Data Collection & Data Quality
Research Proposal
On
Prediction of Heart Disease Using Machine Learning Algorithms

HÖGSKOLAN
DALARNA

Manthri Bala Kiran
H19baman@du.se

Abstract:

A Medical team will be helping the patients with the possible occurrence of the heart disease at the early stages of the disease. Data Mining tools will be used for mining data which is been taken from the medical diagnosis and which is analyzed to help the medical team to understand the disease. Data Mining will play a crucial role in building an efficient model for the medical systems which helps the medical team to predict the possibility of Heart Disease using the patient's data. The data extracted by the medical team will be included with the main risk factors leading to the heart disease. Data Mining is a technique which helps to find the useful patterns from the massive data set which have not been seen before and helps us to understand the unknown relations from the data which would be giving us the knowledge for detecting and preventing the heart disease at the early stages. The available data mining techniques include classification techniques involving Navie Bayes, Decision Tree, Neural Network, Genetic Algorithm and it also include the clustering algorithms like K-NN and Support Vector Machines. As the technology grows we can see the Machine Learning techniques also been applied in identifying the heart disease which improves the accuracy of prediction.

Introduction:

In recent times one of the major cause of deaths is due to the Heart Disease. Heart is considered as the most important organ of the Human Body which weighs about 250-350 grams which is considered as the approximately the size of a fist. In the lifespan of 66-68 years heart will be beating around 2.5 billon times (Ravish, Shanthi, Shenoy, & Nisargh, n.d.). The heart helps to pump the blood to different parts of the human body. If the blood pressure in the body is not sufficient, the organs in the body will be failing and this will be leading to failure of the heart and brain and which leads to death in few minutes. The major factor for heart disease is been associated with these factors they are age, family history, hypertension, tobacco, obesity, diabetes, high cholesterol, smoking, alcohol intake, physical inactivity, poor diet and chest pain.

Chest pain and Fatigue are considered as the symptoms for the heart disease, but most of the times there would be no symptoms until the person is stuck with a heart attack. Sex, age and family history are considered as the risk factors for the heart disease but we can't change them. Some factors in our lifestyle including smoking, high cholesterol, high blood pressure and physical inactivity are considered the risk factors for heart disease which can be modified or eliminated with some medication. Diabetes and obesity can sometimes be prevented with an early change in the lifestyle. For the physicians to diagnose the heart disease requires experience and skill, firstly they will the evaluating the results obtained from the patients and comparing them with the previous decisions made for the other patients who have been examined with the same condition. Which is a complex procedure to follow where it includes to consider number of factors and evaluate results from them (Resul Das, n.d.).

Due to the risk factors associated with the heart disease like diabetes, high blood pressure, high cholesterol, abnormal plus rate and many others predicting the heart disease is very difficult. To find the heart disease many data mining techniques and neural networks have been deployed. The severity of the disease is classified based on methods like K-Nearest Neighbor Algorithm, Decision Tree and Naïve Bayes. The heart disease is need to be handled very careful, not doing so many things which would be affecting the heart or cause premature death. Data investigation of the heart disease using the data mining techniques with classification helps to predict the heart disease at the early stages.

Literature Review:

A promising technique of predicting the heart disease is done with the Association rules. Association rules have been applied on the real data set which contains the medical records of the patients with the heart disease. In the medical terms association rules relate heart perfusion measurements and risk factors to the degree of disease in four specific arteries. The search constraints and test sets validation reduce the number of rules and procedure a set of rules with high predictive accuracy (Ordonez, n.d.)

Development of a multilayer perceptron-based decision support system was done for the diagnosis of heart disease. A proposed coding schemes was encoded with the input layer which includes 40 input variables which are been categorized into four groups. Training of the system was done with an improved backpropagation algorithm. In the Neural Network models an Multilayer perception is used which is known for its good architecture and makes the model important and very easy and simple to understand (Hongmei Yan, n.d.).

There is a risk of rising for the prevalence of the cardiovascular diseases. The adequate risk prediction and identification of its determinants is increasingly important. The study proposed by the Rotterdam was on a prospective population based cohort study started in 1990 in the city of Rotterdam, Netherlands. This was done to identify the determinants and prognosis of the cardiovascular diseases. The clear outcome definitions were depended on the case finding in the epidemiological studies. In this article they talk about the methods performed for the data collection and the up-to-date definitions of the cardiac outcomes based on international guidelines (Maarten J. G. Leening, n.d.).

To monitor and independently validate the coronary heart disease (CHD) events (acute Myocardial Infarction and death due to coronary heart disease) a community surveillance component of Atherosclerosis risk in communities was proposed. And acute decompensated heart failure occurring among residents of four geographically defined communities in the united states in order to evaluate trends in mortality, incidence, case fatality rates and medical care by age, gender, race, community and time (Diseases., n.d.).

