PAPER NAME

Heart Disease Prediction Model 1.docx.pdf

AUTHOR

Dr. Jasleen Gund

WORD COUNT

5166 Words

CHARACTER COUNT

28717 Characters

PAGE COUNT

17 Pages

FILE SIZE

196.9KB

SUBMISSION DATE

Jul 18, 2024 5:07 PM GMT+5:30

REPORT DATE

Jul 18, 2024 5:08 PM GMT+5:30

10% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 7% Internet database
- 5% Publications database
- Crossref database
- Crossref Posted Content database
- 7% Submitted Works database

Excluded from Similarity Report

- Bibliographic material
- Quoted material
- Cited material
- Small Matches (Less then 14 words)

ABSTRACT

Heart disease, or cardiovascular disease, includes a range of conditions that impact the heart and has been a leading cause of death worldwide for many years. It is often associated with risk factors such as high blood pressure, high cholesterol, and lifestyle choices. Managing this disease requires accurate and reliable diagnostic approaches for early detection and prompt management. This research paper focuses on several attributes related to heart disease and evaluates supervised learning algorithms such as Logistic Regression, Support Vector Machines (SVM), and Artificial Neural Networks (ANN). Additionally, it incorporates feature selection techniques using correlation and chi-square tests. The dataset used is sourced from the UCI repository, consisting initially of 76 attributes and 303 instances. For testing purposes, 14 key attributes were selected to evaluate algorithm performance. The study aims to predict the likelihood of developing heart disease in patients by providing output as yes or no (0 and 1). Results indicate that overall, Logistic Regression achieved the highest accuracy score, and after feature selection, SVM and ANN performance improved by approximately 8%

Keywords: Heart disease, Logistic Regression, SVM, ANN, feature selection

CHAPTER-1 INTRODUCTION

Despite advances in diagnosis and treatment, heart disease remains the most comman cause of death globally, responsible for approximately one-third of annual deaths, according to the World Health Organization [1]. “Heart disease” or cardiovascular disease is a broad term encompassing various heart-related problems that affect the heart’s structure and how it works, majorly affect the heart’s blood vessels. It is the leading cause of death. However, many types of heart disease can be prevented and managed with appropriate measures.

The significant increase in heart disease can be attributed to lifestyle choices, lack of exercise, and the consumption of processed foods. Advanced stages of heart disease can lead to heart attacks and pose serious risks to patients' lives, making early detection through advanced and therapeutic methods crucial. One major challenge in diagnosing heart disease is patients' reluctance to participate in clinical trials. Additionally, these trials are costly and time-consuming, leading to limited engagement and Healthcare organizations face the challenge of delivering quality services at affordable costs, necessitating accurate diagnoses and effective treatments. However, in contrast to traditional clinical methods, some approaches can analyse disease patterns by examining data from both patients and healthy individuals.[2][4]

Over the past few years, the application of artificial intelligence technology, particularly Machine Learning (ML), in auxiliary diagnosis has advanced rapidly, leading to significant progress in automatic detection applications. The advantage of ML methods lies in their ability to diagnose diseases, including heart disease, with reasonable accuracy and at a lower cost.[3]

Clinical decisions are often made based on intuition rather than data-driven insights, leading to potential errors and increased costs. Current hospital information systems manage data for billing and simple statistics but fall short in supporting complex clinical decision-making. Integrating clinical decision support with computer-based patient records, as suggested by Wu et al., can reduce errors and improve patient outcomes(Author, 2008). [4]

CHAPTER-2 OBJECTIVE

The objective of studying heart disease and developing a predictive model is to enhance early detection and accurate diagnosis, ultimately reducing the prevalence and impact of the disease. By leveraging advanced technologies like machine learning, aiming to identify patterns and risk factors associated with heart disease, enabling more efficient and cost-effective screening methods. A

predictive model can analyze vast amounts of patient data, offering personalized risk assessments and informing preventative strategies. This proactive approach not only improves patient outcomes but also alleviates the burden on healthcare systems by preventing the progression of heart disease through timely interventions.

