Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Heart Disease Prediction and GUI Interaction based on Machine Learning

Yimin Yang*

Jinan Foreign Language School International Center, Jinan, China

* Corresponding Author Email: minxin15@tzc.edu.cn

Abstract. Machine learning is now being used to detect heart disease. Considering that failure to diagnose a heart disease patient can lead to serious consequences, including delayed treatment, worsening of the condition, and even life-threatening situations, it is crucial to ensure that as many true patients as possible are confirmed. Thus, it is important to consider recall rates while maintaining a focus on accuracy. This article compares the application of four machine learning models, including Decision Trees (DT), Random Forests (RF), K-Nearest Neighbors (KNN), and Support Vector Machines (SVM), in heart disease prediction, and measures the effectiveness of these models by using accuracy, recall rates, and F1 score. The outcomes of the experiment reveal that the SVM model performs the best with a recall rate of 0.97. The balance of the model ensures that it achieves high recall without affecting accuracy. In addition, the author combines it with a Graphical User Interface (GUI) to achieve interactive effects. The model and its interactive functions selected in this experiment can easily avoid missing patients in the first screening and improve the accuracy of disease diagnosis.

Keywords: Heart Disease Prediction; Machine Learning; Recall Rates; Support Vector Machines; Graphical User Interface.

1. Introduction

According to data from the World Health Organization, heart disease is the main cause of death worldwide and 17.9 million people are affected per year [1]. The diagnosis of heart disease is medically critical, but also challenging. So far, early detection and effective treatment remain the primary means of reducing deaths [2]. With the massive accumulation of health data in recent years, gaining useful information from such a big dataset is a daunting task for humans [3]. However, the application of data mining techniques and machine learning algorithms helps to extract key patterns and conclusions from data sets [4]. One practical application of this method is in forecasting the onset of heart disease, which involves evaluating a patient's risk factors, including cholesterol, blood pressure, gender, and age levels to estimate the probability of developing the condition. Chala Beyene and colleagues propose leveraging data mining methods to forecast and scrutinize the prevalence of heart conditions. Their primary objective is to anticipate the occurrence of heart illnesses, enabling prompt and automated detection of the condition promptly [5]. Polaraju and colleagues introduced a method that employs multiple regression models for forecasting heart disease. Their empirical studies demonstrated the suitability of multiple linear regression in estimating the likelihood of heart disease occurrence. The outcomes indicate that the regression-based classification approach outperforms other algorithms in terms of accuracy [6]. S.Prabhavathi et al. have put forth a Neuro-Fuzzy System (DNFS) based on Decision Trees (DT) for the analysis and prediction of different types of heart diseases. Their findings suggest that both neural networks and Support Vector Machines (SVM) are proficient in forecasting heart disease occurrences [7]. S. Seema et al. compared the performance of naive Bayes, artificial neural networks (ANN), DT, and SVM in predicting chronic diseases from historical health data. SVM yielded the highest accuracy in this study, while naive Bayes was the most accurate for diabetes prediction [8]. M. A. Akhil Jabbar, Priti Chandra, and B. L. Deekshatulu used the K-Nearest Neighbors (KNN) algorithm combined with feature subset selection to identify features that contribute more to disease prediction. Their approach indirectly reduces the number of tests patients need to undergo [9]. Madhumita Pal and Smita Parija apply Random Forests (RF) data

mining algorithms to heart disease prediction. The sensitivity was 90.6%. The specificity value was 82.7 and the prediction accuracy was 86.9. The prediction accuracy of the RF algorithm for heart disease was 86.9%, and the diagnosis rate was 93.3% [10].

The purpose of this paper is to introduce machine learning to establish a heart disease prediction model. The core task is to analyze and evaluate the performance of four different models by comparing their properties, and based on this, implement a Graphical User Interface (GUI) based interactive form. This study conducted experiments using the publicly available Framingham Heart Disease dataset. Initially, Univariate Analysis (UA) and Bivariate Analysis (BA) are conducted to delve into the inherent properties of the data and to examine the correlations between each feature and the anticipated outcome. The data is preprocessed and sent to different models for analysis and comparison. Besides, accuracy, recall, F1 score, and other indicators are used to measure the effectiveness of the model, and SVM is combined with GUI to achieve better interaction effects. The experimental results show that SVM has the highest recall rate in the test set, at 97%, and the balance performance of the model ensures that accuracy is not sacrificed in the pursuit of high recall. This model can help doctors maximize the first round of heart disease screening. A high recall rate can reduce the likelihood of missed diagnosis in patients with heart attacks.

