IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 1, January 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

9940 572 462 | 6381 907 438 | ijircce@gmail.com | www.ijircce.com

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 11, Issue 1, January 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1101020 |

An Analytical Study of Feature Selection and Classification Methods for Heart Disease Prediction

Azhar M A

Lecturer in Computer Engineering, Government Polytechnic College, Pala, Kerala, India

ABSTRACT: Feature selection is a pivotal process in machine learning, designed to identify a subset of the most relevant features from a dataset to optimize model construction. By eliminating irrelevant and redundant features, this process reduces model complexity, enhances performance, and accelerates learning. In classification tasks, feature selection is particularly essential for improving accuracy and efficiency, especially when working with large datasets. It not only expedites the learning process but also ensures that classifiers are more generalizable to unseen data.Moreover, feature selection provides valuable insights into the relationships between features and the target variable, offering a deeper understanding of the data. The primary objective is to improve predictive accuracy while simplifying the computational demands of the model.This paper presents an overview of various classification methods, emphasizing techniques that leverage threshold values and benchmark algorithms to identify optimal feature subsets. Additionally, we review and evaluate several feature selection techniques using standard datasets, demonstrating their effectiveness in reducing computational complexity and improving classification accuracy.

KEYWORDS: feature selection, machine learning, classification, feature subset selection, comparative evaluation

I. INTRODUCTION

Feature selection plays a pivotal role in data pre-processing and is an indispensable component of the machine learning process. It serves as one of the most frequently used and crucial techniques, particularly in tasks involving dimensionality reduction, where the goal is to eliminate noisy and redundant features. Dimensionality reduction methods can be broadly categorized into feature extraction, feature transformation, and feature selection.

Feature extraction involves projecting original features into a new lower-dimensional feature space, where the newly constructed features are often combinations of the original ones or transformations of raw data tailored for modeling. Feature transformation, on the other hand, refers to creating new features based on the existing ones, generally to improve model accuracy. A popular technique in this category is Principal Component Analysis (PCA), which applies an orthogonal transformation to produce a set of linearly uncorrelated variables derived from the original feature set, enhancing the accuracy of subsequent algorithms.

In contrast, feature selection aims to identify a smaller subset of features that minimize redundancy while maximizing relevance to the target variable, such as class labels in classification tasks. Unlike feature extraction and transformation, feature selection operates by choosing an optimal set of features based on specific criteria, without altering the original features themselves. This is accomplished through various methods, such as Linear Regression and Decision Trees, among others. As Robert Neuhaus aptly put it, "Feature selection is itself useful, but it mostly acts as a filter, muting out features that are not useful in addition to your existing features."

Feature selection plays a key role in improving predictive model accuracy. Alongside feature extraction and transformation, it contributes to enhancing learning performance, reducing the computational complexity of the data, constructing more generalizable models, and decreasing storage requirements. One of the advantages of feature selection over the other methods is its ability to retain the original features without any transformation, preserving their physical meaning. This feature makes it especially valuable for tasks where interpretability and readability are essential, such as in medical diagnostics or disease detection, where understanding the relationships between features can have significant practical implications.

II. IMPORTANCE OF FEATURE SELECTION

Dimensionality reduction is essential in predictive modeling, as features provide valuable information about the target variable. While more features typically lead to better classification by offering more information, an excessive number of features can result in the "curse of dimensionality," which increases computational complexity and storage

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 11, Issue 1, January 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1101020 |

requirements. Additionally, irrelevant features introduce noise, negatively affecting the performance and accuracy of learning algorithms. Feature selection seeks to identify the most relevant features, denoted as F = {X1, X2, X3, ..., XN}, and to select a subset 'b' where b < n. The optimal value of 'b' varies depending on the specific problem domain and may not always be determined in advance. The need for feature selection arises from several key factors:

Reducing Computational Complexity: By processing fewer features, the model requires less time and resources, improving efficiency.

Removing Noise: Redundant or irrelevant features can act as noise, which can hinder the learning process and degrade the performance of the model.

Gaining Insight: Feature selection enables a clearer understanding of the underlying relationships between features and the target variable, aiding in better model interpretability and decision-making.

Redundant Features

Redundant features provide no additional information and can negatively impact the performance of learning algorithms, contributing to the curse of dimensionality. The presence of too many features leads to slower learning and reduced accuracy. Feature reduction techniques aim to optimize model performance by simplifying the model while either improving or maintaining accuracy.

