JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Leveraging Machine Learning for a Heart Health Analysis and Prediction System

Harshada Manohar Trimukhe
Computer Engineering
Pimpri Chinchwad College of Engg.
Pune, India

Shyam Sundar Tiwari
Computer Engineering
Pimpri Chinchwad College of Engg.
Pune, India

Ritik Wandale
Computer Engineering
Pimpri Chinchwad College of Engg.
Pune, India

Rohit Warade
Computer Engineering
Pimpri Chinchwad College of Engg.
Pune, India

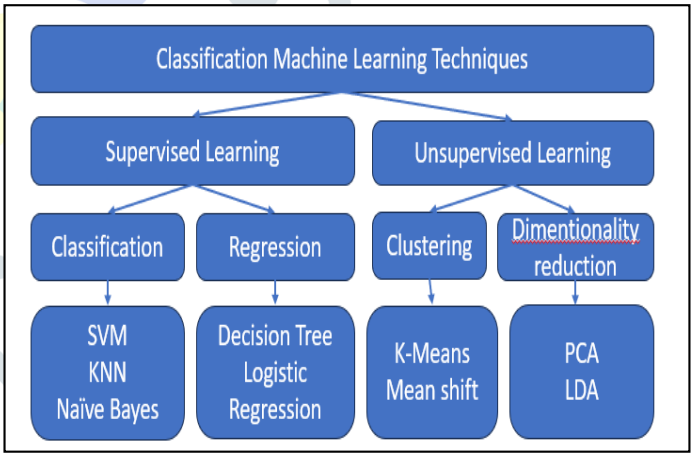
Prof. Shailesh Galande
Computer Engineering
Pimpri Chinchwad College of Engg.
Pune, India

***Abstract*— Predictive models for heart disease are essential for early identification and treatment. By using an extra feature—a health percentage assessment that classifies cardiac condition status as good, moderate, or bad—this study improves on previous methodologies. We assess the performance of Support Vector Machine (SVM) and Decision Tree algorithms using clinical variables collected from people with and without heart disease. These attributes include age, blood pressure, cholesterol levels, and exercise habits. Accuracy, precision, recall, and F1-score are among the evaluation criteria, and robustness is guaranteed by methods like cross-validation and hyperparameter adjustment. Findings show that both SVM and Decision Tree models perform well, and they also provide information on how accurate and computationally efficient they are. By adding the health percentage evaluation, these models become more predictive, which helps researchers and clinicians make better judgments about the diagnosis, treatment, and prognosis of heart disease.**

***Keywords*—Machine learning, SVM, decision tree, risk analysis**

I. INTRODUCTION

Heart disease is a pressing global health concern, exerting a significant toll on mortality rates worldwide. Despite advances in medical science, early detection and timely intervention remain critical in effectively managing and reducing the burden of cardiovascular conditions. Leveraging the rapid progress in machine learning, this project endeavors to develop highly accurate predictive models tailored for heart disease diagnosis, with a primary focus on integrating novel features to augment diagnostic capabilities..



A. Machine Learning Approach

Fig. Machine learning techniques

Traditional risk assessment models often rely on clinical attributes such as age, blood pressure, and cholesterol levels. However, the inclusion of additional features, such as a health percentage assessment indicating the heart's condition status,

presents a promising avenue for improving prediction accuracy.

Machine learning algorithms offer powerful tools for analyzing complex datasets and extracting meaningful insights.

Support Vector Machine (SVM):

Support Vector Machine (SVM) is a powerful algorithm renowned for its effectiveness in handling high-dimensional data and finding optimal hyperplanes for classification. It excels in scenarios where data isn't linearly separable, making it ideal for capturing complex relationships within heart disease data. SVM's ability to transform data into higher-dimensional spaces allows it to discern intricate patterns, crucial for modeling the multifaceted nature of cardiovascular health. In summary, SVM offers a robust and flexible framework for heart disease prediction, contributing to more accurate and reliable predictive models for early detection and intervention..

