Heart Disease Prediction and Risk Factors

Adam Fowler
CSCI-4502: Data Mining
University of Colorado
Boulder CO USA
adfo8311@colorado.edu

Tyler Sanchez
CSCI-4502: Data Mining
University of Colorado
Boulder CO USA
tysa5330@colorado.edu

Jacob Unger
CSCI-4502: Data Mining
University of Colorado
Boulder CO USA
jaun8796@colorado.edu

1. ABSTRACT

Heart disease and heart attacks are very prevalent and are a leading cause of death. This project is going to use data mining techniques to dive into a data set and further analyze what risk factors cause heart disease. When it comes to attributes that are thought to lead to these issues, is there a way to use these attributes to predict if there is a high risk of heart disease based on the attributes that they have? A dataset was obtained with 18 different attributes. The dataset was cleaned and bins were created for each attribute. Using python, pandas, matplotlib, seaborn, and scikit the dataset was then analyzed using multiple techniques and a decision tree was made from the data to predict heart disease. To answer the question of what attributes are the leading cause of heart disease the decision tree found the top attributes to be high blood pressure, age, and general health. After creating a decision tree, the model was trained to be used in a web app where a person can answer certain questions about themselves related to the attributes in the dataset and it would give them an answer if they have a higher risk of heart disease. Using data mining techniques and tools, this project was a able to find attributes that could lead to a high risk of heart disease and then put it into a real world use through the web application in order to help a person see their risk level of heart disease. Overall, this project will show people what they need to change in their life in order to prevent heart disease and show them what in their life is already doing this.

2. INTRODUCTION

Heart disease is the leading cause of death for men, women, and people of most racial and ethnic groups in the United States. About 697,000 people in the United States died from heart disease in 2020—that's 1 in every 5 deaths. Heart disease cost the United States about $229 billion each year from 2017 to 2018. This includes the cost of health care services, medicines, and lost productivity due to death. [1]

The Behavioral Risk Factor Surveillance System (BRFSS) is the nation's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. Established in 1984 with 15 states, BRFSS now collects data in all 50 states as well as the District of Columbia and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the largest continuously conducted health survey system in the world. [2]

Using data from the BRFSS, we intend to determine which risk factors are the strongest indicators of heart disease and look for correlations between risk factors. We will then attempt to predict heart disease in a patient given the presence of indicators using various techniques of classification.

3. LITERATURE REVIEW

Fortunately, heart disease and heart disease indicators are a well-researched topic. Unfortunately, heart disease is a largely prevalent disease that affects many people worldwide. Proposing the question of "how do we better predict heart disease?" helps us narrow down what work we can look at for background information and inspiration in this exploratory study. The prediction of heart disease has far reaching implications for better individual health and societal health, as cardiovascular disease is the biggest cause

of morbidity and mortality. To determine where the current research stands, we looked at papers under search terms such as "heart disease prediction", "predictive factors of heart disease", and "health indicators heart disease".

In one study, by authors Tavia Gordan; William P. Castelli, MD; Marthana C. Hjortland, PhD; et al, titled 'Predicting Coronary Heart Disease in Middle-Aged and Older Persons' [3]. In this study, 2470 people, 1025 men and 1445 women, of ages ranging from 49 to 82 years, were screened in this study. Using factors such as cholesterol in the high- and low-density lipoproteins, systolic blood pressure, left ventricular hypertrophy and diabetes, they were able to generate a function that could be used to classify the chance of coronary heart disease. This in turn found that 25% of men in the highest decile, and 37% of women in the highest decile had coronary heart disease for the sample. This model fit for each specific age group and well as was at least as good as the coronary heart disease risk profile that is generally used when classifying risk for those within a younger population. This study was certainly framed through a more medical perspective, focusing on quantifying health indicators and factors that are obtained through specific medical measurements, whereas out data focuses on binary and largely observable health indicators. One thing we will take away from this study is their application of the model to populations outside of their sample. This provides an interesting insight into how well the model can be applied elsewhere or if the research needs to be retuned and redone for different populations.

