MAT 422 Final Project: Paper Title

By Christopher Annunziato, Saajan Maslanka, Ruth Stowers, and Brian Sweeney

1. Introduction

According to the Centers for Disease Control and Prevention (CDC), the leading cause of death in the US in 2021 was heart disease, killing approximately 695,000 people. In fact, one in every five US deaths is attributed to heart disease, which averages to approximately one death every 33 seconds [1]. This wasn't always the case. At the beginning of the 20th century, heart disease was decidedly less prevalent, though lack of accurate medical diagnoses may be a factor in vital statistics. Heart disease seemed to increase as lifestyles became more sedentary, tobacco use increased, and dietary changes resulted in elevated cholesterol levels. By the 1950s, heart disease was the leading cause of death in the US. Mortality rates from heart disease peaked in the 1960s before plateauing [2]. Despite recent healthcare innovations and improvements, as well as healthy lifestyle initiatives, the percentage of deaths from heart disease has been steadily rising again over the last 30 years [3], and the number of people under 50 with heart disease is also increasing [4].

The term "heart disease" refers to any condition which affects the structure or function of the heart [5]. There are more than 30 different types of heart disease including but not limited to, Coronary Artery Disease, Heart Arrhythmias, Heart Failure, and Pericardial Disease. Some types of heart disease are congenital, meaning occurring from birth, however most heart disease is developed over time and is preventable through healthy diet, regular exercise and the abstinence from harmful chemicals such as tobacco products. While heart disease is incurable, it is treatable through modern medical interventions such as surgery, and through lifestyle changes such as diet and exercise. Early detection is a key factor in the successful treatment of heart disease. The earlier heart disease is detected and treated, the less damage is done to the organ and the greater the life expectancy of the patient.

Traditionally, heart disease detection techniques have been costly and only available in areas of the world where there is ample healthcare availability. X-rays, electrocardiograms (EKGs), and angiograms are examples of some of the tests doctors regularly use to detect heart disease. Despite their accuracy, it is not economically or physically feasible to perform regular tests using such methods on every patient that could potentially have heart disease. Thus, a screening method is needed to predict whether a patient has heart disease or not, filtering the pool of patients for further testing. There are various screening methods that have been developed (some of which are mentioned briefly in the next section.) One of those methods is using a logistic regression model to predict whether a patient has heart disease or not based on a selection of the patient's attributes. As with all predictive models, there is always a margin of error, so the developers of the model must determine how to reduce that

margin as much as possible through appropriate attribute selection and by adjusting the parameters of the model to increase accuracy.

As there are already complex models being used in the screening of patients for heart disease, our project aims to replicate one of the basic predictive methods that others have used previously. Our goal is to create a logistic regression model to predict whether a patient has heart disease or not. We will use previously collected data from a database, select appropriate patient attributes, develop a logistic regression model using python, train and test our model, adjust the parameters as appropriate, report on the performance of the final model, and draw appropriate conclusions from our analysis with suggestions for further study. By doing this, we aim to show how even a basic screening model can be beneficial in predicting heart disease, thus helping healthcare professionals to identify and treat patients who are at risk from this life threatening condition.

2. Related work

There have been many studies focusing on building models with the intent of improving screening for heart disease. Logistic regression is an example of a prediction model that has been used often to predict heart disease. In a study conducted by the East West Institute of Technology in India, a logistic regression model was used to predict cardiovascular disease with an 87.10% accuracy rate [6]. We have based our own project on this model as it is one of the simplest models to create. Other researchers have used more complex methods such as decision trees and neural networks. A study conducted in Poland focused on comparing the performances between these three models, logistic regression, decision trees, and neural networks. They compared the models' sensitivity (the ability to identify a patient with heart disease, given the individual truly has the condition), specificity (the ability to identify a patient without heart disease, given the individual truly does not have the condition) and accuracy (the ability to differentiate the positive and negative cases correctly). The study found that while the three models were similar, Artificial neural networks had the lowest error rate and the highest accuracy, thus the authors concluded that Artificial neural networks was the best technique to use for that particular dataset [7].