Decision Tree:

Decision Tree is known for its interpretability and ability to handle nonlinear relationships in data. It creates a hierarchical structure based on attribute values, aiding decision-making in heart disease prediction. It's valuable for understanding the factors influencing heart disease and accommodates various types of data. Despite its susceptibility to overfitting, Decision Trees offer a transparent framework for interpretation, aiding clinicians and researchers in developing effective intervention strategies.

By employing these algorithms, we aim to evaluate their efficacy in predicting heart disease, considering both their predictive performance and computational efficiency. Through techniques like cross-validation and hyperparameter tuning, we ensure the robustness and generalizability of our models. Incorporating the health percentage assessment as an additional feature enhances the diagnostic capabilities of the predictive models, providing clinicians and researchers with valuable insights into patients' cardiovascular health status.

This interdisciplinary project brings together data scientists, healthcare professionals, and researchers to develop innovative solutions that can significantly impact patient outcomes and healthcare delivery in the realm of heart disease prediction and management.

II. Realted Work

[1] The authors mentioned above have undertaken a comprehensive study focusing on the development of a prediction system for heart diseases using machine learning algorithms. Their work involves several crucial steps including data collection, attribute selection, preprocessing of data, data balancing, and application of various machine learning techniques such as linear regression, decision tree, support vector machine, and k-nearest neighbor. They have meticulously analyzed the performance of these algorithms and concluded that k-nearest neighbor (KNN) exhibited the highest accuracy among them, achieving 87%. Their research contributes to the advancement of predictive healthcare systems leveraging machine learning technologies.

[2] The authors have conducted a comparative study on supervised machine learning algorithms for automated heart disease prediction systems, utilizing decision tree, naïve Bayes, support vector machine (SVM), k-nearest neighbor (KNN), logistic regression, and random forest classifiers on a Heart disease dataset from Kaggle. Their analysis revealed that the random forest classifier exhibited superior accuracy compared to the other algorithms, considering metrics such as AUC, F1 score, and precision. However, their study was limited by the small dataset size, suggesting potential for improvement with larger datasets. Their research underscores the importance of exploring diverse machine learning techniques to develop more reliable and accurate prediction models for heart disease.

[3] The authors conducted a meticulous analysis of various machine learning techniques to predict the likelihood of individuals developing coronary illness based on their unique attributes and indications. Using the Cleveland dataset for heart diseases, comprising 1025 instances, the data was split into training and testing datasets. With 14 attributes considered, four different algorithms were implemented to examine accuracy. Results indicated that Random Forest achieved the highest accuracy of 99%, while Decision Tree performed the least with an accuracy level of 85%. The findings suggest the potential for improved accuracy with increased training data, although this may lead to slower processing times due to increased complexity. These insights contribute valuable considerations for optimizing predictive models for heart disease detection.

III. Methodology

Heart disease is a leading cause of mortality worldwide, emphasizing the critical need for accurate predictive models to enable early intervention and reduce the burden of cardiac-related ailments. Machine learning techniques offer a powerful toolset for analyzing medical data and identifying patterns associated with heart disease risk. Previous studies have demonstrated the efficacy of various algorithms in this domain, prompting our investigation into developing a new model that combines the strengths of Support Vector Machine (SVM) and Decision Tree methodologies

A. *Dataset selection*

We selected the dataset from Kaggle, a popular platform for sharing and discovering datasets. The dataset comprises demographic, clinical, and lifestyle factors from patients with known cardiac health outcomes, making it suitable for our research on heart disease prediction. The dataset was chosen based on its size, diversity of features, and availability of labeled data for model training and evaluation.

