Course Project Report

Predicting Heart Diseases Using Machine Learning and Different Data Classification Techniques

Submitted By

Priyanshu Maniyar (221AI023)
Sachin Choudhary (221AI034)
Rohil Sharma (221AI033)
Tushar Kanda (221AI042)

as part of the requirements of the course

Machine Learning (IT307) [Jul 2024 - Nov 2024]

in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology in Artificial Intelligence

under the guidance of

Dr. Nagamma Patil , Dept of IT, NITK Surathkal

undergone at

DEPARTMENT OF INFORMATION TECHNOLOGY

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA, SURATHKAL

Jul 2024 - Nov 2024

DECLARATION

We hereby declare that the project report entitled **"Predicting Heart Diseases Using Machine Learning and Different Data Classification Techniques"** submitted by us for the course **Machine Learning (IT307)** during the semester **Jul 2024 - Nov 2024**, as part of the partial course requirements for the award of the degree of Bachelor of Technology in Artificial Intelligence at NITK Surathkal is our original work. We declare that the project has not formed the basis for the award of any degree, associateship, fellowship or any other similar titles elsewhere.

Details of Project Group

Name of the Student	Register No.	Signature with Date
Priyanshu Maniyar	221AI023	
Rohil Sharma	221AI033	
Sachin choudhary	221AI034	
Tushar Kanda	221AI042	

Place: NITK, Surathkal
Date: 06/11/2024

Abstract

This study uses advanced machine learning techniques to predict heart disease risk from clinical and demographic data, aiming for accuracy and interpretability that clinicians can trust. Using a cardio dataset with attributes like age, cholesterol, and blood pressure, we applied rigorous preprocessing—handling missing values, scaling, and balancing classes with SMOTE to improve sensitivity to positive cases. We evaluated several classifiers, including SVM, XGBoost, and Random Forest. Feature selection methods like RFE and Lasso helped focus on the most relevant predictors, enhancing model accuracy and interpretability. SHAP values were used to clarify each feature's impact on predictions, with high cholesterol and blood pressure emerging as key indicators. XGBoost excelled in accuracy, while Random Forest provided strong interpretability with SHAP insights. Our results demonstrate the best balance of accuracy and transparency, supporting practical applications of machine learning in clinical heart disease risk assessment.

Table of Content

List of Tables

List of Figures

Introduction

Cardiovascular disease (CVD) is one of the most prevalent and deadly health issues globally, responsible for millions of deaths each year. The impact of CVD spans across all regions and demographics, placing a significant burden on healthcare systems worldwide. Early detection of heart disease can improve treatment outcomes and reduce mortality rates, underscoring the importance of reliable, accurate prediction models that can assist in timely intervention. However, creating a predictive model in healthcare is challenging due to the critical need for both accuracy and interpretability—patients and healthcare providers must understand not only the prediction itself but also the reasoning behind it. This understanding can build trust in model recommendations and enable healthcare professionals to make informed decisions.

The objective of this project is to develop a machine learning model that can accurately predict the likelihood of heart disease using patient data, including various clinical and lifestyle factors. The project evaluates a range of classification algorithms, including Logistic Regression, Random Forests, Support Vector Machines, and neural networks, to find an optimal model that balances accuracy with interpretability. Additionally, feature selection techniques such as ANOVA, Recursive Feature Elimination (RFE), and Lasso regression are applied to identify the most impactful factors influencing heart disease risk. This approach is designed to enhance model performance by removing irrelevant features that could introduce noise and decrease predictive power.

A key challenge in this project is the issue of class imbalance, which is common in medical datasets where the instances of disease presence are significantly fewer than the non-disease cases. Imbalanced data can skew model performance, often leading to models that favor the majority class. To address this, Synthetic Minority Over-sampling Technique (SMOTE) is used to balance the dataset, improving the model's ability to detect true positive cases of heart disease.

Furthermore, interpretability is a primary focus of this study. SHAP (SHapley Additive exPlanations) values are utilized to interpret model predictions, allowing for the identification of key factors that drive the model’s decisions. By explaining individual predictions, SHAP enhances the model’s transparency, making it more practical for healthcare applications. This study's contribution lies in its comparative analysis of feature selection methods, the implementation of SHAP for model interpretation, and an

exploration of balanced and unbalanced data's effects on model accuracy and reliability in the healthcare domain.