Some studies aim to predict the presence of multiple heart diseases such as a study presented at the 2020 Fourth World Conference in London which used multiple complex smaller models in conjunction with each other [8], while other studies focus their models on individual heart diseases such as the study of macrovascular disease in caucasoid diabetic subjects which aimed to identify and analyze common risk variables [9]. Having a broad model encompassing multiple heart diseases may be beneficial as an initial screening for large groups of individuals, while singular models may be helpful on a smaller scale to screen focused groups that may experience additional health issues. Other studies focus on analyzing and improving attribute selection methods as most models include no more than 14 attributes. A study performed in Pakistan used feature selection techniques and algorithms to improve a heart disease prediction model's attribute selection process, resulting in improved model accuracy [10]. The authors did

acknowledge however that model developers would need to account for the ease of obtaining various attributes to decide whether the increase in accuracy is worth the cost or difficulty incurred by obtaining the attribute information, especially in less affluent geographical locations.

3. Proposed methodology

3.1 Dataset Overview

Data for this project was sourced from the UCI Heart Disease Dataset as found in Kaggle [11]. This dataset was originally collected for a paper published in The American Journal of Cardiology by Detrano, et al. in 1989. The dataset, upon being published to Kaggle, had been modified such that attributes categorical in nature were not represented by numerical values.

3.1.1 Attribute Overview

In the original dataset, there were a grand total of 76 attributes, but we will be using the 16 found in the Kaggle dataset. Below are 14 of the aforementioned 16 - excluding the unique id field, and the data origin field.

Variable Name	Role	Type	Description
age	Feature	Integer	Age
sex	Feature	Categorical	Sex
cp	Feature	Categorical	Chest pain type
testbps	Feature	Integer	Resting Blood Pressure (in mm Hg on hospital admission)
chol	Feature	Integer	Serum Cholesterol
fbs	Feature	Categorical	Fasting blood sugar (> 120mg/dl T/F)
restecg	Feature	Categorical	Resting electrocardiographic results
thalach	Feature	Integer	Maximum heart rate achieved
exang	Feature	Categorical	Exercise induced angina
oldpeak	Feature	Integer	ST depression induced by exercise relative to rest
slope	Feature	Categorical	The slope of the peak exercise ST segment
ca	Feature	Integer	Number of major vessels (0-3) colored by flourosopy
thal	Feature	Categorical	Presence of a blood disorder: Thalassemia (T/F)
num	Target	Integer	Diagnosis of heart disease

The original paper used these 14 attributes, using the data origin location verification.

3.2 Data Origin

The data used in this project, and as presented in Detrano, et al. was sourced from multiple physio-geographically disparate regions. Preliminary data was collected from the Cleveland Clinic in Ohio, USA, while follow up verification data was sourced from the Long Beach VA in California, USA, the Hungarian Institute of Cardiology in Hungary,

as well as the Basel and Zurich University Hospitals in Switzerland. Below is a table detailing how the dataset is broken down and some additional geographical metadata.

Location	Sample Count	Heart Disease Prevalence
Cleveland Clinic, Ohio, USA	303	~75%
Long Beach VA , California, USA	200	
Hungarian Institute of Cardiology, Hungary	425	~38%
Basel & Zurich University Hospitals, Switzerland	125	~84%

Though this location data may be helpful in identifying heart disease, training on this extra information may result in a loss of generality. This is because doing so would restrict us from applying predictions to locations outside of these four locations in meaningful ways.

3.3 Proposed Methodology

In order to predict heart disease in patients, we opted to develop a logistic regression model. Below are detailed discussions on attributes trained on, model specifics, and performance criteria.

