Prediction Of Diseases in Smart Healthcare System Using Machine Learning

Nadjem Bailek[1], Mohamed Saber *[2]

[1] Energies and Materials Research Laboratory, Faculty of Sciences and Technology, University of Tamanghasset, Tamanrasset, 10034, Algeria.
[1]Sustainable Development and Computer Science Laboratory, Faculty of Sciences and Technology, Ahmed Draia University of Adrar, Adrar, Algeria
[2] Electronics and Communications Engineering Department, Faculty of Engineering, Delta University for Science and Technology, Gamasa City 11152, Egypt
Emails: bailek.nadjem@univ-adrar.edu.dz; Mohamed.Saber@deltauniv.edu.eg

Abstract

Smart healthcare systems rely heavily on disease prediction because it paves the way for early detection and prompt action, both of which enhance patient outcomes. In this research, we present a machine learning (ML) method for identifying data patterns that might be used to foretell the occurrence of cardiac disease. Our approach entails cleaning the data used for predicting cardiac issues and then using a Support Vector Machine (SVM). Age, sex, chest pain type, blood pressure, cholesterol, and exercise-induced angina are only few of the attributes included in the dataset. Insights into the distributional analysis of categorical and numeric variables, as well as potential connections and trends, are gained through exploratory data analysis (EDA). Cross-validation results show that the SVM model performs exceptionally well, with higher accuracy and AUC than competing models. By utilizing ML methods, our research aids in the development of intelligent healthcare systems. These results add to our understanding of how to forecast diseases and show how machine learning may transform healthcare systems to improve patient outcomes.

Keywords: Machine learning; Smart healthcare; Disease prediction.

1. Introduction

In recent years, the advancement of technology has revolutionized the healthcare industry, paving the way for innovative approaches to patient care and management. Smart healthcare systems, empowered by the integration of various technological components such as Internet of Things (IoT) devices, electronic health records (EHRs), and wearable sensors, hold tremendous potential in transforming healthcare delivery. In parallel, machine learning, a subset of artificial intelligence, has emerged as a powerful tool for analyzing vast amounts of healthcare data and extracting valuable insights. One particularly promising application of machine learning in healthcare is disease prediction, which aims to identify individuals at risk of developing specific medical conditions.

The motivation behind this study stems from the pressing need to enhance disease prediction capabilities within smart healthcare systems. Early identification of diseases can significantly impact patient outcomes, as it allows for timely intervention and personalized treatment plans. By harnessing the power of machine learning algorithms and leveraging the abundance of data generated in smart healthcare environments, we can potentially uncover hidden patterns, risk factors, and predictive indicators that facilitate early disease detection.

Doi: https://doi.org/10.54216/JAIM.030205
Received: August 19, 2022 Revised: November 15, 2022 Accepted: March 25, 2023

This paper aims to provide a comprehensive overview of disease prediction in smart healthcare systems using machine learning techniques. To set the stage, we will begin by delving into the background and rationale for undertaking this research. Subsequently, we will provide an overview of smart healthcare systems and highlight their significance in transforming the way healthcare is delivered. Furthermore, a concise explanation of machine learning and its potential in disease prediction will be presented, emphasizing the role of data-driven algorithms in leveraging healthcare data for improved predictive models.

The primary contribution of this paper lies in the consolidation and synthesis of existing knowledge and research findings related to disease prediction in smart healthcare systems using machine learning. By critically reviewing and analyzing relevant studies, we aim to identify the strengths and limitations of current approaches and address any existing gaps in literature. Additionally, we will discuss the potential implications of our findings, as well as highlight future directions for research in this dynamic field.

The organization of this paper is as follows: the relevant studies are reviewed in the next section. Following this, we will delve into the details of the proposed machine learning algorithms commonly used in this context. After that, we provide the presentation and interpretation of results. Finally, we discuss the main findings of our work.

