research paper

by sai vedant

Submission date: 11-May-2023 03:18AM (UTC-0400)
Submission ID: 2090220411
File name: UPDATED_RESEARCH_PAPER.docx (604.38K)
Word count: 2469
Character count: 14072

Heart Disease Detection and Patient's Sickness Prediction System

Sakshi Singh, Chaitanya Pandey, Tanmay Patil, Gurjeet Singh
(Dept. of Comp. Sci & Engineering, MIT-ADT University, India)

Abstract : - Detecting heart disease and predicting patient illness is an important subject due to the rising number of cases. The timely discovery of such illnesses is crucial and demands accurate and effective identification. The objective of the project is to utilize machine learning algorithms like logistic regression to forecast the probability of a patient developing heart disease, based on their medical attributes and history. The proposed framework exhibited exceptional forecasting abilities and surpassed the performance of previous classifiers, such as Naive Bayes. The Disease Prediction system employs predictive modeling to forecast the likelihood of a disease based on user-provided symptoms, utilizing the Random Forest Classifier algorithm. The objective of this project is to leverage diverse supervised machine learning techniques to predict diseases and identify heart-related illnesses based on symptoms and medical information as input. This demonstrates the possibilities of these algorithms to facilitate the timely identification of high-risk ailments.

Keywords : *Machine Learning , Random Forest, Logistic Regression, Data Preprocessing, Model Evaluation, Streamlit*

I. Introduction

Medical doctors are facing challenges because of the large amount of data. However, Supervised ML algorithms have showcased significant potential in surpassing standard systems for disease diagnosis and aiding medical experts in the early detection of high-risk diseases.

II. Existing Work

Much research and development efforts have been made to develop disease detection and prediction models. Commonly used supervised machine learning algorithms include Naive Bayes (NB), Decision Trees (DT), and K-Nearest Neighbors (KNN). According to the study results, Support Vector Machine (SVM) is suitable for detecting kidney disease and Parkinson's disease, while Logistic Regression (LR) is highly effective in predicting heart disease. In addition, Random Forest (RF) and Convolutional Neural Networks (CNN) successfully predicted breast and common diseases, respectively. Despite advances in computing, doctors still need technology for a variety of purposes, such as surgical imaging and X-ray imaging. However, technology has yet to keep up with doctors' levels of knowledge and experience. This is because many different factors, such as medical records, weather conditions, air, blood pressure, and others, must be considered in order to understand the entire process of labor. Medical decision support systems can help address this challenge by helping doctors make the right decisions. We used machine learning to analyze the hospital's comprehensive data, allowing us to build models that can quickly analyze data and deliver results faster. By leveraging machine learning, physicians can make critical decisions regarding their patient's diagnosis and treatment options, thereby improving patient care services. . The healthcare industry is a great example of how machine learning is revolutionizing the medical field.

To begin with, the ML algorithms used by most systems tend to have lower accuracy when compared to the ones we have used. Decision Tree Classifier and KNN are commonly used in existing systems, but they do not offer the same level of accuracy as our algorithms. Another reason for this difference in accuracy is the use of complex deep learning algorithms, which require a large amount of data to predict a disease, such as medical images taken from different angles, which in turn require medical expertise and other medical diagnosis details.