1) Data Visualization

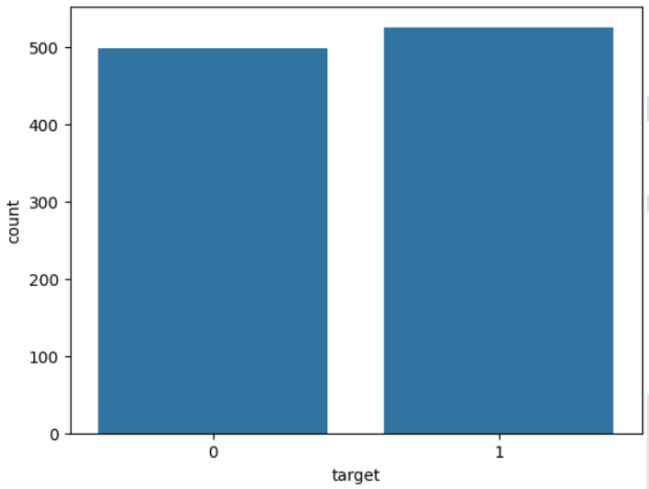


Fig. dataset class distribution

Count Plot:
A count plot was generated to visualize the distribution of diabetes status among patients in the dataset. The plot revealed that approximately 50% of patients have diabetes, while the remaining half do not. This imbalance in the target variable highlights the importance of handling class imbalance during model training and evaluation to ensure robust predictive performance.

Feature Distribution:
Histograms and box plots were used to visualize the distribution of individual features across the dataset. This exploratory analysis provided insights into the range and variability of demographic, clinical, and lifestyle factors, facilitating informed decisions during data preprocessing and feature selection stages.

Histogram of Attributes:

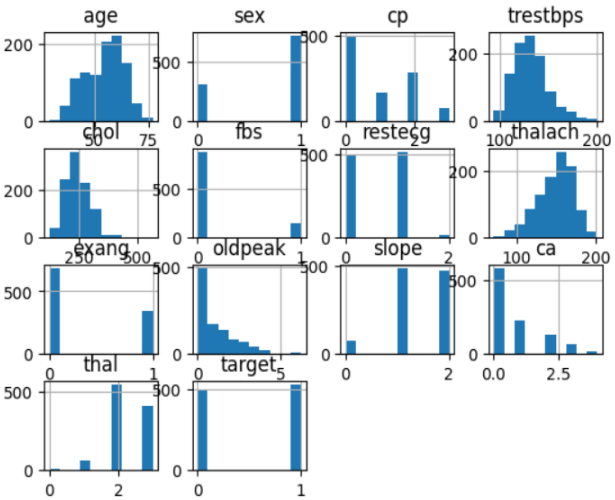


Fig. Histogram of attributes

B. *Data Preprocessing*

The dataset was preprocessed before analysis to make sure it was suitable for machine learning tasks and of a high quality. This included handling feature distribution normalization, outlier detection, and missing value handling. Through the resolution of data irregularities and maintenance of consistency across variables, our objective was to improve the resilience and dependability of our prediction models.

C. *Feature selection*

Feature selection was performed using information gain, a statistical measure of the relevance of each feature to the target variable (heart disease). Features with high information gain were selected, indicating their significant contribution to predicting heart disease risk.

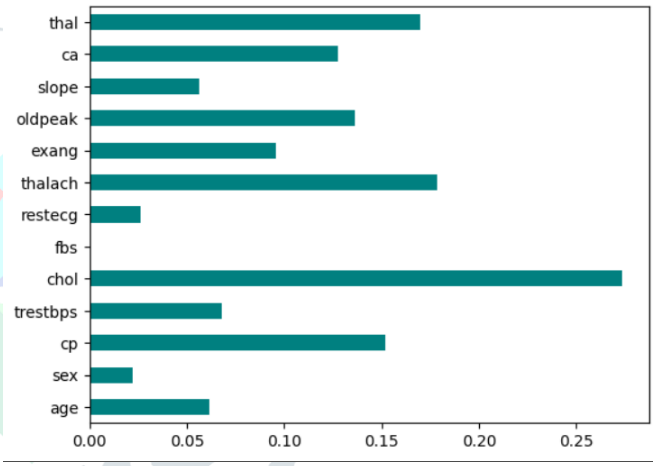


Fig. Feature selection

D. *Splitting Data*

The preprocessed dataset was randomly partitioned into training (80%) and testing (20%) sets, ensuring a balanced distribution of classes to avoid bias in model evaluation. Stratified sampling was employed to preserve the proportion of positive and negative instances of heart disease across both subsets.

