# HEART DISEASE PREDICTION SYSTEM

by  
  
KEMBOI KEVIN  
C026-01-0738/2020

A RESEARCH DOCUMENTATION SUBMITTED TO THE DEPARTMENT OF  
COMPUTER SCIENCE, SCHOOL OF COMPUTER SCIENCE & INFORMATION  
TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS OF AN  
AWARD OF A DEGREE IN BACHELOR OF SCIENCE COMPUTER SCIENCE OF  
DEDAN KIMATHI UNIVERSITY  
  
AUG, 2024

# DECLARATION

I hereby declare that this project is my original work and to the best of my knowledge the project has not formed the basis of any other award.  
  
Student:  
KEMBOI KEVIN  
C026-01-0738/2020  
Signature: ……………… Date: …………………  
  
Personal Contribution Statement:  
1. Proposed the idea and conceptualized the Heart Disease Prediction System.  
2. Developed the machine learning models and integrated them into the system.  
3. Conducted testing and prepared detailed documentation for the project.  
  
Supervisor:  
Mr. Patrick Naivasha  
Signature: ……………… Date: …………………

# ABSTRACT

This project focuses on predicting heart disease using machine learning techniques. Motivated by the global rise in cardiovascular diseases, the project employs algorithms like Gradient Boosting Classifier for high accuracy. The system uses a dataset from the UCI repository and features a user-friendly GUI for prediction. The work achieved accuracy of over 90% and highlights potential areas for future exploration, such as real-time data integration.

# DEDICATION

I dedicate this project to God Almighty, my creator, and my family, who have been my source of inspiration and support throughout this journey.

# ACKNOWLEDGMENT

I express my gratitude to my supervisor, Mr. Patrick Naivasha, for his guidance, and to my family and friends for their unwavering support.

# TABLE OF CONTENTS

1. CHAPTER ONE - INTRODUCTION  
2. CHAPTER TWO - LITERATURE REVIEW  
3. CHAPTER THREE - METHODOLOGY  
4. CHAPTER FOUR - ANALYSIS AND DESIGN  
5. CHAPTER FIVE - TESTING AND IMPLEMENTATION  
6. CHAPTER SIX - CONCLUSION AND RECOMMENDATION  
7. REFERENCES  
8. APPENDIX

# CHAPTER ONE - INTRODUCTION

HEART DESEASE PREDICTION SYSTEM

By

KEMBOI KEVIN

C026-01-0738/2020

A project proposal submitted for the study leading to a project report in partial fulfilment of the requirements for the award of a Bachelor of Science in Computer Science at Dedan Kimathi University.

Supervisor: PATRICK NAIVASHA

DATE

January – April 2024

Declaration

This is to certify that the work being presented in the project entitled “Heart Disease Prediction System” submitted by undersigned student of Bachelors in COMPUTER SCIENCE in the fulfillment for award of Bachelors in Computer Science is a record of my own work carried out by me under guidance and supervision of PATRICK NAIVASHA of the Department of Computer Science and that this work has not submitted elsewhere for award of any other degree.

Name of student: KEMBOI KEVIN

Registration Number: C026-01-0738/2020

Sign:

Approval

This project was done and presented by me before the panel concerned on the 5th February 2024 at Dedan Kimathi University with my approval and that of my supervisor

Supervisor Name: PATRICK NAIVASHA

Signature:

Date:

Acknowledgement

The satisfaction that accompanies that the successful completion of any task would be incomplete without the mention of people whose ceaseless cooperation made it possible, whose constant guidance and encouragement crown all efforts with success. I am very grateful to my project supervisor Patrick Naivasha, for the guidance, inspiration and constructive suggestions that helpful me in the preparation of this project. I won’t forget to also mention my course mates; Chris Sumba ,Joseph Kariuki and Chris Kinyanjui for their wonderful and skillful guidance in assisting me with the necessary support to ensure that my project is a success. I also thank my parents and family at large for their moral and financial support in funding the project to ensure successful completion of the project.

Dedication.

I dedicate this project to God Almighty my creator, my strong pillar, my source of inspiration, Wisdom, knowledge and understanding. He has been the source of my strength throughout this Program and on His wings only have I soared. I also dedicate this work to my friends; Isaac Ikinya who has encouraged me all the way and whose encouragement has made sure that I live it all it takes to finish that which I have started. To my father Kemboi David and all my beloved friends(Chris Sumba ,Joseph Kariuki and Chris Kinyanjui). Who have been affected in every way possible by this quest?

All the work done in coming up with this system is dedicated to my family for being with/part of me in the whole process especially my dear dad and mum who stood by me in all situations even at the times of financial need.

Thank You. My Love for You All Can Never Be Quantified. God Bless You.

Abstract.

As Heart acts a major role in corporeal body part. Heart diseases or Cardiovascular Diseases (CVDs) are some of the main reason for a huge number of death in the world and some of the reports said that over the last few decades heart disease or CVD has emerged as the most life- threatening disease, not only in Kenya but in the whole world.

Therefore, the diseases of heart wants more perfection and exactness for diagnose and analyse. This disease occurs due to various problems such as over pressure, blood sugar,

Chest pain,high blood pressure, Cholesterol etc. in human body . Heart is the next major organ comparing to the brain which has more priority in the Human body. It pumps the blood and supplies it to all organs of the whole body.

In the health care sector, Machine Learning plays an important role in the health care Industry. So, there is a need for a reliable, accurate, and feasible system to diagnose such diseases in time for proper treatment. Machine Learning algorithms and techniques have been applied to various medical datasets to automate the analysis of large and complex data. A system model is capable of several data processing algorithms for the classification of heart disease. Prediction of occurrences of heart diseases in the medical field is significant work. Data analytics is useful for prediction from more information and it helps the medical center to predict various diseases. The correct prediction of heart disease can prevent life threats, and incorrect prediction can prove to be fatal at the same time.

The objective of this project is to detect whether patients have any chance of heart disease or not by giving number of features to patients with having maximum accuracy of above 97%. By Using Machine learning algorithms and deep learning are applied to compare the results and analysis of the UCI Machine Learning Heart Disease dataset. The dataset consists of 14 main attributes used for performing the analysis

Table of Contents

Chapter One.

Introduction

Heart is one of the most indispensable organs in the human body. It is a organ that serves as a pump to circulate the blood. The heart is a muscular organ about the size of a fist, located just behind and slightly left of the breastbone. The heart pumps blood through the network of arteries and veins called the cardiovascular system. Oxygen is distributed through the circulatory system of the body in the blood, and if the heart does not function correctly, the entire circulatory system of the body will fail. So if the heart doesn’t work properly, it could even lead to death.

According to the World Health Organization (WHO), in the last 15 years, an estimated 17 million people die each year from cardiovascular disease, particularly heart attacks and strokes [1]. Heart disease and stroke are the biggest killers. To predict heart disease, Machine Learning can be used for identifying unseen patterns and providing some clinical insights that will assist the physicians in planning and providing care.

According to World Health Organization, heart related diseases are responsible for taking 17.7 million lives every year, 31% of all global deaths. In Kenya too, heart-related diseases have become the leading cause of mortality. Heart diseases have killed 0.3 million Kenyans in 2016, according to the 2016 Global Burden of Disease Report, released on September 15, 2017. Heart-related diseases increase the spending on health care and also reduce the productivity of an individual.

Estimates made by the World Health Organization (WHO), suggest that India has lost up to $237 billion, from 2005- 2015, due to heart-related or cardiovascular diseases. Thus, feasible and accurate prediction of heart-related diseases is very important.

Heart disease is common among both men and women in most countries around the world. Therefore, people should consider heart disease risk factors. Although it plays a genetic role, some lifestyle factors significantly affect heart disease .

The known risk factors for heart disease; radiation therapy for age, gender, family history, smoking, some chemotherapy drugs and cancer, malnutrition, high blood pressure, high blood cholesterol levels, diabetes, obesity, physical mobility, stress, and poor hygiene. These are the various risk factors in which the patient’s exposure towards developing a CVD.

The most common type is coronary artery disease, which can cause a heart attack. Other types of heart disease may involve the valves in the heart, or the heart may not pump well and cause heart failure. Some people are born with heart disease. Anyone, including children, can develop heart disease. It happens when a substance called plaque builds up in your arteries. Smoking, unhealthy eating and lack of exercise increase your risk of heart disease. High cholesterol, high blood pressure or diabetes can also increase your risk of heart disease.

In contrast, the removal or improvement of this factor decreases this risk. This interpretation suggests the causality between the factor and the illness, which means that the risk factor precedes the disease (the notion of anteriority).

Correction of the factor will cause the disease (the idea of reversibility) to decrease its occurrence. Of course, it must be recognized in several different populations and offer a plausible physiopathological explanation of the disease. Strictly speaking, when there is no direct causal relationship, it is a “risk marker,” a witness to a process (e.g., the elevation of micro albuminuria, elevation of C- reactive protein CRP)

.

The main heart disease risk factors such as physiological factors (age, sex, and menopausal status), lifestyle factors (smoking, physical activity, alcohol, stress), metabolic syndrome factors (insulin resistance), dyslipidemia, abdominal obesity, high blood pressure) and dietary factors. A heart disease risk factor is defined as a factor in which the patient’s exposure to this factor increases the risk of developing a CVD. In contrast, the removal or improvement of this factor decreases this risk. The risk factor’s importance is defined by the association’s strength with the disease (expressed by the relative risk observed in the exposed subjects compared to the unexposed) and the gradual association (parallel to the risk factor).

There are several types of heart disease which include:

Coronary Artery Disease (CAD)

Heart Arrhythmias.

Heart Failure.

Heart Valve Disease.

Pericardial Disease.

Cardiomyopathy (Heart Muscle Disease)

Congenital Heart Disease.

To examine the cardiac disease mischance, the particular issues which need to be discussed are those related to the behaviors. Furthermore, patients will undergo extensive examinations, such as blood pressure, glucose, vital signs, chest pain, electrocardiograms, maximum heart rate, and elevated levels of sugar, but the bright side may be that successful treatment is feasible if the disease is easily and early detected and anticipated, but treatment for all of these cardiac patients is depending on clinical studies, the patient history, and the responses to questions by the patient

To deal with this disease, there are several methods of prevention, such us natural methods, like stoping smoking, maintaining a healthy weight, adopting a healthy diet and practicing sports regularly.We also have the scientific methods such as drugs and surgeries. The prediction of this disease before being infected is part of the prevention

Methods, or the computer tools are the most used means in it, more precisely the Machine Learning algorithms. Determining the probability of having cardiac disease manually is hard to depend on as risk factors. Recently, to solve difficult issues, a range of data mining techniques and machine learning techniques are built.

Still, more advanced machine learning will assist us to identify patterns and their useful knowledge. While it has several uses in the medical field, machine learning is mainly utilized to forecast the heart disease.

Machine learning (ML) plays a significant role in disease predicting. It predicts whether the patient has a particular disease type or not based on an efficient learning technique. In making of these projects I have used several supervised learning techniques for predicting the early stage of heart disease by providing them risk factor

Here, I will use the below machine learning algorithms then I will select the best one based on its classification report.

Support Vector Machine

Random Forest

Ada Boost

Gradient Boosting

With using all these techniques the highest accuracy of 90.82% I got in GradientBoostingClassifier.

Background of the Study

Data analysis proves to be crucial in the medical field. It provides a meaningful base to critical decisions. It helps to create a complete study proposal. One of the most important uses of data analysis is that it helps in keeping human bias away from medical conclusion with the help of proper statistical treatment. By use of data mining for exploratory analysis because of nontrivial information in large volumes of data.

The health care industries collect huge amounts of data that contain some hidden information, which is useful for making effective decisions for providing appropriate results and making

Effective decisions on data, some data mining techniques are used to better the experience and conclusion that have been given.

Heart predictor system will use the data mining knowledge to give a user-oriented approach to new and hidden patterns in the data. The knowledge which is implemented can be used by the healthcare experts to get better quality of service and to reduce the extent of adverse medicine effect.

Problem Statement.

Heart disease can be managed effectively with a combination of lifestyle changes, medicine and, in some cases, surgery. With the right treatment, the symptoms of heart disease can be reduced and the functioning of the heart improved. The predicted results can be used to prevent and thus reduce cost for surgical treatment and other expensive.

The overall objective of my work will be to predict accurately with few tests and attributes the presence of heart disease. Attributes considered form the primary basis for tests and give accurate results more or less. Many more input attributes can be taken but our goal is to predict with few attributes and faster efficiency the risk of having heart disease. Decisions are often made based on doctors’ intuition and experience rather than on the knowledge rich data hidden in the data set and databases. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients.

Data mining holds great potential for the healthcare industry to enable health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and reduce costs. According to (Wurz & Takala, 2006) the opportunities to improve care and reduce costs concurrently could apply to as much as 30% of overall healthcare spending. The successful application of data mining in highly visible fields like e-business, marketing and retail has led to its application in other industries and sectors. Among these sectors just discovering is healthcare. The healthcare environment is still „information rich‟ but „knowledge poor‟. There is a wealth of data available within the healthcare systems. However, there is a lack of effective analysis tools to discover hidden relationships and trends in the data for African genres.

Objectives.

Main Objectives.

The main objective of this research is to develop a heart prediction system. The system can discover and extract hidden knowledge associated with diseases from a historical heart data set

Heart disease prediction system aims to exploit data mining techniques on medical data set to assist in the prediction of the heart diseases.

Specific Objectives.

Provides new approach to concealed patterns in the data.

Helps avoid human biasness.

To implement GradientBoosting Classifer that classifies the disease as per the input of the user.

Reduce the cost of medical tests.

Justification.

Clinical decisions are often made based on doctor’s insight and experience rather than on the knowledge rich data hidden in the dataset. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. The proposed system will integrate clinical decision support with computer-based patient records (Data Sets). This will reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge rich environment which can help to significantly improve the quality of clinical decisions.

There are voluminous records in medical data domain and because of this, it has become necessary to use data mining techniques to help in decision support and prediction in the field of healthcare. Therefore, medical data mining contributes to business intelligence which is useful for diagnosing of disease

Scope and Limitation.

Scope.

Here the scope of the project is that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modeling and

analysis tools, e.g., data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions

Limitations.

Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. The automation of the same would be highly beneficial. Clinical decisions are often made based on doctor’s intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Data mining have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions.

CHAPTER TWO: LITERATURE REVIEW

Introduction

Data mining is the process of finding previously unknown patterns and trends in databases and using that information to build predictive models. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases. The World Health Statistics 2012 report enlightens the fact that one in three adults worldwide has raised blood pressure - a condition that causes around half of all deaths from stroke and heart disease. Heart disease, also known as cardiovascular disease (CVD), encloses a number of conditions that influence the heart – not just heart attacks. Heart disease was the major cause of casualties in the different countries including India. Heart disease kills one person every 34 seconds in the United States. Coronary heart disease, Cardiomyopathy and Cardiovascular disease are some categories of heart diseases. The term “cardiovascular disease” includes a wide range of conditions that affect the heart and the blood vessels and the manner in which blood is pumped and circulated through the body. Diagnosis is complicated and important task that needs to be executed accurately and efficiently. The diagnosis is often made, based on doctor’s experience & knowledge. This leads to unwanted results & excessive medical costs of treatments provided to patients. Therefore, an automatic medical diagnosis system would be exceedingly beneficial.

Literature Review

Numerous studies have been done that have focus on diagnosis of heart disease. They have applied different data mining techniques for diagnosis & achieved different probabilities for different methods.

(Polaraju, Durga Prasad, & Tech Scholar, 2017) proposed Prediction of Heart Disease using Multiple Regression Model and it proves that Multiple Linear Regression is appropriate for predicting heart disease chance. The work is performed using training data set consists of 3000 instances with 13 different attributes which has mentioned earlier. The data set is divided into two parts that is 70% of the data are used for training and 30% used for testing.

(Deepika & Seema, 2017) focuses on techniques that can predict chronic disease by mining the data containing in historical health records using Naïve Bayes, Decision tree, Support Vector Machine (SVM) and Artificial Neural Network (ANN). A comparative study is performed on classifiers to measure the better performance on an accurate rate. From this experiment, SVM gives highest accuracy rate, whereas for diabetes Naïve Bayes gives the highest accuracy.

(Beyene & Kamat, 2018) recommended different algorithms like Naive Bayes, Classification Tree, KNN, Logistic Regression, SVM and ANN. The Logistic Regression gives better accuracy compared to other algorithms. (Beyene & Kamat, 2018) suggested Heart Disease Prediction System using Data Mining Techniques. WEKA software used for automatic diagnosis of disease and to give qualities of services in healthcare centers. The paper used various algorithms like SVM, Naïve Bayes, Association rule, KNN, ANN, and Decision Tree. The paper recommended SVM is effective and provides more accuracy as compared with other data mining algorithms.

Chala Beyene recommended Prediction and Analysis the occurrence of Heart Disease Using Data Mining Techniques. The main objective is to predict the occurrence of heart disease for early automatic diagnosis of the disease within result in short time. The proposed methodology is also critical in healthcare organization with experts that have no more knowledge and skill. It uses different medical attributes such as blood sugar and heart rate, age, sex are some of the attributes are included to identify if the person has heart disease or not. Analyses of data set are computed using WEKA software.

(Soni, Ansari, & Sharma, 2011) proposed to use non- linear classification algorithm for heart disease prediction. It is proposed to use bigdata tools such as Hadoop Distributed File System (HDFS), Map reduce along with SVM for prediction of heart disease with optimized attribute set. This work made an investigation on the use of different data mining techniques for predicting heart diseases. It suggests to use HDFS for storing large data in different nodes and executing the prediction algorithm using SVM in more than one node simultaneously using SVM. SVM is used in parallel fashion which yielded better computation time than sequential SVM.

(Science & Faculty, 2009) suggested heart disease prediction using data mining and machine learning algorithm. The goal of this study is to extract hidden patterns by applying data mining techniques. The best algorithm J48 based on UCI data has the highest accuracy rate compared to LMT. (Purushottam, Saxena, & Sharma, 2016) proposed an efficient heart disease prediction system using data mining. This system helps medical practitioner to make effective decision making based on the certain parameter. By testing and training phase a certain parameter, it provides 86.3% accuracy in testing phase and 87.3% in training phase.

(Kirmani, 2017) suggested multi disease prediction using data mining techniques. Nowadays, data mining plays vital role in predicting multiple disease. By using data mining techniques, the number of tests can be reduced. This paper mainly concentrates on predicting the heart disease, diabetes

and breast cancer etc.,

(Sai & Reddy, 2017) proposed Heart disease prediction using ANN algorithm in data mining. Due to increasing expenses of heart disease diagnosis disease, there was a need to develop new system which can predict heart disease. Prediction model is used to predict the condition of the patient after evaluation on the basis of various parameters like heart beat rate, blood pressure, cholesterol etc. The

accuracy of the system is proved in java.

(A & Naik, 2016) recommended to develop the prediction system which will diagnosis the heart disease from patient’s medical data set. 13 risk factors of input attributes have considered to build the system. After analysis of the data from the dataset, data cleaning and data integration was performed. He used k-means and naïve Bayes to predict heart disease. This paper is to build the system using historical heart database that gives diagnosis. 13 attributes have considered for building the system. To extract knowledge from database, data mining techniques such as clustering, classification methods can be used. 13 attributes with total of 300 records were used from the Cleveland Heart Database. This model is to predict whether the patient have heart disease or not based on the values of 13 attributes.

(Sultana, Haider, & Uddin, 2017) proposed an analysis of cardiovascular disease. This paper proposed data mining techniques to predict the disease. It is intended to provide the survey of current techniques to extract information from dataset and it will useful for healthcare practitioners. The performance can be obtained based on the time taken to build the decision tree for the system. The primary objective is to predict the disease with a smaller number of attributes.

Proposed Architecture.

In this system we are implementing effective heart attack prediction system using Supervised Learning techniques. We can give the input as in CSV file or manual entry to the system. After taking input the algorithms apply on that input that is GradientBoosting Classifier. After accessing data set the operation is performed and effective heart attack level is produced.