For instance, if two features, X1 and X2, are linearly related (e.g., X2 = 2X1 - 1), one of them can be removed as it is redundant. Redundancy can also arise from non-linear relationships. For example, if X2 = 2X1^3 - 10X1^2 + 5X1 - 7, X2 can be derived from X1, making it redundant.

The correlation coefficient is commonly used to measure the relationship between features. If two features are independent, their correlation coefficient will be zero. However, it's important to note that a zero correlation coefficient doesn't necessarily imply independence, and thus the correlation coefficient may not always be the ideal measure for determining feature relationships.

The dependency between features can be categorized as either dependent or independent, each of which requires different approaches for handling redundancy in feature selection.

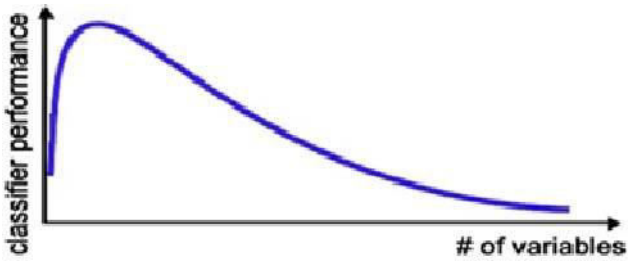

Figure1:Importance of features

III. FEATURE SELECTION PROCESS

A. Basic Steps

Feature selection is a critical step in the machine learning pipeline, aimed at identifying the most relevant features from a given dataset. The process involves four fundamental steps:

1. **Define an Objective Function**: The first step in the feature selection process is to define an objective function that measures the importance of the given collection of features. This function serves as the foundation for selecting the optimal subset of features.
2. **Assign Values to Feature Subsets**: The objective function assigns a value to every subset of features from the dataset. These values represent the relevance and significance of each subset in the context of the given task.
3. **Formulate an Algorithm**: An algorithm is formulated to perform feature selection based on the objective function. This algorithm guides the selection process by evaluating the quality of different subsets of features.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 11, Issue 1, January 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1101020 |

4. **Optimization**: Depending on the nature of the objective function, optimization is performed—either through maximization or minimization. The goal is to identify the subset of features that achieves the best performance according to the optimization criterion.

Let F={X1,X2,X3,…,XN}F = \{X_1, X_2, X_3, \dots, X_N\}F={X1,X2,X3,…,XN} represent the set of features, and let BBB denote the number of elements in a feature subset where B<NB < NB<N. The total number of possible subsets is 2N2^N2N, which grows exponentially, making it computationally infeasible to enumerate every subset. As such, the powerset P(F)P(F)P(F) includes all possible subsets of FFF, and the objective function JJJ assigns a value to each subset.

The goal of the optimization process is to find a subset B0B_0B0 such that for maximization:

J(B0)≥J(B)∀B⊆FJ(B_0) \geq J(B) \quad \forall B \subseteq FJ(B0)≥J(B)∀B⊆F

and for minimization:

J(B0)≤J(B)∀B⊆FJ(B_0) \leq J(B) \quad \forall B \subseteq FJ(B0)≤J(B)∀B⊆F

This process is designed to minimize the misclassification probability, enhancing the model's generalization ability. For example, if N=10N = 10N=10 and B=2B = 2B=2, there are (102)=45\binom{10}{2} = 45(210)=45 possible combinations of feature subsets. Each combination is evaluated by calculating its associated misclassification probability, and the subset with the lowest misclassification rate is selected as the optimal feature subset.

B. Feature Subset Selection and Evaluation Methods

Feature selection techniques can be categorized into three primary methods: optimal methods, heuristic methods, and randomized methods.

- **Optimal Methods**: These methods strive to find the ideal subset of features by conducting exhaustive searches or applying mathematical optimization techniques.
- **Heuristic Methods**: These methods rely on approximation techniques or rules of thumb, guided by empirical results or domain knowledge, to identify the most relevant features.
- **Randomized Methods**: These methods use random search strategies to explore the feature space and identify promising feature subsets through chance-based sampling.

For evaluating feature selection techniques, four commonly used evaluation methods are employed:

1. **Filter Methods**: These methods assess individual features independently of the learning algorithm, typically based on statistical metrics such as correlation or mutual information. Filter methods do not require a machine learning model for evaluation, making them computationally efficient.
2. **Wrapper Methods**: These methods evaluate subsets of features by training and testing a machine learning model on them. The feature subset that results in the best model performance is chosen. Wrapper methods are computationally expensive but tend to yield better feature subsets tailored to the specific model.
3. **Embedded Methods**: These methods perform feature selection during the training process of the model. Embedded methods incorporate feature selection within the model's learning algorithm, optimizing both the model and the feature subset simultaneously.
4. **Hybrid Methods**: These methods combine elements of filter, wrapper, and embedded approaches to leverage the advantages of each. Hybrid methods aim to strike a balance between computational efficiency and model accuracy.

