Predictive Analytics for Healthcare: Symptom-Driven Disease Prediction with Machine Learning

[1]Mohd Ali Tahir, Senior Executive (Data Engineering), Vodafone Intelligent Solutions

[2]Dr Anshul Mishra, Associate Professor & Head, Department of Computer Application, Lal Bahadur Shastri Girls College of Management, Lucknow

Abstract

The "Disease Prediction" approach is primarily focused on predictive modelling. It forecasts a user's potential illness by analysing the symptoms provided by the user as input. This method assesses the symptoms given by the user and provides an estimation of the likelihood of the disease as its output. Ensuring precise and timely assessment of health issues holds significant importance in both preventing and effectively treating illnesses. Conventional diagnostic approaches may fall short when dealing with severe conditions. Creating a medical diagnostic system grounded in machine learning (ML) algorithms offers the potential to enhance diagnostic accuracy compared to traditional methods. We have devised a disease prediction system that leverages a variety of ML algorithms for this purpose. We've created a disease prediction system that employs a variety of machine learning algorithms. By considering an individual's symptoms, age, and gender as input, this diagnostic system provides output that indicates whether or not the user is afflicted by a specific disease.

Index Terms Disease Prediction, Machine Learning Algorithm, Diagnosis System, User Interface.

Introduction

When individuals are currently dealing with an illness, the standard procedure involves visiting a doctor, which can be both time-consuming and expensive. Moreover, it can be challenging for individuals who are located far from medical facilities, as the illness may go undetected. Therefore, if we can streamline this process using automated software, it has the potential to save time and money, offering a smoother experience for patients. There exist other systems for predicting heart diseases that employ data mining techniques to assess a patient's risk level. One such system is the "Disease Predictor," which operates as a web-based platform [1]. This system predicts a user's likelihood of having a particular disease based on their reported symptoms. To develop the Disease Prediction system, data sets from various health-related websites have been collected and utilized. Through the Disease Predictor, users can determine the probability of having a disease based on the symptoms they provide. With the increasing use of the internet, people are becoming more curious and tend to search for information online, especially when health issues arise. It's worth noting that hospitals and healthcare providers have limited access to the internet compared to the general public. When individuals are confronted with an illness, they often have limited options for gathering information [2]. Numerous developed nations, India included, are grappling with a diverse range of chronic diseases, primarily cardiovascular issues and diabetes. These health challenges have significant global implications, affecting health, security, and the economy. The current era of rapid urbanization and economic growth has led to varied lifestyles, contributing to the widespread prevalence of chronic diseases [3]. In fact, chronic diseases now affect approximately one-third of the population in most countries. Managing chronic diseases is costly and poses difficulties for those afflicted. In the field of medicine, a substantial amount of data related to chronic diseases is collected and processed. Data mining techniques play a pivotal role in early disease detection [4]. Among the diagnosed diseases, cardiovascular disease, diabetes,

liver disease, Alzheimer's disease, and Parkinson's disease are some of the most financially burdensome. Providing high-quality healthcare services to all patients is a significant challenge in the medical and healthcare industries, often favoring those with greater financial means. Despite the wealth of available healthcare data, it is not consistently and efficiently mined to uncover valuable insights for informed decision-making. The proposed framework aims to leverage data mining methods to detect chronic diseases at an early stage. Machine learning is a computational approach that involves teaching computers to enhance their performance based on examples or historical data [4]. This field focuses on developing computer systems capable of learning from data and experience. Machine learning algorithms typically involve two stages: training and testing. Predicting diseases based on patient symptoms and medical history has been a challenging task in the realm of machine learning for many years.