2. Organization of the Text

2.1. Dataset Description and Preprocessing

The dataset employed in this research is known as heart.csv, obtained from the Kaggle dataset repository [11]. It includes age, gender (where 'sex' is represented by 1 for true and 0 for false), and the type of chest pain experienced (“cp” classified as typical angina, atypical angina, non-anginal pain, or asymptomatic). The electrocardiographic findings at rest are recorded, with normal results coded as 0, the presence of ST-T wave abnormality as 1, and probable or definite left ventricular hypertrophy as 2 (restecg). Furthermore, the dataset captures the ST depression induced by exercise, compared to the resting state (oldpeak), as well as the slope of the peak exercise ST segment (slope: upsloping = 0, flat = 1, downsloping = 2). The maximum heart rate achieved during exercise (thalach) and the occurrence of exercise-induced angina (exange: 1 for yes, 0 for no) are also noted. Additionally, the dataset delves into physiological measurements such as resting blood pressure (trestbps) and serum cholesterol levels (chol). It also considers the presence of diabetes, indicated by a fasting blood sugar level greater than 120 mg/dL (fbs: 1 for true, 0 for false). The extent of vessel involvement is indicated by the number of major vessels (ca: 0-4) colored by fluoroscopy. A Thallium stress test result provides insight into the heart's response to stress, with normal results coded as 0, fixed defect as 1, reversible defect as 2, and not described as 3 (thal). Finally, the target variable reflects the presence of heart disease, with a status of 0 indicating no disease and a status of 1 indicating the presence of the disease. These detailed health metrics present a rich resource for the development and subsequent evaluation of predictive models. Table 1 showcases some instances from the dataset.

In the data preprocessing, the experiment carries out the following six steps: irrelevant features removal, missing value treatment, outlier treatment, categorical features encoding, feature scaling, and transforming skewed features.

Table 1. Dataset

	age	sex	cp	restecg	oldpeak	slope	thalach	exange	trestbps	chol	fbs	ca	thal	target
4	57	0	0	1	0.6	2	163	1	120	354	0	0	2	1
1	37	1	2	1	3.5	0	187	0	130	250	0	0	2	1
2	41	0	1	0	1.4	2	172	0	130	204	0	0	2	1
3	45	1	3	1	1.2	1	132	0	110	264	0	0	3	0
0	57	1	0	1	1.2	1	115	1	130	131	0	1	3	0
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
298	57	0	0	1	0.2	1	123	1	140	241	0	0	3	0
299	56	1	1	1	0.8	2	178	0	120	236	0	0	2	1
300	68	1	0	1	3.4	1	141	0	144	193	1	2	3	0
301	63	1	3	0	2.3	0	150	0	145	233	1	0	1	1
302	57	0	1	0	0.0	1	174	0	130	236	0	1	2	0

2.2. Proposed Approach

This study compares the effectiveness of four different model performances to select a model with high recall and efficiency. The selected model serves as the foundation for developing interactive GUI forms. Figure 1 shows the systematic approach to data analysis. This process starts with a preliminary exploration of the dataset to reveal patterns, distributions, and correlations. Subsequent analysis focuses on a comprehensive Exploratory Data Analysis (EDA) to study the binary feature's impact on the target. This involves pruning irrelevant features, managing missing values, dealing with outliers, and transforming categorical and skewed numerical data to achieve a more normal distribution. The modeling process then begins with the creation of pipelines tailored for models that require scaling. Various classification models, including KNN, SVM, DT, and RF are subsequently implemented and fine-tuned. Emphasis is placed on achieving a high recall rate for Category 1 to ensure accurate identification of patients with heart disease. Model performance is then assessed and compared using metrics such as accuracy, recall, and F1-score to gauge their effectiveness. The study also develops an interactive GUI.

