Mini Project Report on

HEART DISEASE PREDICTION SYSTEM

Submitted in partial fulfillment of the requirement for the award of the degree of

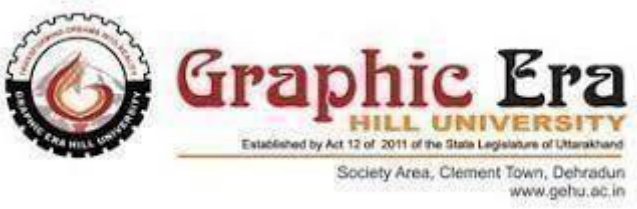
BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE & ENGINEERING

Submitted by:

NAINA **2118815**



Department of Computer Science and Engineering
Graphic Era Hill University

Dehradun, Uttarakhand
January 2024

Table of Contents

.

CHAPTER 1 INTRODUCTION

The integration of artificial intelligence (AI) and machine learning (ML) in healthcare has led to the development of sophisticated disease prediction and recommendation systems. These systems aim to enhance early diagnosis, provide personalized treatment plans, and improve overall patient outcomes. By leveraging vast amounts of data, AI systems can uncover patterns, predict outcomes, and provide insights that are beyond human capabilities.

ROLE OF AI IN HEALTHCARE:

1. Enhanced Diagnostics
   Medical Imaging: AI algorithms, particularly those based on deep learning, ○ have shown remarkable accuracy in interpreting medical images such as Xrays, MRIs, and CT scans. They can detect abnormalities and diseases at an early stage, often with higher accuracy than human radiologists.
   Pathology: AI assists in analyzing pathology slides, identifying cancerous cells, and other tissue abnormalities with high precision.
2. Predictive Analytics
   Disease Prediction: AI models can predict the likelihood of developing certain diseases based on patient data, including genetics, lifestyle, and environmental factors. For example, predictive analytics can identify individuals at high risk for diabetes, cardiovascular diseases, or mental health disorders.
3. Personalized Medicine
   Tailored Treatments: AI enables the customization of treatment plans based on an individual's genetic makeup, medical history, and current health status. This approach increases the effectiveness of treatments and reduces adverse effects.
   Pharmacogenomics: AI helps in understanding how different individuals respond to medications, leading to more effective and personalized drug prescriptions.
4. Operational Efficiency
   Workflow Optimization: AI streamlines administrative tasks such as scheduling, billing, and managing patient records, allowing healthcare professionals to focus more on patient care.
   Resource Management: Predictive analytics optimize the allocation of resources, such as hospital beds, medical staff, and equipment, based on current and projected needs.
5. Robotic Surgery
   Precision and Minimally Invasive Procedures: AI-powered surgical robots assist surgeons in performing precise and minimally invasive procedures, resulting in shorter recovery times and reduced risk of complications.

6. Virtual Health Assistants

Patient Engagement: AI-driven chatbots and virtual assistants provide patients with information, reminders, and support, enhancing patient engagement and adherence to treatment plans.

PROBLEM STATEMENT

Develop a comprehensive AI-driven Disease Prediction and Recommendation System designed to assist healthcare providers in accurately diagnosing diseases and recommending personalized treatment plans. The system aims to utilize vast amounts of healthcare data, including patient records, medical histories, imaging, lab results, and genetic information, to predict the likelihood of diseases and suggest appropriate interventions.

OBJECTIVE

1. Accurate Disease Prediction

Develop Predictive Models: Create machine learning models capable of predicting the likelihood of various diseases based on patient data.

Early Detection: Enable early diagnosis of diseases to improve treatment outcomes and reduce healthcare costs.

Risk Stratification: Identify high-risk patients for proactive monitoring and preventive measures

2. Personalized Recommendations

Tailored Treatment Plans: Provide personalized treatment recommendations based on individual patient data, including medical history, lifestyle, and genetic factors.

MOTIVATION

-Addressing unmet medical needs:

1. Early Detection of Diseases

Many diseases, including cancer, cardiovascular diseases, and diabetes, have better treatment outcomes when detected early. By utilizing AI to predict diseases at an early stage, this system can significantly enhance early intervention strategies.