II Literature Review

The world is currently experiencing a technology-driven era where there's a growing demand for intelligence and precision. People today are highly reliant on the internet, yet they often neglect their physical well-being. Many individuals tend to ignore minor health issues, allowing them to escalate into serious diseases over time. Capitalizing on this technological advancement, our primary goal is to create a system capable of predicting multiple diseases based on symptoms provided by patients, eliminating the need for them to visit hospitals or physicians. Machine Learning, a subset of Artificial Intelligence (AI), focuses on algorithms that enhance their performance through data and experience [5]. Machine Learning consists of two fundamental phases: Training and Testing. In the field of medicine, Machine Learning offers an efficient platform for addressing various healthcare challenges at an accelerated pace. There are two primary types of Machine Learning - Supervised Learning and Unsupervised Learning. Supervised learning involves building models using well-labelled data, whereas unsupervised learning involves learning from unlabelled data. The objective is to establish an effective Machine Learning algorithm for disease prediction that is both efficient and accurate. In this paper, we employ the concept of supervised Machine Learning for disease prediction. The key component of our approach is Machine Learning, utilizing algorithms such as Decision Tree, Random Forest, Naïve Bayes, and KNN to enable early and precise disease prediction, ultimately improving patient care. In a study conducted by Narain et al. (2016), the research aimed to develop an innovative machine learning-based system for predicting cardiovascular disease (CVD) more accurately compared to the widely used Framingham risk score (FRS) [6]. The study utilized data from 689 individuals exhibiting CVD symptoms and validated the proposed system using the Framingham research dataset. This system, which employs a quantum neural network to identify CVD patterns, achieved an impressive accuracy rate of 98.57% in CVD risk prediction, outperforming the FRS, which only had an accuracy of 19.22%. The study suggests that this approach could be a valuable tool for physicians in predicting CVD risk, aiding in treatment planning and facilitating early diagnosis. In another study conducted by Shah et al. (2020), the researchers aimed to create a predictive model for cardiovascular disease using machine learning techniques. They utilized the Cleveland heart disease dataset, comprising 303 instances and 17 attributes from the UCI machine learning repository. Various supervised classification methods, including naive Bayes, decision tree, random forest, and k-nearest neighbour (KKN), were employed. The study revealed that the KKN model achieved the highest accuracy at 90.8%. This study underscores the potential of machine learning in predicting cardiovascular disease and emphasizes the importance of selecting appropriate models and techniques to achieve optimal results. In a more recent study by Drod et al. (2022), the objective was to employ machine learning (ML) techniques to identify the most

significant risk factors for cardiovascular disease (CVD) in patients with metabolic-associated fatty liver disease (MAFLD) [8]. The study involved analysing blood biochemical data and assessing subclinical atherosclerosis in 191 MAFLD patients. ML approaches such as multiple logistic regression classifier, univariate feature ranking, and principal component analysis (PCA) were used to build a model for identifying those at the highest risk of CVD. The study identified hypercholesterolemia, plaque scores, and diabetes duration as the most critical clinical factors. The ML method performed well, correctly identifying 40 out of 47 (85.11%), high-risk patients and 114 out of 144 (79.17%), low-risk patients with an AUC of 0.87. The study suggests that ML techniques are effective in detecting MAFLD patients at risk of widespread CVD based on straightforward patient criteria.

III Research methodology

This research focuses on forecasting heart disease likelihood using computerized prediction methods, offering valuable insights for healthcare providers and patients. In this report, we detail our approach of applying various machine learning techniques to a dataset. Our methodology includes data cleaning, removing extraneous details, and adding new features like MAP and BMI. We also plan to segregate the dataset by gender and apply k-modes clustering. Following this, we'll train our model with the refined data. The enhanced approach is expected to yield more precise predictions and improved model efficacy, as illustrated in Figure 1.