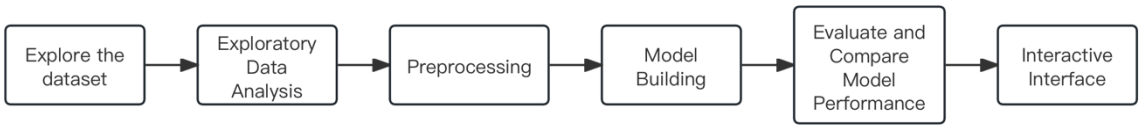

Figure 1. The pipeline of the model

2.2.1. UA and BA

UA and BA are two steps in EDA. These two steps provide insight into the individual characteristics of the data and how each of these characteristics relates to predicting the target variable. UA, a statistical method, analyzes and summarizes single variables or datasets, studying one variable at a time. It examines distribution, central tendency, spread, and other descriptive statistics, often as the initial step in data analysis, providing insights into dataset characteristics before multivariate analysis. BA analyzes the relationship between two variables simultaneously, exploring how changes in one variable relate to changes in another. It delves into cause-and-effect relationships, correlations, and patterns not always evident in UA.

2.2.2. Box-Cox transformation

The Box-Cox transformation was introduced by British statisticians T.A. Browne and L.H. Cox in 1964. The general form of the transformation is:

$$Y = \lambda \cdot \ln(X + \lambda) \tag{1}$$

The variable Y represents the altered form of X, the base variable, with λ serving as a modifiable coefficient to be identified via optimization techniques, such as maximum likelihood estimation or least squares regression. The Box-Cox transformation is a method used to enhance data normality or mitigate heteroscedasticity in non-normal distributions. It is beneficial for skewed or non-zero kurtosis variables, making transformed data resemble normal distributions. This improves statistical analyses like linear regression, enhancing accuracy and efficiency.

2.2.3. DT

DT is a diagrammatic representation where internal nodes signify attribute tests, branches represent test outcomes and leaf nodes hold category labels or values. The paths from the root to the leaves represent classification rules. Each internal node in the DT used to predict heart disease represents a test for a specific attribute or feature in the dataset, such as age, blood pressure, or cholesterol level. As the tree grows, the nodes are divided according to the most informative attributes, with each branch representing the results of these tests. Leaf nodes are the endpoints of these branches and have a category label, or value, used to determine the existence or not of heart disease. The path encoding from the root node to the leaf node guides the classification rules of the prediction process. One of the main advantages of DT is its interpretability, allowing clear, actionable rules to be extracted.

2.2.4. RF

RF is a classifier made up of many DTs. Every tree is created by using algorithm A, training data S, and a random vector θ. θ is chosen independently and at random from a specific distribution. The RF's final prediction is the result that gets the most votes from the predictions of all the trees [12]. The RF algorithm works in the following steps which are shown in Figure 2. Firstly, it picks K data points randomly from the training data and builds a decision tree for them. Then, the N-tree subsets repeat the previous steps. Finally, RF decides the result according to the most votes.

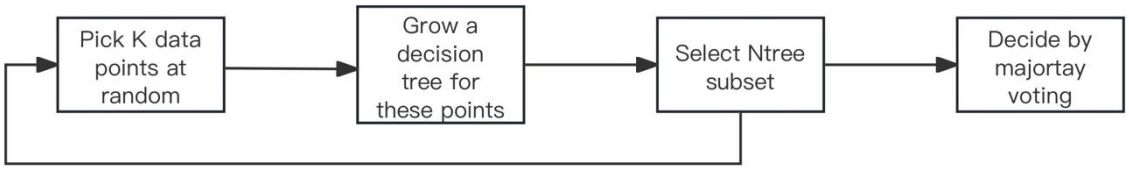

Figure 2. The working steps of the Random Forest algorithm

The RF algorithm for heart disease prediction iteratively builds a collection of DT on different subsets of the training data, each trained on a random sample with replacement. The final prediction is determined by a majority vote from these trees, which collectively improves the model's robustness and accuracy.