3.3.1 Attribute Selection

Out of the 76 total attributes in the original dataset, we chose to use the attributes selected on Kaggle as we have limited experience with the data and do not have as strong a grasp at understanding what is theoretically good for predicting heart disease. Of the 16 attributes present in the Kaggle dataset, we further restricted the attributes used for the logistic regression training. Before training any models, we decided that our model would be attempting to be a "universal" predictor of heart disease. This meant that the location data available in Kaggle Dataset cannot be used as this would restrict meaningful predictions to the given geographical regions. As such we removed the data origin location from the dataset. Additionally, we removed the unique id as this has no meaning when applied to prediction. As discussed below, after creating our initial model, we observed multicollinearity and thus had to remove additional attributes to avoid violating regression assumptions.

3.3.2 Logistic Model

For the logistic regression model, we opted to use the sklearn logistic regression model as provided by their API. We used the liblinear solver for a one-vs-rest classification strategy with balanced class weights. All other parameters were left at their defaults as of sklearn version 1.3.2.

3.3.3 Performance Criteria

Because we utilized a one-vs-rest classification strategy for the multiclass regression problem presented in this paper, we chose accuracy on the test dataset to be the main

indicator of performance. Additionally, we chose to compare the weights assigned to each attribute against existing medical literature to verify the model correctly correlating certain characteristics with heart disease (eg. number of vessels colored under fluoroscopy) over others (eg. age).

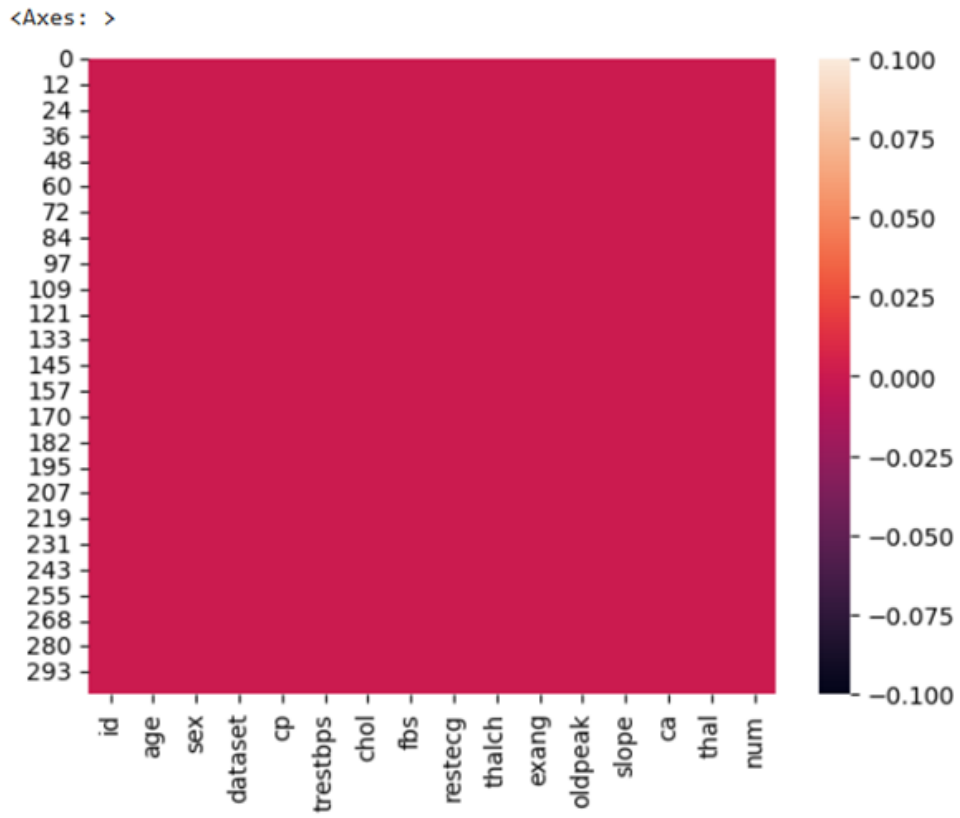
4. Experiment setups and result discussion