The proposed system will add some more parameters significant to heart attack with their gender, age and the priority levels are by consulting expertise doctors and the medical experts. The heart attack prediction system designed to help the identify different risk levels of heart attack like normal, low or high and also giving the prescription details with related to the predicted result.

Flowchart .

The classification tree literally creates a tree with branches, nodes, and leaves that lets us take an unknown data point and move down the tree, applying the attributes of the data point to the tree until a leaf is reached and the unknown output of the data point can be determined. In order to create a good classification tree model, we need to have an existing data set with known output from which we can build our model. We also divide our data set into two parts: a training set, which is used to create the model, and a test set, which is used to verify that the model is accurate and not over fitted.

Project Flow Chart.

This will be the proposed flow chart that the system will look like

Data Flow Diagram

Proposed Model

Chapter 3: Research Methodology

Research Design.

I will be using the experimental type of research design. It is a quantitative research method. Basically, it is a research conducted with a scientific approach, where a set of variables are kept constant while other set of variables are being measured as the subject of the experiment. This is more practically while conducting face recognition and detection as it monitors the behaviours and patterns of a subject to be used to acknowledge whether the subject matches all details presented and cross checked with previous data. It is an effect research method as it is time bound and focuses on the relationship between the variables that give actual results.

System Development Methodology.

The methodology of software development is the method in managing project development. There are many models of the methodology are available such as Waterfall model model, Incremental model, RAD model, Agile model, Iterative model and Spiral model. However, it still need to be considered by developer to decide which is will be used in the project. The methodology model is useful to manage the project efficiently and able to help developer from getting any problem during time of development.

Also, it help to achieve the objective and scope of the projects. In order to build the project, it need to understand the stakeholder requirements.

Methodology provides a framework for undertaking the proposed DM modeling. The methodology is a system comprising steps that transform raw data into recognized data patterns to extract knowledge for users.

There are four phases that involve in the spiral model:

Planning phase

Phase where the requirement are collected and risk is assessed. This phase where the title of the project has been discussed with project supervisor. From that discussion, Heart Prediction System has been proposed. The requirement and risk was assessed after doing study on existing system and do literature review about another existing research.

Risk analysis Phase

Phase where the risk and alternative solution are identified. A prototype are created at the end this phase. If there is any risk during this phase, there will be suggestion about alternate solution.

Engineering phase

At this phase, a software are created and testing are done at the end this phase.

Evaluation phase

At this phase, the user do evaluation toward the software. It will be done after the system are presented and the user do test whether the system meet with their expectation and requirement or not. If there is any error, user can tell the problem about system.

Data Collection and Preprocessing.

The data set for this research was taken from UCI data repository.14 Data accessed from the UCI Machine Learning Repository is freely available. In particular, the Cleveland and Hungarian databases have been used by many researchers and found to be suitable for developing a mining model, because of lesser missing values and outliers. The data is cleaned and preprocessed before it is submitted to the proposed algorithm for training and testing.

The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms.

The overall objective of our work is to predict more accurately the presence of heart disease. In this paper, UCI repository dataset are used to get more accurate results. Two data mining classification techniques were applied namely Decision trees and Naive Bayes

his database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no

presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).

Attributes with categorical values were converted to numerical values since most machine learning algorithms require integer values. Additionally, dummy variables were created for variables with more than two categories. Dummy variables help Neural Networks learn the data more accurately.

Classifiers Used for Experiments.

Model / Libraries Used.

1.12.1.1 FLASK:-

Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre- existing third-party libraries provide common functions. Armin Ronacher, who leads an international group of Python enthusiasts named Pocco, develops it. Flask is based on Werkzeug WSGI toolkit and Jinja2 template engine. Both are Pocco projects.

However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools.

Applications that use the Flask framework include and .

1.12.1.2 NumPy :-

NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices.NumPy was created in 2005 by Travis Oliphant. It is an open source project and you can use it freely.NumPy stands for Numerical Python. NumPy targets the

of Python, which is a non-

optimizing . written for this version of Python often run much slower than equivalents. NumPy addresses the slowness problem partly by providing multidimensional arrays and functions and operators that operate efficiently on arrays; using these requires rewriting some code, mostly , using NumPy.

1.12.1.3 OS :-

The main purpose of the OS module is to interact with your operating system. The primary use I find for it is to create folders, remove folders, move folders, and sometimes change the working directory. You can also access the names of files within a file path by doing listdir(). We do not cover that in this video, but that's an option.The os module is a part of the standard library, or stdlib, within Python 3

1.12.1.4 MATPLOTLIB :-

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK. Matplotlib is a low level graph plotting library in python that serves as a visualization utility.Matplotlib was created by John D. Hunter.Matplotlib is open source and we can use it freely.Matplotlib is mostly written in python, a few segments are written in C, Objective-C and Javascript for Platform compatibility

1.12.1.5 PICKLE :-

Python pickle module is used for serializing and de-serializing a Python object structure. Any object in Python can be pickled so that it can be saved on disk. What pickle does is that it “serializes” the object first before writing it to file. Pickling is a way to convert a python object (list, dict, etc.) into a character stream. The idea is that this character stream contains all the information necessary to reconstruct the object in another python script

Advantages of using Pickle Module:

Recursive objects (objects containing references to themselves):

Object sharing (references to the same object in different places):

User-defined classes and their instances

It is advisable not to unpickle data received from an untrusted source as they may pose security threat. However, the pickle module has no way of knowing or raise alarm while pickling malicious data

1.12.1.6 JOBLIB :-

Joblib is a set of tools to provide lightweight pipelining in Python. In particular:

transparent disk-caching of functions and lazy re-evaluation (memoize pattern)

easy simple parallel computing

Joblib is optimized to be fast and robust on large data in particular and has specific optimizations for numpy arrays

Decision Trees.

The decision tree approach is more powerful for classification problems. There are two steps in this technique building a tree & applying the tree to the dataset. There are many popular decision tree algorithms CART, ID3, C4.5, CHAID, and J48. From these J48 algorithm is used for this system. J48 algorithm uses pruning method to build a tree. Pruning is a technique that reduces size of tree by removing over fitting data, which leads to poor accuracy in predications. The J48 algorithm recursively classifies data until it has been categorized as perfectly as possible. This technique gives maximum accuracy on training data. The overall concept is to build a tree that provides balance of flexibility & accuracy.

Ensemble DM approach.

In order to have more reliable and accurate prediction results, ensemble method is a well-proven approach practiced in research for attaining highly accurate classification of data by hybridizing different classifiers. The improved prediction performance is a well-known in-built feature of ensemble methodology. This study proposes a weighted vote-based classifier ensemble technique, overcoming the limitations of conventional DM techniques by employing the ensemble of two heterogeneous classifiers: Naive Bayesian and classification via decision tree

Tools

JUPYTER NOTEBOOK:-

Project Jupiter is a project and community whose goal is to "develop , , and services for across dozens of programming languages". It was spun off from in 2014 by and Brian Granger. Project Jupyter's name is a reference to the three core programming languages supported by Jupyter, which are , and , and also a to s notebooks recording the discovery of the . Project Jupyter has developed and supported the interactive computing products Jupyter Notebook, JupyterHub, and JupyterLab. Jupyter is financially sponsored by NumFOCUS

VS CODE:

Visual Studio Code is a source-code editor that can be used with a variety of programming languages,

Including , , , , and . It is based on the framework, which is used to develop

that run on the . Visual Studio Code employs the same editor component (codenamed "Monaco") used in

(formerly called Visual Studio Online and Visual Studio Team Services).

GITHUB:-

GitHub, Inc. is a provider of for and using . It offers the and (SCM) functionality of Git, plus its own features. It provides and several collaboration features such as , requests, , and for every project. Headquartered

in , it has been a subsidiary of since 2018.

It is commonly used to host projects. As of November 2021, GitHub reports having over 73 million developers and more than 200 million (including at least 28 million public repositories). It is the largest host as of November 2021.

HEROKU :-

Heroku is a cloud (PaaS) supporting several . One of the first , Heroku has been in development since June 2007, when it supported only the programming language, but now supports , , , , , , and . For this reason, Heroku is said to be a as it has features for

a to build, run and scale applications in a similar manner across most languages. Heroku was acquired by in 2010 for $212 million.

SCIKIT-LEARN :-

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. Extensions or modules for SciPy care conventionally named . As such, the module provides learning algorithms and is named scikit-learn.

The library is built upon the SciPy (Scientific Python) that must be installed before you can use scikit-learn. This stack that includes:

NumPy: Base n-dimensional array package

SciPy: Fundamental library for scientific computing

Matplotlib: Comprehensive 2D/3D plotting

IPython: Enhanced interactive console

Sympy: Symbolic mathematics

Pandas: Data structures and analysis

Software requirements:

Hardware Requirements

For application development, the following Software Requirements are:

Processor: Intel or high RAM: 1024 MB

Space on disk: minimum 100mb For running the application:

Device: Any device that can access the internet Minimum space to execute: 20 MB

The effectiveness of the proposal is evaluated by conducting experiments with a cluster formed by 3 nodes with identical setting, configured with an Intel CORE™ i7-4770 processor (3.40GHZ, 4 Cores, 8GB RAM, running Ubuntu 18.04 LTS with 64-bit Linux 4.31.0 kernel)

Budget.

The budget of completion for developing the heart disease prediction system will require various software and hardware devices. The application is averagely expensive to build but if happens to be as successful as the developer sees it to be it will bring forth enough profit to cover the costs undergone.

The table below explains the planned budget in Kenyan Shillings to develop the system:

Work Plan

Conclusion and Future Work.

The proposed system is GUI-based, user-friendly, scalable, reliable and an expandable system. The proposed working model can also help in reducing treatment costs by providing Initial diagnostics in time. The model can also serve the purpose of training tool for medical students and will be a soft diagnostic tool available for physician and cardiologist. General physicians can utilize this tool for initial diagnosis of cardio-patients. There are many possible improvements that could be explored to improve the scalability and accuracy of this prediction system. As we have developed a generalized system, in future we can use this system for the analysis of different data sets. The performance of the health’s diagnosis can be improved significantly by handling numerous class labels in the prediction process, and it can be another positive direction of research. In DM warehouse, generally, the dimensionality of the heart database is high, so identification and selection of significant attributes for better diagnosis of heart disease are very challenging tasks for future research.

References.

A, A. S., & Naik, C. (2016). Different Data Mining Approaches for Predicting Heart Disease, 277–

281. https://doi.org/10.15680/IJIRSET.2016.0505545

Beyene, C., & Kamat, P. (2018). Survey on prediction and analysis the occurrence of heart disease using data mining techniques. International Journal of Pure and Applied Mathematics, 118(Special Issue 8), 165–173. Retrieved from https:/ 85041895038&partnerID=40&md5=2f0b0c5191a82bc0c3f0daf67d73bc81

Brownlee, J. (2016). Naive Bayes for Machine Learning. Retrieved March 4, 2019, from https://machinelearningmastery.com/naive-bayes-for-machine-learning/

Kirmani, M. (2017). Cardiovascular Disease Prediction using Data Mining Techniques. Oriental Journal of Computer Science and Technology, 10(2), 520–528. https://doi.org/10.13005/ojcst/10.02.38

Polaraju, K., Durga Prasad, D., & Tech Scholar, M. (2017). Prediction of Heart Disease using Multiple Linear Regression Model. International Journal of Engineering Development and Research, 5(4), 2321–9939. Retrieved from

Purushottam, Saxena, K., & Sharma, R. (2016). Efficient Heart Disease Prediction System. In Procedia Computer Science (Vol. 85, pp. 962–969). https://doi.org/10.1016/j.procs.2016.05.288

Sai, P. P., & Reddy, C. (2017). International Journal of Computer Science and Mobile Computing HEART DISEASE PREDICTION USING ANN ALGORITHM IN DATA MINING.

International Journal of Computer Science & Mobile Computing, 6(4), 168–172. Retrieved from

Science, C., & Faculty, G. M. (2009). Heart Disease Prediction Using Machine learning and Data Mining Technique. Ijcsc 0973-7391, 7, 1–9.

Soni, J., Ansari, U., & Sharma, D. (2011). Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers. Heart Disease, 3(6), 2385–2392.

Sultana, M., Haider, A., & Uddin, M. S. (2017). Analysis of data mining techniques for heart isease prediction. In 2016 3rd International Conference n Electrical Engineering and Information and Communication Technology, iCEEiCT 2016 (pp. 1–5).

Artificial, I. T. O. (1995). Chapter 2. Neuron, 36–62. Bozzo, R., Conca, A., & Marangon, F. (2014). Decision support system for city logistics: Literature review, and guidelines for an ex-ante model. Transportation Research Procedia, 3(July), 518–527.

Çela, E. K., & Frasheri, N. (2012a). A literature review of data mining techniques used in healthcare databases. ICT Innovations 2012 Web Proceedings, 577–582.

Çela, E. K., & Frasheri, N. (2012b). A literature review of data mining techniques used in healthcare databases. ICT Innovations 2012 Web Proceedings, 577–582.

David, H. B. F., & Belcy, S. A. (2018). Heart Disease Prediction Using Data Mining Techniques,

6956(October), 1817–1823.

Desai, S. D., Giraddi, S., Narayankar, P., Pudakalakatti, N. R., & Sulegaon, S. (2019). Back- propagation neural network versus logistic regression in heart disease classification. Advances in Intelligent Systems and Computing, 702, 133–144. Kiruthika Devi, S., Krishnapriya, S., & Kalita, D. (2016). Prediction of heart disease using data mining techniques. Indian Journal of Science and Technology, 9(39), 1291–1293.

Lavanya, M., & Gomathi, M. P. M. (2016). Prediction of Heart Disease using Classification Algorithms. International Journal of Advanced Research in Computer Engineering & Technology, 5(7), 2278–1323.

Meenakshi, K., Maragatham, G., Agarwal, N., & Ghosh, I. (2018). A Data mining Technique for Analyzing and Predicting the success of Movie. Journal of Physics: Conference Series, 1000(1).

# CHAPTER TWO - LITERATURE REVIEW

HEART DESEASE PREDICTION SYSTEM

By

KEMBOI KEVIN

C026-01-0738/2020

A project proposal submitted for the study leading to a project report in partial fulfilment of the requirements for the award of a Bachelor of Science in Computer Science at Dedan Kimathi University.

Supervisor: PATRICK NAIVASHA

DATE

January – April 2024

Declaration

This is to certify that the work being presented in the project entitled “Heart Disease Prediction System” submitted by undersigned student of Bachelors in COMPUTER SCIENCE in the fulfillment for award of Bachelors in Computer Science is a record of my own work carried out by me under guidance and supervision of PATRICK NAIVASHA of the Department of Computer Science and that this work has not submitted elsewhere for award of any other degree.

Name of student: KEMBOI KEVIN

Registration Number: C026-01-0738/2020

Sign:

Approval

This project was done and presented by me before the panel concerned on the 5th February 2024 at Dedan Kimathi University with my approval and that of my supervisor

Supervisor Name: PATRICK NAIVASHA

Signature:

Date:

Acknowledgement

The satisfaction that accompanies that the successful completion of any task would be incomplete without the mention of people whose ceaseless cooperation made it possible, whose constant guidance and encouragement crown all efforts with success. I am very grateful to my project supervisor Patrick Naivasha, for the guidance, inspiration and constructive suggestions that helpful me in the preparation of this project. I won’t forget to also mention my course mates; Chris Sumba ,Joseph Kariuki and Chris Kinyanjui for their wonderful and skillful guidance in assisting me with the necessary support to ensure that my project is a success. I also thank my parents and family at large for their moral and financial support in funding the project to ensure successful completion of the project.

Dedication.

I dedicate this project to God Almighty my creator, my strong pillar, my source of inspiration, Wisdom, knowledge and understanding. He has been the source of my strength throughout this Program and on His wings only have I soared. I also dedicate this work to my friends; Isaac Ikinya who has encouraged me all the way and whose encouragement has made sure that I live it all it takes to finish that which I have started. To my father Kemboi David and all my beloved friends(Chris Sumba ,Joseph Kariuki and Chris Kinyanjui). Who have been affected in every way possible by this quest?

All the work done in coming up with this system is dedicated to my family for being with/part of me in the whole process especially my dear dad and mum who stood by me in all situations even at the times of financial need.

Thank You. My Love for You All Can Never Be Quantified. God Bless You.

Abstract.

As Heart acts a major role in corporeal body part. Heart diseases or Cardiovascular Diseases (CVDs) are some of the main reason for a huge number of death in the world and some of the reports said that over the last few decades heart disease or CVD has emerged as the most life- threatening disease, not only in Kenya but in the whole world.

Therefore, the diseases of heart wants more perfection and exactness for diagnose and analyse. This disease occurs due to various problems such as over pressure, blood sugar,

Chest pain,high blood pressure, Cholesterol etc. in human body . Heart is the next major organ comparing to the brain which has more priority in the Human body. It pumps the blood and supplies it to all organs of the whole body.

In the health care sector, Machine Learning plays an important role in the health care Industry. So, there is a need for a reliable, accurate, and feasible system to diagnose such diseases in time for proper treatment. Machine Learning algorithms and techniques have been applied to various medical datasets to automate the analysis of large and complex data. A system model is capable of several data processing algorithms for the classification of heart disease. Prediction of occurrences of heart diseases in the medical field is significant work. Data analytics is useful for prediction from more information and it helps the medical center to predict various diseases. The correct prediction of heart disease can prevent life threats, and incorrect prediction can prove to be fatal at the same time.

The objective of this project is to detect whether patients have any chance of heart disease or not by giving number of features to patients with having maximum accuracy of above 97%. By Using Machine learning algorithms and deep learning are applied to compare the results and analysis of the UCI Machine Learning Heart Disease dataset. The dataset consists of 14 main attributes used for performing the analysis

Table of Contents

Chapter One.

Introduction

Heart is one of the most indispensable organs in the human body. It is a organ that serves as a pump to circulate the blood. The heart is a muscular organ about the size of a fist, located just behind and slightly left of the breastbone. The heart pumps blood through the network of arteries and veins called the cardiovascular system. Oxygen is distributed through the circulatory system of the body in the blood, and if the heart does not function correctly, the entire circulatory system of the body will fail. So if the heart doesn’t work properly, it could even lead to death.

According to the World Health Organization (WHO), in the last 15 years, an estimated 17 million people die each year from cardiovascular disease, particularly heart attacks and strokes [1]. Heart disease and stroke are the biggest killers. To predict heart disease, Machine Learning can be used for identifying unseen patterns and providing some clinical insights that will assist the physicians in planning and providing care.

According to World Health Organization, heart related diseases are responsible for taking 17.7 million lives every year, 31% of all global deaths. In Kenya too, heart-related diseases have become the leading cause of mortality. Heart diseases have killed 0.3 million Kenyans in 2016, according to the 2016 Global Burden of Disease Report, released on September 15, 2017. Heart-related diseases increase the spending on health care and also reduce the productivity of an individual.

Estimates made by the World Health Organization (WHO), suggest that India has lost up to $237 billion, from 2005- 2015, due to heart-related or cardiovascular diseases. Thus, feasible and accurate prediction of heart-related diseases is very important.

Heart disease is common among both men and women in most countries around the world. Therefore, people should consider heart disease risk factors. Although it plays a genetic role, some lifestyle factors significantly affect heart disease .

The known risk factors for heart disease; radiation therapy for age, gender, family history, smoking, some chemotherapy drugs and cancer, malnutrition, high blood pressure, high blood cholesterol levels, diabetes, obesity, physical mobility, stress, and poor hygiene. These are the various risk factors in which the patient’s exposure towards developing a CVD.

The most common type is coronary artery disease, which can cause a heart attack. Other types of heart disease may involve the valves in the heart, or the heart may not pump well and cause heart failure. Some people are born with heart disease. Anyone, including children, can develop heart disease. It happens when a substance called plaque builds up in your arteries. Smoking, unhealthy eating and lack of exercise increase your risk of heart disease. High cholesterol, high blood pressure or diabetes can also increase your risk of heart disease.

In contrast, the removal or improvement of this factor decreases this risk. This interpretation suggests the causality between the factor and the illness, which means that the risk factor precedes the disease (the notion of anteriority).

