Prediction of Cardiovascular Disease using Machine Learning Algorithms

PROJECT REPORT

UNIVERSITY OF NORTH TEXAS

DEPARTMENT OF COMPUTER SCIENCE

SOFTWARE DEVEPMENT FOR ARTIFICIAL INTELLIGENCE

Prediction of Cardiovascular Disease using Machine Learning Algorithms

by

Gopireddy Lakshmi Keerthi (Student ID: 11653876)
LakshmiKeerthiGopireddy@my.unt.edu
Major: Artificial Intelligence
Role: Python programming, Project Documentation,
Domain Understanding & Feature Engineering,
Exploratory Data Analysis, Model Training, and Model Evaluation

Indhu Sri Krishnaraj (Student ID: 11661746)
IndhuSriKrishnaraj@my.unt.edu
Major: Artificial Intelligence
Role: Python programming, Data Visualization,
Exploratory Data Analysis, Data Pre-processing,
Parameter tuning and Model Training

Vishal Rachuri (Student ID: 11658243)
vishalrachuri@my.unt.edu
Major: Artificial Intelligence
Role: Python programming,
Dataset Selection, Parameter tuning and
Model Prediction and Evaluation

Jyothika Vollireddy (Student ID: 11652684)
jyothikavollireddy@my.unt.edu
Major: Cyber Security
Role: Python programming, Data Visualization,
Data Pre-processing, Feature Engineering,
Algorithm Selection and Model Prediction

Pavani Jaya Keerthana Chittedi (Student ID: 11672192)
pavanijayakeerthanchittedi@my.unt.edu
Major: Cyber Security
Role: Python programming, Dataset selection,
Domain Understanding & Feature Engineering,
Parameter tuning and Model training

Meeting Schedule and Collaboration

We are a group of 5 individuals from multiple departments, having different class schedules. Considering the need for team collaboration, we decided to either meet virtually with the help of Microsoft Teams/Google Meet or meet in person at Willis Library depending on the availability at the below-mentioned timings:

Monday: 6:00 PM to 7:00 PM

Thursday: 10:00 AM to 12:00 PM

Friday: 11:00 AM to 1:00 PM

We were using these timings to discuss the progress of individual tasks and to understand the dependency of one task on another. We ensured to share the roles and responsibilities in such a way that each task is handled by 2 or more individuals. In this way, we can brainstorm multiple ideas and techniques which will yield a higher learning curve and exposure.

We are using the below-mentioned platforms for smooth communication and secure information storage:

Outlook Mailing List: sdforai_cardiovascular@groups.students.untsystem.edu

SharePoint: https://myunt.sharepoint.com/:w:/r/sites/SDforAI-Project1

Please note that the links have limited access. Drop an email to any team member for access.

Project Abstract

Did you know that one person dies from cardiovascular disease every 34 seconds in the United States of America? Yes, according to the statistics provided by the centers for disease control and prevention, about 697,000 people in the United States died from heart disease in 2020; that is 1 in every 5 deaths. With the continuous changes in lifestyle and food habits, it is apparent to envision an increase in the number of persons affected with cardiovascular disease over the next years. Early detection of cardiovascular disease can be helpful in timely treatment and can reduce the chances of death. Several machine-learning techniques are being used in the medical field these days for the early prediction of disease. Our study aims to build a machine-learning model for the prediction of cardiovascular disease.

Previously collected demographic and medical information of patients acts as the main source to analyze patterns and predict the disease. There are multiple machine learning algorithms like Random forests, K-Nearest Neighbour (KNN), Gradient Boost Classifier, Support Vector Machines, and Naïve Bayes, which will use the dataset containing demographic information and medical history to build a predictive model which will project the probability of an individual having cardiovascular disease. The model should be trained on a diverse dataset and should be evaluated by considering different evaluation metrics considering the importance of the medical field and its direct relation to human life.

This study is to demonstrate that machine learning can be effectively used for predicting the risk of cardiovascular or heart disease and can be a valuable tool for healthcare providers in making informed decisions for the prevention and treatment of the disease.