IV. SURVEY ON HEART DISEASE PREDICTION MODELS

This section provides a comprehensive survey comparing different data mining techniques aimed at identifying the best approach for predicting heart disease with minimal effort. The analysis focuses on the dependencies between features in a dataset and their impact on the prediction accuracy of various machine learning algorithms. The survey is divided into two sections based on the inclusion or exclusion of feature selection techniques.

1) A. Analysis of Dataset without Feature Selection

The following studies evaluate different approaches for heart disease prediction without the use of feature selection techniques:

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 11, Issue 1, January 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1101020 |

1. **Shouman M. et al. [1]**: This study proposed a combination of k-means clustering and decision trees for heart disease prediction. The authors focused on improving the efficiency of k-means clustering by suggesting different centroid selection methods. Using the Cleveland Clinic Foundation heart disease dataset with thirteen attributes, they evaluated sensitivity, specificity, and accuracy across various centroid selection methods and cluster numbers. Their results demonstrated that combining k-means clustering with decision trees improved accuracy, achieving the best performance of 83.9% with the inlier method and two clusters. However, the study did not incorporate a feature selection method.
2. **Jabbar M. A. et al. [2]**: This paper introduced a new algorithm for mining association rules from medical data using digit sequences and clustering. The dataset was partitioned into equal-sized clusters, each of which was processed individually to calculate frequent itemsets. This approach reduced memory requirements and improved scalability. The study utilized a dataset with fourteen attributes.
3. **Sudha A. et al. [3]**: This study examined the use of classification algorithms such as Naive Bayes, Decision Tree, and Neural Network for predicting stroke diseases. The results showed that Neural Networks outperformed Decision Trees and Naive Bayes in terms of accuracy. The authors also highlighted the importance of data preprocessing, specifically the removal of irrelevant data before mining.
4. **Amin S. U. et al. [4]**: This research developed a hybrid system combining Genetic Algorithms (GA) and Neural Networks for heart disease prediction. The GA optimized the initial weights of the Neural Network, resulting in improved performance. The study achieved a training accuracy of 96.2% and a validation accuracy of 89%. The authors suggested that hybrid data mining techniques could lead to more accurate clinical decision support systems.
5. **Deepika N. et al. [5]**: This study proposed using association rules for classifying heart attack patients. The data warehouse was preprocessed to improve mining efficiency, and association rules were applied to handle missing values. The authors aimed to enhance the accuracy of their heart disease prediction system by exploring different data mining techniques and feature selection methods.

2) B. Analysis of Dataset with Feature Selection

This section highlights studies that incorporate feature selection techniques to improve the accuracy and efficiency of heart disease prediction models. This survey aims to provide a clear understanding of various prediction models in data mining and identify the best model for further research. It compares different techniques, highlighting their accuracy levels in a tabular format. The study suggests that a hybrid approach, combining multiple models, might outperform single-model techniques.

1. **Durga Devi et al. [6]**: This study presents a heart disease prediction system using a Genetic Algorithm for feature selection and a Radial Basis Function (RBF) Network for classification. By reducing the number of attributes through the GA, they achieved better accuracy in predicting heart disease risk. The results showed that the RBF Network combined with feature selection outperformed other data mining techniques like Naive Bayes and J48.
2. **M. Anbarasi et al. [7]**: This research aimed to improve the accuracy of heart disease prediction by reducing the number of attributes. The Genetic Algorithm identified the most significant attributes, reducing the original set of thirteen attributes to six. The performance of Naive Bayes, Classification by Clustering, and Decision Tree classifiers was compared using the reduced attribute set. The Decision Tree classifier showed the best performance, while Naive Bayes maintained consistent accuracy. Classification via clustering performed poorly.
3. **Subanya et al. [8]**: This study aimed to optimize feature selection for cardiovascular disease diagnosis using a metaheuristic algorithm. The Artificial Bee Colony (ABC) algorithm was employed to identify the most relevant features, with ABC being a swarm intelligence-based optimization technique inspired by the foraging behavior of honey bees. The results showed that the ABC-SVM approach, combining ABC with Support Vector Machines (SVM) for classification, outperformed traditional feature selection methods like reverse ranking, achieving good classification accuracy with only seven features. The ABC algorithm ensures fast convergence and is easy to implement due to fewer control parameters, making it effective for heart disease diagnosis.
4. **Haider et al. [9]**: This study investigated the performance of various data mining techniques for heart disease prediction using different datasets. The authors evaluated KStar, J48, SMO, Bayes Net, and Multilayer Perceptron classifiers on both a standard and a collected dataset, measuring performance using predictive accuracy, ROC curves, and AUC values. The study found that Bayes Net and SMO classifiers performed best for heart disease prediction across both datasets, highlighting the importance of selecting the appropriate data mining technique for accurate heart disease prediction.
5. **J. Theor et al. [10]**: This research explored the use of Support Vector Machines (SVMs) for heart disease classification. SVMs work by finding a hyperplane that maximizes the margin between two classes. SVMs are particularly well-suited for high-dimensional and nonlinear datasets, offering good generalization performance and

