Heart disease prediction and recommendation using *machine learning*

Arthika L[1], Ashika J[2], Akhila Judit Nisha S[3], Jegana.R[4]
[1]*Department of Information Technology*
[2,3] *AP/IT*
[4]*Bethlahem Institute of Engineering, Karungal*

Abstract: This heart disease prediction project uses machine learning to identify patients at risk by analyzing medical data with four algorithms: Support Vector Machine (SVM), Naive Bayes, Decision Tree, and XGBoost. The system examines factors like age, sex, chest pain type, blood pressure, cholesterol, and other key indicators to predict heart disease risk. The main goal is to classify patients into those with heart disease (label 1) and those without (label 0). The Decision Tree algorithm performed the best due to its strong accuracy, simplicity, and easy-to-understand structure. The project includes a data preprocessing stage, where missing values are handled, numerical features are scaled, and categorical variables are encoded to prepare the data for training. Feature selection techniques are used to focus on the most important data points, improving model performance and efficiency. The Decision Tree's easy interpretability is a key advantage in healthcare. Its flowchart structure helps doctors understand how the model makes predictions, building trust and transparency. This is especially important because healthcare professionals can use the model's insights to make informed decisions. The system also provides personalized recommendations to help patients manage their heart health. These include advice on medications (under a doctor's supervision), lifestyle changes like stress management and smoking cessation, and a heart-healthy diet and exercise plan. By combining predictive analytics with health recommendations, the project aims to improve preventive care. It not only identifies people at risk early but also educates them on steps they can take to reduce their risk, ultimately improving their quality of life.

I.INTRODUCTION

Heart disease is a leading global health issue and a major cause of death. Early detection and preventive measures are critical for improving patient outcomes. This project aims to tackle this problem by developing a predictive system for heart disease using machine learning. Four algorithms—Support Vector Machine (SVM), Naïve Bayes, Decision Tree, and XGBoost—are used to 4475nalyse patient health data and identify those at risk, aiding healthcare providers in making informed decisions.

The dataset includes various health indicators such as age, sex, chest pain type, blood pressure, cholesterol levels, and other cardiovascular metrics. The system processes this data to classify patients into two groups: those with heart disease and those without. To prepare the data for analysis, preprocessing steps like handling missing values, scaling numerical features, and encoding categorical variables are performed. The system evaluates each algorithm based on performance metrics like accuracy, precision, recall, and F1 score to determine the most effective model.

The Decision Tree model emerged as the best performer due to its simplicity, strong accuracy, and interpretability. One of the key strengths of the Decision Tree is its visual decision-making process, which makes it easy for healthcare professionals to understand how the model arrives at its predictions. This transparency is vital in healthcare, as it allows clinicians to trust the system and make informed decisions based on the model's outputs.

In addition to predicting heart disease risk, the system provides personalized health recommendations. These include advice on medications (under a doctor's supervision), lifestyle changes like stress management, smoking cessation, and regular check-ups, as well as dietary recommendations focused on heart health (such as eating more fruits, vegetables, and whole grains while avoiding high-sodium and high-fat foods). Exercise routines are also tailored to each patient's health condition, encouraging moderate physical activities like walking, swimming, or cycling.

By combining predictive analysis with personalized health guidance, this project offers a comprehensive approach to heart disease prevention and management. It helps healthcare providers identify

patients at risk early, supporting timely interventions and better outcomes. Additionally, the personalized recommendations empower patients to take proactive steps in managing their heart health, potentially reducing the risk of severe complications and hospitalizations. Ultimately, the project contributes to a shift toward more preventive, data-driven, and patient-centered care, improving long-term health outcomes and quality of life for individuals at risk of heart disease.

II.RELATED WORKS

Existing healthcare systems often use Logistic Regression to predict chronic diseases like heart attack, diabetes, breast cancer, and kidney disease. This model is favored for its simplicity and effectiveness in binary classification tasks, which makes it well-suited for assessing medical risks. By analyzing clinical factors such as age, blood pressure, cholesterol levels, glucose levels, and other key health indicators, Logistic Regression categorizes patients into risk groups—typically "low risk" or "high risk"—based on the probability of developing a disease.