LITERATURE SURVEY

Topic	Paper Title	Insight	Methods	Link
Heart Disease Prediction	Heart Disease Prediction Using Machine Learning Algorithms	This paper uses various machine learning models to predict heart disease, comparing models like Decision Trees, Random Forest, and SVM.	Machine Learning (SVM, Decision Trees, Random Forest)	https://arxiv.org/abs/2409.03697
Shap Values	Cardiovascular Disease Prediction Using Supervised Ensemble Machine Learning and Shapley Values	SHAP values are used to explain the predictions of an XGBoost model, showing the impact of features like systolic blood pressure and age.	XGBoost, Shapley values	https://emerginginvestigators.org/articles/23-257
Feature Selection	Improving Heart Disease Prediction Using Feature Selection Methods	Explores feature selection methods like Recursive Feature Elimination (RFE) to improve prediction accuracy by identifying the most relevant features	RFE, Mutual Information	https://www.sciencedirect.com/science/article/pii/S1877050920310936
Chi-Square	Chi-Square and Other Statistical Methods for Heart Disease Prediction	Investigates how chi-square can be used for feature selection by analyzing the relationship between categorical variables and heart disease	Chi-square, Feature Selection	https://ieeexplore.ieee.org/document/8367704
Data Preprocessing	Preprocessing Techniques for Heart Disease Data	Focuses on handling missing values, scaling, and encoding in heart disease datasets to improve model performance	Data Normalization, Imputation, One-Hot Encoding	https://ieeexplore.ieee.org/document/8447890

Table 2.1-Literature Survey

METHODOLOGY

In this project, we aimed to develop a machine learning model to predict the risk of cardiovascular disease, with a strong focus on interpretability and accuracy. Each step of the methodology, from data collection to SHAP analysis, was designed to ensure that the model was robust, clinically useful, and easy to understand for medical professionals. Below, we provide a detailed explanation of each component of our methodology.