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 11, Issue 1, January 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1101020 |

being less prone to overfitting. However, SVMs can be computationally intensive, especially for large datasets, and require careful selection of kernel functions and parameters for optimal performance.

6. **N. Bhatia C. et al. [11]**: This paper surveyed various data mining techniques for heart disease prediction, discussing both structure-based and model-based algorithms. It provided a comprehensive overview of different approaches, helping researchers and practitioners select the most suitable technique for their needs. The survey also explored Nearest Neighbor (NN) techniques for heart disease prediction, categorized into structureless and structure-based methods. Both types aimed to enhance the basic K-Nearest Neighbors (KNN) algorithm, with structure-based methods focusing on reducing computational complexity.
7. **T. M. Lakshmi, A. Martin, R. M. Begum, and V. P. Venkatesan [12]**: This study compared the performance of different decision tree algorithms (ID3, C4.5, and CART) using qualitative data from educational data mining. The results indicated that the CART algorithm, using the Gini Index for splitting, showed higher classification accuracy than ID3 and C4.5, which rely on Information Gain and Gain Ratio, respectively. This study highlights the potential of decision tree algorithms for analyzing educational data and suggests using genetic algorithms for future work to identify key qualitative factors.

V. CLASSIFICATION MODELING

This section describes important machine learning models used for feature selection and classification in the context of heart disease prediction. It also analyzes the accuracy of these algorithms based on various research studies.

UCI DATA SET		
	Accuracy(Withoutfeatureselection)	Accuracy(Withfeatureselection)
NaïveBayes	68	78
LinearModel	78	85
LogisticRegression	79	84
ANN	84.3	88.4
DecisionTree	80.4	87
RandomForest	82.7	88
SupportVectorMachine	79	88

Decision Tree

A Decision Tree is a tree-like classification structure where branches and nodes are constructed based on the characteristics of the data. It is built in a top-down approach, where each node corresponds to a decision based on a specific feature, and the branches represent possible outcomes that lead to further nodes or terminal classifications. This method is simple, computationally efficient, and provides clarity and interpretability in model outcomes. Decision Trees are particularly effective for handling both categorical and numerical data, making them versatile for classification tasks. However, they are prone to overfitting, where the model becomes excessively tailored to the training data, and instability, as minor variations in input data can lead to significantly different tree structures.

Authors	*DecisionTree Accuracyin%*
Zriqatetal2016	97
Palaniappanetal2007	94
Tuet aL2009	81.41
Cheung2001	81.11
BandarageShehani	77
SenthilkumarMohan2019	75.8
Andreeva.p2006	75.73
SitarTautetal2009	60.4
Rajkumaretall 2010	52
sheikAbdulla2017	50.67

- **Description**: A tree-like classification model that builds a structure of branches and nodes based on the data's characteristics. It is a simple and fast method for classification.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 11, Issue 1, January 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1101020 |

- **How it works**: The tree is constructed in a top-down manner, where each node represents a decision based on a specific feature. The branches represent possible outcomes of the decision, leading to further nodes or ultimately to a classification.
- **Advantages**: Easy to understand and interpret, capable of handling both categorical and numerical data.
- **Disadvantages**: Prone to overfitting, can be unstable (small changes in data can result in significantly different tree structures).