2. Literature Review

In this section, we present a comprehensive review of the existing body of work related to disease prediction in smart healthcare systems using machine learning techniques. Numerous studies have explored the application of machine learning algorithms in this domain, aiming to improve disease identification, early detection, and personalized healthcare delivery. Boukenze et al. [1] presented a study on predictive analytics in the healthcare system using data mining techniques. The authors explored the application of data mining algorithms to extract valuable insights from healthcare data, aiming to improve disease prediction and healthcare decision-making. Qureshi et al. [2] proposed an accurate and dynamic predictive model for a smart m-health system using machine learning. Their research focused on developing a robust model that can adapt to changing health conditions, leveraging machine learning algorithms to improve disease prediction and enable personalized healthcare services. Shereen Zaki et al. [3] introduced the use of the interval-valued neutrosophic VIKOR method for assessing green suppliers in the supply chain. While not directly related to disease prediction in healthcare, this study provides insights into the application of advanced computational intelligence techniques in evaluating and selecting relevant stakeholders in healthcare systems. Ray and Chaudhuri [4] discussed smart healthcare disease diagnosis and patient management. Their research highlighted the innovation, improvement, and skill development required in disease diagnosis and patient management within smart healthcare systems, emphasizing the potential impact of advanced technologies, including machine learning. Venkatesh et al. [5] developed a big data predictive analytics model for disease prediction using machine learning techniques. Their study focused on utilizing large-scale healthcare data to build predictive models that can identify individuals at risk of specific diseases, enabling proactive interventions and personalized healthcare. Tuli et al. [6] proposed HealthFog, an ensemble deep learning-based smart healthcare system for automatic diagnosis of heart diseases in integrated IoT and fog computing environments. Their research emphasized the integration of IoT devices, fog computing, and deep learning techniques to enable efficient and accurate diagnosis of heart diseases. Mahmoud and Abdelhafeez [7] presented a computational intelligence approach for biometric gait identification. Although not directly related to disease prediction, this study showcases the application of computational intelligence techniques in healthcare and biometric identification, which can potentially be applied to enhance disease prediction systems. Chhabra and Kumar [8] proposed a smart healthcare system based on the Classifier Dense Net 121 model for detecting multiple diseases. Their research focused on leveraging deep learning techniques for accurate disease detection and diagnosis within a smart healthcare environment. Barak-Corren et al. [9] validated an electronic health record-based suicide risk prediction modeling approach across multiple healthcare systems. Although not specific to disease prediction, this study highlights the potential of leveraging electronic health records and predictive modeling techniques to assess and mitigate mental health risks.

Doi: https://doi.org/10.54216/JAIM.030205
Received: August 19, 2022 Revised: November 15, 2022 Accepted: March 25, 2023

By discussing these related works, your paper can provide a comprehensive overview of the advancements, methodologies, and findings in disease prediction within smart healthcare systems using machine learning techniques.

3. Case study and preparation

Our experiments were conducted on a heart failure prediction dataset, which encompasses a comprehensive set of attributes related to patients' health and diagnostic indicators. The dataset combines information from five previously independent datasets, namely Cleveland, Hungarian, Switzerland, Long Beach VA, and Stalog (Heart) datasets, resulting in a total of 918 observations after removing duplicates. This compilation makes it the largest heart disease dataset available for research purposes. The attribute information of the dataset includes various clinically relevant features. These features consist of the patient's age, sex, chest pain type, resting blood pressure, serum cholesterol level, fasting blood sugar, resting electrocardiogram results, maximum heart rate achieved, exercise-induced angina, ST depression, and the slope of the peak exercise ST segment. The target variable, Heart Disease, indicates the presence or absence of heart disease. To ensure the dataset's reliability and validity, the source data was obtained from reputable sources, specifically the UCI Machine Learning Repository [10]. By combining multiple datasets, we have created a robust and diverse dataset that offers a comprehensive view of heart disease-related attributes. The utilization of this extensive dataset in our experiments allows us to derive meaningful insights and validate the effectiveness of our proposed semantic approach for extracting medical association rules. By analyzing the relationship between these attributes and heart disease, we can provide valuable knowledge to aid in the prediction and understanding of heart failure. In Table 1, we present a sample of the heart disease dataset, showcasing a subset of the observations and their corresponding attribute values. This sample provides a glimpse into the structure and content of the dataset, highlighting the diverse range of variables such as age, sex, chest pain type, resting blood pressure, cholesterol levels, and other relevant features.

Table 1: Samples of heart prediction data in our case study

	Age	Sex	Chest Pain Type	Resting BP	Cholesterol	Fasting BS	Resting ECG	Max HR	Exercise Angina	Old peak	ST_Slope	Heart Disease
0	40	M	ATA	140	289	0	Normal	172	N	0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0	Up	0

In Table 2, we present a descriptive analysis of the heart disease dataset. This analysis includes various statistical measures and summaries that offer insights into the distribution and characteristics of the dataset. For example, we may include measures such as mean, standard deviation, minimum and maximum values, as well as quartiles for numerical attributes.