Correction of the factor will cause the disease (the idea of reversibility) to decrease its occurrence. Of course, it must be recognized in several different populations and offer a plausible physiopathological explanation of the disease. Strictly speaking, when there is no direct causal relationship, it is a “risk marker,” a witness to a process (e.g., the elevation of micro albuminuria, elevation of C- reactive protein CRP)

.

The main heart disease risk factors such as physiological factors (age, sex, and menopausal status), lifestyle factors (smoking, physical activity, alcohol, stress), metabolic syndrome factors (insulin resistance), dyslipidemia, abdominal obesity, high blood pressure) and dietary factors. A heart disease risk factor is defined as a factor in which the patient’s exposure to this factor increases the risk of developing a CVD. In contrast, the removal or improvement of this factor decreases this risk. The risk factor’s importance is defined by the association’s strength with the disease (expressed by the relative risk observed in the exposed subjects compared to the unexposed) and the gradual association (parallel to the risk factor).

There are several types of heart disease which include:

Coronary Artery Disease (CAD)

Heart Arrhythmias.

Heart Failure.

Heart Valve Disease.

Pericardial Disease.

Cardiomyopathy (Heart Muscle Disease)

Congenital Heart Disease.

To examine the cardiac disease mischance, the particular issues which need to be discussed are those related to the behaviors. Furthermore, patients will undergo extensive examinations, such as blood pressure, glucose, vital signs, chest pain, electrocardiograms, maximum heart rate, and elevated levels of sugar, but the bright side may be that successful treatment is feasible if the disease is easily and early detected and anticipated, but treatment for all of these cardiac patients is depending on clinical studies, the patient history, and the responses to questions by the patient

To deal with this disease, there are several methods of prevention, such us natural methods, like stoping smoking, maintaining a healthy weight, adopting a healthy diet and practicing sports regularly.We also have the scientific methods such as drugs and surgeries. The prediction of this disease before being infected is part of the prevention

Methods, or the computer tools are the most used means in it, more precisely the Machine Learning algorithms. Determining the probability of having cardiac disease manually is hard to depend on as risk factors. Recently, to solve difficult issues, a range of data mining techniques and machine learning techniques are built.

Still, more advanced machine learning will assist us to identify patterns and their useful knowledge. While it has several uses in the medical field, machine learning is mainly utilized to forecast the heart disease.

Machine learning (ML) plays a significant role in disease predicting. It predicts whether the patient has a particular disease type or not based on an efficient learning technique. In making of these projects I have used several supervised learning techniques for predicting the early stage of heart disease by providing them risk factor

Here, I will use the below machine learning algorithms then I will select the best one based on its classification report.

Support Vector Machine

Random Forest

Ada Boost

Gradient Boosting

With using all these techniques the highest accuracy of 90.82% I got in GradientBoostingClassifier.

Background of the Study

Data analysis proves to be crucial in the medical field. It provides a meaningful base to critical decisions. It helps to create a complete study proposal. One of the most important uses of data analysis is that it helps in keeping human bias away from medical conclusion with the help of proper statistical treatment. By use of data mining for exploratory analysis because of nontrivial information in large volumes of data.

The health care industries collect huge amounts of data that contain some hidden information, which is useful for making effective decisions for providing appropriate results and making

Effective decisions on data, some data mining techniques are used to better the experience and conclusion that have been given.

Heart predictor system will use the data mining knowledge to give a user-oriented approach to new and hidden patterns in the data. The knowledge which is implemented can be used by the healthcare experts to get better quality of service and to reduce the extent of adverse medicine effect.

Problem Statement.

Heart disease can be managed effectively with a combination of lifestyle changes, medicine and, in some cases, surgery. With the right treatment, the symptoms of heart disease can be reduced and the functioning of the heart improved. The predicted results can be used to prevent and thus reduce cost for surgical treatment and other expensive.

The overall objective of my work will be to predict accurately with few tests and attributes the presence of heart disease. Attributes considered form the primary basis for tests and give accurate results more or less. Many more input attributes can be taken but our goal is to predict with few attributes and faster efficiency the risk of having heart disease. Decisions are often made based on doctors’ intuition and experience rather than on the knowledge rich data hidden in the data set and databases. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients.

Data mining holds great potential for the healthcare industry to enable health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and reduce costs. According to (Wurz & Takala, 2006) the opportunities to improve care and reduce costs concurrently could apply to as much as 30% of overall healthcare spending. The successful application of data mining in highly visible fields like e-business, marketing and retail has led to its application in other industries and sectors. Among these sectors just discovering is healthcare. The healthcare environment is still „information rich‟ but „knowledge poor‟. There is a wealth of data available within the healthcare systems. However, there is a lack of effective analysis tools to discover hidden relationships and trends in the data for African genres.

Objectives.

Main Objectives.

The main objective of this research is to develop a heart prediction system. The system can discover and extract hidden knowledge associated with diseases from a historical heart data set

Heart disease prediction system aims to exploit data mining techniques on medical data set to assist in the prediction of the heart diseases.

Specific Objectives.

Provides new approach to concealed patterns in the data.

Helps avoid human biasness.

To implement GradientBoosting Classifer that classifies the disease as per the input of the user.

Reduce the cost of medical tests.

Justification.

Clinical decisions are often made based on doctor’s insight and experience rather than on the knowledge rich data hidden in the dataset. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. The proposed system will integrate clinical decision support with computer-based patient records (Data Sets). This will reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge rich environment which can help to significantly improve the quality of clinical decisions.

There are voluminous records in medical data domain and because of this, it has become necessary to use data mining techniques to help in decision support and prediction in the field of healthcare. Therefore, medical data mining contributes to business intelligence which is useful for diagnosing of disease

Scope and Limitation.

Scope.

Here the scope of the project is that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modeling and

analysis tools, e.g., data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions

Limitations.

Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. The automation of the same would be highly beneficial. Clinical decisions are often made based on doctor’s intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Data mining have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions.

CHAPTER TWO: LITERATURE REVIEW

Introduction

Data mining is the process of finding previously unknown patterns and trends in databases and using that information to build predictive models. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases. The World Health Statistics 2012 report enlightens the fact that one in three adults worldwide has raised blood pressure - a condition that causes around half of all deaths from stroke and heart disease. Heart disease, also known as cardiovascular disease (CVD), encloses a number of conditions that influence the heart – not just heart attacks. Heart disease was the major cause of casualties in the different countries including India. Heart disease kills one person every 34 seconds in the United States. Coronary heart disease, Cardiomyopathy and Cardiovascular disease are some categories of heart diseases. The term “cardiovascular disease” includes a wide range of conditions that affect the heart and the blood vessels and the manner in which blood is pumped and circulated through the body. Diagnosis is complicated and important task that needs to be executed accurately and efficiently. The diagnosis is often made, based on doctor’s experience & knowledge. This leads to unwanted results & excessive medical costs of treatments provided to patients. Therefore, an automatic medical diagnosis system would be exceedingly beneficial.

Literature Review

Numerous studies have been done that have focus on diagnosis of heart disease. They have applied different data mining techniques for diagnosis & achieved different probabilities for different methods.

(Polaraju, Durga Prasad, & Tech Scholar, 2017) proposed Prediction of Heart Disease using Multiple Regression Model and it proves that Multiple Linear Regression is appropriate for predicting heart disease chance. The work is performed using training data set consists of 3000 instances with 13 different attributes which has mentioned earlier. The data set is divided into two parts that is 70% of the data are used for training and 30% used for testing.

(Deepika & Seema, 2017) focuses on techniques that can predict chronic disease by mining the data containing in historical health records using Naïve Bayes, Decision tree, Support Vector Machine (SVM) and Artificial Neural Network (ANN). A comparative study is performed on classifiers to measure the better performance on an accurate rate. From this experiment, SVM gives highest accuracy rate, whereas for diabetes Naïve Bayes gives the highest accuracy.

(Beyene & Kamat, 2018) recommended different algorithms like Naive Bayes, Classification Tree, KNN, Logistic Regression, SVM and ANN. The Logistic Regression gives better accuracy compared to other algorithms. (Beyene & Kamat, 2018) suggested Heart Disease Prediction System using Data Mining Techniques. WEKA software used for automatic diagnosis of disease and to give qualities of services in healthcare centers. The paper used various algorithms like SVM, Naïve Bayes, Association rule, KNN, ANN, and Decision Tree. The paper recommended SVM is effective and provides more accuracy as compared with other data mining algorithms.

Chala Beyene recommended Prediction and Analysis the occurrence of Heart Disease Using Data Mining Techniques. The main objective is to predict the occurrence of heart disease for early automatic diagnosis of the disease within result in short time. The proposed methodology is also critical in healthcare organization with experts that have no more knowledge and skill. It uses different medical attributes such as blood sugar and heart rate, age, sex are some of the attributes are included to identify if the person has heart disease or not. Analyses of data set are computed using WEKA software.

(Soni, Ansari, & Sharma, 2011) proposed to use non- linear classification algorithm for heart disease prediction. It is proposed to use bigdata tools such as Hadoop Distributed File System (HDFS), Map reduce along with SVM for prediction of heart disease with optimized attribute set. This work made an investigation on the use of different data mining techniques for predicting heart diseases. It suggests to use HDFS for storing large data in different nodes and executing the prediction algorithm using SVM in more than one node simultaneously using SVM. SVM is used in parallel fashion which yielded better computation time than sequential SVM.

(Science & Faculty, 2009) suggested heart disease prediction using data mining and machine learning algorithm. The goal of this study is to extract hidden patterns by applying data mining techniques. The best algorithm J48 based on UCI data has the highest accuracy rate compared to LMT. (Purushottam, Saxena, & Sharma, 2016) proposed an efficient heart disease prediction system using data mining. This system helps medical practitioner to make effective decision making based on the certain parameter. By testing and training phase a certain parameter, it provides 86.3% accuracy in testing phase and 87.3% in training phase.

(Kirmani, 2017) suggested multi disease prediction using data mining techniques. Nowadays, data mining plays vital role in predicting multiple disease. By using data mining techniques, the number of tests can be reduced. This paper mainly concentrates on predicting the heart disease, diabetes

and breast cancer etc.,

(Sai & Reddy, 2017) proposed Heart disease prediction using ANN algorithm in data mining. Due to increasing expenses of heart disease diagnosis disease, there was a need to develop new system which can predict heart disease. Prediction model is used to predict the condition of the patient after evaluation on the basis of various parameters like heart beat rate, blood pressure, cholesterol etc. The

accuracy of the system is proved in java.

(A & Naik, 2016) recommended to develop the prediction system which will diagnosis the heart disease from patient’s medical data set. 13 risk factors of input attributes have considered to build the system. After analysis of the data from the dataset, data cleaning and data integration was performed. He used k-means and naïve Bayes to predict heart disease. This paper is to build the system using historical heart database that gives diagnosis. 13 attributes have considered for building the system. To extract knowledge from database, data mining techniques such as clustering, classification methods can be used. 13 attributes with total of 300 records were used from the Cleveland Heart Database. This model is to predict whether the patient have heart disease or not based on the values of 13 attributes.

(Sultana, Haider, & Uddin, 2017) proposed an analysis of cardiovascular disease. This paper proposed data mining techniques to predict the disease. It is intended to provide the survey of current techniques to extract information from dataset and it will useful for healthcare practitioners. The performance can be obtained based on the time taken to build the decision tree for the system. The primary objective is to predict the disease with a smaller number of attributes.

Proposed Architecture.

In this system we are implementing effective heart attack prediction system using Supervised Learning techniques. We can give the input as in CSV file or manual entry to the system. After taking input the algorithms apply on that input that is GradientBoosting Classifier. After accessing data set the operation is performed and effective heart attack level is produced.

The proposed system will add some more parameters significant to heart attack with their gender, age and the priority levels are by consulting expertise doctors and the medical experts. The heart attack prediction system designed to help the identify different risk levels of heart attack like normal, low or high and also giving the prescription details with related to the predicted result.

Flowchart .

The classification tree literally creates a tree with branches, nodes, and leaves that lets us take an unknown data point and move down the tree, applying the attributes of the data point to the tree until a leaf is reached and the unknown output of the data point can be determined. In order to create a good classification tree model, we need to have an existing data set with known output from which we can build our model. We also divide our data set into two parts: a training set, which is used to create the model, and a test set, which is used to verify that the model is accurate and not over fitted.

Project Flow Chart.

This will be the proposed flow chart that the system will look like

Data Flow Diagram

Proposed Model

Chapter 3: Research Methodology

Research Design.

I will be using the experimental type of research design. It is a quantitative research method. Basically, it is a research conducted with a scientific approach, where a set of variables are kept constant while other set of variables are being measured as the subject of the experiment. This is more practically while conducting face recognition and detection as it monitors the behaviours and patterns of a subject to be used to acknowledge whether the subject matches all details presented and cross checked with previous data. It is an effect research method as it is time bound and focuses on the relationship between the variables that give actual results.

System Development Methodology.

The methodology of software development is the method in managing project development. There are many models of the methodology are available such as Waterfall model model, Incremental model, RAD model, Agile model, Iterative model and Spiral model. However, it still need to be considered by developer to decide which is will be used in the project. The methodology model is useful to manage the project efficiently and able to help developer from getting any problem during time of development.

Also, it help to achieve the objective and scope of the projects. In order to build the project, it need to understand the stakeholder requirements.

Methodology provides a framework for undertaking the proposed DM modeling. The methodology is a system comprising steps that transform raw data into recognized data patterns to extract knowledge for users.

There are four phases that involve in the spiral model:

Planning phase

Phase where the requirement are collected and risk is assessed. This phase where the title of the project has been discussed with project supervisor. From that discussion, Heart Prediction System has been proposed. The requirement and risk was assessed after doing study on existing system and do literature review about another existing research.

Risk analysis Phase

Phase where the risk and alternative solution are identified. A prototype are created at the end this phase. If there is any risk during this phase, there will be suggestion about alternate solution.

Engineering phase

At this phase, a software are created and testing are done at the end this phase.

Evaluation phase

At this phase, the user do evaluation toward the software. It will be done after the system are presented and the user do test whether the system meet with their expectation and requirement or not. If there is any error, user can tell the problem about system.

Data Collection and Preprocessing.

The data set for this research was taken from UCI data repository.14 Data accessed from the UCI Machine Learning Repository is freely available. In particular, the Cleveland and Hungarian databases have been used by many researchers and found to be suitable for developing a mining model, because of lesser missing values and outliers. The data is cleaned and preprocessed before it is submitted to the proposed algorithm for training and testing.

The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms.

The overall objective of our work is to predict more accurately the presence of heart disease. In this paper, UCI repository dataset are used to get more accurate results. Two data mining classification techniques were applied namely Decision trees and Naive Bayes

his database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no

presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).

Attributes with categorical values were converted to numerical values since most machine learning algorithms require integer values. Additionally, dummy variables were created for variables with more than two categories. Dummy variables help Neural Networks learn the data more accurately.

Classifiers Used for Experiments.

Model / Libraries Used.

1.12.1.1 FLASK:-

Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre- existing third-party libraries provide common functions. Armin Ronacher, who leads an international group of Python enthusiasts named Pocco, develops it. Flask is based on Werkzeug WSGI toolkit and Jinja2 template engine. Both are Pocco projects.

However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools.

Applications that use the Flask framework include and .

1.12.1.2 NumPy :-

NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices.NumPy was created in 2005 by Travis Oliphant. It is an open source project and you can use it freely.NumPy stands for Numerical Python. NumPy targets the

of Python, which is a non-

optimizing . written for this version of Python often run much slower than equivalents. NumPy addresses the slowness problem partly by providing multidimensional arrays and functions and operators that operate efficiently on arrays; using these requires rewriting some code, mostly , using NumPy.

1.12.1.3 OS :-

The main purpose of the OS module is to interact with your operating system. The primary use I find for it is to create folders, remove folders, move folders, and sometimes change the working directory. You can also access the names of files within a file path by doing listdir(). We do not cover that in this video, but that's an option.The os module is a part of the standard library, or stdlib, within Python 3

1.12.1.4 MATPLOTLIB :-

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK. Matplotlib is a low level graph plotting library in python that serves as a visualization utility.Matplotlib was created by John D. Hunter.Matplotlib is open source and we can use it freely.Matplotlib is mostly written in python, a few segments are written in C, Objective-C and Javascript for Platform compatibility

1.12.1.5 PICKLE :-

Python pickle module is used for serializing and de-serializing a Python object structure. Any object in Python can be pickled so that it can be saved on disk. What pickle does is that it “serializes” the object first before writing it to file. Pickling is a way to convert a python object (list, dict, etc.) into a character stream. The idea is that this character stream contains all the information necessary to reconstruct the object in another python script

Advantages of using Pickle Module:

Recursive objects (objects containing references to themselves):

Object sharing (references to the same object in different places):

User-defined classes and their instances

It is advisable not to unpickle data received from an untrusted source as they may pose security threat. However, the pickle module has no way of knowing or raise alarm while pickling malicious data

1.12.1.6 JOBLIB :-

Joblib is a set of tools to provide lightweight pipelining in Python. In particular:

transparent disk-caching of functions and lazy re-evaluation (memoize pattern)

easy simple parallel computing

Joblib is optimized to be fast and robust on large data in particular and has specific optimizations for numpy arrays

Decision Trees.

The decision tree approach is more powerful for classification problems. There are two steps in this technique building a tree & applying the tree to the dataset. There are many popular decision tree algorithms CART, ID3, C4.5, CHAID, and J48. From these J48 algorithm is used for this system. J48 algorithm uses pruning method to build a tree. Pruning is a technique that reduces size of tree by removing over fitting data, which leads to poor accuracy in predications. The J48 algorithm recursively classifies data until it has been categorized as perfectly as possible. This technique gives maximum accuracy on training data. The overall concept is to build a tree that provides balance of flexibility & accuracy.

Ensemble DM approach.

In order to have more reliable and accurate prediction results, ensemble method is a well-proven approach practiced in research for attaining highly accurate classification of data by hybridizing different classifiers. The improved prediction performance is a well-known in-built feature of ensemble methodology. This study proposes a weighted vote-based classifier ensemble technique, overcoming the limitations of conventional DM techniques by employing the ensemble of two heterogeneous classifiers: Naive Bayesian and classification via decision tree

Tools

JUPYTER NOTEBOOK:-

Project Jupiter is a project and community whose goal is to "develop , , and services for across dozens of programming languages". It was spun off from in 2014 by and Brian Granger. Project Jupyter's name is a reference to the three core programming languages supported by Jupyter, which are , and , and also a to s notebooks recording the discovery of the . Project Jupyter has developed and supported the interactive computing products Jupyter Notebook, JupyterHub, and JupyterLab. Jupyter is financially sponsored by NumFOCUS

VS CODE:

Visual Studio Code is a source-code editor that can be used with a variety of programming languages,

Including , , , , and . It is based on the framework, which is used to develop

that run on the . Visual Studio Code employs the same editor component (codenamed "Monaco") used in

(formerly called Visual Studio Online and Visual Studio Team Services).

GITHUB:-

GitHub, Inc. is a provider of for and using . It offers the and (SCM) functionality of Git, plus its own features. It provides and several collaboration features such as , requests, , and for every project. Headquartered

in , it has been a subsidiary of since 2018.

It is commonly used to host projects. As of November 2021, GitHub reports having over 73 million developers and more than 200 million (including at least 28 million public repositories). It is the largest host as of November 2021.

HEROKU :-

Heroku is a cloud (PaaS) supporting several . One of the first , Heroku has been in development since June 2007, when it supported only the programming language, but now supports , , , , , , and . For this reason, Heroku is said to be a as it has features for

a to build, run and scale applications in a similar manner across most languages. Heroku was acquired by in 2010 for $212 million.

SCIKIT-LEARN :-

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. Extensions or modules for SciPy care conventionally named . As such, the module provides learning algorithms and is named scikit-learn.