Methodological Considerations:

In this section firstly, I would like to explain about the Heart Disease, next I will be explaining how I choose my sample for this research and I will be talking about my data collection and data quality.

Heart Disease:

In this section I will be explaining the characteristics of heart disease, its symptoms, causes and known diagnostic techniques and the treatment options for the disease (Jesmin Nahar, n.d.).

A) Characteristics of Heart Disease:

They were several researches was done on heart disease and they have established that the heart disease is not a single condition but refers to many conditions in which the heart and blood vessels are injured and do not function properly, which will be resulting in serious and fatal health problems. They came up with different types of heart diseases, the major types are: Atherosclerosis, coronary, rheumatic, congenital, myocarditis, angina and arrhythmia.

B) Symptoms of Heart Disease:

The symptoms of the heart disease will be different to person to person. There will be no early symptoms in majority of cases and the disease is identifiable only in the advance stage. The most common symptoms of heart disease are:

- Cheat Pain
- Strong compressing or flaming sensation in the chest, neck or shoulders
- Discomforts in chest area
- Sweating, Light-headedness, Dizziness, Shortness of Breath
- Cough
- Palpitations
- Fluid Retention

C) Causes of Heart Disease:

The main causes of heart disease are unclear yet, but age, gender, family history, and ethnic background are considered to be the major causes in the different investigations which have been taken place. Other factors like eating habits, fatty foods, lack of exercise, high cholesterol, hypertension, pollution, life style factors, obesity, high blood pressure, stress, diabetes and lack of awareness have also been the main causes of developing a heart disease.

D) Diagnostic techniques of heart disease:

The diagnosis of heart disease patients depends on clinical history and their physical examination, and there also have different diagnostic procedures including:

- **Electrocardiography (ECG)**
- **Stress Testing**
- **Magnetic Resonance Imaging (MRI)**
- **Electrophysiologic Testing**
- **Tilt Table testing**
- **Radiologic Procedures (X-rays)**
- **Angiography**
- **Other Notable procedures are**: Continuous ambulatory electrocardiography (Holter monitor), radionuclide imaging, cardiac catheterization, central venous catheterization. Fluoroscopy is also used to detect heart disease, through the use is generally infrequent. Blood test involving measurement of sugar levels, cholesterol and other substance are also used.

Majority of these procedures have very small risk, the risk intensifies with complex and severe heart disorder.

E) Treatment options for Heart Disease:

Treatment of heart disease depends on the type, patient's age, health condition and the patient's choice. The major types of treatment are:

- **Medications**
- **Balloon Angioplasty**
- **Bypass Graft Surgery**
- **Electrophysiologic Devices (Pacemakers)**

Data Collection:

Sampling: I need to be very selective about my sampling. So I would like to go with Judgement Sampling also known as Selective or Subjective sampling. This technique will be relied on the judgement of the researcher when choosing who to ask to participate. As a researcher I may implicitly choose a "Representative" sample to suit my needs or specifically approach individuals with certain characteristics.

To choose my sample, I would like to choose the people whose age is 30+ and combination of different genders. I can also be specific in choosing the people in specific with their activities like people who smoke and people who don't smoke and drinking habits. People who try to stay activity in their daily life and people who will be inactivity can be chosen. Sample can also be chosen

based on the family history which will be playing an important in the current sample population. Food habits of the sample will also be playing some vital role todays the risk of heart disease.

After the first round of sample, people will be examined physically where some study physicians and various tests will be performed like collecting resting electrocardiogram and echocardiography. With these sample will also be examined on the previous health issues and a keen study will be done on the previous health issues.

Collection of Data: For collecting the habits of the population which helps to choose a sample a questionnaire is prepared. A questionnaire is considered as a research instrument consisting of a series of questions for gathering information from respondents. Questionnaires can be thought of as a kind of written interview. They can be carried out face to face, by telephone, internet or post. Questionnaires provide a relatively cheap, quick and efficient way of obtaining large amounts of information from a large sample of people. Questionnaires can be an effective means of measuring the behavior, attitudes, preferences, opinions and intentions of relatively large number of subjects more cheaply and quickly than other methods. The data collected from the Questionnaire can be considered as the Quantitative data.

All the population will be handed a questionnaire where they will be answering their age, gender, smoking & drinking habits, activity status, food habits and family history. After the population filling the questionnaire when can chose selective sampling putting different combinations on the features and trying to pick the best sample where our best requirements would be matched. After choosing the sample, the sample is being kept into a keen observation where all the other details will be recorded and it will be continued to time over and over again.