Table 2: Descriptive analysis of heart prediction data in our case study

	count	mean	std	min	25%	50%	75%	max
Age	918	53.51	9.43	28	47	54	60	77
Resting BP	918	132.4	18.51	0	120	130	140	200
Cholesterol	918	198.8	109.38	0	173.25	223	267	603
Fasting BS	918	0.23	0.42	0	0	0	0	1
Max. HR	918	136.81	25.46	60	120	138	156	202
Old peak	918	0.89	1.07	-2.6	0	0.6	1.5	6.2
Heart Disease	918	0.55	0.5	0	0	1	1	1

Doi: https://doi.org/10.54216/JAIM.030205
Received: August 19, 2022 Revised: November 15, 2022 Accepted: March 25, 2023

4. Proposed Intelligent Disease Predictor

In our methodology, we employ the Support Vector Machine (SVM) algorithm as a predictive model for heart disease classification. SVM is a powerful machine learning technique widely used for both classification and regression tasks. The goal of SVM is to find an optimal hyperplane that maximally separates the data points of different classes in the feature space. Mathematically, given a training dataset with input samples X and corresponding class labels $\boldsymbol{Y}$, where $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ and $\boldsymbol{Y} = \{y_1, y_2, \ldots, y_n\}$, SVM aims to find a hyperplane defined by the equation $\boldsymbol{w} \cdot \boldsymbol{x} - b = 0$ that effectively separates the samples belonging to different classes [11].

The SVM algorithm operates by transforming the input samples into a higher-dimensional feature space using a kernel function. This transformation allows SVM to find a hyperplane that linearly separates the samples in this transformed space. The choice of the kernel function, such as linear, polynomial, or radial basis function (RBF), influences the SVM's ability to capture complex nonlinear relationships within the data. To predict heart disease using SVM, we first preprocess the heart disease dataset by appropriately scaling the input features and encoding categorical variables. We then split the preprocessed data into training and testing sets, ensuring that both sets maintain a similar distribution of class labels [12].

Next, we train the SVM model on the training set, aiming to find the optimal hyperplane that maximizes the margin between the classes while minimizing the classification error. This optimization problem can be formulated as a quadratic programming task, which involves solving a mathematical optimization problem to determine the optimal values of the model's parameters.

5. Results

In this section, we conducted exploratory data analysis (EDA) to gain insights into the heart disease dataset. In Figure 1, we presented the distributional analysis of categorical features. By visualizing the distributions of these features, we can observe the data follow normal distribution. This analysis allows us to understand the prevalence of different categorical variables and their potential relationship with the target variable.