Additionally, predictive models can facilitate the development of targeted therapies, optimize resource allocation, and support clinical decision-making processes. By integrating these models into routine medical practice, healthcare providers can deliver more precise, patient-centric care, ultimately enhancing the quality of life for individuals at risk of or living with heart disease.

CHAPTER- 3 REVIEW OF LITERATURE

To develop a predictive model accuracy as well as avoidance of overfitting is essential. Ahmad Ayid Ahmad [5] aim to obtain an ML model that can predict heart disease with the highest possible performance using the UCI heart disease dataset. Overfitting was avoided by reducing the dimensional subspace using Jellyfish algorithm. SVM classifier model trained on the dataset along with the Jellyfish algorithm, and obtained Sensitivity of 98.56%, Specificity of 98.37%, Accuracy of 98.47%,, and Area Under Curve of 94.48%.

One of the studies conducted a comparation analysis aiming to enhance predictive accuracy for heart disease risk using ensemble techniques on the Cleveland dataset which has 303 observations. Employing a brute force method, they explored all potential attributes set combinations and trained classifiers accordingly. Their efforts resulted in a significant 7.26% increase in the accuracy of a weak classifier through ensemble algorithms. Ultimately, they achieved an 85.48% accuracy using a majority vote approach with Naive Bayes, Bayesian Networks, Random Forest, and Multilayer Perceptron classifiers, using a set of nine attributes [6]

Predicting survival in coronary heart disease (CHD) patients presents a significant challenge in medical research. This study aims to develop data mining algorithms for this purpose using a dataset of 1000 CHD cases, including clinical observations and a 6-month follow-up period. Survival outcomes were tracked for each case. We employed three prominent data mining algorithms on 502 cases, and assessed their performance using 10-fold cross-validation. Results showed that Support Vector Machines (SVM) achieved the highest accuracy at 92.1% on the holdout sample, followed by artificial neural networks at 91.0%, with decision trees performing least effectively at 89.6%. This comparative analysis offers valuable insights into the predictive capabilities of these models for CHD patient survival.[7][5][16]

Unsupervised learning (Black Box): Artificial neural networks (ANNs) play a crucial role in the prediction of heart disease by mimicking the human brain's ability to select features, classify data, make decisions, and forecast outcomes. Utilizing supervised learning techniques and non-linear mathematical models, ANNs are designed to behave like artificial neurons, enabling them to effectively handle complex medical diagnosis tasks. Recent studies have focused on developing heart disease diagnosis systems (HDDS) using various machine learning techniques, including k-nearest neighbors (kNN), fuzzy rules, support vector machines, and ANNs [8]

Therefore, system will be implementing the following three algorithms:

- Support Vector Machine (SVM)
- Logistic Regression (Base Line model)
- Artificial neural networks (ANNs)

From Research papers , we got to know that Data pre-processing , hyperparameter tuning and feature selection can optimize the classification accuracy of machine learning algorithms which is understood by experimental works.

CHAPTER- 4 METHODOLOGY

4.1 Proposed Methodology

After literature review, the following methodology workflow was opted according to problem statement: Identify and analyze the key risk factors contributing to heart disease. According to Fig. 2, First, The heart disease dataset was obtained in .csv format from the UCI Machine Learning Repository. The next step involved preprocessing the data, which included identifying and addressing missing values by either using a user-defined constant or the mean value, depending on the attribute type, to ensure optimal performance of the machine learning classifiers. After importing the dataset into the software tool, we examined its attributes, types, value ranges, and other statistical details. Subsequently, we performed classification with cross-validation using various machine learning algorithms such as Logistic Regression (LR), (SVM) and ANN using the complete set of attributes. For performing cross-Validation, the dataset was randomly split into k equal-sized groups, with the model trained on k−1 folds and validated on the remaining kth fold where k=5. This process was repeated until each fold served as a test set, and the average performance metric was calculated and used k = 5 for 5-fold cross-validation as depicted in Fig. 1

Moreover, employed attribute evaluators such as correlation-based feature selection, and chi-squared attribute evaluation on the full training set to identify the optimal set of attributes for predicting heart

disease risk. The classifiers were then retrained using cross-validation and the performance evaluation is compared.