2.2.5. KNN

The KNN algorithm stores the entire training data and predicts new cases by examining the labels of the closest examples. It relies on the idea that data features predict labels, suggesting similar objects tend to have the same label. Despite dealing with large datasets, finding the nearest neighbor can often be done quickly.[12]. In mathematical terms, $\rho: \mathrm{X} * \mathrm{X} \rightarrow \mathbb{R}$ where ρ is a function that calculates the distance between two target points of $\mathrm{X}\left(x_{i,} x'_{i}\right)$. The Euclidean distance between these two points can be calculated by the following formula:

$$\rho(x, x') = \left|x - x'\right| \sqrt{\sum_{i=1}^{d} (x_i - x_i')^2} \tag{2}$$

In heart disease prediction, the KNN algorithm is like a person who makes decisions by looking at their neighbors. It measures the similarity between two data points, such as using Euclidean distance, and then finds the closest example to the new data point. KNN makes predictions based on the votes

of those neighbors, and if most of the neighbors say someone has heart disease, the new data point is predicted to have heart disease. This study uses metrics such as accuracy, recall rates, and F1 scores to test KNN's performance on independent datasets. The choice of the k value (the number of neighbors) and the distance measure are hyperparameters in the algorithm that require experimentation and cross-validation to find the best value. The benefit of KNN is that its decision-making process is easy to understand because the decision boundary is implicitly defined by the proximity of the data points, which allows experimenters to see which features are most closely related to the presence of the disease.

2.2.6. SVM

SVM is a supervised machine learning algorithm mainly used for classification and regression analysis. It works by finding a hyperplane that best divides a feature space into two classes. In a complex task like predicting heart disease, SVM can handle a lot of information (high-dimensional data), such as age, gender, and various health indicators. It uses a special technique (kernel function), like converting data into a more understandable form and then finding an optimal dividing line (hyperplane) to Distinguish between people with heart disease and those without. SVM is very good at generalizing to new situations, which means that it is trained to predict data that has not been seen before very well. While the SVM itself is linear, it can use these techniques to capture non-linear relationships in the data, such as the complex link between age and heart disease.

3. Results and Discussion

3.1. EDA

Figure 3 illustrates a consistent age distribution, clustering in the mid-50s with an overall average age of 54.37 years. The data on resting blood pressure reveals a typical range between 120 and 140 mmHg, with a mean of 131.62 mmHg. Cholesterol levels predominantly fall within the 200-300 mg/dL bracket, averaging at 246.26 mg/dL. During stress tests, maximum heart rates typically fall between 140 and 170 bpm, with an average of 149.65 bpm. Most ST depressions prompted by exercise are minimal, close to 0, suggesting that most individuals exhibit little to no significant ST depression during exercise, with an average of 1.04. The histogram of continuous features corresponds well to the provided descriptions, displaying no substantial anomalies or outliers among the continuous variables.