A study titled "Predicting Heart Disease at Early Stages using Machine Learning: A Survey" [4] by authors Rahul Katarya, Polipireddy Srinivas, et al, used some similar data analysis techniques that we aim to use in our analysis, providing a good roadmap for us to look at. These modeling techniques that were used in this study were artificial neural networks, decision tree, random forest, support vector machine, naive Bayes, and k-nearest neighbor algorithm. This study heavily focused on the benefit of predicting heart disease during early stages of the development of the disease as to better focus on treatment and prevention, while also going over where the current knowledge of using machine learning in specifically heart disease stands, highlighting great information such as the various methods that are currently used, there Data Mining schemes that assist in accurate findings and future predictions, and the steps that are currently being used to get to those accurate findings. While being a shorter paper on the whole, the information included is very technical and very helpful in helping us understand how more efficient and large-scale data mining and modeling operations take place, giving us vital information in how to proceed within our exploration of a heart disease prediction data set. By giving us a review of the current statistical modeling that is taking place with data that is similar to what we are working with, we know have a better idea of what modeling we may want to move forward with to get the best result given the limited working time of this project.

Looking at the more technical aspect of studying heart disease using prediction factors, the study titled "Identification of significant features and data mining techniques in predicting heart disease" [5] by authors Mohammad Shafenoor-Aminac, Yin KiaChiama, and Kasturi DewiVarathanb, looks at specific techniques in data mining that are common in this field of study and provides good background information for us as we move into this space. Evaluating different classification algorithms, focusing on efficient data mining techniques for heart disease predictions and determining the performance of classification models for heart disease predictions, allowed the research to develop a prediction model using the significant features of the models evaluated while using hybrid data mining techniques for the best results. These researchers evaluate k-Nearest Neighbor, Naive Bayes, Logistic Regression, Support vector Machine, Neural Network, and Vote, which is a hybrid technique blending Naive Bayes and Logistic Regression. By identifying specific features that shined for each model and data mining technique, they aimed to improve prediction modeling for cardiovascular disease prediction. It was found that the Vote modeling technique (the blend of Naive Bayes and Logistic Regression) was the highest performer, and using the key features determined as applied to mining techniques they were able to achieve an accuracy of 87.4% in heart disease prediction as applied to their sample population. This study has lots of great in-depth information that we can leverage in determine the models we choose to use for our predictive model. While we likely will not

be using a model as complicated as the Vote Model they determined to be the most accurate, choosing a model that had strong results based on the data that we are feeding it will closely follow some of the techniques that were used within this research study.

4. DATASET

The Behavioral Risk Factor Surveillance System (BRFSS) is a health-related telephone survey that is collected annually by the CDC. A sample of the survey and it's questions can be found at https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf. This dataset is reduced and cleaned from the responses given in 2015. It can be found at https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset. Our working copy as well as the version containing derived attributes will be available at https://github.com/cockytrumpet/datamining-group-4.

The data set contains 253680 objects, each with 22 attributes.

Attribute	**Desc**.	**Type**	**Req. Reduc.**
HeartDiseaseorAttack	Heart disease or attack	Binary	
HighBP	High blood pressure	Binary	
HighChol	Cholesterol checked in last 5 years	Binary	
BMI	Body Mass Index	Ordinal	X
Smoker	>99 cigarettes ever smoked	Binary	
Stroke	Ever had stroke	Binary	
Diabetes	Is diabetic	Binary	
PhysActivity	Physical activity in the past 30 days	Binary	
Fruits	Consumes >= 1 time/day	Binary	
Veggies	Consumes >= 1 time/day	Binary	
HvyAlcoholConsump	Men: >14 drinks/week, Women: >7 drinks/week	Binary	
AnyHealthCare	Health coverage	Binary	
NoDocbcCost	Didn’t seek health care in past 12 months due to cost	Binary	
MentHlth	Days in past 30 days where mental health was poor	Ordinal	X
PhysHlth	Days in past 30 days where physical health was poor	Ordinal	X
DiffWalk	Difficulty walking or	Binary	

	climbing stairs		
Sex	Sex	Binary	
Age	Age category	Ordinal	X
Education	Category for highest level completed	Ordinal	
Income	Annual household income	Ordinal	