4.2DATA PRE-PROCESSING

The preprocessing of the data was done through Pandas and Numpy

4.2.1 CHECKING MISSING VALUES:

Using descriptive methods: Displays information about the DataFrame, including the number of columns, column labels, column data types, memory usage, range index, and the number of non-null cells in each column.

.info(): This method provides a concise summary of the DataFrame, detailing data types and the number of non-null values in each column. df.isnull().sum() returns the total number of missing values in each column.

In the UCI data, no missing value is found.

4.2.2 ENCODING VALUES

Encoding values refers to the process of transforming categorical data (data with labels or text) into numerical representations that computers can understand. This is necessary because most machine learning algorithms work with numerical data, not text or labels directly.

4.2.2.1 NUMARICAL DATASET:

It represents measurable quantities with a numerical value.
Following are the numerical data in the dataset:

```
numeric_columns=['age', 'cholesterol', 'st_depression', 'rest_bp', 'max_hr',
'num_vessels']
```

4.2.2.2 NOMINAL DATASET:

It represents qualities or labels, not quantities and categories are distinct and have no inherent order or ranking.

Conversion of object Datatype to int Datatype:

```
columns_to_convert = [‘exercise_angina’, ' fasting_bs','diagnosis’]
for column in columns_to_convert:
    df[column] = df[column].map({'Yes': 1, 'No': 0})
```

4.2.3 OUTLIERS:

Detection of outliers using IQR

```
data=df
columns=['rest_bp', 'age', 'max_hr', 'st_slope', 'cholesterol',
'st_depression',
      'rest_ecg']
Q3 = df[columns].quantile(0.75)
Q1 = df[columns].quantile(0.25)
IQR = Q3 - Q1
upper_bound = Q3 + 1.5 * IQR
lower_bound = Q1 - 1.5 * IQR


outliers = df[df[columns] < lower_bound]
outliers = pd.concat([outliers, df[df[columns] > upper_bound]])

print("Potential outliers:")
print(outliers)
```

Found no Outliers in the dataset.

4.3 DATASET ANALYSIS AND STATISTICS ANALYSIS

4.3.1 STATISTICAL ANALYSIS:

Table 2(a) presents the statistical characteristics of the numeric attributes, including minimum, maximum, mean, standard deviation, missing values, distinct values, and unique values. Notably, there are no missing values in the numeric attributes of the dataset.

4.4 DATA VISTUALISATION (EDA)

4.4.1 EXPLORATORY DATA ANALYSIS:

Exploratory Data Analysis (EDA) is an analytical approach used to identify general patterns in the data, including outliers and unexpected features. EDA is a crucial initial step in any data analysis process.[15]

4.4.2 PROBLEM STATEMENT 1:

What is the proportion of positive and negative cases in the dataset?

ANALYSIS:

For Fasting Blood Sugar (fasting_bs), a higher number of individuals with a fasting blood sugar level