Visual Demonstration - Prediction of Cardiovascular Disease

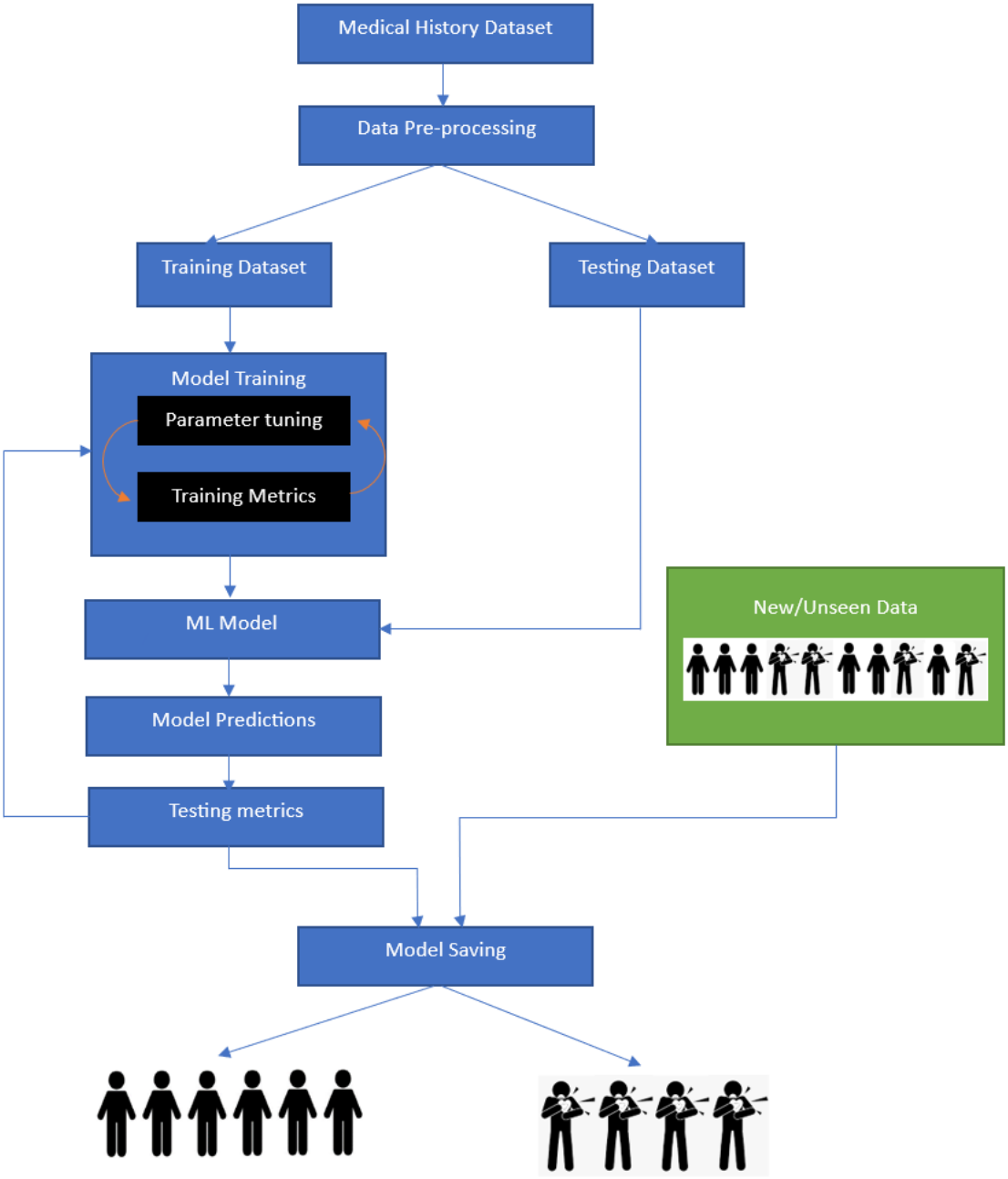


Fig 1: Flow chart for prediction of cardiovascular disease

Data Specification

Source: Kaggle

Link: [Kaggle dataset link]

The dataset consists of health metrics of 70000 people. The health metric values were either collected at the moment of examination through tests or directly from the patient. There are a total of 11 features or health metrics and a target variable that indicates whether the person has cardiovascular disease or not. The below table contains information about each feature:

Feature Name	Feature Description	Feature Type	Data Type	Values
age	Age of the person in days	Objective	integer	Any integer value
height	Height of the person in cm	Objective	integer	Any integer value
weight	Weight of the person in kg	Objective	float	Any decimal value
gender	Gender	Objective	categorical code	1: woman, 2: man
ap_hi	Systolic blood pressure	Examination	int	Any integer value
ap_lo	Diastolic blood pressure	Examination	int	Any integer value
cholesterol	Cholesterol level	Examination	categorical code	1: normal, 2: above normal, 3: well above normal
gluc	Glucose or Sugar level	Examination	categorical code	
smoke	Smoking	Subjective	binary	0: No, 1: Yes
alco	Alcohol intake	Subjective	binary	0: No, 1: Yes
active	Physical activity	Subjective	binary	0: No, 1: Yes

Table 1: Feature descriptions

Feature-wise statistics:

feature	count	mean	min	max
age	70000	19468.87	10798	23713
gender	70000	1.35	1	2
height	70000	164.36	55	250
weight	70000	74.21	10	200
ap_hi	70000	128.82	-150	16020
ap_lo	70000	96.63	-70	11000
cholesterol	70000	1.37	1	3
gluc	70000	1.23	1	3
smoke	70000	0.09	0	1
alco	70000	0.05	0	1
active	70000	0.80	0	1

Table 2: Feature-wise statistics

Project Design

Using the dataset mentioned above, our aim is to build a machine learning model which can predict, given the medical metrics of a new person, whether the person has a chance of getting any cardiovascular disease or not.

As an initial step, we performed exploratory data analysis and tried to understand the statistics of each feature, followed by analyzing the correlation between features.

Considering the results of exploratory data analysis, removing the outliers was an important step in order to remove bad or false data. For this, we have considered a lower 5 percentile and upper 5 percentile of the values detected as outliers and were removed.

```
Feature_name: age-------------
Percentiles: 5th=15069.0, 95th=23259.0, IQR=8190.0
Identified outliers: 0

Feature_name: gender-------------
Percentiles: 5th=1.0, 95th=2.0, IQR=1.0
Identified outliers: 0

Feature_name: height-------------
Percentiles: 5th=152.0, 95th=178.0, IQR=26.0
Identified outliers: 49

Feature_name: weight-------------
Percentiles: 5th=55.0, 95th=100.0, IQR=45.0
Identified outliers: 22

Feature_name: ap_hi-------------
Percentiles: 5th=100.0, 95th=160.0, IQR=60.0
Identified outliers: 50

Feature_name: ap_lo-------------
Percentiles: 5th=70.0, 95th=100.0, IQR=30.0
Identified outliers: 1036
```

```
Feature_name: ap_lo-------------
Percentiles: 5th=70.0, 95th=100.0, IQR=30.0
Identified outliers: 1036

Feature_name: cholesterol-------------
Percentiles: 5th=1.0, 95th=3.0, IQR=2.0
Identified outliers: 0

Feature_name: gluc-------------
Percentiles: 5th=1.0, 95th=3.0, IQR=2.0
Identified outliers: 0

Feature_name: smoke-------------
Percentiles: 5th=0.0, 95th=1.0, IQR=1.0
Identified outliers: 0

Feature_name: alco-------------
Percentiles: 5th=0.0, 95th=1.0, IQR=1.0
Identified outliers: 0

Feature_name: active-------------
Percentiles: 5th=0.0, 95th=1.0, IQR=1.0
Identified outliers: 0
```