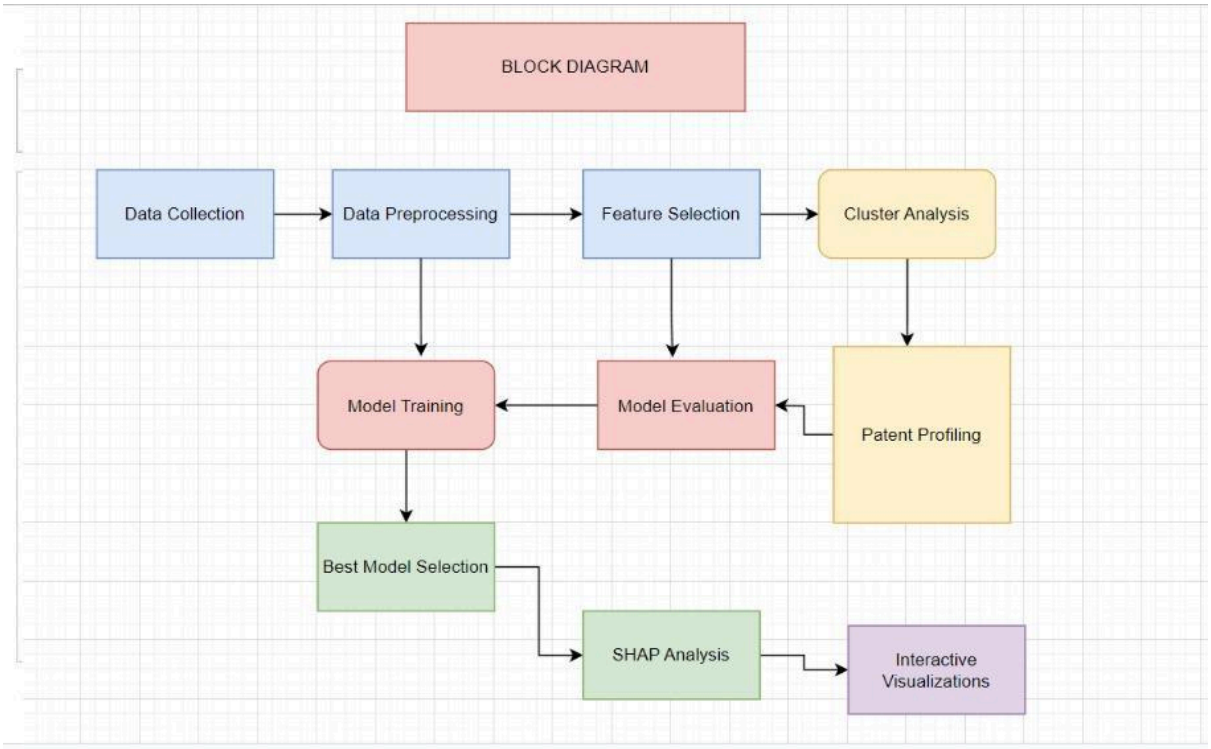


fig 3.1-Block diagram of Methodology

Data Collection & Loading

Our analysis began by sourcing a suitable dataset for cardiovascular disease prediction. We utilized the cardio dataset from Kaggle, which contains various clinical and demographic features such as age, cholesterol levels, blood pressure, and lifestyle factors (like smoking). These variables are key indicators in assessing cardiovascular health, and the dataset's broad coverage made it suitable for building a comprehensive prediction model. The target

variable in this dataset is a binary indicator representing whether a patient is at risk for cardiovascular disease (1 = risk, 0 = no risk).

The cardio dataset was loaded into our analysis environment, where we carried out several preprocessing steps. Properly loading and understanding the data structure was essential for effective preprocessing and feature selection.

Data Preprocessing

Data preprocessing is critical in machine learning, especially for medical datasets, where features can vary significantly in scale and format. We took the following preprocessing steps to ensure data quality and consistency:

1. Age Conversion: In the dataset, age was initially recorded in days. To make this feature more interpretable and clinically relevant, we converted age from days to years. This change aligns with standard medical interpretations, where age is often a key risk factor in cardiovascular disease assessment.

2. Handling Missing Values: Missing data can introduce bias and reduce the predictive power of a model. We checked for missing values in each feature and applied appropriate imputation techniques where necessary. By addressing missing data, we minimized the risk of introducing inaccuracies or skewed results in model training.

3. Standardization: We standardized all numerical features using StandardScaler to ensure they were on a similar scale. StandardScaler transforms features to have a mean of 0 and a standard deviation of 1, which is important for models like Support Vector Machine (SVM) and Logistic Regression, where varying feature scales can impact performance. Standardization also prevents features with larger scales (such as cholesterol levels) from dominating the model's learning process.

4. Addressing Negative Values: For certain feature selection methods, including Chi-Squared analysis, all feature values must be non-negative. In preparation for this, we identified any negative values in the dataset and addressed them by adjusting or filtering the data. Ensuring non-negativity allowed us to proceed with feature selection effectively and without errors.

Feature Selection Techniques

Feature selection is essential in high-dimensional datasets, as it helps reduce noise, enhances model interpretability, and often improves model accuracy. We applied three feature selection techniques to identify the most relevant predictors of cardiovascular risk:

1. Chi-Squared Analysis: The Chi-Squared test measures the independence between each feature and the target variable. Higher Chi-Squared scores indicate a stronger relationship between the feature and the target. This method is particularly useful for categorical and discrete data, allowing us to identify features that significantly impact cardiovascular disease risk. We used Chi-Squared analysis to retain features with the highest relevance, improving our model's efficiency by reducing unnecessary data.

2. ANOVA F-Test: The ANOVA F-Test, which stands for "Analysis of Variance," compares the variance within each feature group to the variance between groups to determine feature importance. The test calculates F-statistics for each feature, with higher F-values indicating a stronger correlation with the target variable. Using ANOVA F-Test helped us identify features with significant explanatory power, enhancing the predictive accuracy of our model.

3. Mutual Information: Mutual Information (MI) measures the degree of dependency between each feature and the target variable. Unlike Chi-Squared and ANOVA, which assume linear relationships, MI can capture non-linear associations. This is especially useful in healthcare data, where risk factors may have complex interactions with disease outcomes. By selecting features with high MI scores, we ensured our model captured nuanced relationships in the data.