of 0 do not have heart disease (diagnosis = 0). When fasting blood sugar is 1, the counts for both diagnoses are more balanced but still slightly higher for diagnosis = 0. Regarding ST Slope (st_slope), a flat slope (st_slope = 1) is more common in individuals with heart disease (diagnosis = 1), while a normal slope (st_slope = 0) is more frequent in those without heart disease. A downward slope (st_slope = 2) shows a more balanced distribution but leans slightly towards diagnosis = 1. For Exercise-Induced Angina (exercise_angina), most individuals without exercise-induced angina (exercise_angina = 0) do not have heart disease. In contrast, the presence of exercise-induced angina (exercise_angina = 1) is more common in individuals with heart disease. The data on Gender shows that a higher count of males (gender = 1) are diagnosed with heart disease compared to females (gender = 0). When examining Chest Pain Type (chest_pain), typical angina (chest_pain = 0) is more common in individuals without heart disease. As chest pain types indicate more severe conditions (chest_pain = 1, 2, 3), the counts for those with heart disease increase. In the Thalassemia vs. Heart Disease Diagnosis analysis, a normal thalassemia level (thalassemia = 0) is more common in individuals without heart disease. Intermediate and high thalassemia levels (thalassemia = 1 and 2) are more prevalent in those with heart disease. For Resting ECG (rest_ecg), normal ECG results (rest_ecg = 0) are more common in individuals without heart disease. Abnormal ECG results (rest_ecg = 1, 2) show a higher prevalence in those with heart disease, particularly for rest_ecg = 2. Lastly, the Number of Major Vessels Colored by Fluoroscopy (num_vessels) shows that as the number of major vessels increases, the likelihood of a heart disease diagnosis (diagnosis = 1) increases significantly. The Thalassemia subplot confirms the same trends observed in the earlier "Thalassemia vs. Heart Disease Diagnosis" analysis.

4.4.3PROBLEM STATEMENT 2:

Which age group is more at risk for heart disease ?

Age group 50-60, shows significant risk for diagnosis for heart disease, from Fig. 3.

4.4.4 PROBLEM STATEMENT 3:

Identify the Distribution for detecting outliners, show central tendency and spread.

ANALYSIS:

The visualization (Fig. 4) provides a comprehensive analysis of various health metrics. For age, the distribution plot reveals a somewhat normal distribution with a peak around the ages of 55-60, while the boxplot indicates a central tendency in the same range and no significant outliers, suggesting most data points fall within this age range. In terms of resting blood pressure (rest_bp), the distribution plot shows a slightly right-skewed normal distribution, with most values falling between 120 and 140. The boxplot for rest_bp highlights a central value around 130-140 and several outliers above 160, indicating some individuals with unusually high resting blood pressure.

Cholesterol levels display a right-skewed distribution, with most values clustering between 200 and 300. The boxplot for cholesterol shows a central tendency around 240-260 and several outliers above 400, suggesting some individuals have exceptionally high cholesterol levels. The maximum heart rate (max_hr) has a normal distribution centered around 150, and its boxplot indicates a central tendency around 150-160, with a few outliers below 100 and above 200, highlighting some individuals with unusual maximum heart rates.

Lastly, the distribution of ST depression (st_depression) is highly skewed to the right, with most values near zero and a long tail extending towards higher values. The boxplot shows a central tendency around 0.5, with several outliers above 4, indicating a few individuals with significantly high ST depression values.

4.5 FEATURE SELECTION:

Feature selection helps in reducing the number of input variables for your model by retaining only the relevant data and eliminating noise. This method involves automatically selecting the most pertinent features for your machine learning model based on the specific problem you aim to solve.

4.5.1 CORRELATION –

Correlation-based Feature Selection (CFS) is a technique that selects subsets of features highly correlated with the target variable while having low correlation with each other.

The principle of CFS is to find a subset of features that offers the maximum amount of information about the target variable and also minimizing redundancy among the features.

With the help of the Fig. , following attributes are highly correlated with Diagnosis (target value):

1)Thalassemia

2)num_vessels

3)st_depression

4)exercise_angina

5)chest_pain

6)st_slope

4.5.2 CHI-SQUARE –

According to Table 3, Features having the highest importance are (p-value < 0.001):

Thalassemia,this feature has the highest Chi-squared score and a p-value much lower than 0.001, indicating a strong association with the target variable. Then, exercise_angina which is similar to thalassemia, the p-value suggests a strong relationship with the target variable.Then comes, num_vessels which has the p-value (1.51E-07) suggests a significant association

Features having moderate importance are (0.001 < p-value < 0.05):

st_depression: The p-value (8.82E-04) indicates a moderate association and might be worth considering depending on the overall number of features. Then comes, st_slop which is similar to st_depression, the p-value suggests a moderate association.