A key benefit of Logistic Regression is its interpretability. The model clearly shows how different variables, like high blood pressure or high glucose levels, influence the likelihood of a disease. This insight helps healthcare providers make informed decisions, as they can identify the most significant risk factors and offer personalized advice or preventive treatments. Additionally, Logistic Regression requires less computational power compared to more complex algorithms, making it a practical choice for healthcare applications and easier to integrate into existing healthcare platforms.

However, Logistic Regression does have some limitations. It primarily works well for linear relationships between the variables and the outcomes, which means it might struggle to handle more complex cases where risk factors interact in non-linear ways. In such cases, the model may require extra steps, like feature scaling or managing outliers, to perform effectively. Despite these limitations, Logistic Regression remains a valuable tool in healthcare, offering simplicity, efficiency, and the ability to support timely, evidence-based decisions that ultimately improve patient outcomes.

III.METHODOLOGY

The methodology for this heart disease prediction project focuses on building and evaluating machine learning models to accurately classify patients at risk. The process begins with data collection and preprocessing. The dataset includes key health indicators like age, sex, chest pain type, blood pressure, and cholesterol. The data is cleaned by handling missing values, scaling numerical features, and encoding categorical variables to ensure consistency and compatibility with machine learning algorithms.

Next, feature selection techniques are used to identify the most important predictors of heart disease, which helps improve model performance and clarity. The selected features are used to train four machine learning algorithms: Support Vector Machine (SVM), Naive Bayes, Decision Tree, and XGBoost. Each model is trained using cross-validation to ensure robust and unbiased results, with the dataset divided into training and testing sets for performance evaluation.

The models are assessed based on accuracy, precision, recall, and F1 score to determine how well they classify patients. The Decision Tree model is selected as the best due to its balance of simplicity and performance, along with its easy-to-understand structure, which is valuable in clinical settings. Hyperparameter tuning, using techniques like grid or random search, is applied to optimize the models' performance. The final Decision Tree model is tested on an independent dataset to confirm its reliability.

In addition to prediction, the project includes a recommendation system that provides personalized health advice to at-risk patients. This includes medication suggestions, lifestyle changes, diet tips, and exercise routines to help improve heart health. By combining predictive analytics with tailored advice, the system aims to assist healthcare professionals in early intervention and better patient care.

IV.DATA PREPARATION

The data preparation for the heart disease prediction project is a comprehensive process that ensures the dataset is clean, consistent, and optimized for accurate model training and predictions. It begins

with data cleaning, where missing values are handled through imputation techniques or removal, and outliers are identified and appropriately addressed to prevent them from distorting the analysis. This ensures the integrity and completeness of the dataset, which is crucial for model accuracy.

Next, scaling numerical features like age, blood pressure, cholesterol, and glucose levels is performed. This step standardizes the values of these features to a similar range, making sure that no single feature disproportionately affects the model, particularly for algorithms sensitive to feature scale, such as Support Vector Machines (SVM) and XGBoost.

Categorical variables (such as chest pain type, sex, and others) are then encoded into numerical formats, typically using techniques like one-hot encoding or label encoding. This allows the machine learning models to process these non-numeric variables and use them in their decision-making.

To further enhance model performance, feature selection is employed to identify the most significant predictors of heart disease. By reducing the dimensionality of the dataset and focusing on the most relevant features, this step improves the model's efficiency, reduces computation time, and increases interpretability. This also helps prevent overfitting by eliminating irrelevant or redundant features.

Finally, the dataset is split into training and testing sets to evaluate the model's performance effectively. Typically, 70-80% of the data is used for training, while the remaining 20-30% is used for testing and validation. This split allows for an unbiased assessment of the model's ability to generalize to new, unseen data. It also ensures that the model's predictive power is tested on real-world data, which is critical for making reliable predictions.

By following these steps—data cleaning, scaling, encoding, feature selection, and data splitting—the dataset is fully prepared for machine learning, ensuring that the models built for heart disease prediction are both accurate and reliable, with the ability to provide actionable insights for healthcare professionals.