5. MAIN TECHNIQUES APPLIED

The first thing that we did for this section is create data bins that will be made from the variables at hand. These will look at variables like BMI, Sex, Age, Physical Health, or any of the other attributes that are listed in our data set. The bins were created for each attribute. As for the data cleaning, it overall looks well organized and complete, it will mostly be checking for any missing data and making sure that all the data lines up with each other structurally. The majority of the data we are using is binary, meaning no manipulation or cleaning was required to modify into a useable format. Of the data that was found in an ordinal format, BMI, Age, Education, Income, MentHlth (Mental Health), and PhysHlth (Physical Health), only some had to be reduced. The main format for reducing the data that we implemented was binning the data into groups to make it easier to work with. Those attributes that were binned were BMI_bin, MentHlth_bin, PhysHlth_bin and Diabetes_bin. The BMI bin was reduced into underweight (BMI of <18.5), Overweight (BMI of 25 – 29.9) and Overweight (BMI >30.0). Mental Health was binned into low representing less than 10 days, medium being 11-20 days, and 21 plus days being binned as high. Physical health was binned in a similar fashion as mental health. Diabetes was turned into a binary of either being yes or no to represent the presence or lack thereof of diabetes.

Once we create the bins from the variables, we then ran a preliminary analysis of the attributes to see if we can create a model to predict heart disease based on the different variables. We created graphs for the different variables to create visuals on how each variable effected the risk of heart disease or attacks. For this we used Python, Matplotlib, and seaborn.

The best classification method was a decision tree, to build the decision tree, scikit was used within python. For this, the data was split into training and testing sets. Using scikit's DecisionTreeClassifier, it was able to be tuned by generating statistics, adjusting the max_depth and min_leaf_samples, then repeating this process. This is how the model was trained.

Looking at past reviews, this is a very studied topic, in some ways this will be a similar study to those done in the past in the case that this will be trying to predict heart disease based on certain variables. Also, when looking at prior studies, it looks like there may be similarities on strategies that we can use as a guide to analyze the data and then create a model for prediction. The difference in this project, is that additional derived attributes will be analyzed to see if it is able to find a better predictor of heart disease.

6. EVALUATION METHODS

Ultimately, the success of our project will be determined by our ability to work with the data set. Can we create a model to make accurate predictions? Although is unlikely that we will be able to achieve the level of accuracy that was obtained from the more advanced methods shown in the literature review, if our results can mirror those results to a significant degree, that would be considered successful. An issue that is often wrestled with in scientific literature is the availability of the information to those outside of the small research ecosystem in which that study exists, therefore a mark of success in our exploration and manipulation of the data is the ability of someone to absorb and apply the information presented in a meaningful and impactful way.

Once the model is in place, it's efficacy can be demonstrated in a real-world scenario via a simple tool that allows a patients data to be entered then classified by the model, providing the patient with

information regarding their heart disease risk given input.

7. TOOLS USED

Python and R will be the main programming frameworks used for data analysis, model generation, and data visualization. Within these frameworks, packages such as Pandas will be used for dataset manipulation to prepare for processing, matplotlib, seaborn, and ggplot will be used to create visualizations of the data. and scikit will be used for analysis and classification of the data.

The web application will be achieved using Flask Python Web Framework, HTML, and CSS as well as the python code that was generated based on the decision tree generated.

Git and GitHub as interacting with through the command line will be the main method of version control.

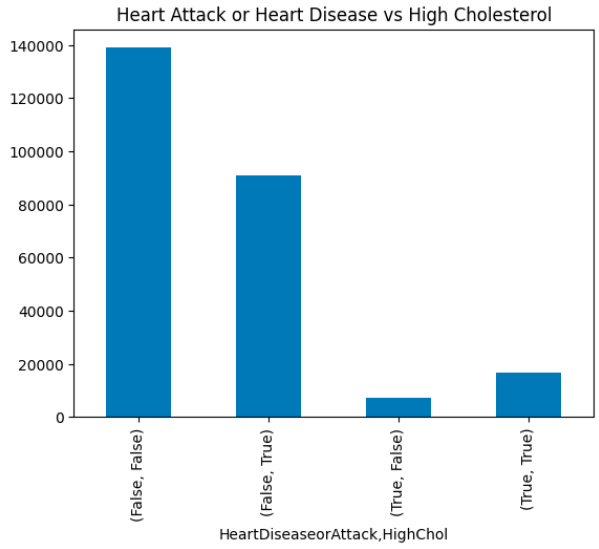
Pandas	https://pandas.pydata.org
Matplotlib	https://matplotlib.org
Seaborn	https://seaborn.pydata.org
Ggplot	https://ggplot2.tidyverse..org
Scikit	https://scikit-learn.org