Support Vector Machine (SVM)

Support Vector Machines (SVM) are supervised learning models designed to identify the optimal hyperplane that separates data points into distinct classes. SVM operates by maximizing the margin between classes, creating a robust and clear boundary for classification. This model is particularly effective in high-dimensional feature spaces and can handle non-linear data using kernel functions, commonly referred to as the "kernel trick." Despite its advantages, SVM can be computationally demanding and requires careful parameter tuning to achieve optimal performance, making it less efficient for very large datasets.

Authors	*SVM Accuracy in%*
daveetal2008	88.9
SASababetal2016	87.79
SenthilkumarMohan2019	86.1
S Radhimeenakshi2016	84.7
Daltoglou&Thefwall2010	82.3
R.ChithraandDr.Vsreenivsan 2018	82
kennedy&Inkpin2006	80-85.9
Pangetal(2002)	77–82.9

Naive Bayes

Naive Bayes is a probabilistic classifier that applies Bayes' theorem under the assumption of feature independence. It calculates the probability of a data point belonging to each class by considering the individual probabilities of its features, assigning the class with the highest probability to the data point. This method is simple and computationally efficient, particularly suited for high-dimensional datasets. However, the independence assumption often does not hold in real-world data, which can limit its accuracy.

Authors	NAÏVEBAYERS Accuracy in %
Palaniappu2007	95
daveetal2008	88.9
Srinivas2010	84.14
Cheung2010	81.48
yeetal2009	80.71-85.14
Pangetal (2002)	77–82.9
SenthilkumarMohan2019	75.8
SitarTaut2009	62.03
Annett&Kondrak2008	>75.00
Andrava2006	78.563

- **Description**: Naive Bayes is a probabilistic classifier based on Bayes' theorem, assuming independence between features.
- **How it works**: It calculates the probability of a data point belonging to each class based on the probabilities of its individual features. The class with the highest probability is assigned to the data point.
- **Advantages**: Simple and fast, works well with high-dimensional data.
- **Disadvantages**: The assumption of feature independence is often unrealistic.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 11, Issue 1, January 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1101020 |

Random Forest

Random Forest is an ensemble learning technique that aggregates the predictions of multiple decision trees to enhance accuracy and reduce overfitting. Each tree in the "forest" is trained on a random subset of the data and features, ensuring diversity in predictions. The final output is determined by averaging or voting on the individual tree predictions. While Random Forest models are robust and accurate, they can be computationally intensive and lack the interpretability of single decision trees.

Authors	RandomForest Accuracyin%
RGSaboji2017	98
M AJabbar etal2016	98
SenthilkumarMohan2019	86.1
T SBrsimietal2018	81.62
sheikAbdulla&RR Rajalaxmi2017	63.3

- **Description**: Random Forest is an ensemble learning method that combines multiple decision trees to improve accuracy and reduce overfitting.
- **How it works**: It creates a "forest" of decision trees, each trained on a random subset of the data and features. The final prediction is made by aggregating the predictions of all the trees.
- **Advantages**: Robust and accurate, less prone to overfitting than individual decision trees.
- **Disadvantages**: Can be computationally expensive, less interpretable than single decision trees.

K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) is a distance-based classification technique that assigns a data point to the majority class among its k-nearest neighbors in the feature space. For each new data point, KNN identifies the closest k data points in the training dataset and predicts the class with the highest frequency among them. While KNN is straightforward and requires no training phase, its computational cost increases with the size of the dataset, and its performance is highly sensitive to the choice of k.

Authors	KNNAccuracy in %
JSingh etal2016	80.1
Bandarage2018	75
S Radhimeenakshi2016	73
BandarageShehani Sanketha2018	45.67
Pangetal(2002)	44.34

- **Description**: KNN is a classification scheme based on distance measures. It classifies a data point based on the majority class among its k-nearest neighbors in the feature space.
- **How it works**: For a new data point, KNN finds the k closest data points in the training set and assigns the new point to the class that is most frequent among its neighbors.
- **Advantages**: Simple to implement, no training period required.
- **Disadvantages**: Can be computationally expensive for large datasets, sensitive to the choice of k.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 11, Issue 1, January 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1101020 |

Artificial Neural Network (ANN)

Artificial Neural Networks (ANNs) consist of layers of interconnected processing units (neurons) that learn to map inputs to outputs by adjusting connection weights through training. Input data passes through multiple layers, undergoing complex transformations and non-linear computations to produce predictions. ANNs are highly adaptable and capable of capturing intricate patterns in data. However, their training requires substantial computational resources and large datasets, and they are prone to overfitting if not regularized properly.