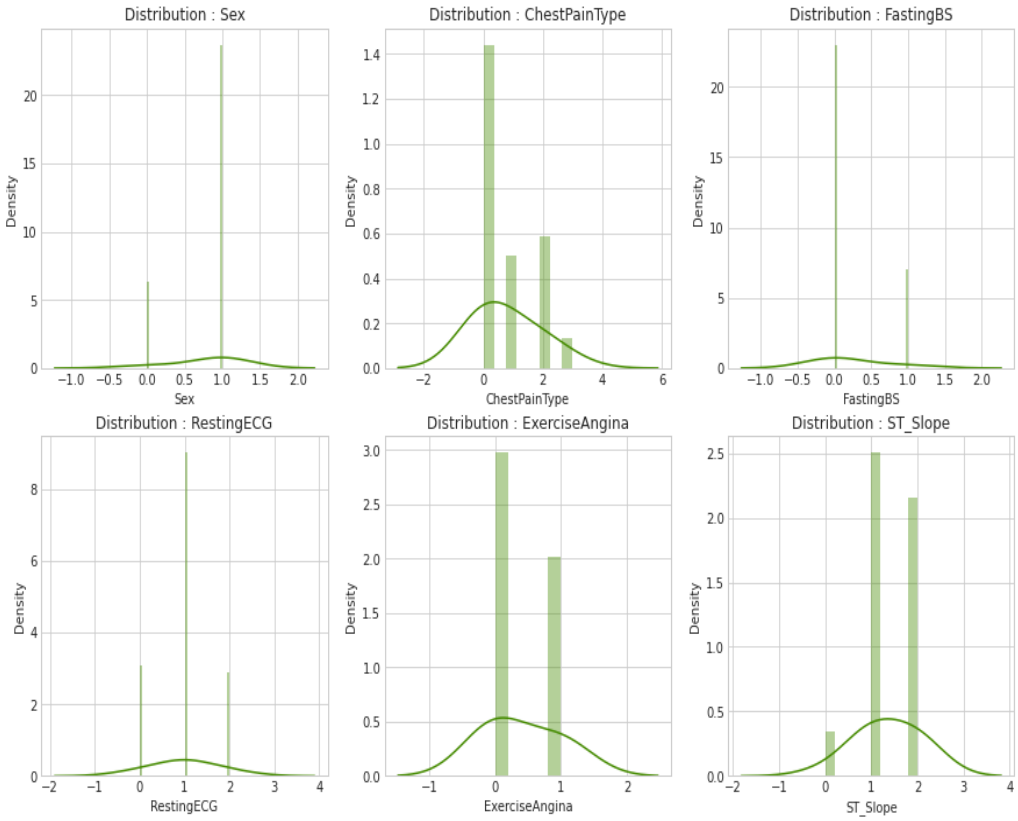

Figure 1: Distributional analysis of categorical features in our case study of heart data

Doi: https://doi.org/10.54216/JAIM.030205
Received: August 19, 2022 Revised: November 15, 2022 Accepted: March 25, 2023

Furthermore, in Figure 2, we performed a distributional analysis of the numerical features in the heart disease dataset. Through visualizations, such as histograms or box plots, we examined the distributions of these numerical variables. This analysis helps us identify key statistical properties such as central tendency, spread, skewness, and potential outliers. By understanding the distributional characteristics of the numerical features, we can gain insights into their potential impact on heart disease prediction.