4.1 Experiment Settings

We carried out our experiment using Google Collab running Python 3.10.12. We used the UCI Heart Disease Dataset available on Kaggle as discussed in section 3.3.1. Using sklearn we created a Logistic Regression model to determine the likelihood of a patient having heart disease given a set of 14 different variables as discussed.

	id	age	sex	dataset	cp	trestbps	chol	fbs	restecg	thalch	exang	oldpeak	slope	ca	thal	num
0	1	63	Male	Cleveland	typical angina	145.0	233.0	True	lv hypertrophy	150.0	False	2.3	downsloping	0.0	fixed defect	0
1	2	67	Male	Cleveland	asymptomatic	160.0	286.0	False	lv hypertrophy	108.0	True	1.5	flat	3.0	normal	2
2	3	67	Male	Cleveland	asymptomatic	120.0	229.0	False	lv hypertrophy	129.0	True	2.6	flat	2.0	reversable defect	1
3	4	37	Male	Cleveland	non-anginal	130.0	250.0	False	normal	187.0	False	3.5	downsloping	0.0	normal	0
4	5	41	Female	Cleveland	atypical angina	130.0	204.0	False	lv hypertrophy	172.0	False	1.4	upsloping	0.0	normal	0

To prepare the data we used the python drop function to remove null values that would impact our model. Null values are missing attributes within a tuple which can be detrimental for regression as you can be attempting to gain insight from unknown values. From the figure below it can be seen that the final dataset we used contained no null values that could influence our model.



As many of the variables are categorical instead of numeric, we used dummy variables to allow for the creation of a logistic regression model. As there are 4 types of heart disease this model only tries to determine if they have one of those four types or no heart disease. Using sklearn we split the data into two groups. The first group we used to train our model. The second we used to test our model's performance. We used Sklearn logistic regression with the liblinear solver to create our model. We chose to use Liblinear as the data would not converge with any other solver types. Our model was able to predict the likelihood of a patient having heart disease with an accuracy of 86.7%.

Next we checked to make sure that the assumptions of logistic regression were met. The observations given by kaggle are independent, the response variable is Binary, There are no extreme outliers, and the sample size is sufficiently large. To check for multicollinearity, we used python to check the statsmodels' variance inflation factor. We found that there was multicollinearity occurring among age, resting blood pressure, cholesterol, and maximum heart rate achieved (see figures below). Due to this we ran a second model without including resting blood pressure, cholesterol, and maximum heart rate achieved. After these variables were removed there was no multicollinearity issue. Our new model did not show a significant difference in accuracy, which was not unexpected as the three attributes we removed did not seem to be a major factor in our model's performance.

	VIF Factor	features
0	40.420617	age
1	63.579270	trestbps
2	27.253806	chol
3	45.700794	thalch
4	2.938659	oldpeak
5	2.270702	ca
6	3.838240	sex_Male
7	1.807639	cp_atypical angina
8	2.137006	cp_non-anginal
9	1.380065	cp_typical angina
10	1.232690	fbs_True
11	2.126754	restecg_normal
12	1.050568	restecg_st-t abnormality
13	2.038235	exang_True
14	5.658240	slope_flat
15	7.556464	slope_upsloping
16	8.370206	thal_normal
17	5.460907	thal_reversable defect