Authors	ANN Accuracy in %
MGudadheetat2010	97.5
TKarayilan&Okilic2017	95.56
Palaniappan2007	93.54
S Radhimeenakshi2016	89.01
S Radhimeenakshi2016	89
R.ChithraandDr.V Sreenivsan2018	88
BandarageShehaniSanketha 2018	87.4
SenthilkumarMohan 2019	86.1
AndreevP2006	82.77
DeBeuleetal	82
Kangwnriyakuletal2010	78.43

- **Description**: ANNs are composed of interconnected units (neurons) organized in layers. They learn by adjusting the weights of the connections between neurons.
- **How it works**: ANNs process input data through multiple layers, performing complex computations and non-linear transformations to arrive at an output.
- **Advantages**: Can learn complex patterns, adaptable to different types of data.
- **Disadvantages**: Can be difficult to train, prone to overfitting, often require large amounts of data.

Hybrid Approaches

Hybrid approaches combine multiple machine learning models to leverage their individual strengths and address their limitations. By integrating different algorithms, these systems can often achieve superior performance compared to any individual model alone. For instance, a hybrid system might first utilize decision trees for feature selection and then apply more complex models such as Random Forests or linear models for classification. These methods often result in improved accuracy, robustness, and generalization ability. However, hybrid models tend to be more computationally intensive and can require careful fine-tuning of parameters to achieve optimal results.

- **Description**: Hybrid approaches combine different machine learning models to capitalize on their respective strengths, thereby improving overall performance.
- **Example**: A hybrid model might use a decision tree to identify relevant features and then employ Random Forests or linear models for the final classification stage.
- **Advantages**: These models can lead to enhanced accuracy, improved robustness, and better generalization.
- **Disadvantages**: Hybrid approaches are generally more complex to implement and require extensive fine-tuning to optimize performance.

This section outlines the steps involved in the construction of a hybrid model for heart disease prediction, utilizing both decision trees and Random Forests for feature selection, followed by classification using a hybrid system of Random Forest and Linear Regression.

Step 1: Data Pre-processing

The first step involves preprocessing the dataset to remove any duplicate, missing, or unknown data. This is crucial to ensure the quality of the data before it is fed into machine learning models. Pre-processing techniques such as imputation, deletion, or interpolation can be applied to handle missing values, while duplicates can be identified and removed to avoid bias in model training.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 11, Issue 1, January 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1101020 |

Step 2: Decision Tree Construction

Once the data is clean, a Decision Tree is constructed on the dataset. The Decision Tree algorithm is employed to model the relationship between the input features and the target variable (e.g., presence or absence of heart disease). This model uses a tree-like structure, where each node represents a decision rule, and branches correspond to the outcomes of these rules.

Step 3: Feature Subset Construction

After the decision tree is built, the next step involves traversing through its branches to construct a subset of features. The features selected by the tree are deemed significant, and they contribute to the decision-making process. These features are then extracted as a potential subset for further analysis.

Step 4: Training Random Forest

The next phase involves training a Random Forest model using the subsets of features identified by the Decision Tree. Random Forest, as an ensemble learning method, aggregates the predictions from multiple decision trees, enhancing accuracy and reducing the risk of overfitting. Each subset of features is used to train an individual tree, and the model performance is evaluated to determine which subset yields the highest accuracy.

Step 5: Hybrid Model Learning and Classification

In the final step, a hybrid model is created by combining the Random Forest classifier with a Linear Regression model. The selected features from the previous step are used for training the hybrid model. Random Forest is used for feature extraction and classification, while Linear Regression helps in refining the classification model by providing linear approximations to the data. This hybrid approach aims to leverage the advantages of both models to achieve improved prediction accuracy.

VI. EVALUATION OF CLASSIFICATION TECHNIQUES FOR HEART DISEASE PREDICTION

This section presents an evaluation of the performance metrics used to assess different classification techniques in predicting heart disease. The key metrics include accuracy, classification error, F-measure, precision, and sensitivity, which provide comprehensive insights into the effectiveness of the models.

Confusion Matrix:

A confusion matrix is an essential tool for evaluating the performance of a classification model. It provides a clear summary of the prediction results by categorizing instances into four categories: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). These categories are defined as follows:

- **True Positive (TP)**: The model correctly predicts a positive case (i.e., the patient has heart disease).
- **True Negative (TN)**: The model correctly predicts a negative case (i.e., the patient does not have heart disease).
- **False Positive (FP)**: The model incorrectly predicts a positive case (i.e., the patient is predicted to have heart disease when they do not).
- **False Negative (FN)**: The model incorrectly predicts a negative case (i.e., the patient is predicted not to have heart disease when they do).