Less Significant Features (p-value > 0.05):

gender, the p-value suggests a weak association. Depending on the number of features and domain knowledge. Then, chest_pain which is similar to gender, the p-value suggests a weak association.. Then rest_ecg which has the p-value indicates a weak association then comes max_hr: The p-value suggests a weak association. age is the lowest Chi-squared score and highest p-value indicate a weak association.

4.5 MACHINE LEARNING ALGORITHUM:

4.5.1 MinMaxScaler is used for scaling the data in the dataset:

MinMaxScaler scales all the data features in the range [0, 1], even if there are negative values present in the dataset.

```
X = df.drop(columns_to_drop, axis=1)
y = df['diagnosis']

from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
X_scaled = scaler.fit_transform(X)
```

4.5.2 The split of dataset for training and testing is 80-20 as following:

```
X_train, X_test, Y_train, Y_test = train_test_split(X_scaled, y,
train_size=0.8, test_size=0.2, random_state=4)
```

4.6.1 LOGISTIC REGRESSION (LR):

In logistic regression, each factor is assigned a weight (coefficient), and these weights are combined to predict the likelihood of something falling into one category or another.[12]

$y=1/1+e-(\beta 0+\beta 1x1+\cdots+\beta nxn)$

,value of y for a single observation x with n features

```
lr_model = LogisticRegression()
lr_model.fit(X_train, Y_train)
```

4.6.2 (SVM)

In SVM, choose the hyperplane whose distance from it to the nearest data point on each side is maximized. If such a hyperplane exists it is known as the maximum-margin hyperplane/hard margin [16].

```
svm_model = SVC()
svm_model.fit(X_train, Y_train)
```

4.6.3 ARTIFICAL NEURAL NETWORK (ANN)

The purpose of the activation function is to transform the weighted sum of input signals of a neuron into the output signal ,which then serves as the input to the next layer

```
ann_model = MLPClassifier(hidden_layer_sizes= (100,), random_state=4,
max_iter=1000)
ann_model.fit(X_train, Y_train)
```

The model consists of one hidden layer with 100 neurons. An ANN can be effective depending on the dataset's complexity. The MLPClassifier by default uses the ReLU activation function for hidden layers and the softmax function for the output layer in multiclass classification or the logistic function for binary classification.

The MLPClassifier is a Multi-Layer Perceptron classifier, a type of feedforward artificial neural

network. The hidden_layer_sizes = 100 ,parameter specifies the architecture of the hidden layers, indicating that there is one hidden layer with 100 neurons. The max_iter = 1000 parameter sets the maximum number of iterations for the optimization algorithm, meaning the training process will stop after 1000 iterations if it hasn't already converged. The random_state = 4 parameter is a seed for the random number generator, ensuring reproducibility of results. Using different seeds results in different initial weights and biases in the network, which can lead to different outcomes.

CHAPTER-5 RESULTS

5.1 PERFORMANCE METRICS

Following, performance metrics used in this work, accuracy, Specificity, precision, F-measure, specificity, and ROC area, are discussed here.

$$\text{Specificity}=\text{True Negatives}+\text{False Positives}/\text{True Negatives}$$

$$\text{Accuracy}=TP+TNTP+FP+TN+FN\times 100\%$$

$$\text{Sensitivity}=TPTP+FN\times 100\%$$

$$\text{Precision}=TPTP+FP\times 100\%$$

$$F-\text{Score} =2\times \text{Precision}\times \text{Recall}/\ \text{Precision}+\text{Recall}$$

ROC Area: The area under the ROC curve quantifies the overall predictive performance of a model across all possible classification thresholds. The ROC curve plots the true positive rate (sensitivity or recall) against the false positive rate (fallout) for each threshold from 0 to 1

Cross-validation: To ensure our model captures the correct patterns from the data and does not overfit to noise, we can't solely rely on fitting the model to the training data. To address this, we use the cross-validation technique. This article explores the process of cross-validation in machine learning.