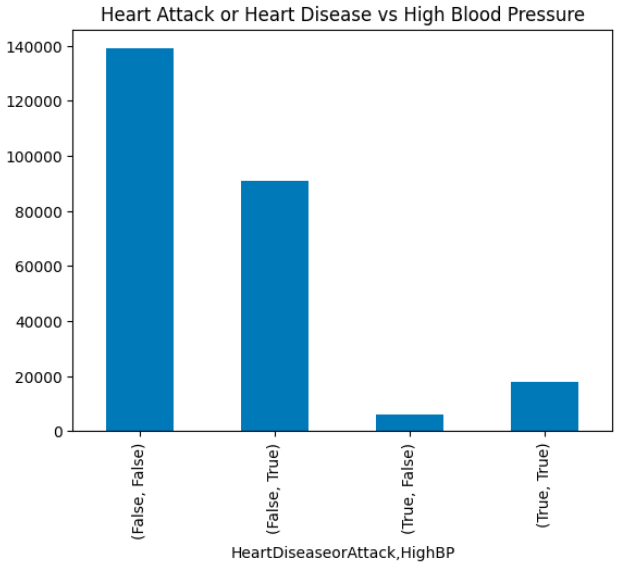
8. MILESTONES

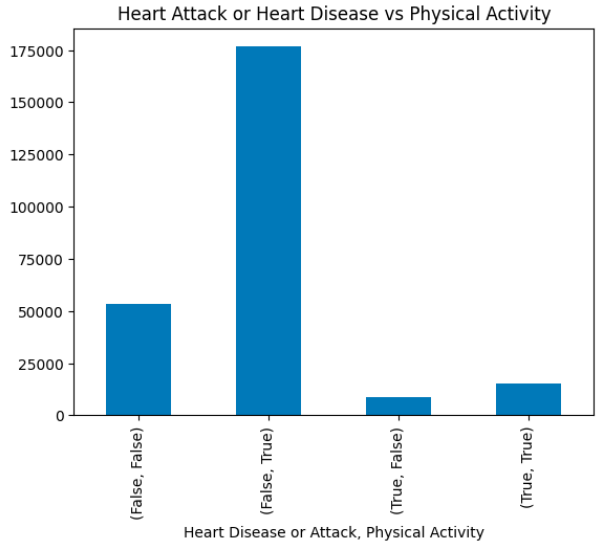
- Oct 31: Data cleaning and attribute binning completed
- Nov 07: Preliminary analysis, classification methodologies chosen
- Nov 14: Model trained
- Nov 21: Report first draft
- Nov 28: Project Part 3 Due – Progress Report
- Nov 28: Report final revisions, Presentation first draft
- Dec 05: Presentation final revisions, Code clean-up
- Dec 08: Project Parts 4-7 Due - Report, Code, Presentation, Evaluations