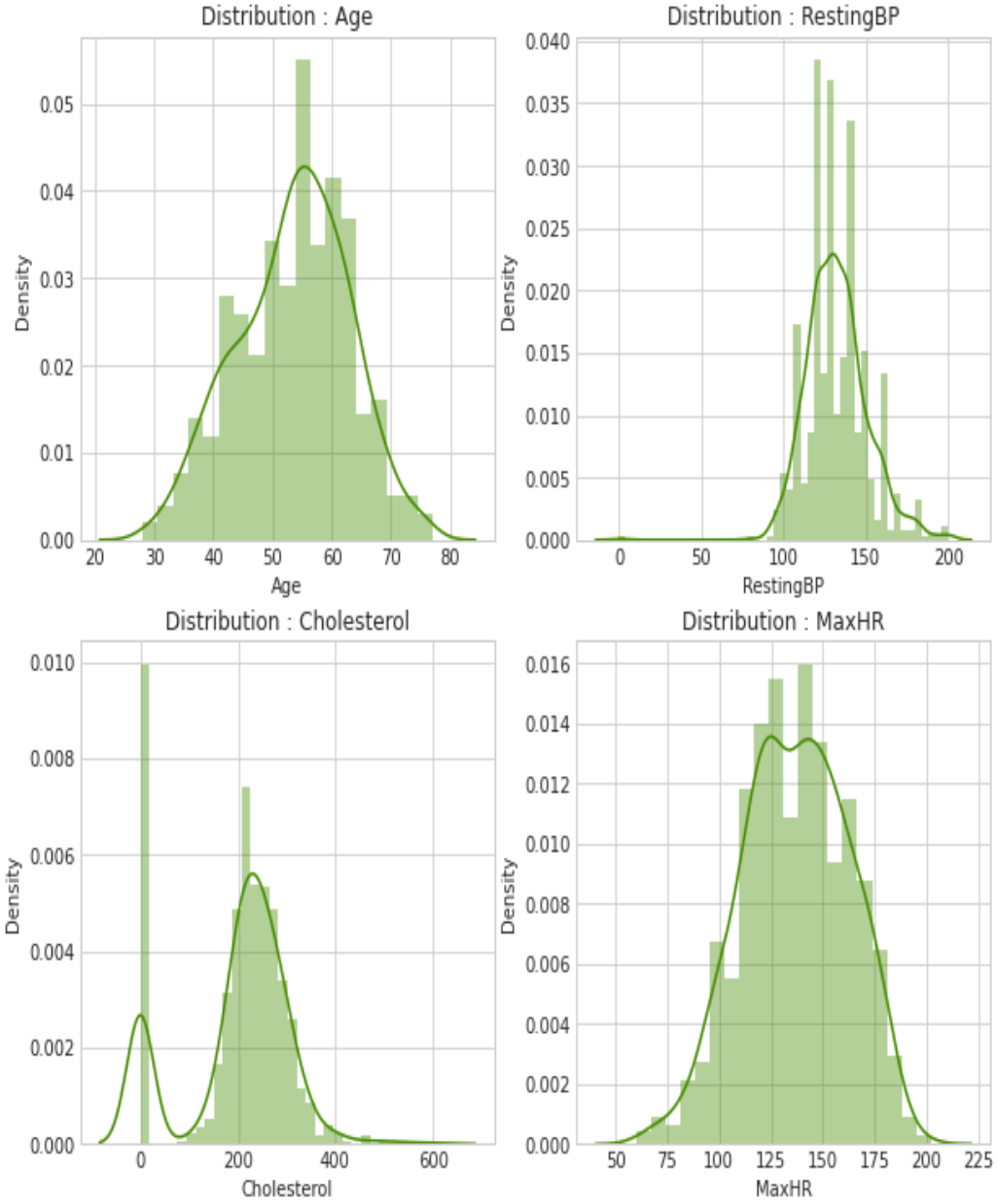

Figure 2: Distributional analysis of numerical features in our case study of heart data

In Figure 3, we presented a visualization of the categorical features against the target variable. This visualization provides a graphical representation of the relationship between the categorical variables and the presence or absence of heart disease. By examining the distribution of the target variable across different categories, we can assess the potential associations or patterns between the categorical features and the occurrence of heart disease.

Doi: https://doi.org/10.54216/JAIM.030205
Received: August 19, 2022 Revised: November 15, 2022 Accepted: March 25, 2023

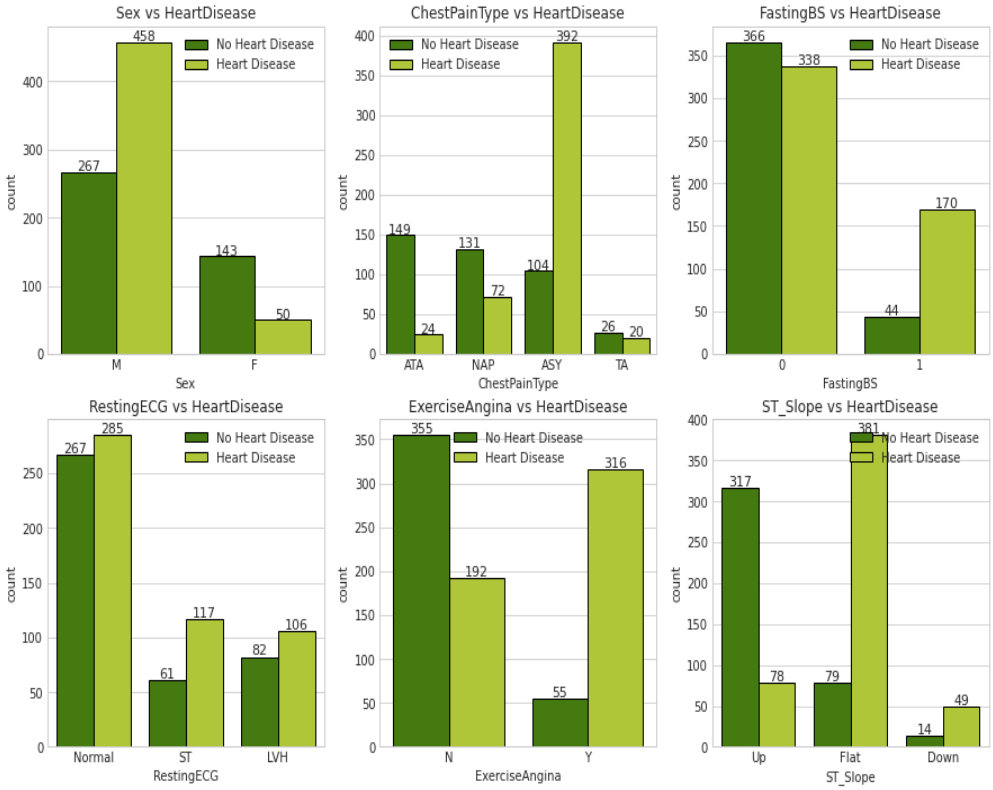

Figure 3: visualization of categorical features vs target variable

In Figure 4, we conducted a visual analysis of the relationship between the categorical features and numerical features with respect to the target variable in our heart disease dataset. This visualization provides a comprehensive view of how the categorical and numerical variables interact and potentially influence the occurrence of heart disease.