Using these feature selection techniques allowed us to refine our dataset to only the most relevant predictors, reducing computational load and making the model more interpretable without sacrificing accuracy.

Data Splitting

To evaluate the model's generalizability, we split the dataset into training and test sets with an 80-20 ratio. The training set is used to train the model, while the test set serves as unseen data for evaluating model performance. This approach minimizes the risk of overfitting, where a model performs well on training data but poorly on new data. By keeping the test set separate, we ensured that the model's performance metrics would reflect its ability to generalize to real-world data.

Model Selection

We experimented with various classification algorithms to determine which model best suited our objectives of accuracy and interpretability. The models tested included:

1**. Support Vector Machine (SVM):** SVMs are effective in high-dimensional spaces and can handle cases where the number of dimensions exceeds the number of samples. We chose SVM because it performs well on binary classification tasks like cardiovascular disease prediction, where it seeks to find the optimal hyperplane separating risk from non-risk cases.

2. **XGBoost:** XGBoost is a gradient boosting algorithm known for its accuracy and efficiency. It constructs an ensemble of decision trees, each focusing on errors from previous trees, thus minimizing overall prediction error. XGBoost is particularly powerful in handling complex datasets with non-linear relationships, making it suitable for our problem.

3. **Random Forest:** Random Forest builds multiple decision trees and combines their outputs for final predictions. This ensemble method helps improve accuracy and reduce overfitting, making Random Forest a strong candidate for our dataset. Additionally, it provides feature importance scores, which aid in identifying influential predictors.

4. **AdaBoost:** Adaptive Boosting, or AdaBoost, adjusts the weights of incorrectly classified samples, making them more likely to be classified correctly in the next iteration. AdaBoost is effective in improving weak models, and its interpretability aligns well with healthcare applications.

5. **Stacking Model:** We used a stacking approach, which combines multiple models to boost accuracy. Stacking involves training a meta-model based on the outputs of several base models, allowing it to learn from their collective strengths. This approach often results in a more accurate model by leveraging the complementary capabilities of different algorithms.

For each classifier, we evaluated accuracy, allowing us to directly compare performance across models and select the best candidates for further analysis.

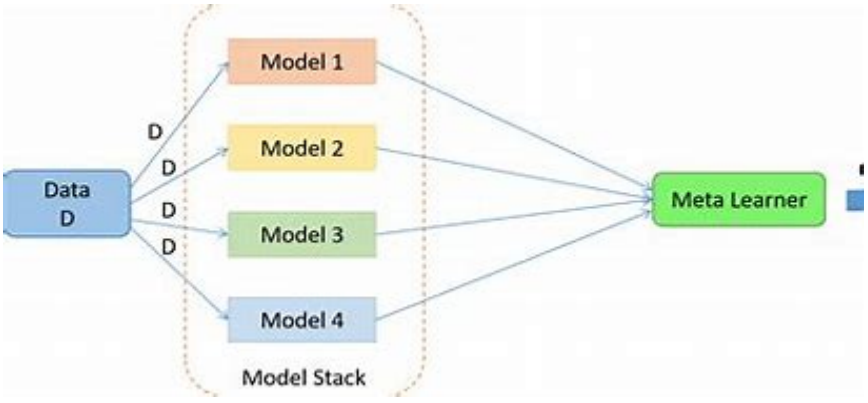


fig 3.2-Stacking of models

SHAP Analysis for Interpretability

In healthcare, model interpretability is crucial for clinician trust and decision-making. We used SHAP (SHapley Additive exPlanations) values to quantify the contribution of each feature to individual predictions. SHAP, based on game theory, provides a consistent way to explain model output by attributing the impact of each feature on the prediction.

1**. Feature Importance**: SHAP allows us to calculate feature importance by averaging the SHAP values across all predictions. This gives us a global perspective on which features are most influential in predicting cardiovascular disease risk, such as age, cholesterol level, and blood pressure. Understanding feature importance is critical for clinical applications, as it highlights key risk factors that practitioners should focus on.

2. **Individual Predictions:** SHAP values can also be used to explain individual predictions, which is especially useful in healthcare. For example, if a model predicts high cardiovascular risk for a particular patient, SHAP can show whether factors like high cholesterol or age contributed most to this prediction. This level of interpretability can help clinicians make informed decisions based on the model's recommendations.

