Heart disease diagnosis

Performance evaluation of supervised machine learning and feature selection techniques

2111CS020337-Pavan Reddy. S

2111CS020338-Pavan Reddy. S

2111CS020339- Pavan Sai .G

2111CS020340-Pavitra Rupa .K

2111CS020341-Poojith Kumar. D

2111CS020342-Poojitha. B

Guided by

Prof: Ravinder

1.Abstract

Heart disease diagnosis is a critical task in modern healthcare, demanding accurate and efficient methods. This study focuses on evaluating the performance of various supervised machine learning algorithms and feature selection techniques for the diagnosis of heart disease. The algorithms considered include Naïve Bayes, Decision Tree, KNN, SVM, and Logistic Regression. Additionally, different feature selection techniques are employed to identify the most relevant features for improved diagnostic accuracy. Furthermore, the study investigates the impact of feature selection on algorithm performance. Feature selection techniques are applied to identify the subset of attributes that contribute most effectively to heart disease diagnosis. The combination of various algorithms and feature selection methods yields insights into which approaches are most suitable for accurate and efficient heart disease diagnosis

Keywords: KNN, SVM, DecisionTree, Naïve Bayes, Logistic Regression, feature selection.

1.1 PROBLEM STATEMENT

Cardiovascular diseases remain a leading cause of mortality worldwide, emphasizing the critical need for accurate and efficient diagnostic tools. In this context, the application of machine learning algorithms for heart disease diagnosis has gained significant attention. However, the performance of these algorithms is contingent on various factors, including the choice of algorithm and the relevance of input features.

1.2 TECHNIQUES

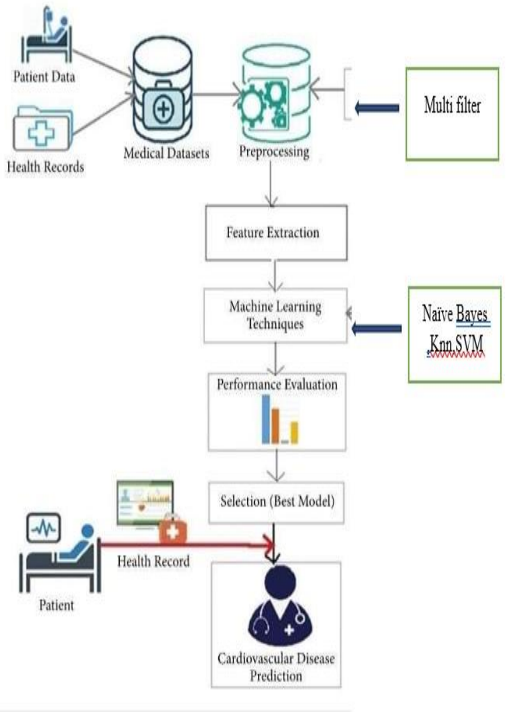
Decision Tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane .It chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

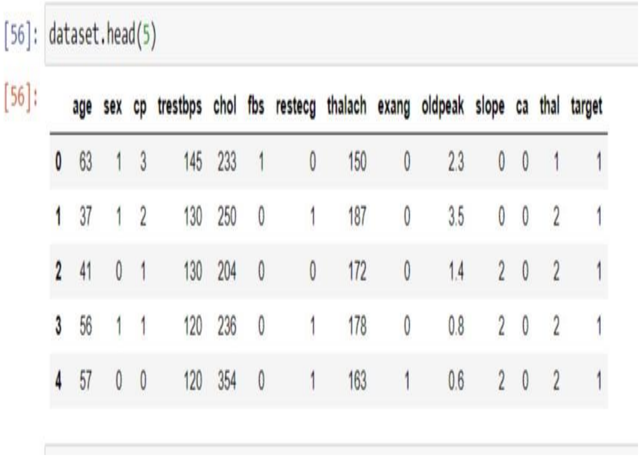
Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. It assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. That stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. This can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

1.3 ARCHITECTURE



1.3 DATASET DESCRIPTION



```
[56]: dataset.head(5)
```

[56]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

This is our dataset which is consisting of 918 rows and 12 columns.

The attributes used are:

- AGE: Age of the person.
- SEX: Whether the person is male or female. If female give 0 and male give 1.
- CP: chest pain type
- TRESTBPS: The person's resting blood pressure
- CHOL: (Cholesterol) The user is requested to mention their cholesterol level.
- FBS: (Fasting Blood Sugar) Elevated levels are associated with diabetes and insulin resistance.