The library is built upon the SciPy (Scientific Python) that must be installed before you can use scikit-learn. This stack that includes:

NumPy: Base n-dimensional array package

SciPy: Fundamental library for scientific computing

Matplotlib: Comprehensive 2D/3D plotting

IPython: Enhanced interactive console

Sympy: Symbolic mathematics

Pandas: Data structures and analysis

Software requirements:

Hardware Requirements

For application development, the following Software Requirements are:

Processor: Intel or high RAM: 1024 MB

Space on disk: minimum 100mb For running the application:

Device: Any device that can access the internet Minimum space to execute: 20 MB

The effectiveness of the proposal is evaluated by conducting experiments with a cluster formed by 3 nodes with identical setting, configured with an Intel CORE™ i7-4770 processor (3.40GHZ, 4 Cores, 8GB RAM, running Ubuntu 18.04 LTS with 64-bit Linux 4.31.0 kernel)

Budget.

The budget of completion for developing the heart disease prediction system will require various software and hardware devices. The application is averagely expensive to build but if happens to be as successful as the developer sees it to be it will bring forth enough profit to cover the costs undergone.

The table below explains the planned budget in Kenyan Shillings to develop the system:

Work Plan

Conclusion and Future Work.

The proposed system is GUI-based, user-friendly, scalable, reliable and an expandable system. The proposed working model can also help in reducing treatment costs by providing Initial diagnostics in time. The model can also serve the purpose of training tool for medical students and will be a soft diagnostic tool available for physician and cardiologist. General physicians can utilize this tool for initial diagnosis of cardio-patients. There are many possible improvements that could be explored to improve the scalability and accuracy of this prediction system. As we have developed a generalized system, in future we can use this system for the analysis of different data sets. The performance of the health’s diagnosis can be improved significantly by handling numerous class labels in the prediction process, and it can be another positive direction of research. In DM warehouse, generally, the dimensionality of the heart database is high, so identification and selection of significant attributes for better diagnosis of heart disease are very challenging tasks for future research.

References.

A, A. S., & Naik, C. (2016). Different Data Mining Approaches for Predicting Heart Disease, 277–

281. https://doi.org/10.15680/IJIRSET.2016.0505545

Beyene, C., & Kamat, P. (2018). Survey on prediction and analysis the occurrence of heart disease using data mining techniques. International Journal of Pure and Applied Mathematics, 118(Special Issue 8), 165–173. Retrieved from https:/ 85041895038&partnerID=40&md5=2f0b0c5191a82bc0c3f0daf67d73bc81

Brownlee, J. (2016). Naive Bayes for Machine Learning. Retrieved March 4, 2019, from https://machinelearningmastery.com/naive-bayes-for-machine-learning/

Kirmani, M. (2017). Cardiovascular Disease Prediction using Data Mining Techniques. Oriental Journal of Computer Science and Technology, 10(2), 520–528. https://doi.org/10.13005/ojcst/10.02.38

Polaraju, K., Durga Prasad, D., & Tech Scholar, M. (2017). Prediction of Heart Disease using Multiple Linear Regression Model. International Journal of Engineering Development and Research, 5(4), 2321–9939. Retrieved from

Purushottam, Saxena, K., & Sharma, R. (2016). Efficient Heart Disease Prediction System. In Procedia Computer Science (Vol. 85, pp. 962–969). https://doi.org/10.1016/j.procs.2016.05.288

Sai, P. P., & Reddy, C. (2017). International Journal of Computer Science and Mobile Computing HEART DISEASE PREDICTION USING ANN ALGORITHM IN DATA MINING.

International Journal of Computer Science & Mobile Computing, 6(4), 168–172. Retrieved from

Science, C., & Faculty, G. M. (2009). Heart Disease Prediction Using Machine learning and Data Mining Technique. Ijcsc 0973-7391, 7, 1–9.

Soni, J., Ansari, U., & Sharma, D. (2011). Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers. Heart Disease, 3(6), 2385–2392.

Sultana, M., Haider, A., & Uddin, M. S. (2017). Analysis of data mining techniques for heart isease prediction. In 2016 3rd International Conference n Electrical Engineering and Information and Communication Technology, iCEEiCT 2016 (pp. 1–5).

Artificial, I. T. O. (1995). Chapter 2. Neuron, 36–62. Bozzo, R., Conca, A., & Marangon, F. (2014). Decision support system for city logistics: Literature review, and guidelines for an ex-ante model. Transportation Research Procedia, 3(July), 518–527.

Çela, E. K., & Frasheri, N. (2012a). A literature review of data mining techniques used in healthcare databases. ICT Innovations 2012 Web Proceedings, 577–582.

Çela, E. K., & Frasheri, N. (2012b). A literature review of data mining techniques used in healthcare databases. ICT Innovations 2012 Web Proceedings, 577–582.

David, H. B. F., & Belcy, S. A. (2018). Heart Disease Prediction Using Data Mining Techniques,

6956(October), 1817–1823.

Desai, S. D., Giraddi, S., Narayankar, P., Pudakalakatti, N. R., & Sulegaon, S. (2019). Back- propagation neural network versus logistic regression in heart disease classification. Advances in Intelligent Systems and Computing, 702, 133–144. Kiruthika Devi, S., Krishnapriya, S., & Kalita, D. (2016). Prediction of heart disease using data mining techniques. Indian Journal of Science and Technology, 9(39), 1291–1293.

Lavanya, M., & Gomathi, M. P. M. (2016). Prediction of Heart Disease using Classification Algorithms. International Journal of Advanced Research in Computer Engineering & Technology, 5(7), 2278–1323.

Meenakshi, K., Maragatham, G., Agarwal, N., & Ghosh, I. (2018). A Data mining Technique for Analyzing and Predicting the success of Movie. Journal of Physics: Conference Series, 1000(1).

# CHAPTER THREE - METHODOLOGY

HEART DESEASE PREDICTION SYSTEM

By

KEMBOI KEVIN

C026-01-0738/2020

A project proposal submitted for the study leading to a project report in partial fulfilment of the requirements for the award of a Bachelor of Science in Computer Science at Dedan Kimathi University.

Supervisor: PATRICK NAIVASHA

DATE

January – April 2024

Declaration

This is to certify that the work being presented in the project entitled “Heart Disease Prediction System” submitted by undersigned student of Bachelors in COMPUTER SCIENCE in the fulfillment for award of Bachelors in Computer Science is a record of my own work carried out by me under guidance and supervision of PATRICK NAIVASHA of the Department of Computer Science and that this work has not submitted elsewhere for award of any other degree.

Name of student: KEMBOI KEVIN

Registration Number: C026-01-0738/2020

Sign:

Approval

This project was done and presented by me before the panel concerned on the 5th February 2024 at Dedan Kimathi University with my approval and that of my supervisor

Supervisor Name: PATRICK NAIVASHA

Signature:

Date:

Acknowledgement

The satisfaction that accompanies that the successful completion of any task would be incomplete without the mention of people whose ceaseless cooperation made it possible, whose constant guidance and encouragement crown all efforts with success. I am very grateful to my project supervisor Patrick Naivasha, for the guidance, inspiration and constructive suggestions that helpful me in the preparation of this project. I won’t forget to also mention my course mates; Chris Sumba ,Joseph Kariuki and Chris Kinyanjui for their wonderful and skillful guidance in assisting me with the necessary support to ensure that my project is a success. I also thank my parents and family at large for their moral and financial support in funding the project to ensure successful completion of the project.

Dedication.

I dedicate this project to God Almighty my creator, my strong pillar, my source of inspiration, Wisdom, knowledge and understanding. He has been the source of my strength throughout this Program and on His wings only have I soared. I also dedicate this work to my friends; Isaac Ikinya who has encouraged me all the way and whose encouragement has made sure that I live it all it takes to finish that which I have started. To my father Kemboi David and all my beloved friends(Chris Sumba ,Joseph Kariuki and Chris Kinyanjui). Who have been affected in every way possible by this quest?

All the work done in coming up with this system is dedicated to my family for being with/part of me in the whole process especially my dear dad and mum who stood by me in all situations even at the times of financial need.

Thank You. My Love for You All Can Never Be Quantified. God Bless You.

Abstract.

As Heart acts a major role in corporeal body part. Heart diseases or Cardiovascular Diseases (CVDs) are some of the main reason for a huge number of death in the world and some of the reports said that over the last few decades heart disease or CVD has emerged as the most life- threatening disease, not only in Kenya but in the whole world.

Therefore, the diseases of heart wants more perfection and exactness for diagnose and analyse. This disease occurs due to various problems such as over pressure, blood sugar,

Chest pain,high blood pressure, Cholesterol etc. in human body . Heart is the next major organ comparing to the brain which has more priority in the Human body. It pumps the blood and supplies it to all organs of the whole body.

In the health care sector, Machine Learning plays an important role in the health care Industry. So, there is a need for a reliable, accurate, and feasible system to diagnose such diseases in time for proper treatment. Machine Learning algorithms and techniques have been applied to various medical datasets to automate the analysis of large and complex data. A system model is capable of several data processing algorithms for the classification of heart disease. Prediction of occurrences of heart diseases in the medical field is significant work. Data analytics is useful for prediction from more information and it helps the medical center to predict various diseases. The correct prediction of heart disease can prevent life threats, and incorrect prediction can prove to be fatal at the same time.

The objective of this project is to detect whether patients have any chance of heart disease or not by giving number of features to patients with having maximum accuracy of above 97%. By Using Machine learning algorithms and deep learning are applied to compare the results and analysis of the UCI Machine Learning Heart Disease dataset. The dataset consists of 14 main attributes used for performing the analysis

Table of Contents

Chapter One.

Introduction

Heart is one of the most indispensable organs in the human body. It is a organ that serves as a pump to circulate the blood. The heart is a muscular organ about the size of a fist, located just behind and slightly left of the breastbone. The heart pumps blood through the network of arteries and veins called the cardiovascular system. Oxygen is distributed through the circulatory system of the body in the blood, and if the heart does not function correctly, the entire circulatory system of the body will fail. So if the heart doesn’t work properly, it could even lead to death.

According to the World Health Organization (WHO), in the last 15 years, an estimated 17 million people die each year from cardiovascular disease, particularly heart attacks and strokes [1]. Heart disease and stroke are the biggest killers. To predict heart disease, Machine Learning can be used for identifying unseen patterns and providing some clinical insights that will assist the physicians in planning and providing care.

According to World Health Organization, heart related diseases are responsible for taking 17.7 million lives every year, 31% of all global deaths. In Kenya too, heart-related diseases have become the leading cause of mortality. Heart diseases have killed 0.3 million Kenyans in 2016, according to the 2016 Global Burden of Disease Report, released on September 15, 2017. Heart-related diseases increase the spending on health care and also reduce the productivity of an individual.

Estimates made by the World Health Organization (WHO), suggest that India has lost up to $237 billion, from 2005- 2015, due to heart-related or cardiovascular diseases. Thus, feasible and accurate prediction of heart-related diseases is very important.

Heart disease is common among both men and women in most countries around the world. Therefore, people should consider heart disease risk factors. Although it plays a genetic role, some lifestyle factors significantly affect heart disease .

The known risk factors for heart disease; radiation therapy for age, gender, family history, smoking, some chemotherapy drugs and cancer, malnutrition, high blood pressure, high blood cholesterol levels, diabetes, obesity, physical mobility, stress, and poor hygiene. These are the various risk factors in which the patient’s exposure towards developing a CVD.

The most common type is coronary artery disease, which can cause a heart attack. Other types of heart disease may involve the valves in the heart, or the heart may not pump well and cause heart failure. Some people are born with heart disease. Anyone, including children, can develop heart disease. It happens when a substance called plaque builds up in your arteries. Smoking, unhealthy eating and lack of exercise increase your risk of heart disease. High cholesterol, high blood pressure or diabetes can also increase your risk of heart disease.

In contrast, the removal or improvement of this factor decreases this risk. This interpretation suggests the causality between the factor and the illness, which means that the risk factor precedes the disease (the notion of anteriority).

Correction of the factor will cause the disease (the idea of reversibility) to decrease its occurrence. Of course, it must be recognized in several different populations and offer a plausible physiopathological explanation of the disease. Strictly speaking, when there is no direct causal relationship, it is a “risk marker,” a witness to a process (e.g., the elevation of micro albuminuria, elevation of C- reactive protein CRP)

.

The main heart disease risk factors such as physiological factors (age, sex, and menopausal status), lifestyle factors (smoking, physical activity, alcohol, stress), metabolic syndrome factors (insulin resistance), dyslipidemia, abdominal obesity, high blood pressure) and dietary factors. A heart disease risk factor is defined as a factor in which the patient’s exposure to this factor increases the risk of developing a CVD. In contrast, the removal or improvement of this factor decreases this risk. The risk factor’s importance is defined by the association’s strength with the disease (expressed by the relative risk observed in the exposed subjects compared to the unexposed) and the gradual association (parallel to the risk factor).

There are several types of heart disease which include:

Coronary Artery Disease (CAD)

Heart Arrhythmias.

Heart Failure.

Heart Valve Disease.

Pericardial Disease.

Cardiomyopathy (Heart Muscle Disease)

Congenital Heart Disease.

To examine the cardiac disease mischance, the particular issues which need to be discussed are those related to the behaviors. Furthermore, patients will undergo extensive examinations, such as blood pressure, glucose, vital signs, chest pain, electrocardiograms, maximum heart rate, and elevated levels of sugar, but the bright side may be that successful treatment is feasible if the disease is easily and early detected and anticipated, but treatment for all of these cardiac patients is depending on clinical studies, the patient history, and the responses to questions by the patient

To deal with this disease, there are several methods of prevention, such us natural methods, like stoping smoking, maintaining a healthy weight, adopting a healthy diet and practicing sports regularly.We also have the scientific methods such as drugs and surgeries. The prediction of this disease before being infected is part of the prevention

Methods, or the computer tools are the most used means in it, more precisely the Machine Learning algorithms. Determining the probability of having cardiac disease manually is hard to depend on as risk factors. Recently, to solve difficult issues, a range of data mining techniques and machine learning techniques are built.

Still, more advanced machine learning will assist us to identify patterns and their useful knowledge. While it has several uses in the medical field, machine learning is mainly utilized to forecast the heart disease.

Machine learning (ML) plays a significant role in disease predicting. It predicts whether the patient has a particular disease type or not based on an efficient learning technique. In making of these projects I have used several supervised learning techniques for predicting the early stage of heart disease by providing them risk factor

Here, I will use the below machine learning algorithms then I will select the best one based on its classification report.

Support Vector Machine

Random Forest

Ada Boost

Gradient Boosting

With using all these techniques the highest accuracy of 90.82% I got in GradientBoostingClassifier.

Background of the Study

Data analysis proves to be crucial in the medical field. It provides a meaningful base to critical decisions. It helps to create a complete study proposal. One of the most important uses of data analysis is that it helps in keeping human bias away from medical conclusion with the help of proper statistical treatment. By use of data mining for exploratory analysis because of nontrivial information in large volumes of data.

The health care industries collect huge amounts of data that contain some hidden information, which is useful for making effective decisions for providing appropriate results and making

Effective decisions on data, some data mining techniques are used to better the experience and conclusion that have been given.

Heart predictor system will use the data mining knowledge to give a user-oriented approach to new and hidden patterns in the data. The knowledge which is implemented can be used by the healthcare experts to get better quality of service and to reduce the extent of adverse medicine effect.

Problem Statement.

Heart disease can be managed effectively with a combination of lifestyle changes, medicine and, in some cases, surgery. With the right treatment, the symptoms of heart disease can be reduced and the functioning of the heart improved. The predicted results can be used to prevent and thus reduce cost for surgical treatment and other expensive.

The overall objective of my work will be to predict accurately with few tests and attributes the presence of heart disease. Attributes considered form the primary basis for tests and give accurate results more or less. Many more input attributes can be taken but our goal is to predict with few attributes and faster efficiency the risk of having heart disease. Decisions are often made based on doctors’ intuition and experience rather than on the knowledge rich data hidden in the data set and databases. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients.

Data mining holds great potential for the healthcare industry to enable health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and reduce costs. According to (Wurz & Takala, 2006) the opportunities to improve care and reduce costs concurrently could apply to as much as 30% of overall healthcare spending. The successful application of data mining in highly visible fields like e-business, marketing and retail has led to its application in other industries and sectors. Among these sectors just discovering is healthcare. The healthcare environment is still „information rich‟ but „knowledge poor‟. There is a wealth of data available within the healthcare systems. However, there is a lack of effective analysis tools to discover hidden relationships and trends in the data for African genres.

Objectives.

Main Objectives.

The main objective of this research is to develop a heart prediction system. The system can discover and extract hidden knowledge associated with diseases from a historical heart data set

Heart disease prediction system aims to exploit data mining techniques on medical data set to assist in the prediction of the heart diseases.

Specific Objectives.

Provides new approach to concealed patterns in the data.

Helps avoid human biasness.

To implement GradientBoosting Classifer that classifies the disease as per the input of the user.

Reduce the cost of medical tests.

Justification.

Clinical decisions are often made based on doctor’s insight and experience rather than on the knowledge rich data hidden in the dataset. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. The proposed system will integrate clinical decision support with computer-based patient records (Data Sets). This will reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge rich environment which can help to significantly improve the quality of clinical decisions.

There are voluminous records in medical data domain and because of this, it has become necessary to use data mining techniques to help in decision support and prediction in the field of healthcare. Therefore, medical data mining contributes to business intelligence which is useful for diagnosing of disease

Scope and Limitation.

Scope.

Here the scope of the project is that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modeling and

analysis tools, e.g., data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions

Limitations.

Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. The automation of the same would be highly beneficial. Clinical decisions are often made based on doctor’s intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Data mining have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions.

CHAPTER TWO: LITERATURE REVIEW

Introduction

Data mining is the process of finding previously unknown patterns and trends in databases and using that information to build predictive models. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases. The World Health Statistics 2012 report enlightens the fact that one in three adults worldwide has raised blood pressure - a condition that causes around half of all deaths from stroke and heart disease. Heart disease, also known as cardiovascular disease (CVD), encloses a number of conditions that influence the heart – not just heart attacks. Heart disease was the major cause of casualties in the different countries including India. Heart disease kills one person every 34 seconds in the United States. Coronary heart disease, Cardiomyopathy and Cardiovascular disease are some categories of heart diseases. The term “cardiovascular disease” includes a wide range of conditions that affect the heart and the blood vessels and the manner in which blood is pumped and circulated through the body. Diagnosis is complicated and important task that needs to be executed accurately and efficiently. The diagnosis is often made, based on doctor’s experience & knowledge. This leads to unwanted results & excessive medical costs of treatments provided to patients. Therefore, an automatic medical diagnosis system would be exceedingly beneficial.

Literature Review

Numerous studies have been done that have focus on diagnosis of heart disease. They have applied different data mining techniques for diagnosis & achieved different probabilities for different methods.

(Polaraju, Durga Prasad, & Tech Scholar, 2017) proposed Prediction of Heart Disease using Multiple Regression Model and it proves that Multiple Linear Regression is appropriate for predicting heart disease chance. The work is performed using training data set consists of 3000 instances with 13 different attributes which has mentioned earlier. The data set is divided into two parts that is 70% of the data are used for training and 30% used for testing.

(Deepika & Seema, 2017) focuses on techniques that can predict chronic disease by mining the data containing in historical health records using Naïve Bayes, Decision tree, Support Vector Machine (SVM) and Artificial Neural Network (ANN). A comparative study is performed on classifiers to measure the better performance on an accurate rate. From this experiment, SVM gives highest accuracy rate, whereas for diabetes Naïve Bayes gives the highest accuracy.

(Beyene & Kamat, 2018) recommended different algorithms like Naive Bayes, Classification Tree, KNN, Logistic Regression, SVM and ANN. The Logistic Regression gives better accuracy compared to other algorithms. (Beyene & Kamat, 2018) suggested Heart Disease Prediction System using Data Mining Techniques. WEKA software used for automatic diagnosis of disease and to give qualities of services in healthcare centers. The paper used various algorithms like SVM, Naïve Bayes, Association rule, KNN, ANN, and Decision Tree. The paper recommended SVM is effective and provides more accuracy as compared with other data mining algorithms.

Chala Beyene recommended Prediction and Analysis the occurrence of Heart Disease Using Data Mining Techniques. The main objective is to predict the occurrence of heart disease for early automatic diagnosis of the disease within result in short time. The proposed methodology is also critical in healthcare organization with experts that have no more knowledge and skill. It uses different medical attributes such as blood sugar and heart rate, age, sex are some of the attributes are included to identify if the person has heart disease or not. Analyses of data set are computed using WEKA software.