In cross-validation for these mode K-fold cross validation is used, which divides the dataset into k subsets, known as folds. The model is trained on k−1 of these folds while the remaining fold is used for evaluation. This process is repeated k times, with each fold serving as the evaluation set exactly each time. For CV, k=5 is kept.

```
def cross_validate_model(model, X, y):
        cv_accuracy = cross_val_score(model, X, y, cv=5, scoring='accuracy').mean()
```

```
    cv_precision = cross_val_score(model, X, y, cv=5, scoring='precision').mean()
    cv_recall = cross_val_score(model, X, y, cv=5, scoring='recall').mean()
    cv_f1 = cross_val_score(model, X, y, cv=5, scoring='f1').mean()
     cv_roc_auc = cross_val_score(model, X, y, cv=5, scoring='roc_auc').mean()
return cv_accuracy, cv_precision, cv_recall, cv_f1, cv_roc_auc
```

5.2 PERFORMANCE EVALUATION

According to Table 4(a), The Logistic Regression model shows good performance with an accuracy of 85% on the test set. It has balanced precision and recall scores, indicating that it correctly identifies both positive and negative cases of heart disease reasonably well. The F1 score, which combines precision and recall, is accurate at 80.85%.The SVM model shows higher training accuracy (89.45%) but lower test accuracy (76.67%), suggesting it might be overfitting the training data. The precision and recall scores are balanced in the test set, but the F1 score is lower compared to Logistic Regression.

The ANN model demonstrates high training accuracy (93.67%) and reasonable test accuracy (76.67%). However, the test precision, recall, and F1 score are lower compared to Logistic Regression, indicating it might not generalize as well to unseen data as Logistic Regression does in this case.

According to Table 4(b), Among the evaluated models, Logistic Regression stands out with the highest cross-validation (CV) accuracy (0.832) and ROC Area (0.910). This suggests excellent overall performance and the ability to effectively distinguish between classes. Additionally, its balanced F1-score (0.806) indicates a good compromise between correctly identifying positive cases (recall) and avoiding false positives (precision).

On the other hand, SVM exhibits significantly lower performance compared to the other models. Its low CV accuracy (0.650) suggests a poor ability to generalize to unseen data. Furthermore, the low F1-score (0.526) indicates both low precision and recall, meaning it struggles to accurately identify both positive and negative cases. This could be due to overfitting or a mismatch between the SVM model and the data being analyzed.

Finally, the performance of the Artificial Neural Network (ANN) falls between the two extremes. While achieving a moderate CV accuracy (0.801), it has a slightly lower ROC Area (0.879) compared to Logistic Regression. However, its balanced F1-score (0.779) suggests a decent trade-off between precision and recall.

According to Fig 6 : Logistic regression model provides the accurate prediction for the heart disease

dataset as compared to other two model (SVM and ANN).

From Fig.7:, the ROC curve in the image deviates from the diagonal line, suggesting the logistic regression model can distinguish between positive and negative cases better than random guessing.

The AUC of 0.88 indicates good overall performance. An AUC of 1 represents a perfect classifier, while 0.5 represents a random classifier. Generally, an AUC above 0.7 is considered acceptable, and above 0.8 is optimum.

5.3 PERFORMANCE EVALUATION BASED ON FEATURE SELECTION BY CORRELATION

After feature selection by correlation Table 5(a) , performance evaluation is received. Models in this table generalize better due to higher testing accuracy for all models (except SVM). This suggests better performance on unseen data.

All three models evaluated (Logistic Regression, SVM, and ANN) exhibited promising performance based on cross-validation. Their accuracy scores ranged from 0.798 to 0.835, suggesting a good ability to generalize to unseen data. Looking at the classifiers individually, Logistic Regression exhibits a slight trade-off between precision and recall. In Table 4(a), its precision (0.857) is higher than recall (0.72), suggesting a better ability to identify positive cases correctly. However, both metrics decrease in Table 5 (precision: 0.826, recall: 0.76), indicating a potential drop in overall performance for this model. SVM, on the other hand, shows a different pattern. In Table 4, its precision (0.783) and recall (0.72) are relatively balanced, meaning it performs similarly when identifying both positive and negative cases. Interestingly, in Table 5(a), both precision (0.8) and recall (0.8) improve, suggesting better handling of the specific data in that scenario. Finally, ANN demonstrates stable performance across both tables. In both cases, its precision (0.857/0.864) is slightly higher than recall (0.72/0.76), highlighting a focus on accurately classifying positive cases.