3. **Visualization:** SHAP values can be visualized using summary plots, dependency plots, and force plots, each providing unique insights into feature contributions. Summary plots show the distribution of SHAP values for each feature, while dependency plots illustrate how changes in a feature affect the prediction. Force plots, meanwhile, display the combined impact of all features for a single prediction, helping to explain individual patient risk.

By incorporating SHAP analysis, we made our model not only accurate but also transparent and interpretable, enhancing its potential for real-world application in clinical settings.

Results and Analysis

In this section, we provide a comprehensive analysis of our findings, including the outcomes of feature selection, model evaluation, SHAP insights, and a clustering analysis to better understand patient segmentation.

Feature Selection Outcomes

We employed three feature selection methods—Chi-Squared, ANOVA F-Test, and Mutual Information—to identify the most relevant predictors of heart disease. Each method selected a slightly different set of features, revealing unique insights into which attributes were statistically significant. While each technique identified different combinations of features, certain variables consistently emerged as key indicators across all methods, such as *blood pressure* and *age*.

1. **Chi-Squared**: This method evaluates the independence of each feature with respect to the target variable. It identified features like blood pressure, cholesterol, and glucose levels as highly relevant, reflecting how these clinical indicators are strongly associated with cardiovascular health.

2. **ANOVA F-Test**: The ANOVA test, which assesses the variance between and within groups, also highlighted blood pressure and age as prominent features but additionally selected variables like body mass index (BMI), suggesting that weight-related factors play a significant role in cardiovascular risk.

3. **Mutual Information:** Mutual Information measures the dependency between each feature and the target. It identified age and smoking status as top features, as well as the number of years with high cholesterol, providing a more nuanced understanding of how lifestyle factors influence heart disease risk.

By using multiple feature selection methods, we not only gained a broader perspective on important predictors but also narrowed down the most influential variables, aiding in efficient model building and improving interpretability.

Model Evaluation

To assess model performance, we trained each classification model using the selected features from each feature selection method. The models included *Support Vector Machine (SVM), XGBoost, Random Forest, AdaBoost, and a Stacking model* combining several classifiers for enhanced predictive power. Each model was trained using a consistent 80-20 data split and evaluated on the test set.

Accuracy Scores

After training, we calculated the accuracy of each model to compare performance across combinations of feature selection methods and classifiers. Here are some key findings:

- *XGBoost and Random Forest* performed consistently well across all feature selection techniques, demonstrating their robustness in handling various data subsets.

- The *Stacking model* yielded the highest accuracy, particularly when trained on features selected by the Chi-Squared method, which aligns with our expectation that ensemble methods benefit from diversified feature selection.

- SVM showed relatively lower accuracy, likely due to the non-linear relationships in the data, which tree-based models like XGBoost and Random Forest are better equipped to handle.

These accuracy scores highlight that model performance depends not only on the choice of classifier but also on the features used, as different combinations yielded variations in accuracy.

SHAP Insights

For interpretability, we conducted a SHAP (SHapley Additive exPlanations) analysis to uncover which features most strongly influenced predictions. SHAP values allowed us to see the contribution of each feature on a local (individual prediction) and global (overall dataset) level, providing valuable interpretive insights:

1. *Age and blood pressure* emerged as dominant predictors of heart disease risk, consistent with clinical findings in cardiology. Older patients and those with elevated systolic blood pressure were shown to have higher probabilities of risk.

2. *Cholesterol levels* also ranked highly, as patients with high cholesterol exhibited increased SHAP values, indicating a higher likelihood of cardiovascular risk.

We generated detailed feature importance charts to visually demonstrate the impact of these features. Additionally, force plots provided individual-level explanations, showing how variations in critical features like age or blood pressure increased or decreased the likelihood of a positive prediction. This interpretability is essential for healthcare settings, where understanding the "why" behind predictions supports informed decision-making.