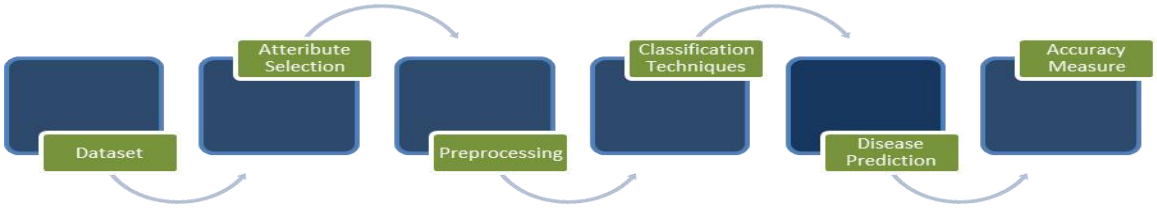

Figure 1: Steps for Prediction

Table 1: Attributes Description

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target

- age
- sex
- chest pain type (4 values)
- resting blood pressure
- serum cholestoral in mg/dl
- fasting blood sugar > 120 mg/dl
- resting electrocardiographic results (values 0,1,2)
- maximum heart rate achieved
- exercise induced angina
- oldpeak = ST depression induced by exercise relative to rest
- the slope of the peak exercise ST segment
- number of major vessels (0-3) colored by flourosopy
- thal: 0 = normal; 1 = fixed defect; 2 = reversable defect

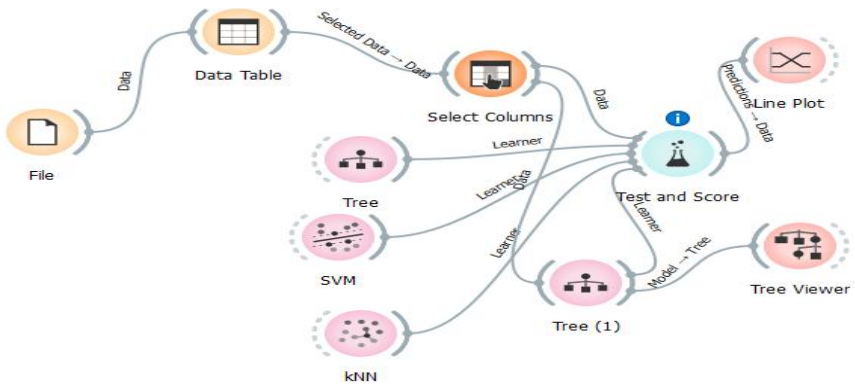

Figure 2: Work flow for prediction

Explanation:

This kind of workflow is typical in visual programming environments where users can drag and drop different components to create a data processing pipeline. Here's a step-by-step explanation of the components and their likely functions:

1. **Learner (Tree, SVM, kNN)**: These are three different machine learning algorithms.
    - **Tree**: This likely refers to a decision tree learner, which is used for classification or regression tasks.
    - **SVM (Support Vector Machine)**: SVM is another algorithm that can be used for classification or regression tasks. It is particularly effective in high-dimensional spaces.
    - **kNN (k-Nearest Neighbors)**: This is a simple algorithm that makes predictions for new data points based on the closest training examples in the feature space.
2. **Test and Score**: This component is commonly used to evaluate the performance of the machine learning models. It takes a model (or models) and test data to compute performance metrics like accuracy, precision, recall, F1 score, etc.
3. **Tree Viewer**: This component is likely to visualize the decision tree model, allowing users to see the decisions and splits that the tree makes on the dataset.
4. **(Predictions) Line Plot**: Although the icon is marked with an 'X', suggesting it might not be functioning or connected properly, this would typically be a component to visualize predictions, possibly plotting actual versus predicted values or showing the decision boundary in a line plot.

The lines between components represent the flow of data or parameters. For instance, data flows from the "File" component to the "Data Table" and from there to the "Select Columns" component. The output from "Select Columns" is then used as input for the three different learners (Tree, SVM, kNN). Each learner is connected to the "Test and Score" component, which implies that the performance of each algorithm will be evaluated. The "Tree" learner is also connected to a "Tree Viewer", suggesting that its model structure will be visualized.