Evaluation Metrics:

Several metrics derived from the confusion matrix are used to evaluate model performance:

	1	0
1	TP	FN
0	FP	TN

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 11, Issue 1, January 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1101020 |

- **Accuracy**: This is the proportion of correctly classified instances (both positive and negative) out of the total instances.
  - **Formula**: Accuracy=TP+TNTP+TN+FP+FN\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}Accuracy=TP+TN+FP+FNTP+TN
- **Precision**: This metric represents the proportion of true positive predictions among all positive predictions made by the model. It is also known as the positive predictive value.
  - **Formula**: Precision=TPTP+FP\text{Precision} = \frac{TP}{TP + FP}Precision=TP+FPTP
- **Sensitivity (Recall)**: This metric measures the proportion of true positive predictions among all actual positive instances. It is also known as the true positive rate.
  - **Formula**: Sensitivity=TPTP+FN\text{Sensitivity} = \frac{TP}{TP + FN}Sensitivity=TP+FNTP
- **F-measure**: This is the harmonic mean of precision and recall, providing a balanced measure of a model's accuracy by considering both false positives and false negatives.
  - **Formula**: F-measure=2×Precision×RecallPrecision+Recall\text{F-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}F-measure=2×Precision+RecallPrecision×Recall

Models	Accuracy	ClassificationError	Precision	Sensitivity
NaïveBayers	75.8	24.2	90.5	79.8
GeneralizedLinearModel	85.1	14.9	88.8	94.9
DecisionTree	85	15	86	98.8
RandomForest	86.1	13.9	87.1	98.8
SupportVector Machine	86.1	13.9	86.1	100
HRFLM(Basepaper)	88.4	11.6	90.1	92.8

Table Results of various models with hybrid model

General Analysis

This section addresses the challenges and limitations encountered in feature selection within machine learning, particularly in the context of heart disease prediction.

Challenges of Feature Selection:

Feature selection is a critical step in machine learning model construction. However, it poses several challenges, including:

- **Difficulty in Finding the Optimal Subset**: The vast number of possible feature combinations makes it difficult to determine the best subset.
- **Computational Infeasibility**: An exhaustive search (evaluating all possible subsets of features) can be computationally expensive and often infeasible, especially with large datasets.
- **No Universal Solution**: There is no single feature selection algorithm that guarantees optimal results for every dataset or task. The choice of feature selection method is often dependent on the specific characteristics of the data.

"The Best Two Independent Measurements Are Not the Two Best":

This principle highlights an important aspect of feature selection: selecting the two features with the highest individual performance scores may not necessarily yield the best overall performance. The interaction and combined effect of features often have a greater impact on model accuracy than their individual contributions. Therefore, feature selection should consider not just the individual strength of each feature but also how features interact with one another in the context of the overall prediction task.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 11, Issue 1, January 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1101020 |

VII. INFERENCES

The following inferences have been drawn based on the evaluation and analysis of feature selection techniques and classification models for heart disease prediction:

Correlation and Feature Relationships:

The correlation coefficient is a widely used tool to measure the linear relationship between two features. While it can provide insights into the degree of association between features, it is not always the definitive measure of feature relationships. Features that are statistically independent might still exhibit a non-zero correlation coefficient due to underlying non-linear interactions or other factors. Consequently, relying solely on correlation for feature selection might not be sufficient, as it could overlook complex dependencies between features that affect the model's performance. Therefore, a deeper understanding of the data and its inherent structure is essential for effective feature selection.

Hybrid Feature Selection:

Hybrid methods that combine multiple feature selection techniques can significantly enhance the effectiveness of feature selection. By leveraging the strengths of different methods, hybrid approaches are better suited to capture complex interrelationships among features. These methods can help identify a more representative subset of features, which in turn leads to improved performance in machine learning algorithms. Hybrid approaches address the limitations of individual techniques and offer a more comprehensive view of the feature space, resulting in higher accuracy and robustness in predictive models.

VIII. CONCLUSIONS

The following conclusions have been drawn from the study of feature selection techniques and their application to heart disease prediction models:

No Universal Feature Selection Technique:

There is no single feature selection technique that consistently outperforms others across all types of datasets. The effectiveness of a feature selection method is heavily dependent on the specific dataset being used and the inherent characteristics of the features. What works well for one dataset may not necessarily be effective for another, highlighting the need for tailored approaches to feature selection.