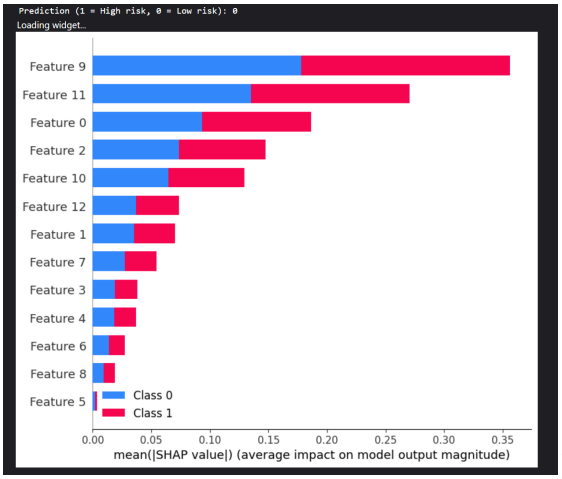
Clustering Analysis

To further explore patient segmentation, we applied *KMeans* and *DBSCAN* clustering on selected features, such as age and blood pressure. Clustering provides an additional layer of insight by grouping patients based on similar characteristics, potentially identifying high-risk subgroups within the population:

- ***KMeans Clustering*** revealed three distinct clusters. One group consisted predominantly of older patients with high blood pressure, representing a high-risk cluster. Another included younger individuals with normal blood pressure, indicating a lower risk group. This clustering can help target interventions more effectively based on patient profiles.

- ***DBSCAN Clustering*** was used as a secondary approach to validate KMeans results and to account for any outliers. DBSCAN identified similar clusters, reinforcing the presence of a high-risk group. The model's density-based approach also highlighted potential anomalies, such as patients with extreme blood pressure values, warranting further investigation.

These clustering results suggest that age and blood pressure are effective for creating meaningful risk segments within the population. Such segmentation can help medical practitioners design more personalized prevention and treatment plans, ensuring high-risk patients receive the necessary attention.



]

Conclusion & Future work

Our study confirms that machine learning models, when combined with effective feature selection techniques, can provide accurate predictions for heart disease risk.

The stacking ensemble model was the most robust classifier, and SHAP values facilitated model interpretability.

Implications:

The model, with interpretability from SHAP analysis, could be valuable for clinicians in identifying at-risk individuals based on key health metrics.

Cluster analysis showed potential in categorizing patients by risk segments, possibly aiding personalized treatment approaches.

Future Work:

Explore larger and more diverse datasets, consider real-time prediction applications, and further integrate explainability techniques to improve trust in machine learning models within the medical community.

References

"Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization" - This study used six machine learning algorithms, including Random Forest and AdaBoost, and employed techniques like GridSearchCV for hyperparameter tuning to optimize prediction accuracy. It achieved significant results on the Cleveland dataset by using an ensemble classifier for improved accuracy, achieving up to 95% accuracy in prediction. This work highlights the use of soft voting ensemble classifiers to enhance predictive power in heart disease diagnosis
MDPI
.

"Effective Heart Disease Prediction Using Machine Learning Techniques" - This research applied multiple models like Decision Tree, XGBoost, and Random Forest on a dataset of 70,000 instances. It used techniques such as k-modes clustering and multilayer perceptron for model accuracy improvement. The study also incorporated cross-validation, achieving a high accuracy rate and demonstrating the potential of machine learning in heart disease classification
MDPI
.

"Machine Learning for Heart Disease Prediction: A Review" - This review discusses various machine learning methods such as logistic regression, support vector machines, and neural networks used in predicting heart disease. It highlights how each method performs differently on diverse datasets and offers insights into the strengths and limitations of different algorithms. This is useful for understanding the foundational approaches for heart disease prediction using data-driven techniques.

"SHAP Analysis for Interpretable Heart Disease Prediction Models" - This paper focuses on using SHAP (SHapley Additive exPlanations) values for interpretability in heart disease prediction models. It emphasizes the importance of model transparency and explains how SHAP can help identify which features contribute most significantly to predictions, thus assisting healthcare providers in understanding patient-specific risks better.

"Feature Selection Techniques in Cardiovascular Disease Prediction" - This study explores the use of feature selection techniques, including ANOVA, Chi-square tests, and mutual information (MI), to improve model accuracy. Selecting the most relevant features helps in reducing complexity and enhancing the model's predictive performance, which is especially useful when working with large medical datasets.