2. Reducing Diagnostic Errors

Diagnostic errors are a substantial issue in healthcare, leading to inappropriate treatments and adverse patient outcomes. AI can assist in reducing these errors by providing more accurate and data-driven diagnoses.

-Enhancing medical requirement: 1.

Personalized Medicine

Every patient is unique, and treatments that are effective for one person may not work for another. This system aims to tailor recommendations based on individual patient data, leading to more effective and personalized healthcare.

2. Preventive Healthcare

By identifying individuals at risk of developing certain conditions, the system can recommend preventive measures, potentially reducing the incidence of chronic diseases and improving overall public health.

-Contribution to medical research:

1. Advancing Medical Knowledge

The data and insights generated by the system can contribute to medical research, helping to uncover new patterns and associations in disease development and treatment.

TOOLS

1. Programming Languages:
   Python: Pandas, Numpy, Scipy
2. Feature Selection and Creation:
   Python Libraries: Scikit-learn, Feature-engine
3. Dimensionality Reduction:
   Python Libraries: Scikit-learn. 4.

Machine Learning Frameworks:

Scikit-learn: General-purpose machine learning library TensorFlow: Deep learning framework.

PyTorch: Deep learning framework

5. Integrated Development Environments (IDEs):

Jupter Notebook

PyCharm 6.

Recommendation Algorithms:

Collaborative Filtering: Surprise library.
ContentBased Filtering: Scikit-learn

Chapter II. METHODOLOGY

Database System Module

The diseases and the symptoms related to them are contained in first dataset. The motive is to predict the ailment the person might be suffering based on the symptoms that are shown by them. The attributes are 'Symptoms', 'Precautions', 'Medications', 'Workout' ,' Disease Name ', and 'Diets'. 'Symptoms ' helps in identifying each disease uniquely.. Once the ailment has been predicted, the patient has been diagnosed correctly; the next step is to aid the recovery process by prescribing the right kind of medication. After the predictions is done the system will tell you the name of disease patient is suffering from plus medication required for the specific disease plus precautions need to done and workout and diet patient need to change to cure from disease.. Hence a new dataset has to be created which has only single disease and single symptom in every row. Since in machine learning algorithms dataset should be processed fast we convert the disease names and symptoms into numerical values. This dataframe has all the possible symptoms as its column values or attributes. The rows values indicate the corresponding disease. If that disease is associated with a particular symptom, the respective entry in the data frame is given the value of '1' and '0' otherwise. Now the dataset is ready.

Disease Prediction Module

Disease Prediction Module deals with choosing the suitable algorithm to be deployed on the data. The data which has been cleaned is fragmented into two parts. The first part of the dataset is called the training data and it is used for developing the model. The second part of the dataset is called the test data, and is used as a reference to test the model. Models like *Logistic Regression, Random Forest, Gradient Boosting Classifier, Naive Bayes and SVC* have been applied on the training dataset. This was followed by testing of the model by applying it on the testing dataset. Later the accuracy was given. But, the accuracy turned out to be zero. The reason why this particular model behaved that way was due to the reason that it does not have any first-hand knowledge of disease prior to the testing of it. Three different classification algorithms are imported from the specific libraries and deployed on the dataset. After this, the model is evaluated on the testing dataset. The algorithm with highest accuracy is used to predict the diseases.

Logistic regression

Logistic regression is a widely used statistical and machine learning algorithm for binary classification tasks, where the goal is to predict one of two possible outcomes based on a set of input features. Unlike linear regression, which predicts continuous values, logistic regression predicts probabilities that map to discrete classes using a logistic function, also known as the sigmoid function. This function outputs values between 0 and 1, representing the probability of the default class (typically labeled as 1). If the probability is greater than a threshold (usually 0.5), the instance is classified as belonging to the default class; otherwise,

it is classified as the other class (labeled as 0). The model is trained by maximizing the likelihood of the observed data using techniques such as maximum likelihood estimation. Logistic regression is prized for its simplicity, interpretability, and efficiency, making it suitable for various practical applications, including medical diagnosis, spam detection, and credit scoring.