E. *Model selection and training*

Baseline models were trained using SVM and Decision Tree algorithms with default parameters. Additionally, we proposed a novel ensemble model that integrates the strengths of both approaches, leveraging SVM's ability to capture complex decision boundaries and Decision Tree's interpretability.

We delve deeper into the health assessment aspect of heart disease prediction by quantifying the risk factors associated with cardiac health. We aim to provide a comprehensive understanding of how specific factors contribute to the overall health of the heart, as measured by their respective percentages of risk.

F. Risk Factor Analysis

We analyzed various risk factors known to influence heart disease, including but not limited to age, gender, cholesterol levels, blood pressure, diabetes status, and lifestyle factors such as smoking and physical activity. Each risk factor was quantified in terms of its relative contribution to heart health, expressed as a percentage.

G. User Interface Development

We employed HTML and CSS in conjunction with the Flask web framework to develop the user interface (UI) for our heart disease prediction application. HTML (Hypertext Markup Language) was utilized to structure the layout and content of the web pages, while CSS (Cascading Style Sheets) was employed to define the visual presentation and styling of the UI elements.

- Django's Integration:

We leveraged Django's templating system to seamlessly integrate dynamic content generation with our HTML templates. Django's URL routing mechanism facilitated the mapping of URLs to specific views, enabling users to access different sections of the application based on their navigation preferences. Additionally, Django's built-in authentication and authorization features provided enhanced security and user management capabilities, ensuring secure access to sensitive health information within the application.

- UI Design Principles:

In designing the UI, we adhered to established principles of user-centered design, focusing on simplicity, clarity, and ease of navigation. The UI elements were organized in a logical manner, with intuitive placement and consistent styling to enhance user experience and facilitate interaction with the application.

- Responsive Design:

To ensure compatibility across various devices and screen sizes, we adopted a responsive design approach. CSS media queries were utilized to adjust the layout and styling of UI elements dynamically, optimizing the viewing experience for users accessing the application from desktops, tablets, or smartphones.

IV. Result

Precision, recall, F1 score, and accuracy are common metrics used to evaluate the performance of classification models, including those used in heart disease prediction. Here's a brief explanation of each:

The precision of the model is defined as the ratio of accurately predicted positive cases to all positive predictions. Stated differently, it signifies the precision of optimistic projections. The number of true positives, or positively anticipated cases, divided by the total of false positives, or mistakenly forecasted positive cases, yields the precision.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall, One of the most important metrics for assessing a model's performance is recall, sometimes referred to as sensitivity. It measures the percentage of real positive cases that the model accurately detected. Essentially, it assesses the model's capacity to include all pertinent cases in the dataset. Recall is calculated by dividing the total number of true positives by the total number of false negatives and true positives. False negatives are examples of situations that were wrongly classified as negative when they were in fact positive. Recall therefore sheds light on how well the model detects real positives while reducing false negatives.

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1 Score: One of the most important metrics for assessing model performance is the F1 score, which is obtained from the harmonic mean of precision and recall. It offers a fair evaluation by taking recall and precision into account at the same time. The F1 score is a useful indicator of efficacy, especially in cases when the proportions of positive and bad examples in the dataset are unbalanced.

