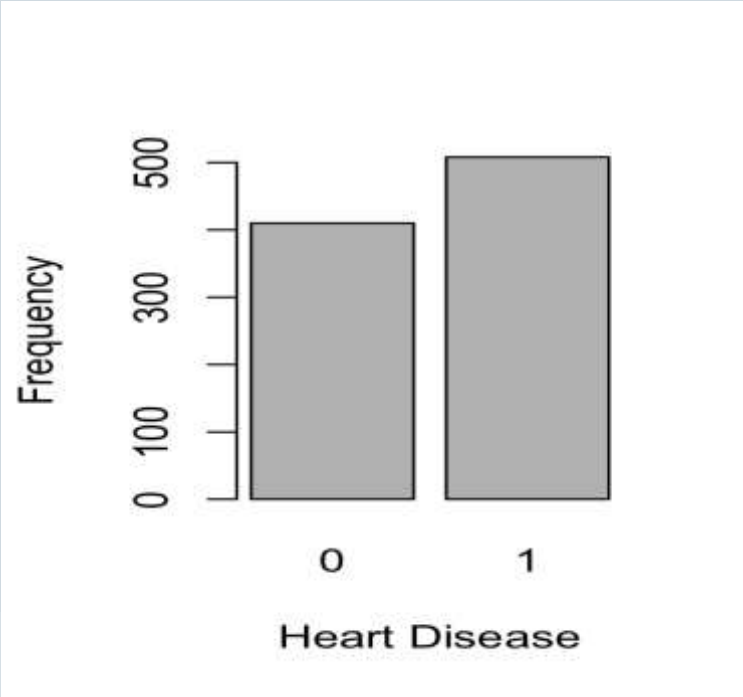
A	B	C	D	E	F	G	H	I	J	K	L	M
Age	Sex	ChestPainTy	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngi	Oldpeak	ST_Slope	HeartDisease	
40	M	ATA	140	289	0	Normal	172	N	0	Up	0	
49	F	NAP	160	180	0	Normal	156	N	1	Flat	1	
37	M	ATA	130	283	0	ST	98	N	0	Up	0	
48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1	
54	M	NAP	150	195	0	Normal	122	N	0	Up	0	
39	M	NAP	120	339	0	Normal	170	N	0	Up	0	
45	F	ATA	130	237	0	Normal	170	N	0	Up	0	
54	M	ATA	110	208	0	Normal	142	N	0	Up	0	
37	M	ASY	140	207	0	Normal	130	Y	1.5	Flat	1	
48	F	ATA	120	284	0	Normal	120	N	0	Up	0	



Summary Statistics of Numeric Features

S.No.	Columns	Min	Mean	Median	Max
1	Age	28	53.62	54	77
2	RestingBP	0	132.26	130	200
3	Cholesterol	0	198.80	223	603
4	FastingBS	0	0.23	0	1
5	MaxHR	60	136.81	138	202
6	Oldpeak	-3	0.89	0.6	6.2
7	HeartDisease	0	0.55	1	1

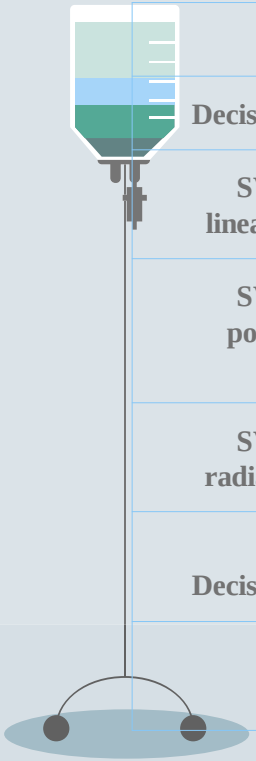
Dependent Variable – Heart Disease



Data Cleaning and Partition

- No missing value or outlier found
- Dataset is balanced – both target classes well represented in dataset
- No specific cleaning was performed
- Data split into 80% for training models and 20% for testing the models
- A ten-fold cross validation is used in all the models to minimize the bias

MODEL BUILDING



Model	Accuracy	Precision	Sensitivity
Decision Tree	87.43%	90.41%	80.49%
SVM with linear kernel	89.07%	87.80%	87.80%
SVM with polynomial kernel	89.62%	87.95%	89.02%
SVM with radial kernel	90.16%	90%	87.80%
Bagged Decision Tree	88.52%	88.61%	85.37%
Random Forest	91.26%	91.25%	89.02%

- Used ten-fold cross validation
- Trained following models on 80% training dataset
  - Decision tree
  - SVM with linear, polynomial and radial kernel
  - Bagged Trees
  - Random Forest
- Model was evaluated on test dataset

CONCLUSIONS

- Relationship exists between human health metrics and Cardiovascular disease.
- Various classifiers can be used to prewarn human to make changes in lifestyle.
- In our study, Random Forest provides us the best accuracy of 92%.
- This can save millions of lives and billions of dollar of healthcare spending.

THANKS