Hybrid Methods for Improved Performance:

Hybrid feature selection techniques offer a promising approach for improving machine learning model performance. By combining the strengths of multiple feature selection methods, hybrid approaches are more adaptable and can better handle the unique characteristics of different datasets. These methods are particularly effective in capturing complex feature interactions that individual techniques might overlook.

Reduced Feature Subsets for Better Prediction:

Using a reduced set of relevant features, derived through feature selection, can enhance the performance of predictive models. When non-relevant or redundant features are eliminated, models can focus on the most significant attributes, leading to improved accuracy and reduced computational complexity. This demonstrates the importance of efficient feature selection in developing robust heart disease prediction models.

Stability of Hybrid Methods:

Hybrid methods tend to exhibit greater stability and are less sensitive to variations in the data compared to individual feature selection techniques. By combining different methods, these approaches are less prone to overfitting and are more capable of maintaining high performance across varying datasets, ensuring robustness in real-world applications.

REFERENCES

[1] Jabbar MA, Chandra P, Deekshatulu BL. Cluster-based association rule mining for heart attack prediction. Journal of Theoretical and Applied Information Technology. 2011; 32(2):197–201.

[2] Sudha A, Gayathiri P, Jaisankar N. Effective analysis and predictive model of stroke disease using classification methods. International Journal of Computer Applications. 2012; 43(14):26–31.

[3] Amin SU, Agarwal K, Beg R. Genetic neural network-based data mining in prediction of heart disease using risk factor. Proceeding of IEEE Conference on Information and Communication Technologies (ICT); 2013 Apr. p. 1227–31.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 11, Issue 1, January 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1101020 |

[4] Deepika N, Chandrashekar K. Association rule for classification of Heart Attack Patients. International Journal of Advanced Engineering Science and Technologies. 2011; 11(2):253–57.
[5] Sellappan Palaniappan and Rafiah Awang. Intelligent Heart Disease Prediction System Using Data Mining Techniques. IEEE, 2008.
[6] M Anbarasi, E Anupriya, S Iyengar, Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm. 2012.
[7] B Subanya, Dr. R. R. Rajalaxmi. Feature Selection using Artificial Bee Colony for Cardiovascular Disease Classification. 2014.
[8] M Sultana, A Haider, M S. Uddin. "Analysis of data mining techniques for heart disease prediction," 2016-2017.
[9] D. K. Srivastava, L. Bhambhu. "Data classification using support vector machine," J. Theor. Appl. Inf. Technol., 2009.
[10] N. Bhatia, C. Author. "Survey of Nearest Neighbor Techniques," IJCSIS Int. J. Comput. Sci. Inf. Secur., vol. 8, no. 2, pp. 302–305, 2010.
[11] T. M. Lakshmi, A. Martin, R. M. Begum, V. P. Venkatesan. "An Analysis on Performance of Decision Tree Algorithms using Student's Qualitative Data," Int. J. Mod. Educ. Comput. Sci., vol. 5, no. 5, pp. 18–27, 2013.
[12] Simon Fong, Raymond Wong, and Athanasios V. Vasilakos. "Accelerated PSO SwarmSearch Feature Selection for Data Stream Mining Big Data."
[13] T. Peter, K. Sonausundaram. "An empirical study on prediction of heart disease using classification data mining techniques," IEEE International Conference and Management, 2012.
[14] A. Khemphila, V. Boonjing. "Comparing Performance of logistic regression decision trees and neural networks for classifying heart disease patients," International Conference on Computer Information Systems and Industrial Management Applications.
[15] G. Purusothaman, P. Krishnakumari. "Data Mining Techniques on Risk Prediction: Heart Disease," Indian Journal of Science and Technology, 2015.
[16] Thomas M. Cover. "The Best Two Independent Measurements Are Not the Two Best," 1974.
[17] A. Durga Devi. "Enhanced Prediction of Heart Disease by Genetic Algorithm and RBF Network," 2015.
[18] Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava. "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques."
[19] M. Gjoreski, A. Gradišek, M. Gams, M. Simjanoska, A. Peterlin, G. Poglajen. "Chronic heart failure detection from heart sounds using a stack of machine-learning classifiers," Proceedings - 2017 13th International Conference on Intelligent Environments, IE 2017, pp. 14–19, 2017.
[20] G. Savarese, L. Lund. "Global Public Health Burden of Heart Failure," Card. Fail. Rev., vol. 3, no. 1, 2017.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details