(Soni, Ansari, & Sharma, 2011) proposed to use non- linear classification algorithm for heart disease prediction. It is proposed to use bigdata tools such as Hadoop Distributed File System (HDFS), Map reduce along with SVM for prediction of heart disease with optimized attribute set. This work made an investigation on the use of different data mining techniques for predicting heart diseases. It suggests to use HDFS for storing large data in different nodes and executing the prediction algorithm using SVM in more than one node simultaneously using SVM. SVM is used in parallel fashion which yielded better computation time than sequential SVM.

(Science & Faculty, 2009) suggested heart disease prediction using data mining and machine learning algorithm. The goal of this study is to extract hidden patterns by applying data mining techniques. The best algorithm J48 based on UCI data has the highest accuracy rate compared to LMT. (Purushottam, Saxena, & Sharma, 2016) proposed an efficient heart disease prediction system using data mining. This system helps medical practitioner to make effective decision making based on the certain parameter. By testing and training phase a certain parameter, it provides 86.3% accuracy in testing phase and 87.3% in training phase.

(Kirmani, 2017) suggested multi disease prediction using data mining techniques. Nowadays, data mining plays vital role in predicting multiple disease. By using data mining techniques, the number of tests can be reduced. This paper mainly concentrates on predicting the heart disease, diabetes

and breast cancer etc.,

(Sai & Reddy, 2017) proposed Heart disease prediction using ANN algorithm in data mining. Due to increasing expenses of heart disease diagnosis disease, there was a need to develop new system which can predict heart disease. Prediction model is used to predict the condition of the patient after evaluation on the basis of various parameters like heart beat rate, blood pressure, cholesterol etc. The

accuracy of the system is proved in java.

(A & Naik, 2016) recommended to develop the prediction system which will diagnosis the heart disease from patient’s medical data set. 13 risk factors of input attributes have considered to build the system. After analysis of the data from the dataset, data cleaning and data integration was performed. He used k-means and naïve Bayes to predict heart disease. This paper is to build the system using historical heart database that gives diagnosis. 13 attributes have considered for building the system. To extract knowledge from database, data mining techniques such as clustering, classification methods can be used. 13 attributes with total of 300 records were used from the Cleveland Heart Database. This model is to predict whether the patient have heart disease or not based on the values of 13 attributes.

(Sultana, Haider, & Uddin, 2017) proposed an analysis of cardiovascular disease. This paper proposed data mining techniques to predict the disease. It is intended to provide the survey of current techniques to extract information from dataset and it will useful for healthcare practitioners. The performance can be obtained based on the time taken to build the decision tree for the system. The primary objective is to predict the disease with a smaller number of attributes.

Proposed Architecture.

In this system we are implementing effective heart attack prediction system using Supervised Learning techniques. We can give the input as in CSV file or manual entry to the system. After taking input the algorithms apply on that input that is GradientBoosting Classifier. After accessing data set the operation is performed and effective heart attack level is produced.

The proposed system will add some more parameters significant to heart attack with their gender, age and the priority levels are by consulting expertise doctors and the medical experts. The heart attack prediction system designed to help the identify different risk levels of heart attack like normal, low or high and also giving the prescription details with related to the predicted result.

Flowchart .

The classification tree literally creates a tree with branches, nodes, and leaves that lets us take an unknown data point and move down the tree, applying the attributes of the data point to the tree until a leaf is reached and the unknown output of the data point can be determined. In order to create a good classification tree model, we need to have an existing data set with known output from which we can build our model. We also divide our data set into two parts: a training set, which is used to create the model, and a test set, which is used to verify that the model is accurate and not over fitted.

Project Flow Chart.

This will be the proposed flow chart that the system will look like

Data Flow Diagram

Proposed Model

Chapter 3: Research Methodology

Research Design.

I will be using the experimental type of research design. It is a quantitative research method. Basically, it is a research conducted with a scientific approach, where a set of variables are kept constant while other set of variables are being measured as the subject of the experiment. This is more practically while conducting face recognition and detection as it monitors the behaviours and patterns of a subject to be used to acknowledge whether the subject matches all details presented and cross checked with previous data. It is an effect research method as it is time bound and focuses on the relationship between the variables that give actual results.

System Development Methodology.

The methodology of software development is the method in managing project development. There are many models of the methodology are available such as Waterfall model model, Incremental model, RAD model, Agile model, Iterative model and Spiral model. However, it still need to be considered by developer to decide which is will be used in the project. The methodology model is useful to manage the project efficiently and able to help developer from getting any problem during time of development.

Also, it help to achieve the objective and scope of the projects. In order to build the project, it need to understand the stakeholder requirements.

Methodology provides a framework for undertaking the proposed DM modeling. The methodology is a system comprising steps that transform raw data into recognized data patterns to extract knowledge for users.

There are four phases that involve in the spiral model:

Planning phase

Phase where the requirement are collected and risk is assessed. This phase where the title of the project has been discussed with project supervisor. From that discussion, Heart Prediction System has been proposed. The requirement and risk was assessed after doing study on existing system and do literature review about another existing research.

Risk analysis Phase

Phase where the risk and alternative solution are identified. A prototype are created at the end this phase. If there is any risk during this phase, there will be suggestion about alternate solution.

Engineering phase

At this phase, a software are created and testing are done at the end this phase.

Evaluation phase

At this phase, the user do evaluation toward the software. It will be done after the system are presented and the user do test whether the system meet with their expectation and requirement or not. If there is any error, user can tell the problem about system.

Data Collection and Preprocessing.

The data set for this research was taken from UCI data repository.14 Data accessed from the UCI Machine Learning Repository is freely available. In particular, the Cleveland and Hungarian databases have been used by many researchers and found to be suitable for developing a mining model, because of lesser missing values and outliers. The data is cleaned and preprocessed before it is submitted to the proposed algorithm for training and testing.

The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms.

The overall objective of our work is to predict more accurately the presence of heart disease. In this paper, UCI repository dataset are used to get more accurate results. Two data mining classification techniques were applied namely Decision trees and Naive Bayes

his database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no

presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).

Attributes with categorical values were converted to numerical values since most machine learning algorithms require integer values. Additionally, dummy variables were created for variables with more than two categories. Dummy variables help Neural Networks learn the data more accurately.

Classifiers Used for Experiments.

Model / Libraries Used.

1.12.1.1 FLASK:-

Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre- existing third-party libraries provide common functions. Armin Ronacher, who leads an international group of Python enthusiasts named Pocco, develops it. Flask is based on Werkzeug WSGI toolkit and Jinja2 template engine. Both are Pocco projects.

However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools.

Applications that use the Flask framework include and .

1.12.1.2 NumPy :-

NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices.NumPy was created in 2005 by Travis Oliphant. It is an open source project and you can use it freely.NumPy stands for Numerical Python. NumPy targets the

of Python, which is a non-

optimizing . written for this version of Python often run much slower than equivalents. NumPy addresses the slowness problem partly by providing multidimensional arrays and functions and operators that operate efficiently on arrays; using these requires rewriting some code, mostly , using NumPy.

1.12.1.3 OS :-

The main purpose of the OS module is to interact with your operating system. The primary use I find for it is to create folders, remove folders, move folders, and sometimes change the working directory. You can also access the names of files within a file path by doing listdir(). We do not cover that in this video, but that's an option.The os module is a part of the standard library, or stdlib, within Python 3

1.12.1.4 MATPLOTLIB :-

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK. Matplotlib is a low level graph plotting library in python that serves as a visualization utility.Matplotlib was created by John D. Hunter.Matplotlib is open source and we can use it freely.Matplotlib is mostly written in python, a few segments are written in C, Objective-C and Javascript for Platform compatibility

1.12.1.5 PICKLE :-

Python pickle module is used for serializing and de-serializing a Python object structure. Any object in Python can be pickled so that it can be saved on disk. What pickle does is that it “serializes” the object first before writing it to file. Pickling is a way to convert a python object (list, dict, etc.) into a character stream. The idea is that this character stream contains all the information necessary to reconstruct the object in another python script

Advantages of using Pickle Module:

Recursive objects (objects containing references to themselves):

Object sharing (references to the same object in different places):

User-defined classes and their instances

It is advisable not to unpickle data received from an untrusted source as they may pose security threat. However, the pickle module has no way of knowing or raise alarm while pickling malicious data

1.12.1.6 JOBLIB :-

Joblib is a set of tools to provide lightweight pipelining in Python. In particular:

transparent disk-caching of functions and lazy re-evaluation (memoize pattern)

easy simple parallel computing

Joblib is optimized to be fast and robust on large data in particular and has specific optimizations for numpy arrays

Decision Trees.

The decision tree approach is more powerful for classification problems. There are two steps in this technique building a tree & applying the tree to the dataset. There are many popular decision tree algorithms CART, ID3, C4.5, CHAID, and J48. From these J48 algorithm is used for this system. J48 algorithm uses pruning method to build a tree. Pruning is a technique that reduces size of tree by removing over fitting data, which leads to poor accuracy in predications. The J48 algorithm recursively classifies data until it has been categorized as perfectly as possible. This technique gives maximum accuracy on training data. The overall concept is to build a tree that provides balance of flexibility & accuracy.

Ensemble DM approach.

In order to have more reliable and accurate prediction results, ensemble method is a well-proven approach practiced in research for attaining highly accurate classification of data by hybridizing different classifiers. The improved prediction performance is a well-known in-built feature of ensemble methodology. This study proposes a weighted vote-based classifier ensemble technique, overcoming the limitations of conventional DM techniques by employing the ensemble of two heterogeneous classifiers: Naive Bayesian and classification via decision tree

Tools

JUPYTER NOTEBOOK:-

Project Jupiter is a project and community whose goal is to "develop , , and services for across dozens of programming languages". It was spun off from in 2014 by and Brian Granger. Project Jupyter's name is a reference to the three core programming languages supported by Jupyter, which are , and , and also a to s notebooks recording the discovery of the . Project Jupyter has developed and supported the interactive computing products Jupyter Notebook, JupyterHub, and JupyterLab. Jupyter is financially sponsored by NumFOCUS

VS CODE:

Visual Studio Code is a source-code editor that can be used with a variety of programming languages,

Including , , , , and . It is based on the framework, which is used to develop

that run on the . Visual Studio Code employs the same editor component (codenamed "Monaco") used in

(formerly called Visual Studio Online and Visual Studio Team Services).

GITHUB:-

GitHub, Inc. is a provider of for and using . It offers the and (SCM) functionality of Git, plus its own features. It provides and several collaboration features such as , requests, , and for every project. Headquartered

in , it has been a subsidiary of since 2018.

It is commonly used to host projects. As of November 2021, GitHub reports having over 73 million developers and more than 200 million (including at least 28 million public repositories). It is the largest host as of November 2021.

HEROKU :-

Heroku is a cloud (PaaS) supporting several . One of the first , Heroku has been in development since June 2007, when it supported only the programming language, but now supports , , , , , , and . For this reason, Heroku is said to be a as it has features for

a to build, run and scale applications in a similar manner across most languages. Heroku was acquired by in 2010 for $212 million.

SCIKIT-LEARN :-

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. Extensions or modules for SciPy care conventionally named . As such, the module provides learning algorithms and is named scikit-learn.

The library is built upon the SciPy (Scientific Python) that must be installed before you can use scikit-learn. This stack that includes:

NumPy: Base n-dimensional array package

SciPy: Fundamental library for scientific computing

Matplotlib: Comprehensive 2D/3D plotting

IPython: Enhanced interactive console

Sympy: Symbolic mathematics

Pandas: Data structures and analysis

Software requirements:

Hardware Requirements

For application development, the following Software Requirements are:

Processor: Intel or high RAM: 1024 MB

Space on disk: minimum 100mb For running the application:

Device: Any device that can access the internet Minimum space to execute: 20 MB

The effectiveness of the proposal is evaluated by conducting experiments with a cluster formed by 3 nodes with identical setting, configured with an Intel CORE™ i7-4770 processor (3.40GHZ, 4 Cores, 8GB RAM, running Ubuntu 18.04 LTS with 64-bit Linux 4.31.0 kernel)

Budget.

The budget of completion for developing the heart disease prediction system will require various software and hardware devices. The application is averagely expensive to build but if happens to be as successful as the developer sees it to be it will bring forth enough profit to cover the costs undergone.

The table below explains the planned budget in Kenyan Shillings to develop the system:

Work Plan

Conclusion and Future Work.

The proposed system is GUI-based, user-friendly, scalable, reliable and an expandable system. The proposed working model can also help in reducing treatment costs by providing Initial diagnostics in time. The model can also serve the purpose of training tool for medical students and will be a soft diagnostic tool available for physician and cardiologist. General physicians can utilize this tool for initial diagnosis of cardio-patients. There are many possible improvements that could be explored to improve the scalability and accuracy of this prediction system. As we have developed a generalized system, in future we can use this system for the analysis of different data sets. The performance of the health’s diagnosis can be improved significantly by handling numerous class labels in the prediction process, and it can be another positive direction of research. In DM warehouse, generally, the dimensionality of the heart database is high, so identification and selection of significant attributes for better diagnosis of heart disease are very challenging tasks for future research.

References.

A, A. S., & Naik, C. (2016). Different Data Mining Approaches for Predicting Heart Disease, 277–

281. https://doi.org/10.15680/IJIRSET.2016.0505545

Beyene, C., & Kamat, P. (2018). Survey on prediction and analysis the occurrence of heart disease using data mining techniques. International Journal of Pure and Applied Mathematics, 118(Special Issue 8), 165–173. Retrieved from https:/ 85041895038&partnerID=40&md5=2f0b0c5191a82bc0c3f0daf67d73bc81

Brownlee, J. (2016). Naive Bayes for Machine Learning. Retrieved March 4, 2019, from https://machinelearningmastery.com/naive-bayes-for-machine-learning/

Kirmani, M. (2017). Cardiovascular Disease Prediction using Data Mining Techniques. Oriental Journal of Computer Science and Technology, 10(2), 520–528. https://doi.org/10.13005/ojcst/10.02.38

Polaraju, K., Durga Prasad, D., & Tech Scholar, M. (2017). Prediction of Heart Disease using Multiple Linear Regression Model. International Journal of Engineering Development and Research, 5(4), 2321–9939. Retrieved from

Purushottam, Saxena, K., & Sharma, R. (2016). Efficient Heart Disease Prediction System. In Procedia Computer Science (Vol. 85, pp. 962–969). https://doi.org/10.1016/j.procs.2016.05.288

Sai, P. P., & Reddy, C. (2017). International Journal of Computer Science and Mobile Computing HEART DISEASE PREDICTION USING ANN ALGORITHM IN DATA MINING.

International Journal of Computer Science & Mobile Computing, 6(4), 168–172. Retrieved from

Science, C., & Faculty, G. M. (2009). Heart Disease Prediction Using Machine learning and Data Mining Technique. Ijcsc 0973-7391, 7, 1–9.

Soni, J., Ansari, U., & Sharma, D. (2011). Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers. Heart Disease, 3(6), 2385–2392.

Sultana, M., Haider, A., & Uddin, M. S. (2017). Analysis of data mining techniques for heart isease prediction. In 2016 3rd International Conference n Electrical Engineering and Information and Communication Technology, iCEEiCT 2016 (pp. 1–5).

Artificial, I. T. O. (1995). Chapter 2. Neuron, 36–62. Bozzo, R., Conca, A., & Marangon, F. (2014). Decision support system for city logistics: Literature review, and guidelines for an ex-ante model. Transportation Research Procedia, 3(July), 518–527.

Çela, E. K., & Frasheri, N. (2012a). A literature review of data mining techniques used in healthcare databases. ICT Innovations 2012 Web Proceedings, 577–582.

Çela, E. K., & Frasheri, N. (2012b). A literature review of data mining techniques used in healthcare databases. ICT Innovations 2012 Web Proceedings, 577–582.

David, H. B. F., & Belcy, S. A. (2018). Heart Disease Prediction Using Data Mining Techniques,

6956(October), 1817–1823.

Desai, S. D., Giraddi, S., Narayankar, P., Pudakalakatti, N. R., & Sulegaon, S. (2019). Back- propagation neural network versus logistic regression in heart disease classification. Advances in Intelligent Systems and Computing, 702, 133–144. Kiruthika Devi, S., Krishnapriya, S., & Kalita, D. (2016). Prediction of heart disease using data mining techniques. Indian Journal of Science and Technology, 9(39), 1291–1293.

Lavanya, M., & Gomathi, M. P. M. (2016). Prediction of Heart Disease using Classification Algorithms. International Journal of Advanced Research in Computer Engineering & Technology, 5(7), 2278–1323.

Meenakshi, K., Maragatham, G., Agarwal, N., & Ghosh, I. (2018). A Data mining Technique for Analyzing and Predicting the success of Movie. Journal of Physics: Conference Series, 1000(1).

# CHAPTER FOUR - ANALYSIS AND DESIGN

HEART DESEASE PREDICTION SYSTEM

By

KEMBOI KEVIN

C026-01-0738/2020

A project proposal submitted for the study leading to a project report in partial fulfilment of the requirements for the award of a Bachelor of Science in Computer Science at Dedan Kimathi University.

Supervisor: PATRICK NAIVASHA

DATE

January – April 2024

Declaration

This is to certify that the work being presented in the project entitled “Heart Disease Prediction System” submitted by undersigned student of Bachelors in COMPUTER SCIENCE in the fulfillment for award of Bachelors in Computer Science is a record of my own work carried out by me under guidance and supervision of PATRICK NAIVASHA of the Department of Computer Science and that this work has not submitted elsewhere for award of any other degree.

Name of student: KEMBOI KEVIN

Registration Number: C026-01-0738/2020

Sign:

Approval

This project was done and presented by me before the panel concerned on the 5th February 2024 at Dedan Kimathi University with my approval and that of my supervisor

Supervisor Name: PATRICK NAIVASHA

Signature:

Date:

Acknowledgement

The satisfaction that accompanies that the successful completion of any task would be incomplete without the mention of people whose ceaseless cooperation made it possible, whose constant guidance and encouragement crown all efforts with success. I am very grateful to my project supervisor Patrick Naivasha, for the guidance, inspiration and constructive suggestions that helpful me in the preparation of this project. I won’t forget to also mention my course mates; Chris Sumba ,Joseph Kariuki and Chris Kinyanjui for their wonderful and skillful guidance in assisting me with the necessary support to ensure that my project is a success. I also thank my parents and family at large for their moral and financial support in funding the project to ensure successful completion of the project.

Dedication.

I dedicate this project to God Almighty my creator, my strong pillar, my source of inspiration, Wisdom, knowledge and understanding. He has been the source of my strength throughout this Program and on His wings only have I soared. I also dedicate this work to my friends; Isaac Ikinya who has encouraged me all the way and whose encouragement has made sure that I live it all it takes to finish that which I have started. To my father Kemboi David and all my beloved friends(Chris Sumba ,Joseph Kariuki and Chris Kinyanjui). Who have been affected in every way possible by this quest?

All the work done in coming up with this system is dedicated to my family for being with/part of me in the whole process especially my dear dad and mum who stood by me in all situations even at the times of financial need.

Thank You. My Love for You All Can Never Be Quantified. God Bless You.

Abstract.

As Heart acts a major role in corporeal body part. Heart diseases or Cardiovascular Diseases (CVDs) are some of the main reason for a huge number of death in the world and some of the reports said that over the last few decades heart disease or CVD has emerged as the most life- threatening disease, not only in Kenya but in the whole world.

Therefore, the diseases of heart wants more perfection and exactness for diagnose and analyse. This disease occurs due to various problems such as over pressure, blood sugar,

Chest pain,high blood pressure, Cholesterol etc. in human body . Heart is the next major organ comparing to the brain which has more priority in the Human body. It pumps the blood and supplies it to all organs of the whole body.