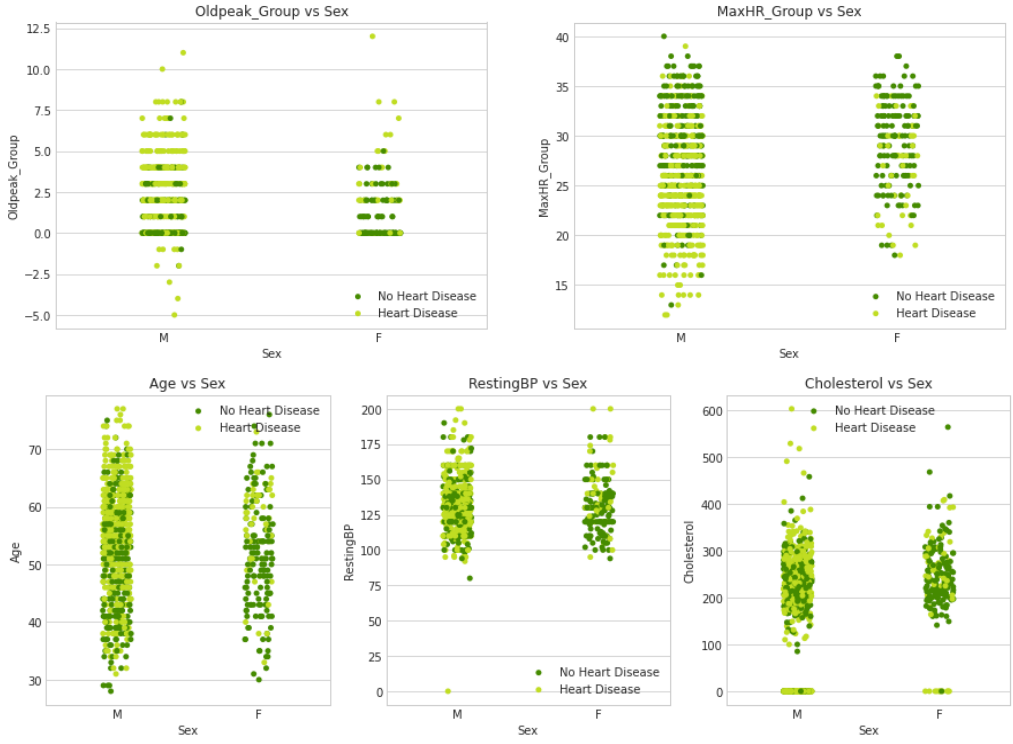

Figure 4: visualization of categorical features vs numerical ones w.r.t target variable

Doi: https://doi.org/10.54216/JAIM.030205
Received: August 19, 2022 Revised: November 15, 2022 Accepted: March 25, 2023

We may see whether there are any trends or patterns between the categorical features and the target variable by graphing them against one another. By conducting this study, we can better understand the relationships between the various feature categories and their influence on heart disease prediction. This study clarifies the relationship between the categorical and numeric variables, providing insight into the role played by various predictors in the development of heart disease. It permits us to discover any trends or patterns within the information, which in turn guides future analysis and may affect the evolution of our heart disease prediction model.

In Table 3, we present the results of cross-validation for evaluating the performance of our proposed SVM-based predictor in predicting heart disease. Cross-validation is a widely used technique in machine learning that helps assess the model's performance by partitioning the dataset into multiple subsets and iteratively training and testing the model on different combinations of these subsets. Based on the results presented in Table 3, it is evident that our proposed SVM-based predictor achieves the best performance among the evaluated models. The high values of accuracy indicate the predictor's ability to accurately classify instances of heart disease and effectively discriminate between positive and negative cases. The superior performance of our SVM-based predictor can be attributed to its ability to capture complex patterns and relationships within the heart disease dataset. By leveraging the mathematical principles and optimization techniques of SVM, our predictor can find an optimal hyperplane that maximally separates the instances belonging to different classes, leading to accurate and reliable predictions.