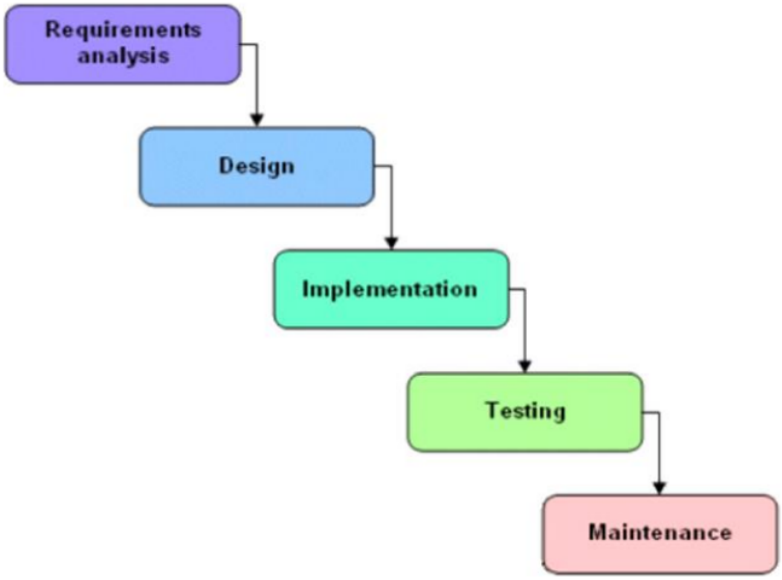
III. Motivation

Identifying and predicting diseases is critical to preventing their severity and fatal consequences. In India, the majority of deaths are due to heart attacks, especially among elderly people affected by cardiovascular diseases. To solve this problem, our prediction system uses various machine learning algorithms to predict the risk level of these diseases. The advent of artificial intelligence (AI) has enabled computer systems to act intelligently, like humans, by perceiving, thinking, and making decisions. AI encompasses multidisciplinary areas such as machine learning, computer vision, deep learning, and natural language processing. By applying optimization, statistical, and probabilistic techniques to past data, ML algorithms can learn and assist healthcare professionals in decision making, contributing to treat the patient appropriately.

IV. Objectives

1. Goal is to use supervised ML algorithms to improve healthcare through accuracy and early detection of diseases that become harmful at a later stage.
2. ML models will be used to predict diseases ranging from common to severe, just to name a few that are located in the heart, kidney, breast and brain.
3. For sickness prediction, we used Random Forest Classification and logistic regression to detect heart disease.

V. Project Plan



We used the waterfall method to develop the system. This picture shows a plan that we can use to get what we need. The Annexure includes some guesses or calculations. We thought about the stages in a waterfall model when making calculations. First, we looked at each part individually and then we calculated the necessary guesses.

VI. Methodology

Heart Disease Prediction

A methodology for predicting heart disease using machine learning typically involves the following steps:

1. Information collection: Collecting the important information for heart disease prediction, which could include medical history, patient attributes such as age, gender, etc., and any other relevant information.
2. Data preprocessing: Cleaning and processing the collected data, dealing with missing values, and normalizing the data to make it suitable for use with machine learning algorithms.
3. Feature selection: Selecting the most relevant features that are likely to contribute to the accurate prediction of heart disease.
4. Model training: Training a machine learning model on the preprocessed and feature-selected data. The model is typically trained on a portion of the data, known as the training set.
5. Model evaluation: Evaluating the trained model using various execution measurements such as precision, accuracy, review, and F1-score, among others. The model is evaluated on another portion of the data, known as the test set.
6. Model optimization: Tuning the model's hyperparameters and optimizing the performance of the model by selecting the best combination of parameters and features.
7. Deployment: Deploying the trained and optimized model in the clinical setting to assist doctors and medical professionals in the accurate prediction of heart disease in patients.

Heart Disease Dataset

S.No	Attributes	Value type
1.	age	Numerical
2.	sex	Nominal
3.	cp	Nominal
4.	trestbps	Numerical
5.	cho	Numerical
6.	fbs	Nominal
7.	restecg	Nominal
8.	thalach	Numerical
9.	ca	Nominal
10.	target	Nominal

Logistic Regression is a Machine Learning Algorithm that Calculates the output of a categorical dependent variable. As a result, the outcome must be categorical or discrete. The answer can be Yes or No, 0 or 1, True or False, etc. The probabilistic values between 0 and 1 are provided instead of the exact values of 0 and 1. There are many similarities between Logistic Regression and Linear Regression, except for the way they are applied. Linear Regression is used for solving regression problems, whereas Logistic regression is used for solving the classification problems.Logistic regression involves the following data and operations:[1]

1. Gather columns
2. Splitting Data
3. Normalization
4. Fitting into Model
5. Prediction
6. Model Evaluation

Patient's Sickness Prediction

An Excel sheet was created from an open-source dataset, listing all the symptoms for respective diseases.The dataset contained approximately 230 diseases with over 1000 unique symptoms. The symptoms of an individual were used as inputs for various machine learning algorithms.