In the health care sector, Machine Learning plays an important role in the health care Industry. So, there is a need for a reliable, accurate, and feasible system to diagnose such diseases in time for proper treatment. Machine Learning algorithms and techniques have been applied to various medical datasets to automate the analysis of large and complex data. A system model is capable of several data processing algorithms for the classification of heart disease. Prediction of occurrences of heart diseases in the medical field is significant work. Data analytics is useful for prediction from more information and it helps the medical center to predict various diseases. The correct prediction of heart disease can prevent life threats, and incorrect prediction can prove to be fatal at the same time.

The objective of this project is to detect whether patients have any chance of heart disease or not by giving number of features to patients with having maximum accuracy of above 97%. By Using Machine learning algorithms and deep learning are applied to compare the results and analysis of the UCI Machine Learning Heart Disease dataset. The dataset consists of 14 main attributes used for performing the analysis

Table of Contents

Chapter One.

Introduction

Heart is one of the most indispensable organs in the human body. It is a organ that serves as a pump to circulate the blood. The heart is a muscular organ about the size of a fist, located just behind and slightly left of the breastbone. The heart pumps blood through the network of arteries and veins called the cardiovascular system. Oxygen is distributed through the circulatory system of the body in the blood, and if the heart does not function correctly, the entire circulatory system of the body will fail. So if the heart doesn’t work properly, it could even lead to death.

According to the World Health Organization (WHO), in the last 15 years, an estimated 17 million people die each year from cardiovascular disease, particularly heart attacks and strokes [1]. Heart disease and stroke are the biggest killers. To predict heart disease, Machine Learning can be used for identifying unseen patterns and providing some clinical insights that will assist the physicians in planning and providing care.

According to World Health Organization, heart related diseases are responsible for taking 17.7 million lives every year, 31% of all global deaths. In Kenya too, heart-related diseases have become the leading cause of mortality. Heart diseases have killed 0.3 million Kenyans in 2016, according to the 2016 Global Burden of Disease Report, released on September 15, 2017. Heart-related diseases increase the spending on health care and also reduce the productivity of an individual.

Estimates made by the World Health Organization (WHO), suggest that India has lost up to $237 billion, from 2005- 2015, due to heart-related or cardiovascular diseases. Thus, feasible and accurate prediction of heart-related diseases is very important.

Heart disease is common among both men and women in most countries around the world. Therefore, people should consider heart disease risk factors. Although it plays a genetic role, some lifestyle factors significantly affect heart disease .

The known risk factors for heart disease; radiation therapy for age, gender, family history, smoking, some chemotherapy drugs and cancer, malnutrition, high blood pressure, high blood cholesterol levels, diabetes, obesity, physical mobility, stress, and poor hygiene. These are the various risk factors in which the patient’s exposure towards developing a CVD.

The most common type is coronary artery disease, which can cause a heart attack. Other types of heart disease may involve the valves in the heart, or the heart may not pump well and cause heart failure. Some people are born with heart disease. Anyone, including children, can develop heart disease. It happens when a substance called plaque builds up in your arteries. Smoking, unhealthy eating and lack of exercise increase your risk of heart disease. High cholesterol, high blood pressure or diabetes can also increase your risk of heart disease.

In contrast, the removal or improvement of this factor decreases this risk. This interpretation suggests the causality between the factor and the illness, which means that the risk factor precedes the disease (the notion of anteriority).

Correction of the factor will cause the disease (the idea of reversibility) to decrease its occurrence. Of course, it must be recognized in several different populations and offer a plausible physiopathological explanation of the disease. Strictly speaking, when there is no direct causal relationship, it is a “risk marker,” a witness to a process (e.g., the elevation of micro albuminuria, elevation of C- reactive protein CRP)

.

The main heart disease risk factors such as physiological factors (age, sex, and menopausal status), lifestyle factors (smoking, physical activity, alcohol, stress), metabolic syndrome factors (insulin resistance), dyslipidemia, abdominal obesity, high blood pressure) and dietary factors. A heart disease risk factor is defined as a factor in which the patient’s exposure to this factor increases the risk of developing a CVD. In contrast, the removal or improvement of this factor decreases this risk. The risk factor’s importance is defined by the association’s strength with the disease (expressed by the relative risk observed in the exposed subjects compared to the unexposed) and the gradual association (parallel to the risk factor).

There are several types of heart disease which include:

Coronary Artery Disease (CAD)

Heart Arrhythmias.

Heart Failure.

Heart Valve Disease.

Pericardial Disease.

Cardiomyopathy (Heart Muscle Disease)

Congenital Heart Disease.

To examine the cardiac disease mischance, the particular issues which need to be discussed are those related to the behaviors. Furthermore, patients will undergo extensive examinations, such as blood pressure, glucose, vital signs, chest pain, electrocardiograms, maximum heart rate, and elevated levels of sugar, but the bright side may be that successful treatment is feasible if the disease is easily and early detected and anticipated, but treatment for all of these cardiac patients is depending on clinical studies, the patient history, and the responses to questions by the patient

To deal with this disease, there are several methods of prevention, such us natural methods, like stoping smoking, maintaining a healthy weight, adopting a healthy diet and practicing sports regularly.We also have the scientific methods such as drugs and surgeries. The prediction of this disease before being infected is part of the prevention

Methods, or the computer tools are the most used means in it, more precisely the Machine Learning algorithms. Determining the probability of having cardiac disease manually is hard to depend on as risk factors. Recently, to solve difficult issues, a range of data mining techniques and machine learning techniques are built.

Still, more advanced machine learning will assist us to identify patterns and their useful knowledge. While it has several uses in the medical field, machine learning is mainly utilized to forecast the heart disease.

Machine learning (ML) plays a significant role in disease predicting. It predicts whether the patient has a particular disease type or not based on an efficient learning technique. In making of these projects I have used several supervised learning techniques for predicting the early stage of heart disease by providing them risk factor

Here, I will use the below machine learning algorithms then I will select the best one based on its classification report.

Support Vector Machine

Random Forest

Ada Boost

Gradient Boosting

With using all these techniques the highest accuracy of 90.82% I got in GradientBoostingClassifier.

Background of the Study

Data analysis proves to be crucial in the medical field. It provides a meaningful base to critical decisions. It helps to create a complete study proposal. One of the most important uses of data analysis is that it helps in keeping human bias away from medical conclusion with the help of proper statistical treatment. By use of data mining for exploratory analysis because of nontrivial information in large volumes of data.

The health care industries collect huge amounts of data that contain some hidden information, which is useful for making effective decisions for providing appropriate results and making

Effective decisions on data, some data mining techniques are used to better the experience and conclusion that have been given.

Heart predictor system will use the data mining knowledge to give a user-oriented approach to new and hidden patterns in the data. The knowledge which is implemented can be used by the healthcare experts to get better quality of service and to reduce the extent of adverse medicine effect.

Problem Statement.

Heart disease can be managed effectively with a combination of lifestyle changes, medicine and, in some cases, surgery. With the right treatment, the symptoms of heart disease can be reduced and the functioning of the heart improved. The predicted results can be used to prevent and thus reduce cost for surgical treatment and other expensive.

The overall objective of my work will be to predict accurately with few tests and attributes the presence of heart disease. Attributes considered form the primary basis for tests and give accurate results more or less. Many more input attributes can be taken but our goal is to predict with few attributes and faster efficiency the risk of having heart disease. Decisions are often made based on doctors’ intuition and experience rather than on the knowledge rich data hidden in the data set and databases. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients.

Data mining holds great potential for the healthcare industry to enable health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and reduce costs. According to (Wurz & Takala, 2006) the opportunities to improve care and reduce costs concurrently could apply to as much as 30% of overall healthcare spending. The successful application of data mining in highly visible fields like e-business, marketing and retail has led to its application in other industries and sectors. Among these sectors just discovering is healthcare. The healthcare environment is still „information rich‟ but „knowledge poor‟. There is a wealth of data available within the healthcare systems. However, there is a lack of effective analysis tools to discover hidden relationships and trends in the data for African genres.

Objectives.

Main Objectives.

The main objective of this research is to develop a heart prediction system. The system can discover and extract hidden knowledge associated with diseases from a historical heart data set

Heart disease prediction system aims to exploit data mining techniques on medical data set to assist in the prediction of the heart diseases.

Specific Objectives.

Provides new approach to concealed patterns in the data.

Helps avoid human biasness.

To implement GradientBoosting Classifer that classifies the disease as per the input of the user.

Reduce the cost of medical tests.

Justification.

Clinical decisions are often made based on doctor’s insight and experience rather than on the knowledge rich data hidden in the dataset. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. The proposed system will integrate clinical decision support with computer-based patient records (Data Sets). This will reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge rich environment which can help to significantly improve the quality of clinical decisions.

There are voluminous records in medical data domain and because of this, it has become necessary to use data mining techniques to help in decision support and prediction in the field of healthcare. Therefore, medical data mining contributes to business intelligence which is useful for diagnosing of disease

Scope and Limitation.

Scope.

Here the scope of the project is that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modeling and

analysis tools, e.g., data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions

Limitations.

Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. The automation of the same would be highly beneficial. Clinical decisions are often made based on doctor’s intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Data mining have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions.

CHAPTER TWO: LITERATURE REVIEW

Introduction

Data mining is the process of finding previously unknown patterns and trends in databases and using that information to build predictive models. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases. The World Health Statistics 2012 report enlightens the fact that one in three adults worldwide has raised blood pressure - a condition that causes around half of all deaths from stroke and heart disease. Heart disease, also known as cardiovascular disease (CVD), encloses a number of conditions that influence the heart – not just heart attacks. Heart disease was the major cause of casualties in the different countries including India. Heart disease kills one person every 34 seconds in the United States. Coronary heart disease, Cardiomyopathy and Cardiovascular disease are some categories of heart diseases. The term “cardiovascular disease” includes a wide range of conditions that affect the heart and the blood vessels and the manner in which blood is pumped and circulated through the body. Diagnosis is complicated and important task that needs to be executed accurately and efficiently. The diagnosis is often made, based on doctor’s experience & knowledge. This leads to unwanted results & excessive medical costs of treatments provided to patients. Therefore, an automatic medical diagnosis system would be exceedingly beneficial.

Literature Review

Numerous studies have been done that have focus on diagnosis of heart disease. They have applied different data mining techniques for diagnosis & achieved different probabilities for different methods.

(Polaraju, Durga Prasad, & Tech Scholar, 2017) proposed Prediction of Heart Disease using Multiple Regression Model and it proves that Multiple Linear Regression is appropriate for predicting heart disease chance. The work is performed using training data set consists of 3000 instances with 13 different attributes which has mentioned earlier. The data set is divided into two parts that is 70% of the data are used for training and 30% used for testing.

(Deepika & Seema, 2017) focuses on techniques that can predict chronic disease by mining the data containing in historical health records using Naïve Bayes, Decision tree, Support Vector Machine (SVM) and Artificial Neural Network (ANN). A comparative study is performed on classifiers to measure the better performance on an accurate rate. From this experiment, SVM gives highest accuracy rate, whereas for diabetes Naïve Bayes gives the highest accuracy.

(Beyene & Kamat, 2018) recommended different algorithms like Naive Bayes, Classification Tree, KNN, Logistic Regression, SVM and ANN. The Logistic Regression gives better accuracy compared to other algorithms. (Beyene & Kamat, 2018) suggested Heart Disease Prediction System using Data Mining Techniques. WEKA software used for automatic diagnosis of disease and to give qualities of services in healthcare centers. The paper used various algorithms like SVM, Naïve Bayes, Association rule, KNN, ANN, and Decision Tree. The paper recommended SVM is effective and provides more accuracy as compared with other data mining algorithms.

Chala Beyene recommended Prediction and Analysis the occurrence of Heart Disease Using Data Mining Techniques. The main objective is to predict the occurrence of heart disease for early automatic diagnosis of the disease within result in short time. The proposed methodology is also critical in healthcare organization with experts that have no more knowledge and skill. It uses different medical attributes such as blood sugar and heart rate, age, sex are some of the attributes are included to identify if the person has heart disease or not. Analyses of data set are computed using WEKA software.

(Soni, Ansari, & Sharma, 2011) proposed to use non- linear classification algorithm for heart disease prediction. It is proposed to use bigdata tools such as Hadoop Distributed File System (HDFS), Map reduce along with SVM for prediction of heart disease with optimized attribute set. This work made an investigation on the use of different data mining techniques for predicting heart diseases. It suggests to use HDFS for storing large data in different nodes and executing the prediction algorithm using SVM in more than one node simultaneously using SVM. SVM is used in parallel fashion which yielded better computation time than sequential SVM.

(Science & Faculty, 2009) suggested heart disease prediction using data mining and machine learning algorithm. The goal of this study is to extract hidden patterns by applying data mining techniques. The best algorithm J48 based on UCI data has the highest accuracy rate compared to LMT. (Purushottam, Saxena, & Sharma, 2016) proposed an efficient heart disease prediction system using data mining. This system helps medical practitioner to make effective decision making based on the certain parameter. By testing and training phase a certain parameter, it provides 86.3% accuracy in testing phase and 87.3% in training phase.

(Kirmani, 2017) suggested multi disease prediction using data mining techniques. Nowadays, data mining plays vital role in predicting multiple disease. By using data mining techniques, the number of tests can be reduced. This paper mainly concentrates on predicting the heart disease, diabetes

and breast cancer etc.,

(Sai & Reddy, 2017) proposed Heart disease prediction using ANN algorithm in data mining. Due to increasing expenses of heart disease diagnosis disease, there was a need to develop new system which can predict heart disease. Prediction model is used to predict the condition of the patient after evaluation on the basis of various parameters like heart beat rate, blood pressure, cholesterol etc. The

accuracy of the system is proved in java.

(A & Naik, 2016) recommended to develop the prediction system which will diagnosis the heart disease from patient’s medical data set. 13 risk factors of input attributes have considered to build the system. After analysis of the data from the dataset, data cleaning and data integration was performed. He used k-means and naïve Bayes to predict heart disease. This paper is to build the system using historical heart database that gives diagnosis. 13 attributes have considered for building the system. To extract knowledge from database, data mining techniques such as clustering, classification methods can be used. 13 attributes with total of 300 records were used from the Cleveland Heart Database. This model is to predict whether the patient have heart disease or not based on the values of 13 attributes.

(Sultana, Haider, & Uddin, 2017) proposed an analysis of cardiovascular disease. This paper proposed data mining techniques to predict the disease. It is intended to provide the survey of current techniques to extract information from dataset and it will useful for healthcare practitioners. The performance can be obtained based on the time taken to build the decision tree for the system. The primary objective is to predict the disease with a smaller number of attributes.

Proposed Architecture.

In this system we are implementing effective heart attack prediction system using Supervised Learning techniques. We can give the input as in CSV file or manual entry to the system. After taking input the algorithms apply on that input that is GradientBoosting Classifier. After accessing data set the operation is performed and effective heart attack level is produced.

The proposed system will add some more parameters significant to heart attack with their gender, age and the priority levels are by consulting expertise doctors and the medical experts. The heart attack prediction system designed to help the identify different risk levels of heart attack like normal, low or high and also giving the prescription details with related to the predicted result.

Flowchart .

The classification tree literally creates a tree with branches, nodes, and leaves that lets us take an unknown data point and move down the tree, applying the attributes of the data point to the tree until a leaf is reached and the unknown output of the data point can be determined. In order to create a good classification tree model, we need to have an existing data set with known output from which we can build our model. We also divide our data set into two parts: a training set, which is used to create the model, and a test set, which is used to verify that the model is accurate and not over fitted.

Project Flow Chart.

This will be the proposed flow chart that the system will look like

Data Flow Diagram

Proposed Model

Chapter 3: Research Methodology

Research Design.

I will be using the experimental type of research design. It is a quantitative research method. Basically, it is a research conducted with a scientific approach, where a set of variables are kept constant while other set of variables are being measured as the subject of the experiment. This is more practically while conducting face recognition and detection as it monitors the behaviours and patterns of a subject to be used to acknowledge whether the subject matches all details presented and cross checked with previous data. It is an effect research method as it is time bound and focuses on the relationship between the variables that give actual results.

System Development Methodology.

The methodology of software development is the method in managing project development. There are many models of the methodology are available such as Waterfall model model, Incremental model, RAD model, Agile model, Iterative model and Spiral model. However, it still need to be considered by developer to decide which is will be used in the project. The methodology model is useful to manage the project efficiently and able to help developer from getting any problem during time of development.

Also, it help to achieve the objective and scope of the projects. In order to build the project, it need to understand the stakeholder requirements.

Methodology provides a framework for undertaking the proposed DM modeling. The methodology is a system comprising steps that transform raw data into recognized data patterns to extract knowledge for users.

There are four phases that involve in the spiral model:

Planning phase

Phase where the requirement are collected and risk is assessed. This phase where the title of the project has been discussed with project supervisor. From that discussion, Heart Prediction System has been proposed. The requirement and risk was assessed after doing study on existing system and do literature review about another existing research.

Risk analysis Phase

Phase where the risk and alternative solution are identified. A prototype are created at the end this phase. If there is any risk during this phase, there will be suggestion about alternate solution.

Engineering phase

At this phase, a software are created and testing are done at the end this phase.

Evaluation phase

At this phase, the user do evaluation toward the software. It will be done after the system are presented and the user do test whether the system meet with their expectation and requirement or not. If there is any error, user can tell the problem about system.

Data Collection and Preprocessing.

The data set for this research was taken from UCI data repository.14 Data accessed from the UCI Machine Learning Repository is freely available. In particular, the Cleveland and Hungarian databases have been used by many researchers and found to be suitable for developing a mining model, because of lesser missing values and outliers. The data is cleaned and preprocessed before it is submitted to the proposed algorithm for training and testing.

The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms.

The overall objective of our work is to predict more accurately the presence of heart disease. In this paper, UCI repository dataset are used to get more accurate results. Two data mining classification techniques were applied namely Decision trees and Naive Bayes

his database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no

presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).

Attributes with categorical values were converted to numerical values since most machine learning algorithms require integer values. Additionally, dummy variables were created for variables with more than two categories. Dummy variables help Neural Networks learn the data more accurately.

Classifiers Used for Experiments.

Model / Libraries Used.

1.12.1.1 FLASK:-

Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre- existing third-party libraries provide common functions. Armin Ronacher, who leads an international group of Python enthusiasts named Pocco, develops it. Flask is based on Werkzeug WSGI toolkit and Jinja2 template engine. Both are Pocco projects.

However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools.

Applications that use the Flask framework include and .

1.12.1.2 NumPy :-

NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices.NumPy was created in 2005 by Travis Oliphant. It is an open source project and you can use it freely.NumPy stands for Numerical Python. NumPy targets the

of Python, which is a non-

optimizing . written for this version of Python often run much slower than equivalents. NumPy addresses the slowness problem partly by providing multidimensional arrays and functions and operators that operate efficiently on arrays; using these requires rewriting some code, mostly , using NumPy.

1.12.1.3 OS :-

The main purpose of the OS module is to interact with your operating system. The primary use I find for it is to create folders, remove folders, move folders, and sometimes change the working directory. You can also access the names of files within a file path by doing listdir(). We do not cover that in this video, but that's an option.The os module is a part of the standard library, or stdlib, within Python 3

1.12.1.4 MATPLOTLIB :-

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK. Matplotlib is a low level graph plotting library in python that serves as a visualization utility.Matplotlib was created by John D. Hunter.Matplotlib is open source and we can use it freely.Matplotlib is mostly written in python, a few segments are written in C, Objective-C and Javascript for Platform compatibility

1.12.1.5 PICKLE :-

Python pickle module is used for serializing and de-serializing a Python object structure. Any object in Python can be pickled so that it can be saved on disk. What pickle does is that it “serializes” the object first before writing it to file. Pickling is a way to convert a python object (list, dict, etc.) into a character stream. The idea is that this character stream contains all the information necessary to reconstruct the object in another python script

Advantages of using Pickle Module:

Recursive objects (objects containing references to themselves):

Object sharing (references to the same object in different places):

User-defined classes and their instances

It is advisable not to unpickle data received from an untrusted source as they may pose security threat. However, the pickle module has no way of knowing or raise alarm while pickling malicious data

1.12.1.6 JOBLIB :-

Joblib is a set of tools to provide lightweight pipelining in Python. In particular:

transparent disk-caching of functions and lazy re-evaluation (memoize pattern)

easy simple parallel computing

Joblib is optimized to be fast and robust on large data in particular and has specific optimizations for numpy arrays

Decision Trees.