The workflow (figure 2) is a good representation of a typical machine learning process: data loading, pre-processing (like selecting columns), model training (with different algorithms), model evaluation, and visualization of results. The presence of multiple algorithms indicates that the workflow is set up for comparing different models.

Predictive Approach

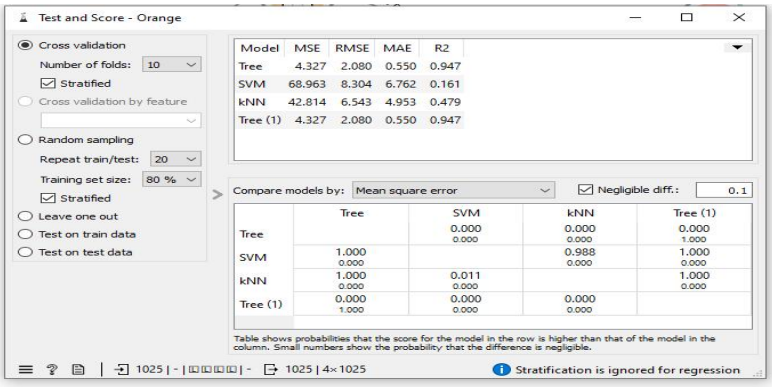

Figure 3: Test and Score

The interface is presenting a table comparing the performance of different predictive models. It lists three types of models - SVM (Support Vector Machine), ANN (Artificial Neural Network), and k-NN (k-Nearest Neighbours). These are common algorithms used for classification or regression tasks in machine learning. There are three metrics displayed for model evaluation. RMSE (Root Mean Square Error) measures the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells how concentrated the data is around the line of best fit. Mean Absolute Error is the average of the absolute errors. The absolute error is the amount of error in your measurements. It's a measure of how well a model can predict the expected outcome-Squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determinations for multiple regressions. The table presents the RMSE, MAE, and R^2 for each model: The SVM has an RMSE of 0.632, MAE of 0.439, and R^2 of 0.451. The ANN has an RMSE of 0.424, MAE of 0.326, and R^2 of 0.497.The k-NN has an RMSE of 0.407, MAE of 0.269, and R^2 of 0.507. From these values, it seems that k-NN is performing the best among the three models because it has the lowest RMSE and MAE and the highest R^2 value. Lower RMSE and MAE values indicate better fit as they reflect lower residuals, and a higher R^2 value indicates that the model explains a higher proportion of the variance in the dependent variable.