Random Forest Classifier

The Random Forest, so named for its composition of numerous decision trees functioning in conjunction, is a prominent machine learning algorithm. In the random forest approach, each constituent tree generates a discrete prediction for the given observation, and the class that receives the highest number of votes is considered as the model's eventual classification prediction. Such a procedure is known to enhance the accuracy and generalize ability of the predictive model..[2]

Patient's Sickness Dataset

Sickness Prediction Dataset

1) fever
2) Sweating
3) Chills and shivering
4) Headache
5) Muscle aches
6) Loss of appetite
7) irritability
8) Dehydration
9) general weakness
10) Cold
11) Vomiting
12) feeling uneasy

Working

1. The website will collect input data from users and utilize a training dataset to determine the outcome.
2. When the user clicks the result button, it will trigger a request to the Streamlit server containing their inputs, which the server will subsequently restructure.
3. Subsequently, the inputs will be fed into a trained model.
4. The model will analyze the provided data and generate a forecasted output.
5. The server will transmit the forecasted output to the web application as a response, and the web application will exhibit the projected outcome

VII. Usage Scneario

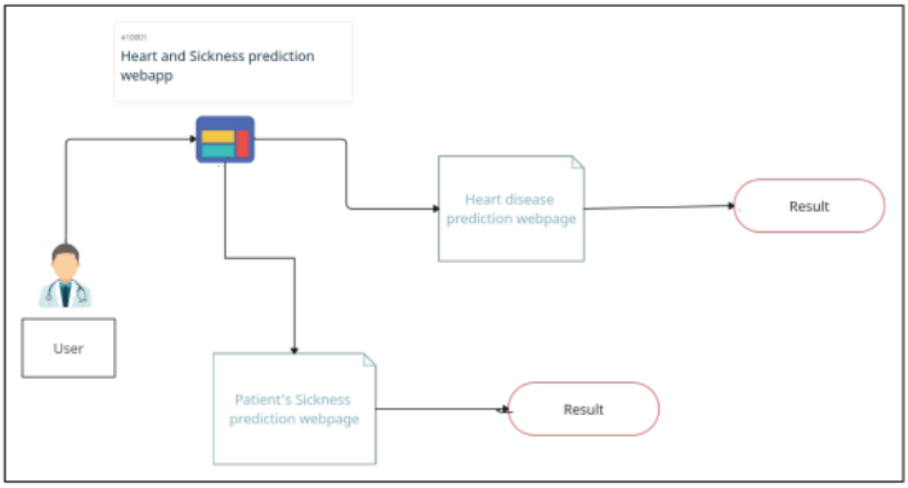
Heart disease Prediction

Early Detection: It can predict the likelihood of heart disease in a patient before the disease becomes severe. This can help in early detection and timely intervention to prevent or manage the disease.
Risk Assessment: Healthcare providers can use model to assess the risk of developing heart disease in patients based on their medical history, lifestyle factors, and other risk factors. This can help in developing personalized treatment plans and preventive strategies.

Patient's Sickness Prediction

Suppose a provider wants to figure out how likely it is for a patient to have a specific sickness according to their current symptoms. The provider collects patient's current symptoms.[3].
It can be used to guess if a person will get sick with a certain illness by looking at their current symptoms. This information can be used by healthcare professionals to identify high-risk patients and provide early interventions or treatment plans. Additionally, the model can be continuously updated with new data to improve its accuracy over time.