Table 3: The results of Experimental comparison among different ML algorithms

Sr. No.	ML Algorithm	Accuracy	Cross Validation Score	ROC AUC Score
1	Logistic Regression	87.50%	91.12%	87.55%
2	Support Vector Classifier	87.50%	92.53%	88.81%
3	Decision Tree Classifier	84.78%	89.09%	84.62%
4	Random Forest Classifier	84.24%	91.91%	84.06%
5	K-Nearest Neighbors Classifier	81.52%	89.34%	81.36%

6. Conclusion

In this study, we successfully predicted cases of disease using Support Vector Machine (SVM) as the predictive model. Our approach to using SVM for heart disease prediction included preparing the dataset, engaging in EDA, and deploying SVM. The EDA's distributional analysis of categorical and numeric variables revealed possible correlations and trends, which was a noteworthy discovery. The features included in the case study for heart failure prediction included age, sex, chest pain kind, blood pressure, cholesterol level, and exercise-induced angina. In this research, we show that our semantic approach combined with SVM greatly improves the accuracy with which we can anticipate cardiac issues. The results shed light on the utility of association rule mining methods for analyzing healthcare data and making decisions.

Funding: "This research received no external funding"

Conflicts of Interest: "The authors declare no conflict of interest."

Reference

[1] Boukenze, Basma, Hajar Mousannif, and Abdelkrim Haqiq, Predictive analytics in healthcare system using data mining techniques. Comput Sci Inf Technology, 1, 1-9, 2016.
[2] Qureshi, Kashif Naseer, Sadia Din, Gwanggil Jeon, and Francesco Piccialli, An accurate and dynamic predictive model for a smart M-Health system using machine learning. Information Sciences, 538, 486-502, 2020.

Doi: https://doi.org/10.54216/JAIM.030205
Received: August 19, 2022 Revised: November 15, 2022 Accepted: March 25, 2023

[3] Shereen Zaki, Mahmoud M. Ibrahim , Mahmoud M. Ismail, Interval Valued Neutrosophic VIKOR Method for Assessment Green Suppliers in Supply Chain. Journal of International Journal of Advances in Applied Computational Intelligence, 2 (1), 15-22, 2022.
[4] Ray, Arkadip, and Avijit Kumar Chaudhuri, Smart healthcare disease diagnosis and patient management: Innovation, improvement and skill development. Machine Learning with Applications, 3, 2021.
[5] Venkatesh R., C. Balasubramanian, and Madasamy Kaliappan, Development of big data predictive analytics model for disease prediction using machine learning technique. Journal of medical systems, 43, 1-8, 2019.
[6] Tuli Shreshth, et al., Health Fog: An ensemble deep learning based Smart Healthcare System for Automatic Diagnosis of Heart Diseases in integrated IoT and fog computing environments. Future Generation Computer Systems, 104, 187-200, 2020.
[7] Hadeer Mahmoud, Ahmed Abdelhafeez, Computational Intelligence Approach for Biometric Gait Identification. Journal of International Journal of Advances in Applied Computational Intelligence, 2(1), 36-43, 2023.
[8] Chhabra, Mohit, and Rajneesh Kumar, A Smart Healthcare System Based on Classifier DenseNet 121 Model to Detect Multiple Diseases. In Mobile Radio Communications and 5G Networks: Proceedings of Second MRCN 2021 Singapore: Springer Nature Singapore, 297-312, 2022.
[9] Barak Corren et al., Validation of an electronic health record–based suicide risk prediction modeling approach across multiple health care systems. JAMA network open 3(3), 2020.
[10] Heba R. Abdelhady, Mahmoud M. Ismail, Cardiovascular Diseases Forecasting using Machine Learning Models. International Journal of Advances in Applied Computational Intelligence, 1(2), 56-62, 2022.
[11] Mohamed Saber, A novel design and Implementation of FBMC transceiver for low power applications, Indonesian Journal of Electrical Engineering and Informatics (IJEEI), 8(1), 83-93, 2020.
[12] Abouelatta, Mohamed A., Sayed A. Ward, Ahmad M. Sayed, Karar Mahmoud, Matti Lehtonen, and Mohamed MF Darwish, Measurement and assessment of corona current density for HVDC bundle conductors by FDM integrated with full multigrid technique. Electric Power Systems Research 199, 2021.

Doi: https://doi.org/10.54216/JAIM.030205
Received: August 19, 2022 Revised: November 15, 2022 Accepted: March 25, 2023