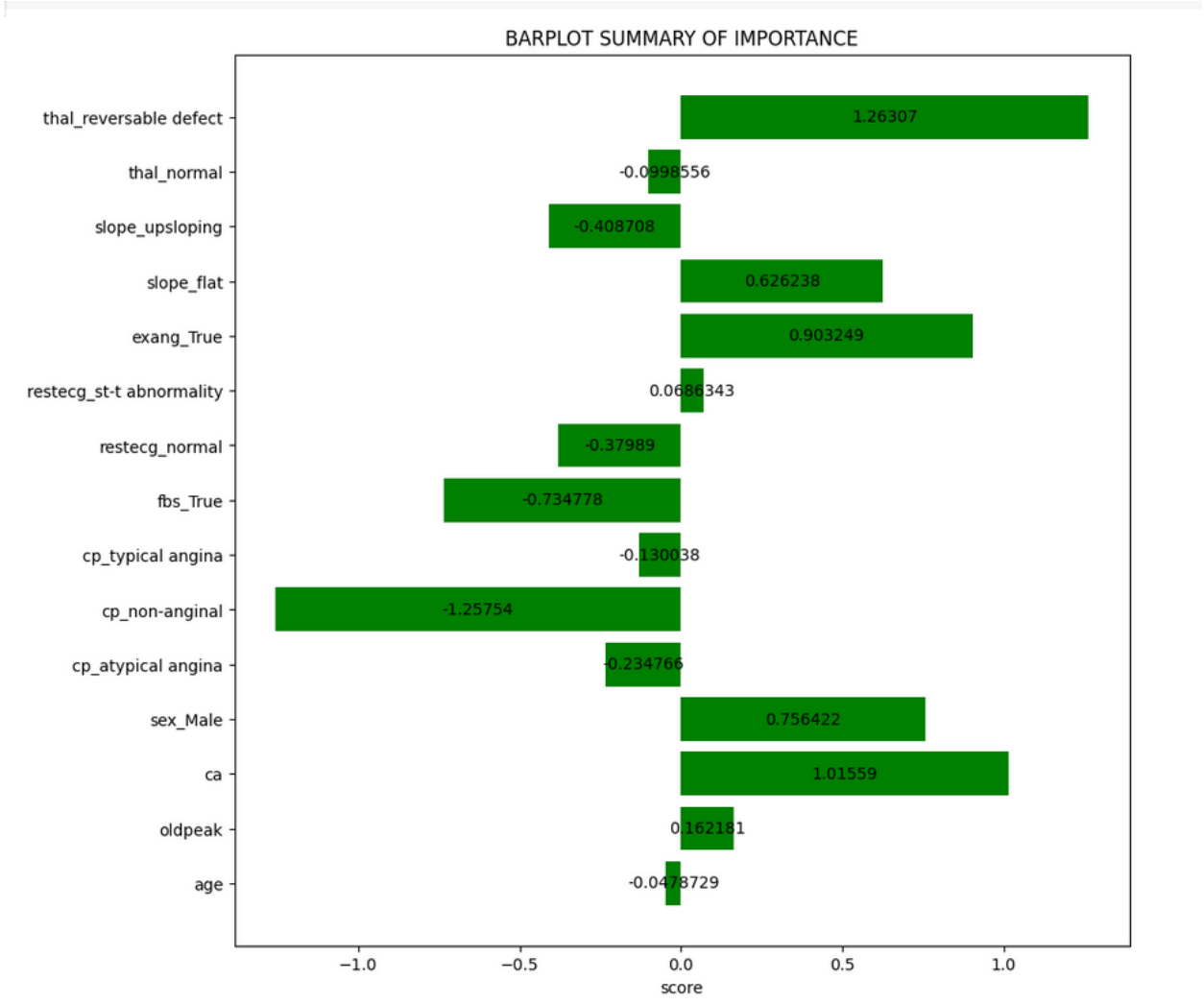
Multicollinear Variables included

	VIF Factor	features
0	4.916071	age
1	2.873386	oldpeak
2	1.985829	ca
3	3.557646	sex_Male
4	1.741927	cp_atypical angina
5	2.092731	cp_non-anginal
6	1.346868	cp_typical angina
7	1.222700	fbs_True
8	2.084919	restecg_normal
9	1.038566	restecg_st-t abnormality
10	2.008849	exang_True
11	5.644332	slope_flat
12	7.420370	slope_upsloping
13	7.738217	thal_normal
14	5.333908	thal_reversable defect