The decision tree approach is more powerful for classification problems. There are two steps in this technique building a tree & applying the tree to the dataset. There are many popular decision tree algorithms CART, ID3, C4.5, CHAID, and J48. From these J48 algorithm is used for this system. J48 algorithm uses pruning method to build a tree. Pruning is a technique that reduces size of tree by removing over fitting data, which leads to poor accuracy in predications. The J48 algorithm recursively classifies data until it has been categorized as perfectly as possible. This technique gives maximum accuracy on training data. The overall concept is to build a tree that provides balance of flexibility & accuracy.

Ensemble DM approach.

In order to have more reliable and accurate prediction results, ensemble method is a well-proven approach practiced in research for attaining highly accurate classification of data by hybridizing different classifiers. The improved prediction performance is a well-known in-built feature of ensemble methodology. This study proposes a weighted vote-based classifier ensemble technique, overcoming the limitations of conventional DM techniques by employing the ensemble of two heterogeneous classifiers: Naive Bayesian and classification via decision tree

Tools

JUPYTER NOTEBOOK:-

Project Jupiter is a project and community whose goal is to "develop , , and services for across dozens of programming languages". It was spun off from in 2014 by and Brian Granger. Project Jupyter's name is a reference to the three core programming languages supported by Jupyter, which are , and , and also a to s notebooks recording the discovery of the . Project Jupyter has developed and supported the interactive computing products Jupyter Notebook, JupyterHub, and JupyterLab. Jupyter is financially sponsored by NumFOCUS

VS CODE:

Visual Studio Code is a source-code editor that can be used with a variety of programming languages,

Including , , , , and . It is based on the framework, which is used to develop

that run on the . Visual Studio Code employs the same editor component (codenamed "Monaco") used in

(formerly called Visual Studio Online and Visual Studio Team Services).

GITHUB:-

GitHub, Inc. is a provider of for and using . It offers the and (SCM) functionality of Git, plus its own features. It provides and several collaboration features such as , requests, , and for every project. Headquartered

in , it has been a subsidiary of since 2018.

It is commonly used to host projects. As of November 2021, GitHub reports having over 73 million developers and more than 200 million (including at least 28 million public repositories). It is the largest host as of November 2021.

HEROKU :-

Heroku is a cloud (PaaS) supporting several . One of the first , Heroku has been in development since June 2007, when it supported only the programming language, but now supports , , , , , , and . For this reason, Heroku is said to be a as it has features for

a to build, run and scale applications in a similar manner across most languages. Heroku was acquired by in 2010 for $212 million.

SCIKIT-LEARN :-

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. Extensions or modules for SciPy care conventionally named . As such, the module provides learning algorithms and is named scikit-learn.

The library is built upon the SciPy (Scientific Python) that must be installed before you can use scikit-learn. This stack that includes:

NumPy: Base n-dimensional array package

SciPy: Fundamental library for scientific computing

Matplotlib: Comprehensive 2D/3D plotting

IPython: Enhanced interactive console

Sympy: Symbolic mathematics

Pandas: Data structures and analysis

Software requirements:

Hardware Requirements

For application development, the following Software Requirements are:

Processor: Intel or high RAM: 1024 MB

Space on disk: minimum 100mb For running the application:

Device: Any device that can access the internet Minimum space to execute: 20 MB

The effectiveness of the proposal is evaluated by conducting experiments with a cluster formed by 3 nodes with identical setting, configured with an Intel CORE™ i7-4770 processor (3.40GHZ, 4 Cores, 8GB RAM, running Ubuntu 18.04 LTS with 64-bit Linux 4.31.0 kernel)

Budget.

The budget of completion for developing the heart disease prediction system will require various software and hardware devices. The application is averagely expensive to build but if happens to be as successful as the developer sees it to be it will bring forth enough profit to cover the costs undergone.

The table below explains the planned budget in Kenyan Shillings to develop the system:

Work Plan

Conclusion and Future Work.

The proposed system is GUI-based, user-friendly, scalable, reliable and an expandable system. The proposed working model can also help in reducing treatment costs by providing Initial diagnostics in time. The model can also serve the purpose of training tool for medical students and will be a soft diagnostic tool available for physician and cardiologist. General physicians can utilize this tool for initial diagnosis of cardio-patients. There are many possible improvements that could be explored to improve the scalability and accuracy of this prediction system. As we have developed a generalized system, in future we can use this system for the analysis of different data sets. The performance of the health’s diagnosis can be improved significantly by handling numerous class labels in the prediction process, and it can be another positive direction of research. In DM warehouse, generally, the dimensionality of the heart database is high, so identification and selection of significant attributes for better diagnosis of heart disease are very challenging tasks for future research.

References.

A, A. S., & Naik, C. (2016). Different Data Mining Approaches for Predicting Heart Disease, 277–

281. https://doi.org/10.15680/IJIRSET.2016.0505545

Beyene, C., & Kamat, P. (2018). Survey on prediction and analysis the occurrence of heart disease using data mining techniques. International Journal of Pure and Applied Mathematics, 118(Special Issue 8), 165–173. Retrieved from https:/ 85041895038&partnerID=40&md5=2f0b0c5191a82bc0c3f0daf67d73bc81

Brownlee, J. (2016). Naive Bayes for Machine Learning. Retrieved March 4, 2019, from https://machinelearningmastery.com/naive-bayes-for-machine-learning/

Kirmani, M. (2017). Cardiovascular Disease Prediction using Data Mining Techniques. Oriental Journal of Computer Science and Technology, 10(2), 520–528. https://doi.org/10.13005/ojcst/10.02.38

Polaraju, K., Durga Prasad, D., & Tech Scholar, M. (2017). Prediction of Heart Disease using Multiple Linear Regression Model. International Journal of Engineering Development and Research, 5(4), 2321–9939. Retrieved from

Purushottam, Saxena, K., & Sharma, R. (2016). Efficient Heart Disease Prediction System. In Procedia Computer Science (Vol. 85, pp. 962–969). https://doi.org/10.1016/j.procs.2016.05.288

Sai, P. P., & Reddy, C. (2017). International Journal of Computer Science and Mobile Computing HEART DISEASE PREDICTION USING ANN ALGORITHM IN DATA MINING.

International Journal of Computer Science & Mobile Computing, 6(4), 168–172. Retrieved from

Science, C., & Faculty, G. M. (2009). Heart Disease Prediction Using Machine learning and Data Mining Technique. Ijcsc 0973-7391, 7, 1–9.

Soni, J., Ansari, U., & Sharma, D. (2011). Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers. Heart Disease, 3(6), 2385–2392.

Sultana, M., Haider, A., & Uddin, M. S. (2017). Analysis of data mining techniques for heart isease prediction. In 2016 3rd International Conference n Electrical Engineering and Information and Communication Technology, iCEEiCT 2016 (pp. 1–5).

Artificial, I. T. O. (1995). Chapter 2. Neuron, 36–62. Bozzo, R., Conca, A., & Marangon, F. (2014). Decision support system for city logistics: Literature review, and guidelines for an ex-ante model. Transportation Research Procedia, 3(July), 518–527.

Çela, E. K., & Frasheri, N. (2012a). A literature review of data mining techniques used in healthcare databases. ICT Innovations 2012 Web Proceedings, 577–582.

Çela, E. K., & Frasheri, N. (2012b). A literature review of data mining techniques used in healthcare databases. ICT Innovations 2012 Web Proceedings, 577–582.

David, H. B. F., & Belcy, S. A. (2018). Heart Disease Prediction Using Data Mining Techniques,

6956(October), 1817–1823.

Desai, S. D., Giraddi, S., Narayankar, P., Pudakalakatti, N. R., & Sulegaon, S. (2019). Back- propagation neural network versus logistic regression in heart disease classification. Advances in Intelligent Systems and Computing, 702, 133–144. Kiruthika Devi, S., Krishnapriya, S., & Kalita, D. (2016). Prediction of heart disease using data mining techniques. Indian Journal of Science and Technology, 9(39), 1291–1293.

Lavanya, M., & Gomathi, M. P. M. (2016). Prediction of Heart Disease using Classification Algorithms. International Journal of Advanced Research in Computer Engineering & Technology, 5(7), 2278–1323.

Meenakshi, K., Maragatham, G., Agarwal, N., & Ghosh, I. (2018). A Data mining Technique for Analyzing and Predicting the success of Movie. Journal of Physics: Conference Series, 1000(1).

# CHAPTER FIVE - TESTING AND IMPLEMENTATION

HEART DESEASE PREDICTION SYSTEM

By

KEMBOI KEVIN

C026-01-0738/2020

A project proposal submitted for the study leading to a project report in partial fulfilment of the requirements for the award of a Bachelor of Science in Computer Science at Dedan Kimathi University.

Supervisor: PATRICK NAIVASHA

DATE

January – April 2024

Declaration

This is to certify that the work being presented in the project entitled “Heart Disease Prediction System” submitted by undersigned student of Bachelors in COMPUTER SCIENCE in the fulfillment for award of Bachelors in Computer Science is a record of my own work carried out by me under guidance and supervision of PATRICK NAIVASHA of the Department of Computer Science and that this work has not submitted elsewhere for award of any other degree.

Name of student: KEMBOI KEVIN

Registration Number: C026-01-0738/2020

Sign:

Approval

This project was done and presented by me before the panel concerned on the 5th February 2024 at Dedan Kimathi University with my approval and that of my supervisor

Supervisor Name: PATRICK NAIVASHA

Signature:

Date:

Acknowledgement

The satisfaction that accompanies that the successful completion of any task would be incomplete without the mention of people whose ceaseless cooperation made it possible, whose constant guidance and encouragement crown all efforts with success. I am very grateful to my project supervisor Patrick Naivasha, for the guidance, inspiration and constructive suggestions that helpful me in the preparation of this project. I won’t forget to also mention my course mates; Chris Sumba ,Joseph Kariuki and Chris Kinyanjui for their wonderful and skillful guidance in assisting me with the necessary support to ensure that my project is a success. I also thank my parents and family at large for their moral and financial support in funding the project to ensure successful completion of the project.

Dedication.

I dedicate this project to God Almighty my creator, my strong pillar, my source of inspiration, Wisdom, knowledge and understanding. He has been the source of my strength throughout this Program and on His wings only have I soared. I also dedicate this work to my friends; Isaac Ikinya who has encouraged me all the way and whose encouragement has made sure that I live it all it takes to finish that which I have started. To my father Kemboi David and all my beloved friends(Chris Sumba ,Joseph Kariuki and Chris Kinyanjui). Who have been affected in every way possible by this quest?

All the work done in coming up with this system is dedicated to my family for being with/part of me in the whole process especially my dear dad and mum who stood by me in all situations even at the times of financial need.

Thank You. My Love for You All Can Never Be Quantified. God Bless You.

Abstract.

As Heart acts a major role in corporeal body part. Heart diseases or Cardiovascular Diseases (CVDs) are some of the main reason for a huge number of death in the world and some of the reports said that over the last few decades heart disease or CVD has emerged as the most life- threatening disease, not only in Kenya but in the whole world.

Therefore, the diseases of heart wants more perfection and exactness for diagnose and analyse. This disease occurs due to various problems such as over pressure, blood sugar,

Chest pain,high blood pressure, Cholesterol etc. in human body . Heart is the next major organ comparing to the brain which has more priority in the Human body. It pumps the blood and supplies it to all organs of the whole body.

In the health care sector, Machine Learning plays an important role in the health care Industry. So, there is a need for a reliable, accurate, and feasible system to diagnose such diseases in time for proper treatment. Machine Learning algorithms and techniques have been applied to various medical datasets to automate the analysis of large and complex data. A system model is capable of several data processing algorithms for the classification of heart disease. Prediction of occurrences of heart diseases in the medical field is significant work. Data analytics is useful for prediction from more information and it helps the medical center to predict various diseases. The correct prediction of heart disease can prevent life threats, and incorrect prediction can prove to be fatal at the same time.

The objective of this project is to detect whether patients have any chance of heart disease or not by giving number of features to patients with having maximum accuracy of above 97%. By Using Machine learning algorithms and deep learning are applied to compare the results and analysis of the UCI Machine Learning Heart Disease dataset. The dataset consists of 14 main attributes used for performing the analysis

Table of Contents

Chapter One.

Introduction

Heart is one of the most indispensable organs in the human body. It is a organ that serves as a pump to circulate the blood. The heart is a muscular organ about the size of a fist, located just behind and slightly left of the breastbone. The heart pumps blood through the network of arteries and veins called the cardiovascular system. Oxygen is distributed through the circulatory system of the body in the blood, and if the heart does not function correctly, the entire circulatory system of the body will fail. So if the heart doesn’t work properly, it could even lead to death.

According to the World Health Organization (WHO), in the last 15 years, an estimated 17 million people die each year from cardiovascular disease, particularly heart attacks and strokes [1]. Heart disease and stroke are the biggest killers. To predict heart disease, Machine Learning can be used for identifying unseen patterns and providing some clinical insights that will assist the physicians in planning and providing care.

According to World Health Organization, heart related diseases are responsible for taking 17.7 million lives every year, 31% of all global deaths. In Kenya too, heart-related diseases have become the leading cause of mortality. Heart diseases have killed 0.3 million Kenyans in 2016, according to the 2016 Global Burden of Disease Report, released on September 15, 2017. Heart-related diseases increase the spending on health care and also reduce the productivity of an individual.

Estimates made by the World Health Organization (WHO), suggest that India has lost up to $237 billion, from 2005- 2015, due to heart-related or cardiovascular diseases. Thus, feasible and accurate prediction of heart-related diseases is very important.

Heart disease is common among both men and women in most countries around the world. Therefore, people should consider heart disease risk factors. Although it plays a genetic role, some lifestyle factors significantly affect heart disease .

The known risk factors for heart disease; radiation therapy for age, gender, family history, smoking, some chemotherapy drugs and cancer, malnutrition, high blood pressure, high blood cholesterol levels, diabetes, obesity, physical mobility, stress, and poor hygiene. These are the various risk factors in which the patient’s exposure towards developing a CVD.

The most common type is coronary artery disease, which can cause a heart attack. Other types of heart disease may involve the valves in the heart, or the heart may not pump well and cause heart failure. Some people are born with heart disease. Anyone, including children, can develop heart disease. It happens when a substance called plaque builds up in your arteries. Smoking, unhealthy eating and lack of exercise increase your risk of heart disease. High cholesterol, high blood pressure or diabetes can also increase your risk of heart disease.

In contrast, the removal or improvement of this factor decreases this risk. This interpretation suggests the causality between the factor and the illness, which means that the risk factor precedes the disease (the notion of anteriority).

Correction of the factor will cause the disease (the idea of reversibility) to decrease its occurrence. Of course, it must be recognized in several different populations and offer a plausible physiopathological explanation of the disease. Strictly speaking, when there is no direct causal relationship, it is a “risk marker,” a witness to a process (e.g., the elevation of micro albuminuria, elevation of C- reactive protein CRP)

.

The main heart disease risk factors such as physiological factors (age, sex, and menopausal status), lifestyle factors (smoking, physical activity, alcohol, stress), metabolic syndrome factors (insulin resistance), dyslipidemia, abdominal obesity, high blood pressure) and dietary factors. A heart disease risk factor is defined as a factor in which the patient’s exposure to this factor increases the risk of developing a CVD. In contrast, the removal or improvement of this factor decreases this risk. The risk factor’s importance is defined by the association’s strength with the disease (expressed by the relative risk observed in the exposed subjects compared to the unexposed) and the gradual association (parallel to the risk factor).

There are several types of heart disease which include:

Coronary Artery Disease (CAD)

Heart Arrhythmias.

Heart Failure.

Heart Valve Disease.

Pericardial Disease.

Cardiomyopathy (Heart Muscle Disease)

Congenital Heart Disease.

To examine the cardiac disease mischance, the particular issues which need to be discussed are those related to the behaviors. Furthermore, patients will undergo extensive examinations, such as blood pressure, glucose, vital signs, chest pain, electrocardiograms, maximum heart rate, and elevated levels of sugar, but the bright side may be that successful treatment is feasible if the disease is easily and early detected and anticipated, but treatment for all of these cardiac patients is depending on clinical studies, the patient history, and the responses to questions by the patient

To deal with this disease, there are several methods of prevention, such us natural methods, like stoping smoking, maintaining a healthy weight, adopting a healthy diet and practicing sports regularly.We also have the scientific methods such as drugs and surgeries. The prediction of this disease before being infected is part of the prevention

Methods, or the computer tools are the most used means in it, more precisely the Machine Learning algorithms. Determining the probability of having cardiac disease manually is hard to depend on as risk factors. Recently, to solve difficult issues, a range of data mining techniques and machine learning techniques are built.

Still, more advanced machine learning will assist us to identify patterns and their useful knowledge. While it has several uses in the medical field, machine learning is mainly utilized to forecast the heart disease.

Machine learning (ML) plays a significant role in disease predicting. It predicts whether the patient has a particular disease type or not based on an efficient learning technique. In making of these projects I have used several supervised learning techniques for predicting the early stage of heart disease by providing them risk factor

Here, I will use the below machine learning algorithms then I will select the best one based on its classification report.

Support Vector Machine

Random Forest

Ada Boost

Gradient Boosting

With using all these techniques the highest accuracy of 90.82% I got in GradientBoostingClassifier.

Background of the Study

Data analysis proves to be crucial in the medical field. It provides a meaningful base to critical decisions. It helps to create a complete study proposal. One of the most important uses of data analysis is that it helps in keeping human bias away from medical conclusion with the help of proper statistical treatment. By use of data mining for exploratory analysis because of nontrivial information in large volumes of data.

The health care industries collect huge amounts of data that contain some hidden information, which is useful for making effective decisions for providing appropriate results and making

Effective decisions on data, some data mining techniques are used to better the experience and conclusion that have been given.

Heart predictor system will use the data mining knowledge to give a user-oriented approach to new and hidden patterns in the data. The knowledge which is implemented can be used by the healthcare experts to get better quality of service and to reduce the extent of adverse medicine effect.

Problem Statement.

Heart disease can be managed effectively with a combination of lifestyle changes, medicine and, in some cases, surgery. With the right treatment, the symptoms of heart disease can be reduced and the functioning of the heart improved. The predicted results can be used to prevent and thus reduce cost for surgical treatment and other expensive.

The overall objective of my work will be to predict accurately with few tests and attributes the presence of heart disease. Attributes considered form the primary basis for tests and give accurate results more or less. Many more input attributes can be taken but our goal is to predict with few attributes and faster efficiency the risk of having heart disease. Decisions are often made based on doctors’ intuition and experience rather than on the knowledge rich data hidden in the data set and databases. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients.

Data mining holds great potential for the healthcare industry to enable health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and reduce costs. According to (Wurz & Takala, 2006) the opportunities to improve care and reduce costs concurrently could apply to as much as 30% of overall healthcare spending. The successful application of data mining in highly visible fields like e-business, marketing and retail has led to its application in other industries and sectors. Among these sectors just discovering is healthcare. The healthcare environment is still „information rich‟ but „knowledge poor‟. There is a wealth of data available within the healthcare systems. However, there is a lack of effective analysis tools to discover hidden relationships and trends in the data for African genres.

Objectives.

Main Objectives.

The main objective of this research is to develop a heart prediction system. The system can discover and extract hidden knowledge associated with diseases from a historical heart data set

Heart disease prediction system aims to exploit data mining techniques on medical data set to assist in the prediction of the heart diseases.

Specific Objectives.

Provides new approach to concealed patterns in the data.

Helps avoid human biasness.

To implement GradientBoosting Classifer that classifies the disease as per the input of the user.

Reduce the cost of medical tests.

Justification.

Clinical decisions are often made based on doctor’s insight and experience rather than on the knowledge rich data hidden in the dataset. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. The proposed system will integrate clinical decision support with computer-based patient records (Data Sets). This will reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge rich environment which can help to significantly improve the quality of clinical decisions.