A hospital can also be contacted if they would be having the details of the patients who have been recognized with the heart disease currently and we can collect the basic information of the age, gender, smoking and drinking habits, food habits and activity with the questionnaire and they can also be requested to participate in the physical experiment where their ECG and Echocardiography details are been collected. By doing this we can have sample which is a mixture of people with and without the heart disease. The data collected by the physical experiment is considered as the qualitative data.

Trustworthiness in Quantitative Research:

For quantitative research it is referred to a validity and reliability.

Reliability: In our case this stands with the machinery which is been used to collect the data at the second part of the experiment for this we need to check with the machines whether they would be

giving the same result continuously for the same experiment. We also need to train our physician well to understand the readings from the machines and collect the data efficiently.

Validity: For the experiment need to be valid we need to sit back and prepare the questions on the point which implies to cover the major cause of the heart disease at the first stage of the research. We need to get all the possibility information which helps us to take a proper sample from the first part of the research.

Trustworthiness in Qualitative Research:

Trustworthiness is made up of four criteria, each of which has an equivalent criterion in quantitative research:

1) Credibility
2) Transferability
3) Dependability
4) Confirmability

Credibility: For the later part of the research we would be collecting the data of heart disease in the previously recognized patients from hospital and we will be coming that with the current research so we can find the blind spots in the current research. We also check the research data and the analytical process with an expect to find where we can improve the research and find the gaps in the research to improve the research and predication quality.

Transferability: It is how the qualitative researcher demonstrates that the research study's findings are applicable to other contexts. In this case, we hope our data can be used for the more study for the new people who are starting out. This can help them to understand the inside situation of the research and use the data which will have the best effects on the research.

Confirmability: it is a degree of neutrality in the research study's findings. In other words, we can say that the findings should be based on the responses given by the participants and the data collected from the physicians during the research. This means that the researchers should not be making any interpretation during the research and should not influence the participants for making necessary changes to answers the questions to achieve certain goals and motives.

Dependability: We need to report our findings in a way where it can be used by the new research team again who wants to replicate or implement the study with the new data. All the findings should well properly document and analyzed well to ensure the findings are accurate. It will be better if we ask a third person who is not in part of our current research and see whether he will be understanding the research and analyze how well he can follow through our research.

Data Collection and Data Quality issues:

This process of data collection is time taking but we would be getting the best results when we consider this approach. Time and money should be invested accordingly for the best results. Selecting the sample will be considered as the most important step as everything will be followed after the sample is selected. But the trustworthiness of the people filling out the questionnaire is on the take as the sample will be chosen based on the questionnaire. All the physicians participating should be every keep and attentive while recoding the measurements of the people which also will be playing a role for detecting the heart disease.

In the first part of the research the data quality will be depended on the answers given in the questionnaire. How genuine the answers were given and it also depends up on the person filling the questionnaire (means is the right person filling the questionnaire). These things can be handled by sending the questionnaire personally and asking him to be genuine towards the answers which were asked. In the second part the quality of the data depends on how the machines perform and how the physicians would be noticing or collecting the data from the equipment. This can be controlled by installing the proper machinery and test with the machines with all the basic tests which it needs to pass. And the physicians should be given a proper training how to record the values and collect the data from the machines which would be resulting for a good data quality.

Ethical Considerations:

It is important to follow ethical considerations during the data collection and data mining. They are (peer, n.d.):

1) Informed Consent: All the participants should be informed about the research during our part of the research where we will be collecting the information of the participants with the questionnaire. And it would be best if we include the purpose of the research in the questionnaire so that the participants can see the purpose before they are filling out the questionnaire.

2) Voluntary Participation: All the participants should be volunteered to fill the questionnaire and allow them self's to be in the part of the research. The participants will be given a right to leave the research whenever they feel to move out of it without any pressure forced on them.

3) Do no harm: We will be clearly explaining our participants that there will be no harm in both physical and psychological and therefore can be in in form of: stress, pain, anxiety, diminishing self-esteem or an invasion of privacy.

4) Confidentiality: All the identification which will be collected from the participants will also be kept hidden from the outside world. All the published reports of the research will be done by removing the personal details of the participants.

5) Anonymity: The details of the participants will be hidden from the research team so there will be no risk or influence on the participants from the research team.

6) Only assess relevant components: During our research will be on point in the collecting the data or treating the participants. We will be focused on the intention of the research and what the data gathered will be useful for.