Figure 2: Feature-wise outliers

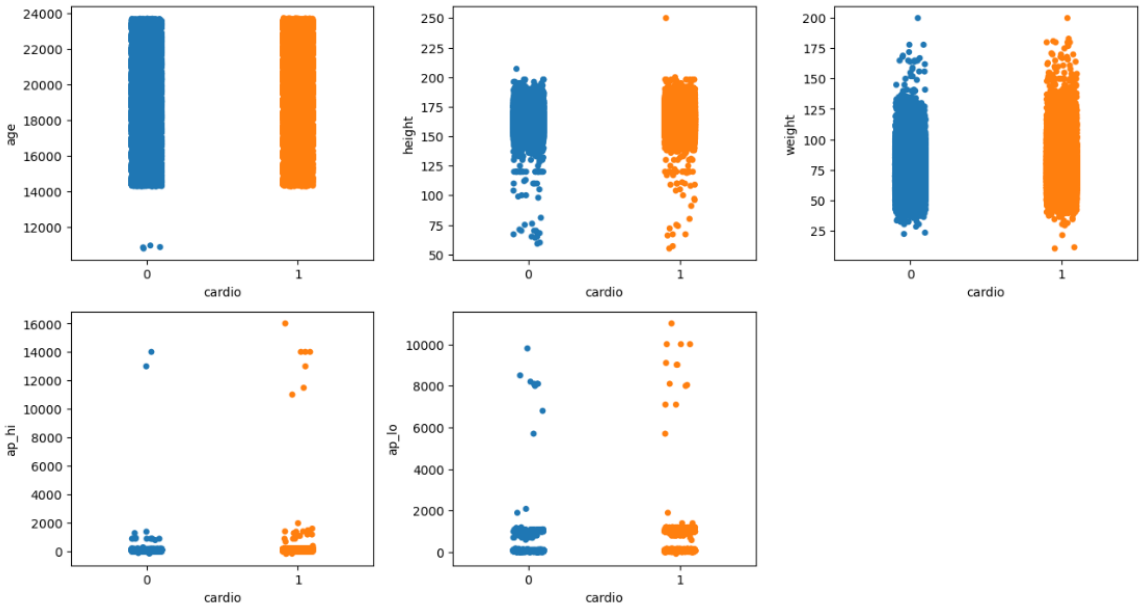


Figure 3: Strip plot of numerical features

From the correlation matrix, we have observed a higher correlation of 0.7 between ap_hi and ap_lo. Thus, we dropped the ap_lo feature to avoid data redundancy.

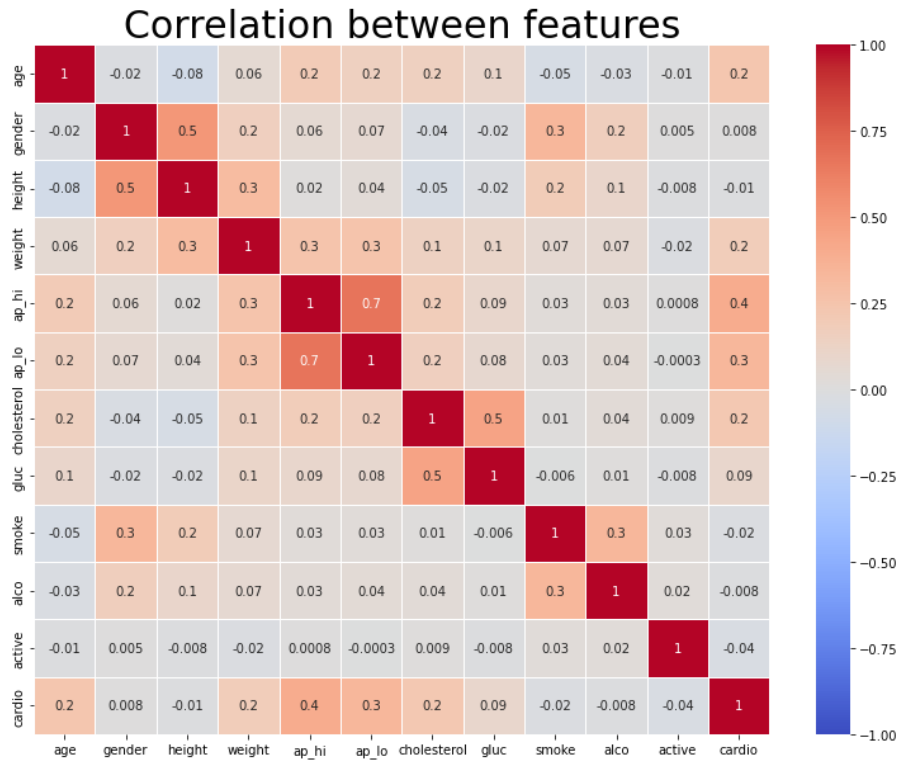


Figure 4: Correlation between features

We have multiple numerical features in the data like age, height, weight, etc. The ranges of these features are completely dissimilar from each other. For example, the age of the person is mentioned in days ranging from 10798 to 23713 and the weight of the person is in kilograms ranging from 10 to 200. As the value of numbers is very high for age, this feature might influence the output. To avoid this situation, we are standardizing the values of all numerical features i.e., age, height, weight, and ap_hi using the StandardScalar() function.