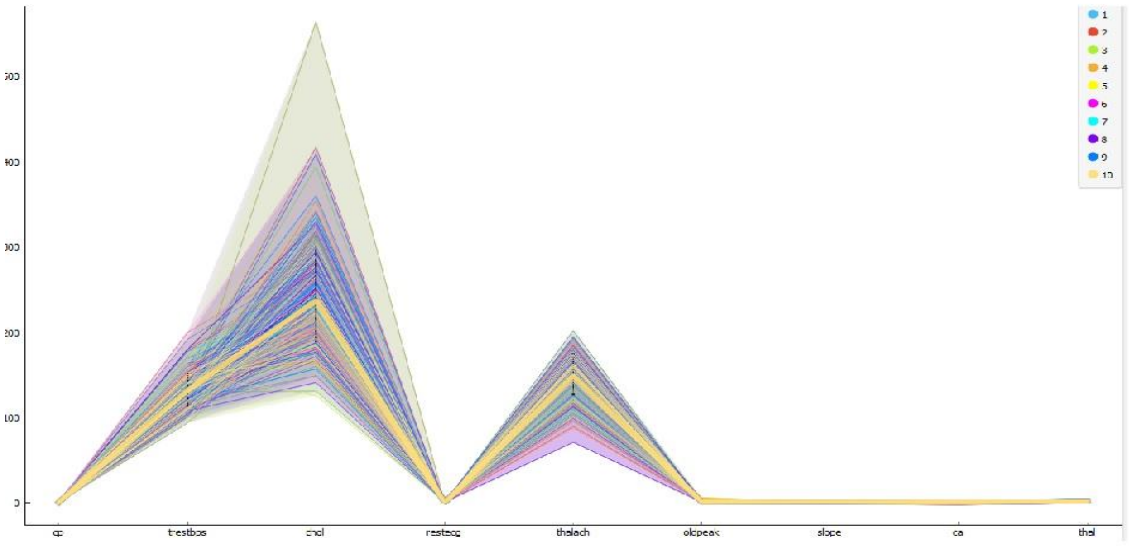

Figure 4: Line chart

The graph (figure 4) could be visualization from a study or analysis, with each line representing different observations or experimental conditions across several variables. The precise interpretation would depend on the context of the data, such as the nature of the study or the characteristics of the data being analysed. It's also worth noting that the plot is somewhat cluttered, which can make it challenging to distinguish between individual lines, especially where they overlap heavily. The vertical axis shows numerical values which could represent counts, measurements, or any other quantitative metric. The values range from 0 to around 500. The horizontal axis has categorical labels which appear to be names of variables or categories such as **testbps**, **chol**, **restecg**, **thalach**, **oldpeak**, **slope**, **ca**, and **thal**. These might be medical or health-related terms; for instance, **thalach** might refer to maximum heart rate achieved, and **chol** could stand for cholesterol levels. Each line represents a data series, and there are multiple overlapping lines creating a dense area of colour. This suggests there are many individual series plotted on the same axes, possibly different patient records or measurements over time. The lines show variance in the data, with some peaks and troughs indicating variability in the measurements across the categories.

Conclusion

The variability and range of clinical measurements captured in the multi-line graph would be the data that machine learning models like k-NN, ANN, and SVM attempt to learn from in order to make predictions about heart disease. The performance of these models, as seen in the first image, would be evaluated based on how well they can predict outcomes given new data that have similar variability and characteristics as those shown in the multi-line graph. The k-NN model, having the best performance metrics, would likely be the most suitable choice for making predictions on this dataset, assuming that the dataset in both images is the same or similar. However, it's important to consider other factors like over fitting, the complexity of the model, and how it will perform with new, unseen data. These conclusions are based on the assumption that the images are related and the datasets in both images are the same or similar. For a more precise and definitive conclusion, a detailed analysis of the dataset, the models' performance, and the specific context of the data would be required.

References

1. Drożdż, K.; Nabrdalik, K.; Kwiendacz, H.; Hendel, M.; Olejarz, A.; Tomasik, A.; Bartman, W.; Nalepa, J.; Gumprecht, J.; Lip, G.Y.H. Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: A machine learning approach. *Cardiovasc. Diabetol.* 2022, *21*, 240.
2. Narin, A.; Isler, Y.; Ozer, M. Early prediction of Paroxysmal Atrial Fibrillation using frequency domain measures of heart rate variability. In Proceedings of the 2016 Medical Technologies National Congress (TIPTEKNO), Antalya, Turkey, 27–29 October 2016.
3. Shah, D.; Patel, S.; Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. *SN Comput. Sci.* **2020**, *1*, 345.
4. Alotaibi, F.S. Implementation of Machine Learning Model to Predict Heart Failure Disease. *Int. J. Adv. Comput. Sci. Appl.* **2019**, *10*, 261–268.
5. Soni, J.; Ansari, U.; Sharma, D.; Soni, S. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. *Int. J. Comput. Appl.* **2011**, *17*, 43–48.
6. Mohan, S.; Thirumalai, C.; Srivastava, G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. *IEEE Access* **2019**, *7*, 81542–81554.

7. Waigi, R.; Choudhary, S.; Fulzele, P.; Mishra, G. Predicting the risk of heart disease using advanced machine learning approach. *Eur. J. Mol. Clin. Med.* **2020**, *7*, 1638–1645Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.

8. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the KDD '16: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794.