Multicollinear variables removed

4.2 Results

Using the settings outlined in 4.1, the output below was generated showing the impact a variable has on the probability of a patient having heart disease. A value close to zero does not have an impact on predicting heart disease. A negative value reduces the likelihood that a patient has heart disease and a positive value increases the likelihood that a patient has heart disease. Using the test data as outlined above the model was able to predict heart disease in patients with an accuracy of 86.7%. This accuracy score is high and shows that the model is effective at predicting heart disease.



Looking at the output the most impactful variables that increase the likelihood of a patient having heart disease are a reversible defect maximum heart rate with an increase of 1.3, having exercised induced angina with an increase of 0.9, being male with an increase of .8, and every unit increase in number of major vessels colored by fluoroscopy increases the chance of heart disease by a factor of 0.1.0. The most

impactful variables that decrease the likelihood of a patient having heart disease are chest pain type non angina, and having a fasting blood pressure greater than 120 mg/dl.

5. Comparison

References	Dataset	Methods	Best Results	Accuracy
Ambrish et al. [6]	UCI ML Repository	Machine Learning	LR	0.871
Khemphila, Boonjing [7]	Patient data from VA Medical Center in Long Beach California	Deep Learning	ANN	0.802
Ambesange et al. [8]	UCI ML Repository	Machine Learning	LR	0.9032
Bashir et al. [10]	UCI ML Repository	Machine Learning	SVM	0.8485
Proposed Method	UCI ML Repository	Machine Learning	LR	0.867

This section compares the proposed method and results from this study to previous studies that performed similar modeling and training for detecting heart disease. The above table shows the comparison in accuracy between all the studies. In some studies, logistic regression was the only technique used. Accuracy can vary for multiple reasons, including things like the ratio of the train-test split, weight assignment, and other configurations/specifications. Our logistic regression model, when compared to the four other methods, falls above the average in terms of accuracy. Using the UCI heart disease dataset, our model generates an accuracy of 0.867.

6. Conclusion

In this study, we created a logistic regression model based on the UCI Heart Disease Dataset from the UCI ML Repository to predict heart disease in patients, using accuracy as the primary indicator of performance. After creating our model, we noticed there was multicollinearity between age, cholesterol levels, resting blood pressure, and the presence of Thalassemia. As these were not able to be combined, we removed cholesterol levels, resting blood pressure, and the presence of Thalassemia from our model. This left only ten attributes in our model. As we know there are many other attributes available in the larger dataset, we feel a deeper study would benefit from

exploring the inclusion of other attributes and testing to see which ones would be the most beneficial to a prediction model. Even with our model only having 10 attributes, and only using basic logistic regression, we were able to achieve an accuracy of 86.7%. With this being the only measure of performance, when compared to other models which used this dataset, our model was of comparable performance. This is significant as our model used very basic logistic regression techniques, thus we can conclude that even a very basic prediction model could be beneficial to healthcare screenings.

In a future study, it could be beneficial to use multiple datasets from multiple locations to test how well a model predicts heart disease across diverse settings. Another avenue for study would be to use other methods or techniques to generate different models to compare as was done in the study by A. Khemphila and V. Boonjing [7]. Other metrics like an F1-Score and ROC/AUC could be used in order to provide more insight into the data.

Overall, our findings show that machine learning models, even when not overly complex, can be used to help screen patients for heart disease, which is important as early detection aids health professionals in treating patients successfully through life-saving interventions and treatments.

Coding File

The code used by the authors can be found here:
https://github.com/RuthStowers/MAT422/blob/ffc59795ce59072b21a40a29dc551bd17d3eeb25/ProjectLog.ipynb

Acknowledgments

The authors would like to thank Dr. Haiyan Wang, Professor at ASU West's School of Mathematical and Natural Sciences for his guidance and support of this project.