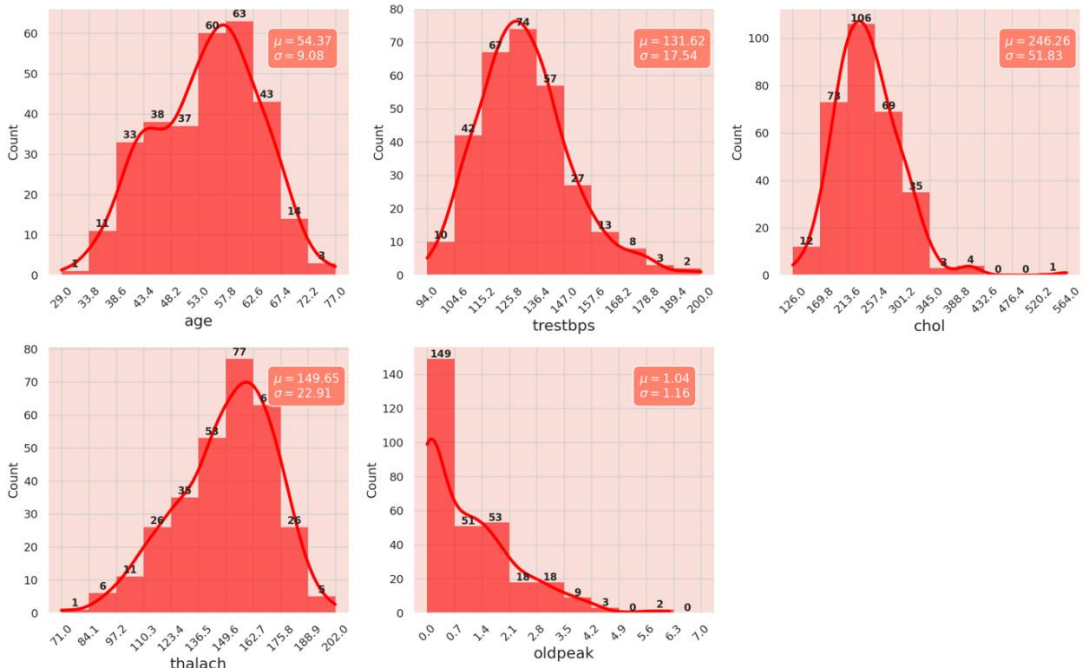

Figure 3. Distribution of Continuous Variables

Figure 4 reveals that typical angina is the most common chest pain type, most patients have fasting blood sugar levels below 120 mg/dL, and resting electrocardiograph results vary with some types being more prevalent. Exercise-induced angina is infrequent, and the slope of the peak exercise ST segment has a frequently occurring type. Most patients have few major blood vessel lesions detected by fluoroscopy, and Thallium stress test results include a range of outcomes, with one being the most common. The presence of heart disease is evenly split among patients.

Figure 4. Distribution of Categorical Variable

Figure 5 suggests that patients with heart disease tend to be younger on average compared to those without. Resting blood pressure ("trestbps") and serum cholesterol levels ("chol") show overlapping distributions between the groups, indicating these features may not strongly predict heart disease. However, the maximum heart rate during stress tests ("thalach") is significantly higher in heart disease patients. Exercise-induced ST depression ("oldpeak") is also lower in these patients, with their peak distribution close to zero compared to a more spread-out distribution in the non-disease category. "Thalach" appears to be the most distinguishing feature between the two groups, followed by "oldpeak" and "age".