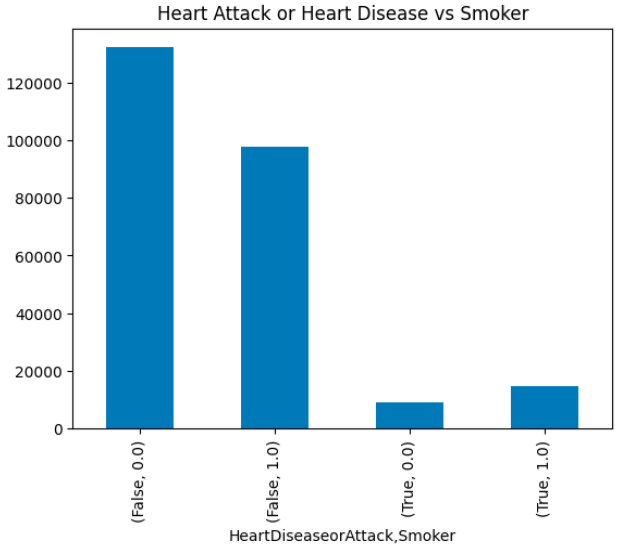
9. RESULTS SO FAR

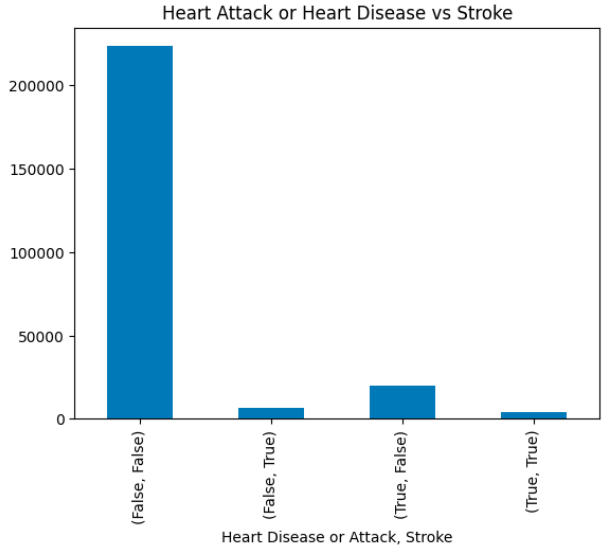
Through manipulative exploration of the data and modeling using various models, our ideas of a “healthy lifestyle” decreasing the risk of heart disease and attacks have been reinforced. Bringing back into light the original goal of this exploration, we set out to look at the use of modeling and how it can be applied to determine the risk factors and co-morbidities. This research has been done in a variety of ways and a variety of methods yet had not been replicated on this data set. Below you will see some preliminary graphs that helped reinforce the results of the decision tree and our choice in other models, such as Naïve Bayes, Adaptive Boost, Multilayer Perceptron, and a Decision tree, both with the original Confusion Matrix and a Balanced Confusion Matrix.

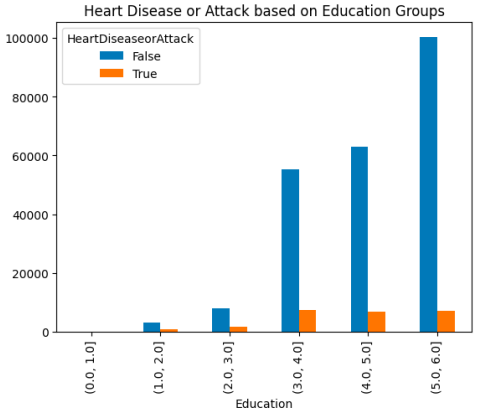



Heart Attack or Heart Disease vs High Blood Pressure
140000
120000
100000
80000
60000
40000
20000
0
(False, False)
(False, True)
(True, False)
(True, True)
HeartDiseaseorAttack,HighBP


Heart Attack or Heart Disease vs Physical Activity
175000
150000
125000
100000
75000
50000
25000
0
(False, False)
(False, True)
(True, False)
(True, True)
Heart Disease or Attack, Physical Activity

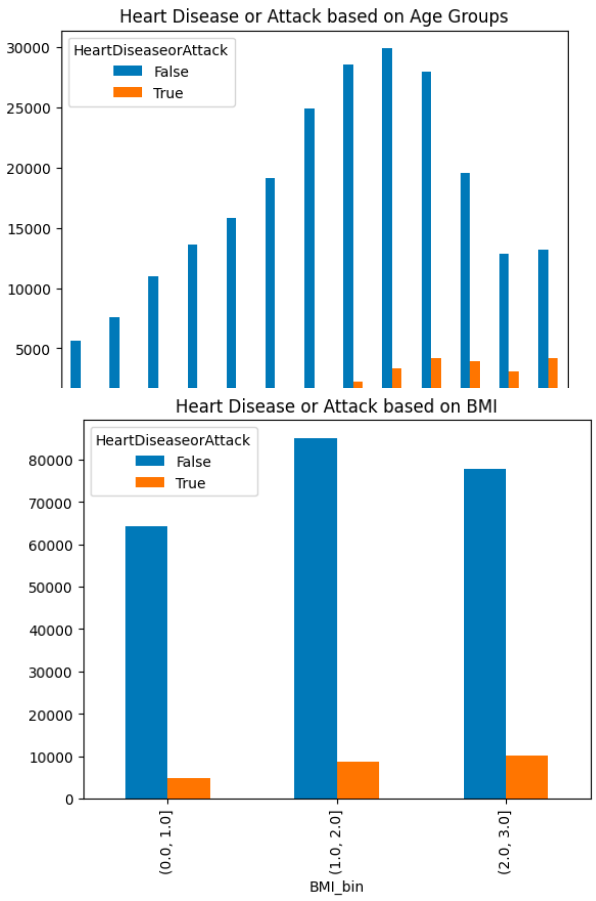

Heart Attack or Heart Disease vs Smoker
120000
100000
80000
60000
40000
20000
0
(False, 0.0)
(False, 1.0)
(True, 0.0)
(True, 1.0)
HeartDiseaseorAttack,Smoker