MODIFICATION ON THE PRE-TRAINED RANDOM FOREST ALGORITHM FOR HEART DISEASE PREDICTION

To enhance the pre-trained Random Forest algorithm for heart disease prediction, it begins with refining the data by addressing missing values, handling outliers, and encoding categorical variables. These preprocessing steps ensure that the dataset is clean and suitable for the model. Next, hyperparameters such as the number of trees, tree depth, and minimum samples for splits or leaves can be fine-tuned to optimize the algorithm's performance. Methods like grid search or random search are valuable for finding the best combination of parameters.

The model's feature importance is evaluated to highlight the most relevant predictors of heart disease, allowing for better feature selection. This ensures the model focuses on the most impactful variables while reducing noise from less significant features. For handling class imbalances, techniques such as adjusting class weights or using SMOTE for resampling can help improve the model's ability to predict the minority class.

To assess the model's robustness, cross-validation is applied to ensure reliable performance across different data subsets. Key performance metrics like accuracy, precision, recall, F1 score, and ROC-AUC are used to gauge the effectiveness of the model. For greater transparency, SHAP values or feature importance plots can be utilized to interpret the model's decision-making process.

Finally, the modified model can be deployed for real-time heart disease prediction, with ongoing monitoring to ensure continued accuracy. Additionally, the model may need retraining as new data becomes available to maintain its reliability and predictive power.

V.THE PROPOSED SYSYTEM

The proposed system employs a Support Vector Machine (SVM) algorithm to predict heart disease by analyzing key health indicators such as age, sex, cholesterol levels, blood pressure, exercise capacity, and other relevant factors. The SVM is well-suited for medical applications because of its ability to identify complex patterns within data, making it a reliable method for predicting diseases with high accuracy. By focusing on these important health indicators, the system aims to assist healthcare professionals in evaluating a patient's risk of developing heart disease, thereby supporting early detection and timely medical intervention.

The first stage of the system involves collecting patient data from diverse sources such as electronic health records (EHRs), clinical assessments, and patient questionnaires. These data sources provide valuable insights into a patient's health history, lifestyle, and medical conditions. Once the data is collected, it undergoes a thorough preprocessing phase to ensure its quality and reliability. This includes cleaning the data to handle missing values, normalizing numerical features like age and cholesterol levels for consistency, and encoding categorical variables such as chest pain type and sex to make them compatible with the algorithm. By standardizing and transforming the raw data into a structured format, the model can more effectively learn from the dataset and produce accurate predictions.

Following preprocessing, the SVM model is trained on the cleaned and structured data. During this training phase, the algorithm analyzes the relationships between the input features (e.g., cholesterol, age, and exercise capacity) and the target variable (heart disease or no heart disease). SVM works by finding the optimal hyperplane that separates the data into two categories, maximizing the margin between the classes. To further improve the model's performance, hyperparameters are fine-tuned using cross-validation. This process ensures that the model generalizes well to unseen data and avoids overfitting, which is crucial for making reliable predictions on new patient information.

Once the model is trained and evaluated, it is integrated into an intuitive web-based interface, making it accessible to healthcare professionals. The user-friendly design of the web interface allows medical staff to easily input patient information and receive instant predictions on the likelihood of heart disease. This interface is designed to be simple to navigate, requiring minimal technical expertise, so healthcare providers can focus on making informed clinical decisions quickly. The system provides healthcare professionals with a valuable tool to assess patient risk, supporting more accurate diagnoses and helping to prioritize patients who need immediate care.

In addition to predicting the likelihood of heart disease, the system also offers personalized health recommendations tailored to each patient's specific profile. Based on the results of the prediction, the system generates customized advice regarding lifestyle changes, medications, dietary adjustments, and exercise routines. For instance, if a patient is at high risk for heart disease, the system might recommend medications such as statins, lifestyle modifications like smoking cessation or stress management, and suggest a heart-healthy diet rich in fruits, vegetables, and lean proteins. Additionally, exercise plans can be recommended based on the patient's fitness level and health status, emphasizing activities that improve cardiovascular health without overexertion.

This dual approach, combining both predictive analytics and personalized health recommendations, empowers both healthcare providers and patients. It not only helps identify individuals who are at risk of heart disease but also provides actionable steps for managing and improving their heart health. The system's personalized advice helps patients understand their health risks and motivates them to adopt healthier behaviors, potentially reducing the likelihood of developing severe heart disease complications in the future.