Interestingly, all models displayed a slight bias towards correctly identifying positive cases, as evidenced by slightly higher precision compared to recall. F1-scores, which provide a balanced view of these metrics, hovered around 0.78-0.81, indicating a trade-off between precision and recall. Comparing Table 4(b) to Table 5(b), Logistic Regression and ANN exhibit consistent performance. They maintain good CV accuracy (around 0.83 for Logistic Regression and 0.8 for ANN) and F1-score (around 0.8 for both) in both tables, indicating balanced performance. In both cases, their precision (around 0.85 for Logistic Regression and 0.8 for ANN) is slightly higher than recall (around 0.78 for Logistic Regression and 0.77 for ANN), suggesting a focus on correctly identifying positive cases. SVM stands out for its significant improvement between the two tables. Its CV accuracy jumps from a low 0.65 in Table 1 to a much stronger 0.818 in Table 5(b). This suggests that SVM might be more

sensitive to the specific data distribution or training parameters used. While all other metrics (precision, recall, F1-score, ROC area) also improve considerably in Table 5(b), SVM's performance remains more variable compared to Logistic Regression and ANN.

The other two model(SVM and ANN) shows significant improvement as compared to Logistic Regression in Fig. 8.

From Fig. 9:, the ROC curve in the image deviates from the diagonal line, suggesting the logistic regression model can distinguish between positive and negative cases better than random guessing.

The AUC of 0.87 indicates good overall performance. An AUC of 1 represents a perfect classifier, and 0.5 represents a random classifier. Generally, an AUC above 0.7 is considered acceptable, and above 0.8 is optimum.

5.4 PERFORMANCE EVALUATION BASED ON FEATURE SELECTION BY CHI-SQUARE TESTING:

Comparing Table 6(a) to Table 5(a), Analyzing the performance variations across the two tables reveals some interesting insights. Logistic Regression exhibits a slight decline in performance metrics like testing accuracy, precision, recall, and F1-score in Table 5(a). This dip could be attributed to two factors: overfitting or data variability. Overfitting occurs when the model memorizes the training data in Table 6(a) too well, leading to a performance drop on unseen data in Table 5(a). Alternatively, the data distribution itself might be slightly different between the tables, causing the model to perform less effectively. Whereas, SVM shows a slight improvement in testing accuracy and F1-score in Table 5(a). This could be due to SVM's inherent resistance to overfitting compared to Logistic Regression, resulting in more stable performance across datasets. Additionally, the specific data characteristics or features in Table 5(a)might better align with SVM's decision boundaries, leading to improved performance. Then, ANN exhibits a contrasting trend. While testing accuracy shows a slight improvement in Table 5(a), the ROC area metric, which often correlates with accuracy, shows a decrease. This unexpected outcome could be due to the stochastic nature of ANN training. Random variations during training runs can lead to minor performance diffrence, it might be prone to overfitting on the specific training data used in Table 5(a), explaining the dip in ROC area.

Comparing Table 6(b) and Table 5(b), Logistic Regression and ANN demonstrate generally good performance across the two datasets, with Logistic Regression showing slightly more consistency. Both models maintain good CV accuracy (around 0.83 for Logistic Regression and 0.8 for ANN) and F1-score (around 0.8 for both) in both tables. Similar to the previous comparison, their precision (around 0.84 for Logistic Regression and 0.83 for ANN) is slightly higher than recall (around 0.79 for

Logistic Regression and 0.77 for ANN), suggesting a focus on correctly identifying positive cases. SVM, however, exhibits a remarkable improvement in the second table. Its CV accuracy jumps significantly, from a low 0.65 in Table 6(b) to a much stronger 0.811 in Table 5(b). This suggests that SVM might be more sensitive to the specific data distribution or training parameters used. While all other metrics also improve considerably in Table 5(b), SVM's performance remains more variable compared to Logistic Regression and ANN

CHAPTER-6 DISCUSSION

Our exploration of these classifiers reveals some key trends.