Heart Attack or Heart Disease vs Stroke
200000
150000
100000
50000
0
(False, False)
(False, True)
(True, False)
(True, True)
Heart Disease or Attack, Stroke


Heart Disease or Attack based on Education Groups
HeartDiseaseorAttack
False
True
100000
80000
60000
40000
20000
0
(0.0, 1.0]
(1.0, 2.0]
(2.0, 3.0]
(3.0, 4.0]
(4.0, 5.0]
(5.0, 6.0]
Education

The graphs above show the comparison of certain attributes compared to whether or not a personal has or does not have heart disease. It shows one of four outcomes: (False,False), (False,True), (True,False), or (True,True). Of these the first one is always the true or false of heart disease or attack and the second is always the true or false of the attribute.

In these graphs it is very easy to see when attributes help to avoid heart problems when looking at things like eating fruits and veggies or physical activity, it shows that if those attributes are true, then it is more likely that heart problems are false. On the other side there are certain things that show if they are false then, heart disease is more likely to be false. The biggest example of this is the stroke attribute, where it shows that if the does not have a stroke, they are more likely to also not have heart problems.



The above graphs show other attributes compared to the number of heart problems that people have. On all of them the blue is number of people without heart problems with that attribute and the orange is the number of those with heart problems. These show where an increased number of heart problems happen for example when it comes to the mental health groups the higher mental health problems tend to have higher heart problems. When looking at age the percentage of people in that age group with heart problems go up as the age goes up. Through these graphs it is easy to compare these different categories of certain attributes and the number of heart problems that happen.

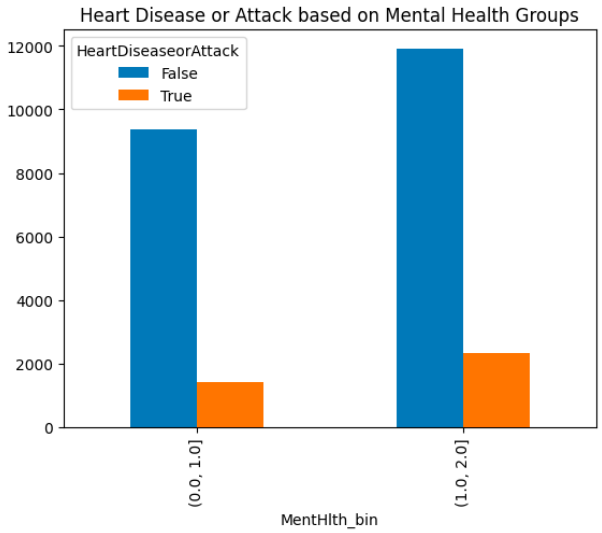
10. REFERENCES

[1] Centers for Disease Control and Prevention, https://www.cdc.gov/heartdisease/facts.html.

[2] Centers for Disease Control and Prevention, https://www.cdc.gov/brfss/index.html.

[3] Gordon T, Castelli WP, Hjortland MC, Kannel WB, Dawber TR. Predicting Coronary Heart Disease in Middle-Aged and Older Persons: The Framington Study.JAMA.1977;238(6):497499.doi:10.1001/jama.1977.03280060041018DOI: https://doi.org/10.1007/3-540-09237-4.

[4] R. Katarya and P. Srinivas, "Predicting Heart Disease at Early Stages using Machine Learning: A Survey," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020, pp. 302-305, doi: 10.1109/ICESC48915.2020.9155586.



[5] Mohammad Shafenoor Amin, Yin Kia Chiam, Kasturi Dewi Varathan, Identification of significant features and data mining techniques in predicting heart disease, Telematics and Informatics, Volume 36, 2019, Pages 82-93, ISSN 0736-5853, https://doi.org/10.1016/j.tele.2018.11.007.