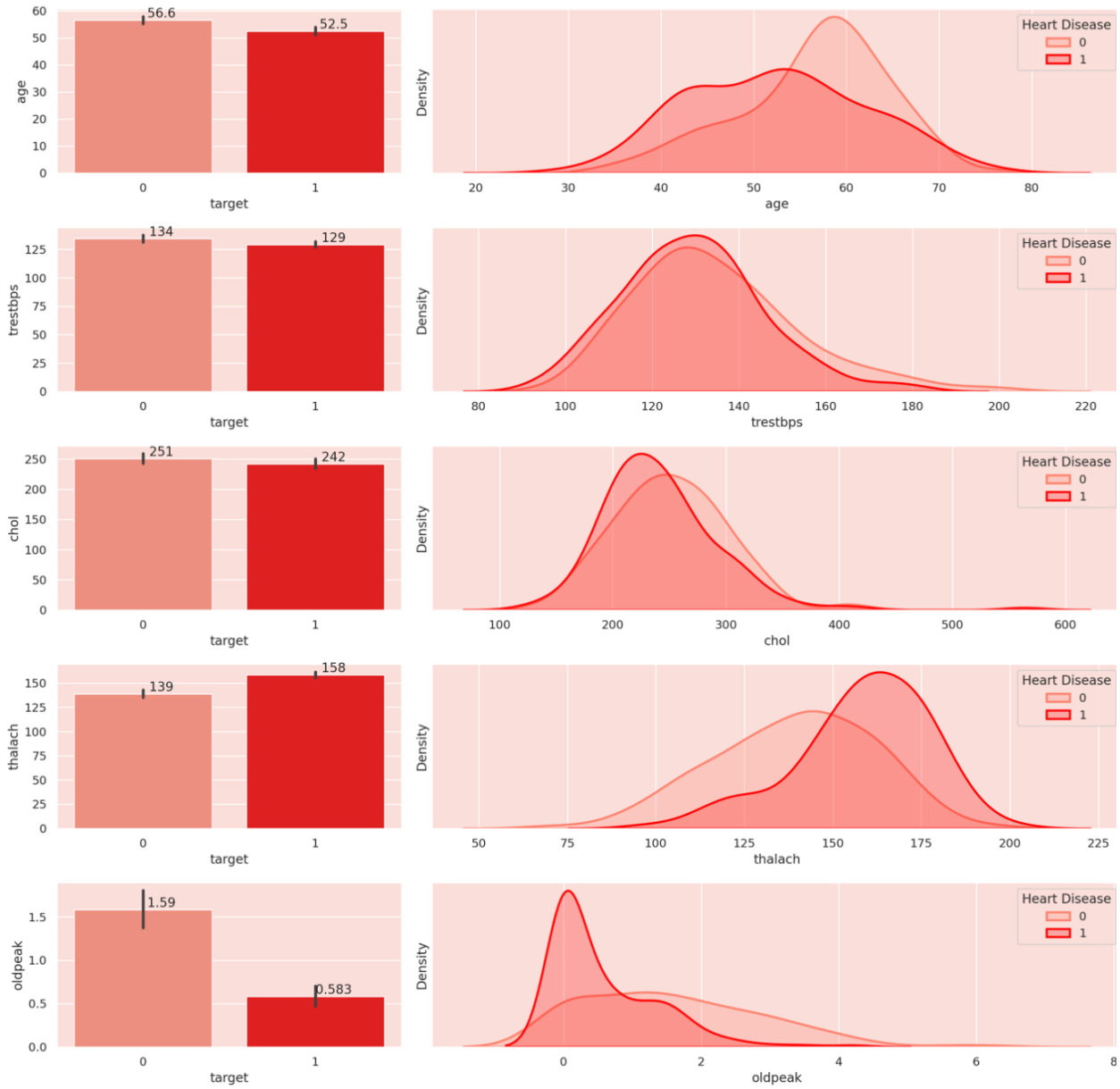

Figure 5. Continuous Features vs Target Distribution

As Figure 6 shows, people with heart disease usually have few large blood vessels stained by fluoroscopy, and those without large blood vessels are at higher risk. Types 1, 2, and 3 chest pain are more associated with heart disease than type 0 chest pain. Men who do not exercise have higher rates of angina and heart disease. Fasting blood glucose levels had a limited effect on predicting heart disease. Resting electrocardiograph type 1 and slope type 2 are more prevalent in patients with heart disease, indicating their impact. The reversible defect category in thallium stress testing is also associated with a higher risk. In summary, the main predictors of heart disease are the presence of large vessel disease, gender, experiencing angina during exercise, the character of chest pain, the slope of the ST segment, and the outcomes of thallium stress tests. Meanwhile, fasting blood sugar levels and resting electrocardiograms have a lesser influence on the prediction.