Author contributions

The sections of this project were completed as follows:
Introduction: Ruth Stowers
Related Work: Ruth Stowers
Proposed methodology: Saajan Maslanka
Experiment setups and result discussion: Christopher Annunziato
Comparison: Brian Sweeney
Conclusion: Brian Sweeney

Ethical standard

This Project does not contain any studies with human participants or animals

performed by any of the authors.

Data availability

The authors declare that all data supporting the findings of this study are available at https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data and https://archive.ics.uci.edu/dataset/45/heart+disease.

References

[1] Centers for Disease Control and Prevention. (2023, May 15). Heart disease facts. Centers for Disease Control and Prevention. https://www.cdc.gov/heartdisease/facts.htm

[2] Dalen JE, Alpert JS, Goldberg RJ, Weinstein RS. The epidemic of the 20(th) century: coronary heart disease. Am J Med. 2014 Sep;127(9):807-12. doi: 10.1016/j.amjmed.2014.04.015. Epub 2014 May 5. PMID: 24811552.

[3] U.S. Department of Health and Human Services. (2021, February 3). *Cardiovascular disease is on the rise, but we know how to curb it. We've done it before*. National Heart Lung and Blood Institute. https://www.nhlbi.nih.gov/news/2021/cardiovascular-disease-rise-we-know-how-curb-it-weve-done-it

[4] Aggarwal R, Yeh RW, Joynt Maddox KE, Wadhera RK. Cardiovascular Risk Factor Prevalence, Treatment, and Control in US Adults Aged 20 to 44

Years, 2009 to March 2020. JAMA. 2023;329(11):899–909. doi:10.1001/jama.2023.2307

[5] U.S. Department of Health and Human Services. (n.d.). *Know the differences: Cardiovascular disease, heart ... - NHLBI, NIH*. National Heart Blood and Lung Institute. https://www.nhlbi.nih.gov/sites/default/files/publications/FactSheetKnowDiffDesign2020V4a.pdf

[6] Ambrish G, Bharathi Ganesh, Anitha Ganesh, Chetana Srinivas, Dhanraj, Kiran Mensinkal, Logistic regression technique for prediction of cardiovascular disease, Global Transitions Proceedings, Volume 3, Issue 1, 2022, Pages 127-130, ISSN 2666-285X, https://doi.org/10.1016/j.gltp.2022.04.008.

[7] A. Khemphila and V. Boonjing, "Comparing performances of logistic regression, decision trees, and neural networks for classifying heart disease patients," 2010 International Conference on Computer Information Systems and Industrial Management Applications (CISIM), Krakow, Poland, 2010, pp. 193-198, doi: 10.1109/CISIM.2010.5643666.

[8] S. Ambesange, A. Vijayalaxmi, S. Sridevi, Venkateswaran and B. S. Yashoda, "Multiple Heart Diseases Prediction using Logistic Regression with Ensemble and Hyper Parameter tuning Techniques," 2020 Fourth World Conference on Smart Trends in

Systems, Security and Sustainability (WorldS4), London, UK, 2020, pp. 827-832, doi: 10.1109/WorldS450073.2020.9210404.

[9] Welborn, T.A., Knuiman, M., McCann, V. et al. Clinical macrovascular disease in Caucasoid diabetic subjects: logistic regression analysis of risk variables. Diabetologia 27, 568–573 (1984). https://doi.org/10.1007/BF00276969

[10] S. Bashir, Z. S. Khan, F. Hassan Khan, A. Anjum and K. Bashir, "Improving Heart Disease Prediction Using Feature Selection Approaches," 2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST), Islamabad, Pakistan, 2019, pp. 619-623, doi: 10.1109/IBCAST.2019.8667106.

[11] Janosi,Andras, Steinbrunn,William, Pfisterer,Matthias, and Detrano,Robert. (1988). Heart Disease. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X.