Logistic Regression was giving consistent performance across datasets, balancing precision and recall effectively. This makes it a strong model. Even after Feature selection based on both Chi-square as well as Correlation. The accuracy of the model stayed ~85% both for testing data and training data. This suggests that no overfitting or underfitting of the model was present.

SVM, on the other hand, can be quite sensitive to the data and training settings. While it achieved impressive results in one scenario, SVM with full set of attributes was overfitted as it had accuracy of training data as 89.5% and testing data as 76.7%.. After feature selection the accuracy for training data 85.7% and testing data is 83.3% showing drastic improvement in the model as well as suggest no overfitting in the model.

ANN offers flexibility but shows some variation in performance across datasets. Initially, with full set of attributes, has training data as 93.7% and testing data as 0.767 which suggest overfitting. After feature selection following as 85% for training dataset as well as testing dataset.

Currently, LR is often used as a standalone model. However, the future might see it collaborating with other algorithms like Random Forests or Gradient Boosting. This approach, known as ensemble learning, leverages the strengths of each model. LR would contribute interpretability and simplicity, while other algorithms handle complex relationships in the data, leading to potentially better overall performance. The future might see LR combined with non-logistic components to create even more versatile models. For example, combining LR with neural networks could allow for incorporating non-linearities while maintaining LR's interpretability. This flexibility caters to the diverse demands of real-world applications.[14]

CHAPTER-7 CONCLUSION

In conclusion, following feature are responsible for heart disease according to UCI dataset, Thalassemia(Maximum heart rate achieved),num_vessels (Number of major vessels colored by fluoroscopy),st_depression(induced by exercise relative to rest) , exercise_angina, chest_pain , st_slope(The slope of the peak exercise ST segment).

Research on Heart Disease Prediction helps in healthcare systems can use risk prediction data to optimize resource allocation for preventive care programs and interventions aimed at reducing the overall burden of heart disease. Moreover, Healthcare providers can leverage these applications to identify individuals at high risk of heart disease within a population. This allows for targeted screening programs, focusing resources on those who might benefit the most from early intervention

10% Overall Similarity

Top sources found in the following databases:

- 7% Internet database
- Crossref database
- 7% Submitted Works database
- 5% Publications database
- Crossref Posted Content database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	mdpi.com Internet	2%
2	ncbi.nlm.nih.gov Internet	1%
3	medium.com Internet	<1%
4	fastercapital.com Internet	<1%
5	University of Sydney on 2023-03-01 Submitted works	<1%
6	Youness Hakam, Ahmed Gaga, Mohamed Tabaa, Benachir El Hadadi. "I... Crossref	<1%
7	d-nb.info Internet	<1%
8	Sunway Education Group on 2021-11-19 Submitted works	<1%

9	University of Bolton on 2023-05-15 Submitted works	<1%
10	link.springer.com Internet	<1%
11	University of Strathclyde on 2024-07-06 Submitted works	<1%
12	De Montfort University on 2023-05-12 Submitted works	<1%
13	Universidad Politecnica Salesiana del Ecuador on 2024-06-29 Submitted works	<1%
14	University of Surrey on 2024-04-26 Submitted works	<1%
15	National School of Business Management NSBM, Sri Lanka on 2024-0... Submitted works	<1%
16	University of Queensland on 2023-11-10 Submitted works	<1%
17	dspace.daffodilvarsity.edu.bd:8080 Internet	<1%
18	Xinkang Li, Feng Zhang, Liangzhen Zheng, Jingjing Guo. "Advancing Ec... Crossref	<1%