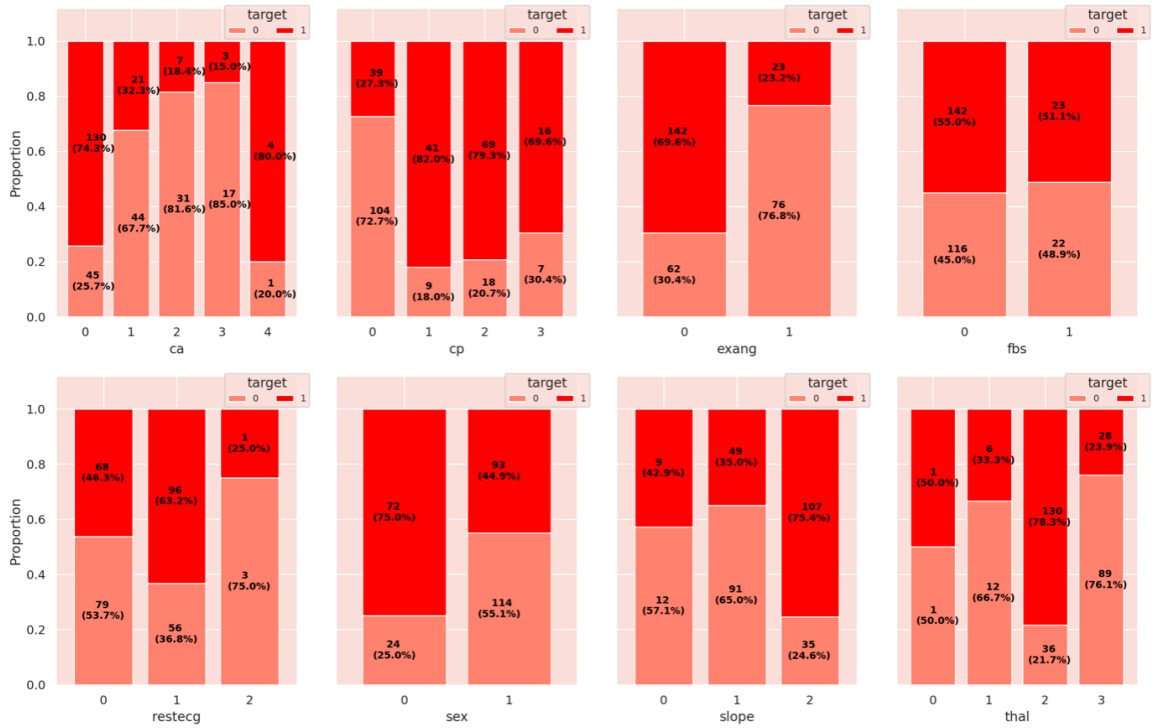

Figure 6. Categorical Features vs Target Distribution

3.2. Multi-Model Performance Comparison

Table 2 presents a comprehensive performance evaluation of four classification models: SVM, RRF, DT, and KNN. The table captures a suite of nine performance metrics for each model, including precision_0 and precision_1 for class 0 and class 1, respectively, recall_0 and recall_1 for class 0 and class 1, respectively, f1_0 and f1_1 for class 0 and class 1, respectively, as well as macro_avg_precision, macro_avg_recall, and macro_avg_f1, which represent the average precision, recall, and F1 score across all classes. These metrics collectively illustrate the all-round performance of each model.

Table 2. Multi-Model Performance Comparison

	Precision_0	Precision_1	recall-_0	recall-_1	f1_0	f1_1	macro_avg_precision	macro-_avg_recall	macro-_avg_f1	accuracy
SVM	0.94	0.73	0.57	0.97	0.71	0.83	0.77	0.77	0.77	0.79
RF	0.85	0.83	0.79	0.88	0.81	0.84	0.83	0.83	0.83	0.84
DT	0.80	0.78	0.71	0.85	0.75	0.79	0.78	0.78	0.78	0.79
KNN	0.82	0.85	0.82	0.85	0.82	0.83	0.83	0.83	0.83	0.84

The comparable performance of RF, DT, and KNN on both training and testing datasets suggests that these models exhibit low levels of overfitting. In contrast, the SVM model exhibits a high recall of 0.97 for class 1, indicating that nearly all true positive cases (i.e. patients with heart disease) are accurately identified. It is also noteworthy that the SVM model maintains a balanced performance, as evidenced by its F1-score of 0.83 for class 1. This indicates that the model does not solely prioritize maximizing the recall rate at the expense of precision.

When models are sorted based on values of 'recall_1', which is shown in Figure 7, it can be easily found that The SVM model demonstrates a commendable capability in recognizing potential heart patients. With a recall of 0.97 for class 1, it's evident that almost all patients with heart disease are correctly identified. This is of paramount importance in a medical setting. Also, the model's balanced performance ensures that while aiming for high recall, it doesn't compromise on precision, thereby not overburdening the system with unnecessary alerts.

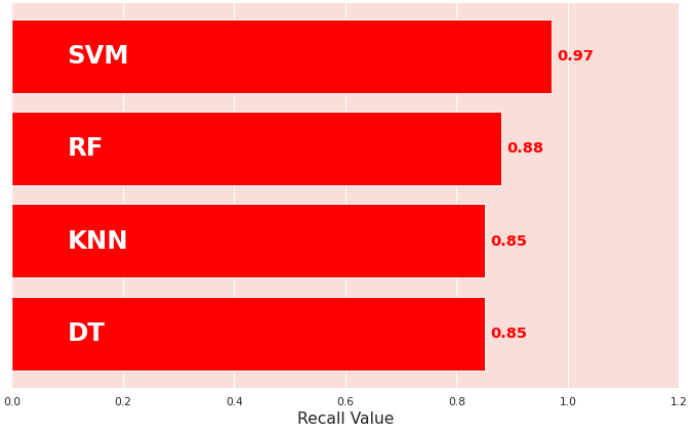

Figure 7. Recall for Positive Class across Models

3.3. GUI Development for SVM Model

GUI, which is shown in Figure 8 develops based on the SVM model to facilitate user interaction and implementation of the classification algorithm. The GUI allows users to input data, view the model's predictions, and interpret the results in a user-friendly visual format, thereby enhancing the accessibility and practical application of the SVM model in real-world scenarios.