	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	4	17474	1	156	56.0	100	60	1	1	0	0	0	0

Figure 5: Data values before scaling

	age	gender	height	weight	ap_hi	cholesterol	gluc	smoke	alco	active	cardio
0	-0.436268	2	0.450841	-0.848595	-0.925881	1	1	0	0	1	0
1	0.307538	1	-1.048253	0.750291	0.736653	3	1	0	0	1	1
2	-0.248189	1	0.076067	-0.709562	0.182475	3	1	0	0	0	1
3	-0.748383	2	0.575765	0.541741	1.290831	1	1	0	0	1	1
4	-0.808779	1	-1.048253	-1.265696	-1.480059	1	1	0	0	0	0

Figure 6: Data values after scaling

70% of the entire population was utilized for training, with the remaining 30% used to test the model. As it is tabular data, we started with training the data on multiple basic machine learning algorithms such as Logistic Regression, Decision tree and K-Nearest Neighbour followed by boosting techniques like Gradient Boosting Classifier. The below table indicates the accuracy, sensitivity, and objective value of each model:

Model	Accuracy	Sensitivity	Objective Value
Logistic Regression	0.730189	0.674918	0.898918
Linear Discriminant Analysis	0.726762	0.666617	0.893417
Decision Tree	0.737406	0.681589	0.907803
Gradient Boosting Classifier	0.736458	0.695079	0.910228
K-Nearest Neighbour	0.736677	0.676994	0.905926
Quadratic Discriminant Analysis	0.71991	0.588793	0.867108
Random Forest	0.735365	0.694041	0.908875
Support Vector Classifier	0.734126	0.660243	0.899187

Table 3: Model-wise metrics

For the prediction of any disease, the sensitivity of the model plays a vital role in model selection as the number of false negatives should be as low as possible in health-related problems. Considering the same, from the above metrics we can observe that Gradient Boosting Classifier has better sensitivity and thus is appropriate for the problem.

Gradient Boosting is an ensemble method that uses multiple weak learners and creates a strong learner. Gradient Boosting Classifier constructs an additive model in stages, allowing for the optimization of any differentiable loss functions in a forward stage-wise fashion. In each stage, n classes regression trees are fitted to the negative gradient of loss function creating a stronger classifier.

Further, to get the best estimator, necessary parameter tuning was performed using GridSearchCV taking recall to evaluate the performance of the model. Using this function, the best estimator output parameters are max_depth=5, and n_estimators=90.

The final model was having an **accuracy** of **~73%** and a **sensitivity** of **~69%.**