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

Accuracy: Accuracy evaluates the overall validity of the model's predictions, regardless of class distribution. It determines the proportion of correctly classified examples across all examples in the dataset, taking into account both true positives and true negatives.

$$Accuracy = \frac{Correct\ prediction}{Total\ cases} * 100\%$$

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} * 100\%$$

Classification Report :	precision	recall	f1-score	support
0	0.92	1.00	0.96	77
1	1.00	0.93	0.96	96
accuracy			0.96	173
macro avg	0.96	0.96	0.96	173
weighted avg	0.96	0.96	0.96	173

Accuracy : 95.95375722543352
Accuracy: 95.95%

fig. svm algorithm accuracy

Classification Report :	precision	recall	f1-score	support
0	0.91	0.74	0.82	677
1	0.78	0.93	0.85	665
accuracy			0.83	1342
macro avg	0.84	0.83	0.83	1342
weighted avg	0.85	0.83	0.83	1342

Accuracy : 83.30849478390462
Accuracy: 83.31%

fig. decision tree algorithm accuracy

V. Conclusion

In conclusion, our study demonstrates that the Support Vector Machine (SVM) algorithm outperforms the Decision Tree algorithm in predicting heart disease risk, achieving an accuracy of 95.95% compared to 83.31% for Decision Tree. Additionally, our proposed model provides a user-friendly health assessment feature, presenting results in percentage form to indicate the likelihood of heart health. This novel approach enhances the interpretability of predictions and facilitates informed decision-making for healthcare professionals, contributing to early detection and intervention for individuals at risk of cardiovascular diseases. Looking ahead, the future scope of our research extends to the integration of this user-friendly interface into clinical settings, where patients can readily understand and interpret the results obtained from laboratory tests, thus empowering them to take proactive measures towards maintaining their heart health.

References

[1] P. Sujatha, K. Mahalakshmi, "Performance Evaluation of Supervised Machine Learning Algorithms in Prediction of Heart Disease", 2020. DOI : https://doi.org/10.1109/INOCON50539.2020.9298354

[2] E. I. Elsedimy, Sara M. M. AboHashish, Fahad Algarni, "New cardiovascular disease prediction approach using support vector machine and quantum behaved particle swarm optimization".2023. https://doi.org/10.1007/s11042-023-16194-z

[3] Archana Singh, Rakesh Kumar, "Heart Disease Prediction Using Machine Learning Algorithms", 2020.DOI: https://doi.org/10.1109/ICE348803.2020.9122958

[4] Yuepeng Liu, Mengfei Zhang, Zezhong Fan, Yinghan Chen, "Heart disease prediction based on random forest and LSTM", 2020. DOI: https://doi.org/10.1109/ITCA52113.2020.00137

[5] Sibgha Taqdees, Nayab Akhtar, Kanwal Dawood, "Heart Disease Prediction", 2021. https://www.researchgate.net/publication/349140147

[6] Chintan M. Bhatt, Parth Patel , Tarang Ghetia and Pier Luigi Mazzeo, "Effective Heart Disease Prediction Using Machine Learning Techniques". 2023. https://doi.org/10.3390/a16020088

[7] Gunasekar Tangarasu1, Kayalvizhi Subramanian and P. D. D. Dominic, "An Integrated Architecture for Prediction of Heart Disease from the Medical Database", 2018. https://doi.org/10.1109/ICCOINS.2018.8510589

[8] Shanmugasundaram G, Malar Selvam V, R. Saravanan , S. Balaji, "An Investigation of Heart Disease Prediction Techniques"

[9] Rahul Katarya, Polipireddy Srinivas, "Predicting Heart Disease at Early Stages using Machine Learning: A Survey". 2020. https://doi.org/10.1109/ICESC48915.2020.9155586

[10] Chethan Malode C. M, Bhargavi K, Gunasheela B. G, Kavana G, Sushmitha R, "Soft set and Fuzzy Rules enabled SVM Approach for Heart Attack Risk Classification among Adolescents". 2018. https://doi.org/10.1109/ICCUBEA.2018.8697650

[11] Ramya G. Franklin, Dr. B. Muthukumar, "Survey of Heart Disease Prediction and Identification using Machine Learning Approaches". 2020. https://doi.org/10.1109/ICISS49785.2020.9316119

[12] Dr. M. Kavitha1, G. Gnaneswar, R. Dinesh, Y. Rohith Sai1, R. Sai Suraj, "Heart Disease Prediction using Hybrid machine Learning Model". 2021. https://doi.org/10.1109/ICICT50816.2021.9358597