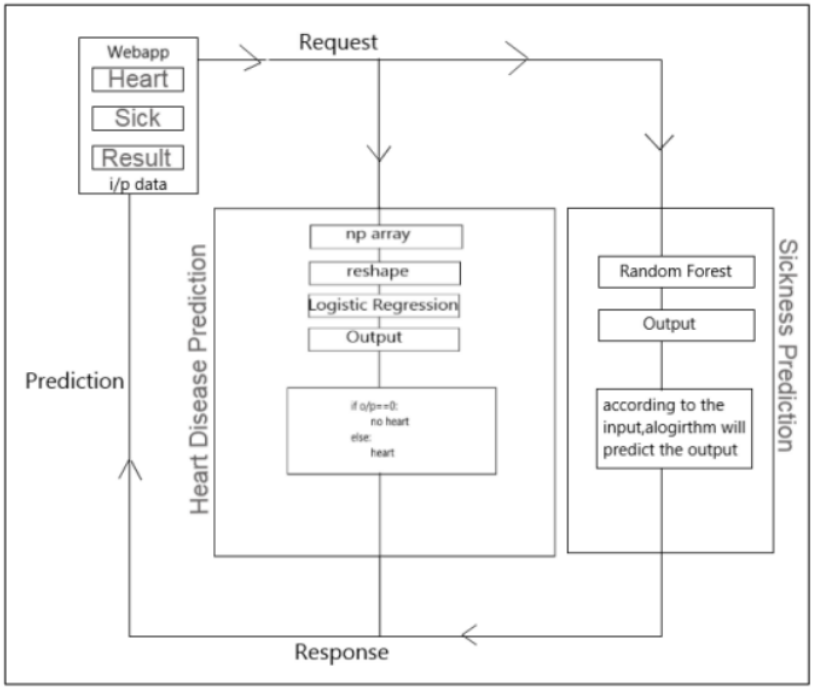
VIII. Use Case



IX. System Architecture

A system with an i5 quad-core processor, 6 GB of RAM, and software packages such as pandas, python, SciPy, StatsModels, and Matplotlib is required. Test analysis was performed in the Jupyter and Spyder web application environments. The analysis is done at two levels: first, the dataset is cleaned using Pandas tool, and second, the cleaned data is passed through classifiers to predict heart disease.

• Python: It is utilized as the back-end programming language to implement the prediction models.
• Streamlit: It is an open-source Python library used for building interactive web applications for data science and machine learning projects. It simplifies the process of creating a web application by allowing developers to create a user interface with simple Python scripts. We have used this for the frontend development
• Jupyter Notebook: It is often used in data science and machine learning projects for data exploration, analysis, visualization and for training and testing data and for creating models.[4]
• Spyder: This is a free computer program that helps people write and run scientific programs in the Python language. This thing helps you write computer instructions, run them, and find mistakes.
• Libraries : NumPy to deal with huge data-set in numerical form, Pandas to analyze the data, Pickle for creating model.sav files, Stream-lit and streamlitoptionmenu for creating website, Sklearn for traintestsplit , logistics regression for accuracy and Sklearn for random forest classifier



X. Project Scope

The incidence of heart disease is steadily increasing around the world, including in our country. To address this, we used logistic regression to detect and predict heart disease, achieving 81 accuracy. Although this accuracy is satisfactory, it can still be improved by exploring alternative methods for attribute selection and increasing the size of the dataset to avoid overfitting and improve performance. Alternatively, we may consider leveraging the power of deep learning algorithms to achieve even more accurate results in the future.

XI. Future Work

The project aids in disease prediction at both personal and public levels. It provides information about the potential risk factors, thus reducing the need for unnecessary tests and associated costs. Moreover, the system can extract and analyze data from patients to reveal hidden patterns, which can be useful for future research and medical advancements.