There are voluminous records in medical data domain and because of this, it has become necessary to use data mining techniques to help in decision support and prediction in the field of healthcare. Therefore, medical data mining contributes to business intelligence which is useful for diagnosing of disease

Scope and Limitation.

Scope.

Here the scope of the project is that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modeling and

analysis tools, e.g., data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions

Limitations.

Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. The automation of the same would be highly beneficial. Clinical decisions are often made based on doctor’s intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Data mining have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions.

CHAPTER TWO: LITERATURE REVIEW

Introduction

Data mining is the process of finding previously unknown patterns and trends in databases and using that information to build predictive models. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases. The World Health Statistics 2012 report enlightens the fact that one in three adults worldwide has raised blood pressure - a condition that causes around half of all deaths from stroke and heart disease. Heart disease, also known as cardiovascular disease (CVD), encloses a number of conditions that influence the heart – not just heart attacks. Heart disease was the major cause of casualties in the different countries including India. Heart disease kills one person every 34 seconds in the United States. Coronary heart disease, Cardiomyopathy and Cardiovascular disease are some categories of heart diseases. The term “cardiovascular disease” includes a wide range of conditions that affect the heart and the blood vessels and the manner in which blood is pumped and circulated through the body. Diagnosis is complicated and important task that needs to be executed accurately and efficiently. The diagnosis is often made, based on doctor’s experience & knowledge. This leads to unwanted results & excessive medical costs of treatments provided to patients. Therefore, an automatic medical diagnosis system would be exceedingly beneficial.

Literature Review

Numerous studies have been done that have focus on diagnosis of heart disease. They have applied different data mining techniques for diagnosis & achieved different probabilities for different methods.

(Polaraju, Durga Prasad, & Tech Scholar, 2017) proposed Prediction of Heart Disease using Multiple Regression Model and it proves that Multiple Linear Regression is appropriate for predicting heart disease chance. The work is performed using training data set consists of 3000 instances with 13 different attributes which has mentioned earlier. The data set is divided into two parts that is 70% of the data are used for training and 30% used for testing.

(Deepika & Seema, 2017) focuses on techniques that can predict chronic disease by mining the data containing in historical health records using Naïve Bayes, Decision tree, Support Vector Machine (SVM) and Artificial Neural Network (ANN). A comparative study is performed on classifiers to measure the better performance on an accurate rate. From this experiment, SVM gives highest accuracy rate, whereas for diabetes Naïve Bayes gives the highest accuracy.

(Beyene & Kamat, 2018) recommended different algorithms like Naive Bayes, Classification Tree, KNN, Logistic Regression, SVM and ANN. The Logistic Regression gives better accuracy compared to other algorithms. (Beyene & Kamat, 2018) suggested Heart Disease Prediction System using Data Mining Techniques. WEKA software used for automatic diagnosis of disease and to give qualities of services in healthcare centers. The paper used various algorithms like SVM, Naïve Bayes, Association rule, KNN, ANN, and Decision Tree. The paper recommended SVM is effective and provides more accuracy as compared with other data mining algorithms.

Chala Beyene recommended Prediction and Analysis the occurrence of Heart Disease Using Data Mining Techniques. The main objective is to predict the occurrence of heart disease for early automatic diagnosis of the disease within result in short time. The proposed methodology is also critical in healthcare organization with experts that have no more knowledge and skill. It uses different medical attributes such as blood sugar and heart rate, age, sex are some of the attributes are included to identify if the person has heart disease or not. Analyses of data set are computed using WEKA software.

(Soni, Ansari, & Sharma, 2011) proposed to use non- linear classification algorithm for heart disease prediction. It is proposed to use bigdata tools such as Hadoop Distributed File System (HDFS), Map reduce along with SVM for prediction of heart disease with optimized attribute set. This work made an investigation on the use of different data mining techniques for predicting heart diseases. It suggests to use HDFS for storing large data in different nodes and executing the prediction algorithm using SVM in more than one node simultaneously using SVM. SVM is used in parallel fashion which yielded better computation time than sequential SVM.

(Science & Faculty, 2009) suggested heart disease prediction using data mining and machine learning algorithm. The goal of this study is to extract hidden patterns by applying data mining techniques. The best algorithm J48 based on UCI data has the highest accuracy rate compared to LMT. (Purushottam, Saxena, & Sharma, 2016) proposed an efficient heart disease prediction system using data mining. This system helps medical practitioner to make effective decision making based on the certain parameter. By testing and training phase a certain parameter, it provides 86.3% accuracy in testing phase and 87.3% in training phase.

(Kirmani, 2017) suggested multi disease prediction using data mining techniques. Nowadays, data mining plays vital role in predicting multiple disease. By using data mining techniques, the number of tests can be reduced. This paper mainly concentrates on predicting the heart disease, diabetes

and breast cancer etc.,

(Sai & Reddy, 2017) proposed Heart disease prediction using ANN algorithm in data mining. Due to increasing expenses of heart disease diagnosis disease, there was a need to develop new system which can predict heart disease. Prediction model is used to predict the condition of the patient after evaluation on the basis of various parameters like heart beat rate, blood pressure, cholesterol etc. The

accuracy of the system is proved in java.

(A & Naik, 2016) recommended to develop the prediction system which will diagnosis the heart disease from patient’s medical data set. 13 risk factors of input attributes have considered to build the system. After analysis of the data from the dataset, data cleaning and data integration was performed. He used k-means and naïve Bayes to predict heart disease. This paper is to build the system using historical heart database that gives diagnosis. 13 attributes have considered for building the system. To extract knowledge from database, data mining techniques such as clustering, classification methods can be used. 13 attributes with total of 300 records were used from the Cleveland Heart Database. This model is to predict whether the patient have heart disease or not based on the values of 13 attributes.

(Sultana, Haider, & Uddin, 2017) proposed an analysis of cardiovascular disease. This paper proposed data mining techniques to predict the disease. It is intended to provide the survey of current techniques to extract information from dataset and it will useful for healthcare practitioners. The performance can be obtained based on the time taken to build the decision tree for the system. The primary objective is to predict the disease with a smaller number of attributes.

Proposed Architecture.

In this system we are implementing effective heart attack prediction system using Supervised Learning techniques. We can give the input as in CSV file or manual entry to the system. After taking input the algorithms apply on that input that is GradientBoosting Classifier. After accessing data set the operation is performed and effective heart attack level is produced.

The proposed system will add some more parameters significant to heart attack with their gender, age and the priority levels are by consulting expertise doctors and the medical experts. The heart attack prediction system designed to help the identify different risk levels of heart attack like normal, low or high and also giving the prescription details with related to the predicted result.

Flowchart .

The classification tree literally creates a tree with branches, nodes, and leaves that lets us take an unknown data point and move down the tree, applying the attributes of the data point to the tree until a leaf is reached and the unknown output of the data point can be determined. In order to create a good classification tree model, we need to have an existing data set with known output from which we can build our model. We also divide our data set into two parts: a training set, which is used to create the model, and a test set, which is used to verify that the model is accurate and not over fitted.

Project Flow Chart.

This will be the proposed flow chart that the system will look like

Data Flow Diagram

Proposed Model

Chapter 3: Research Methodology

Research Design.

I will be using the experimental type of research design. It is a quantitative research method. Basically, it is a research conducted with a scientific approach, where a set of variables are kept constant while other set of variables are being measured as the subject of the experiment. This is more practically while conducting face recognition and detection as it monitors the behaviours and patterns of a subject to be used to acknowledge whether the subject matches all details presented and cross checked with previous data. It is an effect research method as it is time bound and focuses on the relationship between the variables that give actual results.

System Development Methodology.

The methodology of software development is the method in managing project development. There are many models of the methodology are available such as Waterfall model model, Incremental model, RAD model, Agile model, Iterative model and Spiral model. However, it still need to be considered by developer to decide which is will be used in the project. The methodology model is useful to manage the project efficiently and able to help developer from getting any problem during time of development.

Also, it help to achieve the objective and scope of the projects. In order to build the project, it need to understand the stakeholder requirements.

Methodology provides a framework for undertaking the proposed DM modeling. The methodology is a system comprising steps that transform raw data into recognized data patterns to extract knowledge for users.

There are four phases that involve in the spiral model:

Planning phase

Phase where the requirement are collected and risk is assessed. This phase where the title of the project has been discussed with project supervisor. From that discussion, Heart Prediction System has been proposed. The requirement and risk was assessed after doing study on existing system and do literature review about another existing research.

Risk analysis Phase

Phase where the risk and alternative solution are identified. A prototype are created at the end this phase. If there is any risk during this phase, there will be suggestion about alternate solution.

Engineering phase

At this phase, a software are created and testing are done at the end this phase.

Evaluation phase

At this phase, the user do evaluation toward the software. It will be done after the system are presented and the user do test whether the system meet with their expectation and requirement or not. If there is any error, user can tell the problem about system.

Data Collection and Preprocessing.

The data set for this research was taken from UCI data repository.14 Data accessed from the UCI Machine Learning Repository is freely available. In particular, the Cleveland and Hungarian databases have been used by many researchers and found to be suitable for developing a mining model, because of lesser missing values and outliers. The data is cleaned and preprocessed before it is submitted to the proposed algorithm for training and testing.

The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms.

The overall objective of our work is to predict more accurately the presence of heart disease. In this paper, UCI repository dataset are used to get more accurate results. Two data mining classification techniques were applied namely Decision trees and Naive Bayes

his database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no

presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).

Attributes with categorical values were converted to numerical values since most machine learning algorithms require integer values. Additionally, dummy variables were created for variables with more than two categories. Dummy variables help Neural Networks learn the data more accurately.

Classifiers Used for Experiments.

Model / Libraries Used.

1.12.1.1 FLASK:-

Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre- existing third-party libraries provide common functions. Armin Ronacher, who leads an international group of Python enthusiasts named Pocco, develops it. Flask is based on Werkzeug WSGI toolkit and Jinja2 template engine. Both are Pocco projects.

However, Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools.

Applications that use the Flask framework include and .

1.12.1.2 NumPy :-

NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices.NumPy was created in 2005 by Travis Oliphant. It is an open source project and you can use it freely.NumPy stands for Numerical Python. NumPy targets the

of Python, which is a non-

optimizing . written for this version of Python often run much slower than equivalents. NumPy addresses the slowness problem partly by providing multidimensional arrays and functions and operators that operate efficiently on arrays; using these requires rewriting some code, mostly , using NumPy.

1.12.1.3 OS :-

The main purpose of the OS module is to interact with your operating system. The primary use I find for it is to create folders, remove folders, move folders, and sometimes change the working directory. You can also access the names of files within a file path by doing listdir(). We do not cover that in this video, but that's an option.The os module is a part of the standard library, or stdlib, within Python 3

1.12.1.4 MATPLOTLIB :-

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK. Matplotlib is a low level graph plotting library in python that serves as a visualization utility.Matplotlib was created by John D. Hunter.Matplotlib is open source and we can use it freely.Matplotlib is mostly written in python, a few segments are written in C, Objective-C and Javascript for Platform compatibility

1.12.1.5 PICKLE :-

Python pickle module is used for serializing and de-serializing a Python object structure. Any object in Python can be pickled so that it can be saved on disk. What pickle does is that it “serializes” the object first before writing it to file. Pickling is a way to convert a python object (list, dict, etc.) into a character stream. The idea is that this character stream contains all the information necessary to reconstruct the object in another python script

Advantages of using Pickle Module:

Recursive objects (objects containing references to themselves):

Object sharing (references to the same object in different places):

User-defined classes and their instances

It is advisable not to unpickle data received from an untrusted source as they may pose security threat. However, the pickle module has no way of knowing or raise alarm while pickling malicious data

1.12.1.6 JOBLIB :-

Joblib is a set of tools to provide lightweight pipelining in Python. In particular:

transparent disk-caching of functions and lazy re-evaluation (memoize pattern)

easy simple parallel computing

Joblib is optimized to be fast and robust on large data in particular and has specific optimizations for numpy arrays

Decision Trees.

The decision tree approach is more powerful for classification problems. There are two steps in this technique building a tree & applying the tree to the dataset. There are many popular decision tree algorithms CART, ID3, C4.5, CHAID, and J48. From these J48 algorithm is used for this system. J48 algorithm uses pruning method to build a tree. Pruning is a technique that reduces size of tree by removing over fitting data, which leads to poor accuracy in predications. The J48 algorithm recursively classifies data until it has been categorized as perfectly as possible. This technique gives maximum accuracy on training data. The overall concept is to build a tree that provides balance of flexibility & accuracy.

Ensemble DM approach.

In order to have more reliable and accurate prediction results, ensemble method is a well-proven approach practiced in research for attaining highly accurate classification of data by hybridizing different classifiers. The improved prediction performance is a well-known in-built feature of ensemble methodology. This study proposes a weighted vote-based classifier ensemble technique, overcoming the limitations of conventional DM techniques by employing the ensemble of two heterogeneous classifiers: Naive Bayesian and classification via decision tree

Tools

JUPYTER NOTEBOOK:-

Project Jupiter is a project and community whose goal is to "develop , , and services for across dozens of programming languages". It was spun off from in 2014 by and Brian Granger. Project Jupyter's name is a reference to the three core programming languages supported by Jupyter, which are , and , and also a to s notebooks recording the discovery of the . Project Jupyter has developed and supported the interactive computing products Jupyter Notebook, JupyterHub, and JupyterLab. Jupyter is financially sponsored by NumFOCUS

VS CODE:

Visual Studio Code is a source-code editor that can be used with a variety of programming languages,

Including , , , , and . It is based on the framework, which is used to develop

that run on the . Visual Studio Code employs the same editor component (codenamed "Monaco") used in

(formerly called Visual Studio Online and Visual Studio Team Services).

GITHUB:-

GitHub, Inc. is a provider of for and using . It offers the and (SCM) functionality of Git, plus its own features. It provides and several collaboration features such as , requests, , and for every project. Headquartered

in , it has been a subsidiary of since 2018.

It is commonly used to host projects. As of November 2021, GitHub reports having over 73 million developers and more than 200 million (including at least 28 million public repositories). It is the largest host as of November 2021.

HEROKU :-

Heroku is a cloud (PaaS) supporting several . One of the first , Heroku has been in development since June 2007, when it supported only the programming language, but now supports , , , , , , and . For this reason, Heroku is said to be a as it has features for

a to build, run and scale applications in a similar manner across most languages. Heroku was acquired by in 2010 for $212 million.

SCIKIT-LEARN :-

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. Extensions or modules for SciPy care conventionally named . As such, the module provides learning algorithms and is named scikit-learn.

The library is built upon the SciPy (Scientific Python) that must be installed before you can use scikit-learn. This stack that includes:

NumPy: Base n-dimensional array package

SciPy: Fundamental library for scientific computing

Matplotlib: Comprehensive 2D/3D plotting

IPython: Enhanced interactive console

Sympy: Symbolic mathematics

Pandas: Data structures and analysis

Software requirements:

Hardware Requirements

For application development, the following Software Requirements are:

Processor: Intel or high RAM: 1024 MB

Space on disk: minimum 100mb For running the application:

Device: Any device that can access the internet Minimum space to execute: 20 MB

The effectiveness of the proposal is evaluated by conducting experiments with a cluster formed by 3 nodes with identical setting, configured with an Intel CORE™ i7-4770 processor (3.40GHZ, 4 Cores, 8GB RAM, running Ubuntu 18.04 LTS with 64-bit Linux 4.31.0 kernel)

Budget.

The budget of completion for developing the heart disease prediction system will require various software and hardware devices. The application is averagely expensive to build but if happens to be as successful as the developer sees it to be it will bring forth enough profit to cover the costs undergone.

The table below explains the planned budget in Kenyan Shillings to develop the system:

Work Plan

Conclusion and Future Work.

The proposed system is GUI-based, user-friendly, scalable, reliable and an expandable system. The proposed working model can also help in reducing treatment costs by providing Initial diagnostics in time. The model can also serve the purpose of training tool for medical students and will be a soft diagnostic tool available for physician and cardiologist. General physicians can utilize this tool for initial diagnosis of cardio-patients. There are many possible improvements that could be explored to improve the scalability and accuracy of this prediction system. As we have developed a generalized system, in future we can use this system for the analysis of different data sets. The performance of the health’s diagnosis can be improved significantly by handling numerous class labels in the prediction process, and it can be another positive direction of research. In DM warehouse, generally, the dimensionality of the heart database is high, so identification and selection of significant attributes for better diagnosis of heart disease are very challenging tasks for future research.

References.

A, A. S., & Naik, C. (2016). Different Data Mining Approaches for Predicting Heart Disease, 277–

281. https://doi.org/10.15680/IJIRSET.2016.0505545

Beyene, C., & Kamat, P. (2018). Survey on prediction and analysis the occurrence of heart disease using data mining techniques. International Journal of Pure and Applied Mathematics, 118(Special Issue 8), 165–173. Retrieved from https:/ 85041895038&partnerID=40&md5=2f0b0c5191a82bc0c3f0daf67d73bc81

Brownlee, J. (2016). Naive Bayes for Machine Learning. Retrieved March 4, 2019, from https://machinelearningmastery.com/naive-bayes-for-machine-learning/

Kirmani, M. (2017). Cardiovascular Disease Prediction using Data Mining Techniques. Oriental Journal of Computer Science and Technology, 10(2), 520–528. https://doi.org/10.13005/ojcst/10.02.38

Polaraju, K., Durga Prasad, D., & Tech Scholar, M. (2017). Prediction of Heart Disease using Multiple Linear Regression Model. International Journal of Engineering Development and Research, 5(4), 2321–9939. Retrieved from

Purushottam, Saxena, K., & Sharma, R. (2016). Efficient Heart Disease Prediction System. In Procedia Computer Science (Vol. 85, pp. 962–969). https://doi.org/10.1016/j.procs.2016.05.288

Sai, P. P., & Reddy, C. (2017). International Journal of Computer Science and Mobile Computing HEART DISEASE PREDICTION USING ANN ALGORITHM IN DATA MINING.

International Journal of Computer Science & Mobile Computing, 6(4), 168–172. Retrieved from

Science, C., & Faculty, G. M. (2009). Heart Disease Prediction Using Machine learning and Data Mining Technique. Ijcsc 0973-7391, 7, 1–9.

Soni, J., Ansari, U., & Sharma, D. (2011). Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers. Heart Disease, 3(6), 2385–2392.

Sultana, M., Haider, A., & Uddin, M. S. (2017). Analysis of data mining techniques for heart isease prediction. In 2016 3rd International Conference n Electrical Engineering and Information and Communication Technology, iCEEiCT 2016 (pp. 1–5).

Artificial, I. T. O. (1995). Chapter 2. Neuron, 36–62. Bozzo, R., Conca, A., & Marangon, F. (2014). Decision support system for city logistics: Literature review, and guidelines for an ex-ante model. Transportation Research Procedia, 3(July), 518–527.

Çela, E. K., & Frasheri, N. (2012a). A literature review of data mining techniques used in healthcare databases. ICT Innovations 2012 Web Proceedings, 577–582.

Çela, E. K., & Frasheri, N. (2012b). A literature review of data mining techniques used in healthcare databases. ICT Innovations 2012 Web Proceedings, 577–582.

David, H. B. F., & Belcy, S. A. (2018). Heart Disease Prediction Using Data Mining Techniques,

6956(October), 1817–1823.

Desai, S. D., Giraddi, S., Narayankar, P., Pudakalakatti, N. R., & Sulegaon, S. (2019). Back- propagation neural network versus logistic regression in heart disease classification. Advances in Intelligent Systems and Computing, 702, 133–144. Kiruthika Devi, S., Krishnapriya, S., & Kalita, D. (2016). Prediction of heart disease using data mining techniques. Indian Journal of Science and Technology, 9(39), 1291–1293.

Lavanya, M., & Gomathi, M. P. M. (2016). Prediction of Heart Disease using Classification Algorithms. International Journal of Advanced Research in Computer Engineering & Technology, 5(7), 2278–1323.

Meenakshi, K., Maragatham, G., Agarwal, N., & Ghosh, I. (2018). A Data mining Technique for Analyzing and Predicting the success of Movie. Journal of Physics: Conference Series, 1000(1).