Alternative Case/ Research:

As the process is a huge time taking one and continuous research and processes of data collection should be followed throughout the research. For a short term research, we can use publicly available datasets where the researchers have followed all the steps with respective of your research and your requirements. One of the example of such dataset is UCI heart disease dataset (Jesmin Nahar, n.d.). Where the dataset consists of total of 76 attributes, where majority of the studies used maximum of 14 attributes which are considerably linked to the heart disease. They are:

1) Age: Which is Numeric
2) Sex: Nominal – Has 2 values: Male, Female
3) Chest pain type: which is Nominal – has 4 values: typical angina, atypical angina, non-anginal pain, asymptomatic.
4) Trestbps: Which is Numeric, and indicates resting blood pressure on admission.
5) Chol: Which is Numeric, and indicates Serum cholesterol in mg/dl.
6) FBS: Which is Nominal – has 2 values: True, False, indicates whether fasting blood sugar is greater than 120 mg/dl.
7) Restecg: Which is Nominal – has 4 values: normal, abnormal, ST-T wave abnormality, ventricular hypertrophy- Indicates resting electrocardiographic outcomes.
8) Thalach: Which is Numeric, and indicates maximum heart rate achieved.
9) Exang: Which is Nominal – has 2 Values: Yes, or No – highlights existence of exercise induced angina.
10) Old Peak: Which is Numeric: ST depression induced by exercise relative to rest.
11) Slope: Which is Nominal – has 3 values: upsloping, flat, downsloping – the slope characteristics of the peak exercise ST segment.
12) CA: Which is Numeric – number of fluoroscopy colored major vessels (0-3).
13) Thal: Which is Nominal – and has 3 values, Normal, fixed defect, reversible defect- the heart status.

14) The class attribute: Value is either healthy or existence of heart disease (sick type: 1,2,3 and 4).

References

Banu, N. K., & Swamy, S. (n.d.). *Prediction of heart disease at early stage using data mining and big data analytics: A survey*. (IEEE Exolore) Retrieved 10 26, 2020, from https://ieeexplore-ieee-org.www.bibproxy.du.se/document/7955226

Diseases., I. o. (n.d.). *A Nationwide Framework for Surveillance of Cardiovascular and Chronic Lung Diseases.* (NCBI NLM) Retrieved 10 26, 2020, from https://www.ncbi.nlm.nih.gov/books/NBK83162/

Hongmei Yan, Y. J. (n.d.). *A multilayer perceptron-based medical decision support system for heart disease diagnosis*. (Science Direct) Retrieved 10 26, 2020, from https://www.sciencedirect.com/science/article/pii/S095741740500142 9?casa_token=EqxXwgOuIjMAAAAA:oJg-ls1E1-rD-cpNc-u7Jusv6RHNZsKAn-QOG4p1sRpT-cyxqXWYI2ipslYUx296Dqdbnfh32W0

Jesmin Nahar, T. I.-P. (n.d.). *Association rule mining to detect factors which contribute to heart disease in males and females*. (Science Direct) Retrieved 10 26, 2020, from https://www.sciencedirect.com/science/article/pii/S095741741200989X

Maarten J. G. Leening, M. K.-v.-R. (n.d.). *Methods of data collection and definitions of cardiac outcomes in the Rotterdam Study*. (NCBI NLM) Retrieved 10 26, 2020, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3319884/

Ordonez, C. (n.d.). *Association rule discovery with the train and test approach for heart disease prediction*. (IEEE Explore) Retrieved 10 26, 2020, from https://ieeexplore-ieee-org.www.bibproxy.du.se/document/1613959

peer, M. (n.d.). *Ethicial Considerations*. (my peer) Retrieved 10 26, 2020, from http://mypeer.org.au/monitoring-evaluation/ethical-considerations/

Ravish, D. K., Shanthi, K., Shenoy, N. R., & Nisargh, S. (n.d.). *Heart function monitoring, prediction and prevention of Heart Attacks: Using Artificial Neural Networks*. (IEEE Explore) Retrieved 10 26, 2020, from https://ieeexplore-ieee-org.www.bibproxy.du.se/document/7019580

Resul Das, I. T. (n.d.). *Effective diagnosis of heart disease through neural networks ensembles*. (Science Direct) Retrieved 10 26, 2020, from https://www.sciencedirect.com/science/article/pii/S095741740800657X?casa_token=tF3NXI3eDW0AAAAA:lR3m0E-THH1QLq9R-F9RmKG5FZTGNtjAjqts4tZACqUGY08Z4A2aJ9yD0PAYOgfC0weMND56DRw

Solutions, S. (n.d.). *What is Trustworthiness in Qualitative Research?* (Statustics Solutions) Retrieved 10 26, 2020, from https://www.statisticssolutions.com/what-is-trustworthiness-in-qualitative-research/