During the training and testing phase of our project, we evaluated multiple machine learning models for predicting common diseases. The Random Forest Classifier model emerged as the most effective and accurate model with a precision of 95 percent. To forecast and identify heart disease, we smartly employed Logistic Regression and achieved an accuracy rate of 81 percent. There are several opportunities for enhancing the precision of this model in the upcoming days. The Heart disease detection model using Logistic Regression and the sickness prediction model using Random Forest Classifier showed superior results. These algorithms are not only more accurate but also cost-efficient and faster compared to the algorithms used by previous researchers. The Random Forest Classifier achieved a maximum accuracy of 95 percent, and Logistic Regression accuracy was 81 percent, which is either greater or almost equal to the accuracies obtained from previous research. We can conclude that our accuracy has improved by using additional medical attributes from the dataset we used and can also improve more in future updates.

There are several ways in which we can further enhance the accuracy of our system. For instance, we can explore the utilization of deep learning algorithms,[5] consider alternative techniques for attribute selection, and potentially expand the size of the dataset in the future to address issues related to overfitting and to improve overall performance.

XII. Conclusion

As a collaborative unit, we obtained valuable insights into the ways in which skilled programmers operate within the field. There is perpetual scope for enhancement and our devised application is no exception to it. This is primarily because we had a time constraint due to our involvement in other projects, quizzes, and exams. The main goal of our system is to help detect any illness a person may be carrying early, allowing doctors to keep the treatment on track. This is especially helpful for people with heart problems because they can determine if they are at risk for heart disease. By using our system, individuals can benefit from early detection and treatment, which ultimately leads to better health outcomes.

Acknowledgements

It gives us great pleasure in presenting the project report on 'Heart Disease Detection and Patient's Sickness Prediction System'. We would like to take this opportunity to thank my internal guide Prof. Shahin Shoukat Makubhai for giving me all the help and guidance I needed. I am really grateful to them for their kind support. Their valuable suggestions were very helpful. We are also grateful to Dr. Shraddha Phansalkar, Head of Computer Science & Engineering indispensable support, suggestions. We are also grateful to our technology experts Prof. Reena Gunjan and Prof. Suvarna Pawar for their help, support and suggestion.

References

Journal Papers:

[1] *Lynne Connelly. Logistic regression. Medsurg Nursing, 29(5):353–354, 2020.*

[2] *Aakash Parmar, Rakesh Katariya, and Vatsal Patel. A review on random forest: An ensemble classifier. In International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018, pages 758–763. Springer, 2019.*

[3] *P Hamsagayathri and S Vigneshwaran. Symptoms based disease prediction using machine learning techniques. In 2021 Third international conference on intelligent communication technologies and virtual mobile networks (ICICV), pages 747–752. IEEE, 2021.*

[4] *Jiawei Wang, Li Li, and Andreas Zeller. Better code, better sharing: on the need of analyzing jupyter notebooks. In Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering: New Ideas and Emerging Results, pages 53–56, 2020.*

[5] *Sumit Sharma and Mahesh Parmar. Heart diseases prediction using deep learning neural network model. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 9(3):2244–2248, 2020.*

research paper

ORIGINALITY REPORT

14%	12%	2%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.researchgate.net Internet Source	2%
2	www.coursehero.com Internet Source	2%
3	Submitted to Symbiosis International University Student Paper	1%
4	cse.anits.edu.in Internet Source	1%
5	bitswithbits.com Internet Source	1%
6	www.ijraset.com Internet Source	1%
7	Submitted to Aston University Student Paper	1%
8	towardsdatascience.com Internet Source	1%
9	Submitted to Kingston University Student Paper	1%

10	Submitted to De Montfort University Student Paper	1 %
11	bolton.ac.uk Internet Source	1 %
12	webapps.cs.umu.se Internet Source	1 %
13	scholarworks.uaeu.ac.ae Internet Source	<1 %
14	Submitted to Liverpool John Moores University Student Paper	<1 %
15	link.springer.com Internet Source	<1 %
16	tanthiamhuat.files.wordpress.com Internet Source	<1 %
17	"Third International Conference on Image Processing and Capsule Networks", Springer Science and Business Media LLC, 2022 Publication	<1 %

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		