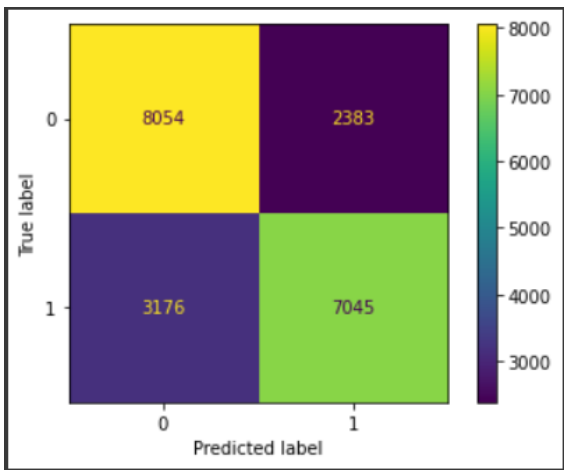
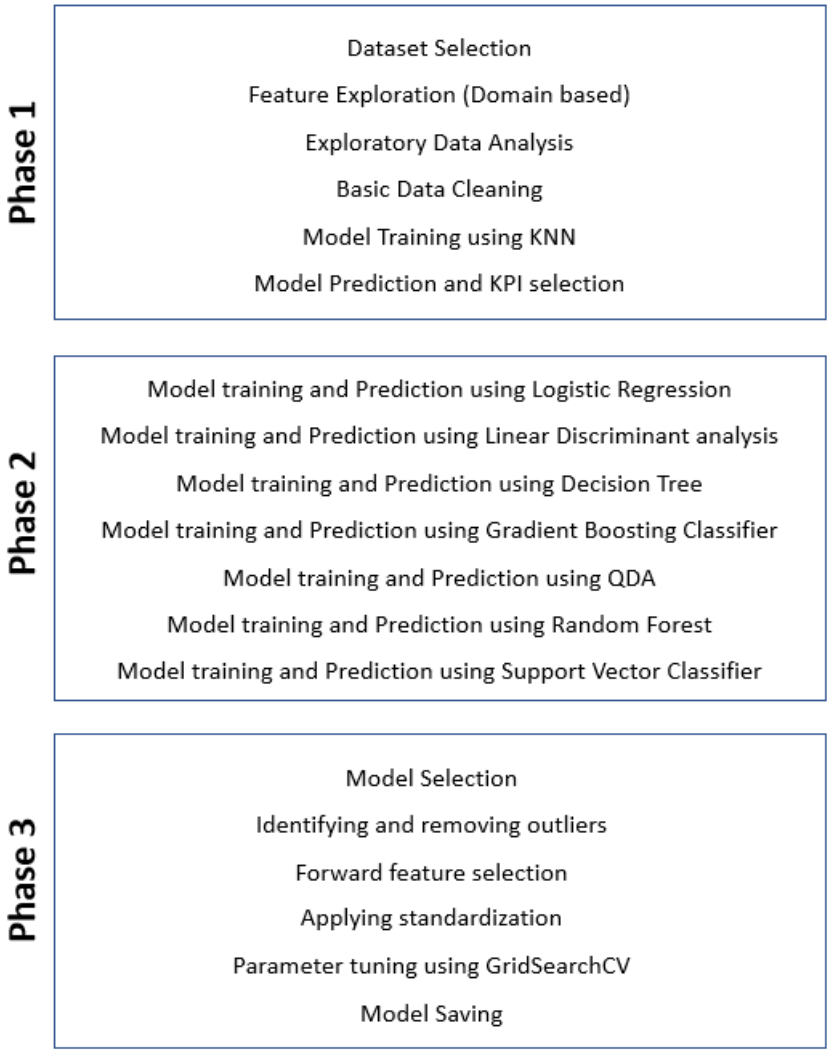


Figure 7: Confusion Matrix of Gradient Boosting Classifier

Project Milestones



Code Repository

Final Code PDF: https://drive.google.com/file/d/1ttfgfHnF301vxsYC075k47UcJY4d_bTV/view?usp=share_link

Code Repository Link: https://drive.google.com/drive/folders/1KCAkxcnq9sbInlsl328ViC7d_sf-bt2M?usp=sharing

References

Datasets:

https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset?resource=download
https://github.com/Syed-Owais-Noor/ML_Project_on_Heart_Disease_Dataset
https://github.com/Bakar31/Heart-Disease/blob/master/data/heart.csv
https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data
https://www.kaggle.com/datasets/bhadaneeraj/cardio-vascular-disease-detection
https://archive.ics.uci.edu/ml/datasets/heart+disease
https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data

Research Papers:

https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012072/pdf
https://www.researchgate.net/publication/341870785_Heart_Disease_Prediction_using_Machine_Learning
https://www.ijeat.org/wp-content/uploads/papers/v9i3/B3986129219.pdf
https://ieeexplore.ieee.org/abstract/document/9333574
https://towardsdatascience.com/should-i-look-at-precision-recall-or-specificity-sensitivity-3946158aace1
https://machinelearningmastery.com/how-to-use-statistics-to-identify-outliers-in-data/
https://machinelearningmastery.com/introduction-to-function-optimization/#:~:text=In%20machine%20learning%2C%20the%20objective,define%2C%20although%20expensive%20to%20evaluate.

Algorithm Papers:

https://eudl.eu/pdf/10.4108/eai.13-7-2017.2270596
https://www.jmlr.org/papers/volume2/tong01a/tong01a.pdf
https://www.kdd.org/kdd2016/papers/files/rfp0697-chenAemb.pdf
https://www.cs.toronto.edu/~rgrosse/courses/csc2515_2019/tutorials/tut4/tut4_boost.pdf
https://towardsdatascience.com/gradient-boosting-classification-explained-through-python-60cc980eeb3d
https://towardsdatascience.com/quadratic-discriminant-analysis-ae55d8a8148a#:~:text=Quadratic%20Discriminant%20Analysis%20(QDA)%20is,that%20belong%20to%20the%20class.
https://www.knowledgehut.com/blog/data-science/linear-discriminant-analysis-for-machine-learning