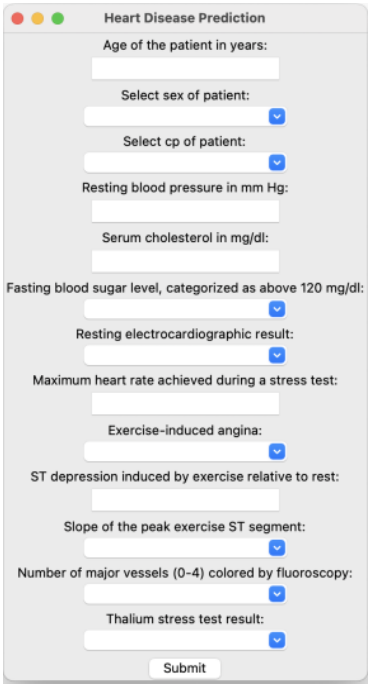

Figure 8. GUI

4. Conclusion

This research aims to find a heart disease prediction model that has the highest recall rate with a good balanced performance among DT, RF, KNN, and SVM, and make a GUI for the best model. UA and BA are used to examine the binary relationship with the target variable and Box-Cox transformation is proposed to transform features to be more normal-like primarily which can help in mitigating the impact of outliers. The study uses metrics like F1 scores, accuracy, and recall rates to evaluate and compare the performance and effectiveness of different machine learning models. Experimental results show that SVM achieved the highest recall rate in the test set, at 97%, and the balanced performance of the models ensured that accuracy is not sacrificed in pursuit of high recall rates. This model's performance is promising for medical diagnostics, especially when prioritizing the accurate identification of patients with heart disease without overburdening the system with false alarms. Since

SVM's performance in terms of interaction effects, it is chosen for making GUI to achieve better interaction effects. In the future, finding larger and better data sets, as well as using web authorities to improve interactivity and display results as web pages will be considered as the research objective for the next stage.

References

[1] Kanwal, Amna, K. Tehseen Ahmad, and N. Aslam. Detection of Heart Disease Using Supervised Machine Learning. 2022.

[2] Ali, M. Mamun, et al. Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. Computers in Biology and Medicine 136, 2021, p. 104672.

[3] M. Aljanabi, H. Mahmoud. Qutqut, and M. Hijjawi. Machine learning classification techniques for heart disease prediction: a review. International Journal of Engineering & Technology 7(4), 2018, pp. 5373-5379.

[4] M. Marimuthu, et al. A review on heart disease prediction using machine learning and data analytics approach. International Journal of Computer Applications, 181(18), 2018, pp. 20-25.

[5] Beyene, Chala, and P. Kamat. Survey on prediction and analysis the occurrence of heart disease using data mining techniques. International Journal of Pure and Applied Mathematics, 118(8), 2018, pp. 165-174.

[6] K. Polaraju and D. Durga Prasad. Prediction of heart disease using multiple linear regression model. International Journal of Engineering Development and Research Development, 5(4), 2017, pp. 1419-1425.

[7] S. Prabhavathi and D.M. Chitra. Analysis and prediction of various heart diseases using DNFS techniques. International journal of innovations in scientific and engineering research, 2(1), 2016, pp. 1-7.

[8] Deepika, Kumari, and S. Seema. Predictive analytics to prevent and control chronic diseases. international conference on applied and theoretical computing and communication technology (iCATccT). IEEE, 2016.

[9] M. Jabbar, Akhil, B.L. Deekshatulu, and P. Chandra. Heart disease classification using nearest neighbor classifier with feature subset selection. Seria Informatica 11, 2013, pp. 47-54.

[10] Pal, Madhumita, and S. Parija. Prediction of heart diseases using random forest. Journal of Physics: Conference Series. 1817(1), 2021.

[11] Information on: https://www.kaggle.com/datasets/arezaei81/heartcsv.

[12] Shalev-Shwartz, Shai, and S. Ben-David. Understanding machine learning: From theory to algorithms. Cambridge university press, 2014.