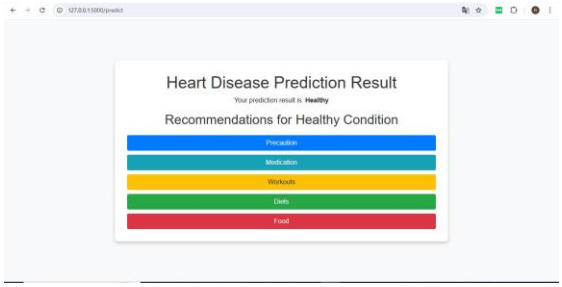
By integrating advanced machine learning techniques like SVM with practical health management recommendations, the proposed system offers a comprehensive tool for heart disease prevention. It not only aids healthcare professionals in making better-informed decisions but also contributes to improving patient outcomes by promoting preventive care and healthier lifestyle choices. Furthermore, as the system is continuously updated with new patient data, it can adapt and evolve over time, maintaining its relevance and improving its accuracy with each iteration.

RESULTS

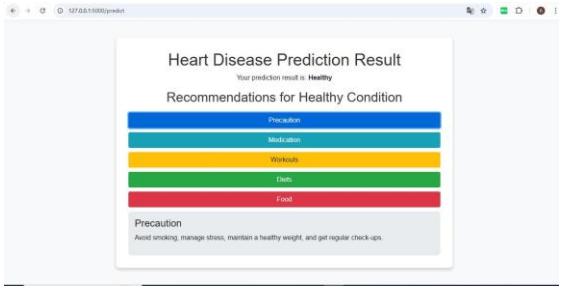
INPUT

OUTPUT:

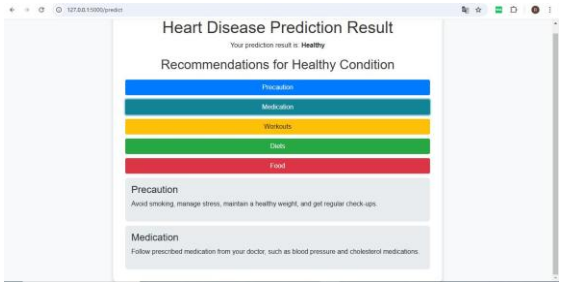
HEALTHY HEART CONDITON



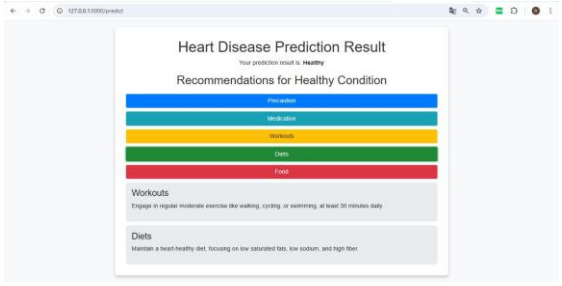
HEALTHY OF PRECAUTION



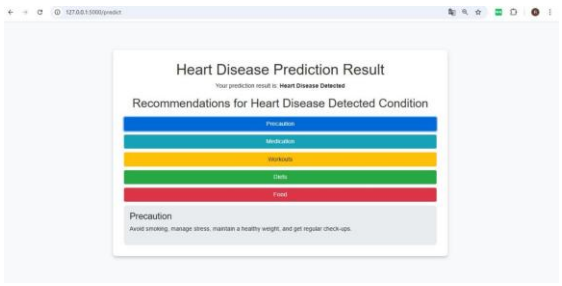
HEALTHY OF MEDICATION



HEALTHY OF WORKOUTS AND DIETS



PRECAUTION OF HEART DISEASE



CONCLUSION

In conclusion, the heart disease prediction project uses machine learning, specifically the Support Vector Machine (SVM) algorithm, to improve heart health management. By collecting and preparing data, the project creates a reliable model to predict heart disease risk. The model uses key health indicators to identify individuals at risk, helping healthcare professionals provide timely care. It also offers personalized recommendations on medications, lifestyle changes, and diet, empowering both patients and healthcare providers to make informed decisions. This project demonstrates the potential of data-driven solutions in healthcare, promoting better patient outcomes and preventive care strategies. As technology continues to evolve, projects like this can help create more personalized and effective treatments for chronic conditions like heart disease.