Random Forest:

The fundamental principle that governs random forest prediction is that there is wisdom in crowds: a large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models (Yiu, 2019).

Support Vector Machines

A Support Vector Machine (SVM) is a supervised machine learning algorithm that can be employed for both classification and regression purposes. Support Vector Machine algorithms have been used effectively to predict cardiovascular disease in several studies. Alty uses a simple digital volume pulse to effectively predict (over 85% accuracy) CVD risk (Alty et al., 2003).

SVM in linear non-separable cases

In the linearly separable case, SVM is trying to find the hyperplane that maximizes the margin, with the condition that both classes are classified correctly. But in reality, datasets are probably never linearly separable, so the condition of 100% correctly classified by a hyperplane will never be met (Bambrick, 2022).

We decided to move forward using RandomForest algorithm to build our optimized model. The RandomForest Classification algorithm provides one of the highest accuracies among all classification methods. To build a model, we collected our data set from the UCI repository, which had 303 patients, with 14 features. We imported the data as a CSV file to PyCharm. With the help of NumPy, Pandas, and Scikit-Learn libraries in Python, we can clean, extract features, split into training and test datasets, and then train the model using the RandomForest algorithm. To optimize the model, we change the hyperparameters, the results of each hyperparameter change will be shown in the results section. Next chapter we will start by introducing our dataset and then a detailed analysis and results .

Gradient Boosting Classifier

Gradient Boosting Classifier is a powerful ensemble machine learning technique that builds a predictive model in a stage-wise fashion by combining the strengths of weak learners, typically decision trees, into a strong predictive model. Unlike traditional decision trees that are built sequentially, gradient boosting works by sequentially adding models to an ensemble, where each new model minimizes the errors made by the previous ones. This iterative process focuses on improving the predictive accuracy of the model by optimizing a loss

function, such as the logarithmic loss for classification tasks or mean squared error for regression tasks, through gradient descent. Each tree built in the ensemble corrects the errors of its predecessor, thereby reducing bias and variance and improving overall model performance. Gradient Boosting Classifier is known for its robustness against overfitting and ability to capture complex relationships in data, making it widely used in various domains, including recommendation systems, financial forecasting, and healthcare diagnostics.

CHAPTER V DISCUSSION, CONCLUSION

Discussion

The research questions are: What Machine learning algorithms are used in the diagnosis of heart disease? How can Machine Learning techniques be used to minimize misdiagnosis (additional tests, and wrong treatment all resulting in greater monetary impact to the patient)? How can Machine Learning be used to detect early abnormalities, thus benefiting both patients and the healthcare system?

What follows is the discussion of the findings and conclusion, followed by suggestions for areas for further study.

The findings and conclusion for each question are:

1)Machine learning algorithms used in predicting heart disease are Naïve Bayes, Decision Trees, Support Vector Machine, Gradient Boosting classifier, Logistic Regression and Random Forest, concluding that these algorithms can achieve high accuracy in predicting heart disease.

2) Machine learning algorithms can analyze a large amount of data to assist medical professionals in making more informed decisions cost-effectively.

3) Machine Learning algorithms allowed us to analyze clinical data, draw relationships between diagnostic variables, design the predictive model, and tests it against the new case. The predictive model achieved an maximum accuracy percent using SV Classifier's default setting to predict disease and to recommend the precautions taken to cure the disease.

Machine learning and data mining techniques are a major turning point in medical diagnosis and this project has shown how important information from medical records can be utilized to diagnose heart disease patients. The project's objective to explore how machine learning algorithms can be used in the diagnosis of disease has been achieved by identifying 5 different algorithms covered in Chapter 2, additionally developing an optimized model with one of them. Finally, the model we built to predict disease can save enormous medical bills, improve diagnosis capability on large scale, and most importantly save lives.

Conclusion

Disease is a life-threatening disease affecting millions of people around the world every year (Asadi et al., 2021). Hence, early prediction of heart disease can benefit patients and healthcare professionals by providing the information they need to minimize death and reduce costs. Since medical big data has been increasing daily and data storage costs decreasing, machine learning algorithms can play an important part in processing these medical data and predicting diseases.