-RESTECG: (electrocardiogram) An ECG is a simple test that can be used to check your heart's rhythm and electrical activity.

-THALACH: It is the person's maximum heartrate received.

- EXANG: Exercise induced angina (1 = yes; 0 = no)

-OLDPEAK: Asymptomatic chest pain

-SLOPE: The ST segment shift relative to exercise-induced increments in heart rate.

-CA: The calcium present in your body.

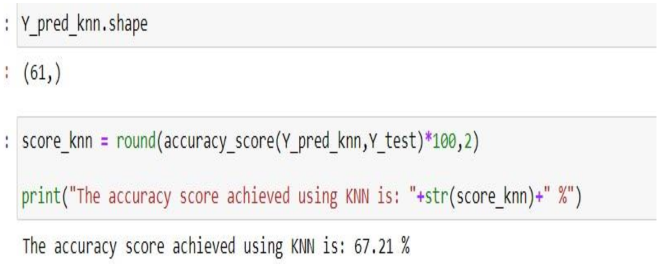
-THAL: A blood disorder called thalassemia (3 = normal; 6 = fixed defect; 7 = reversable defect)

-TARGET: Heart disease (0 = no, 1 = yes)

1.5 MODEL EVALUATION METRICS

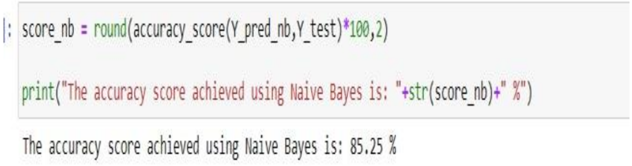
Evaluating each model on the testing set using metrics like accuracy

KNN:

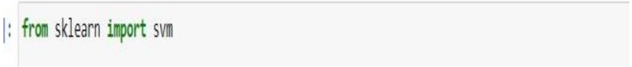


```
: Y_pred_knn.shape
: (61,)
: score_knn = round(accuracy_score(Y_pred_knn,Y_test)*100,2)

  print("The accuracy score achieved using KNN is: "+str(score_knn)+" %")

  The accuracy score achieved using KNN is: 67.21 %
```

NAÏVE BAYES:

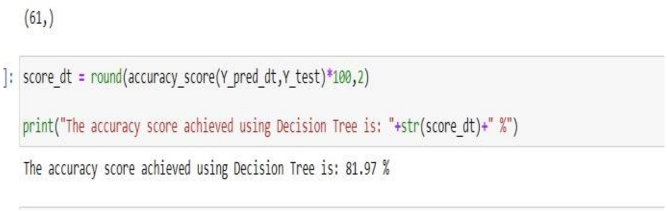


```
]: score_nb = round(accuracy_score(Y_pred_nb,Y_test)*100,2)

   print("The accuracy score achieved using Naive Bayes is: "+str(score_nb)+" %")

   The accuracy score achieved using Naive Bayes is: 85.25 %
```



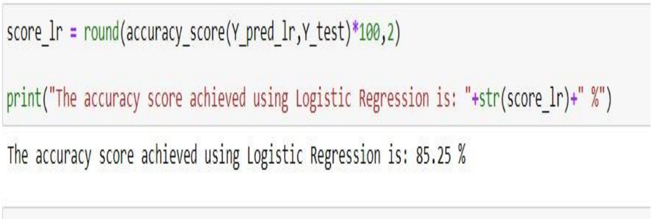
```
]: from sklearn import svm
```

DECISION TREE:



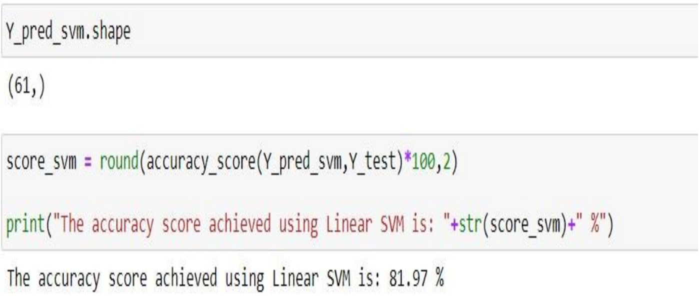
```
(61,)

]: score_dt = round(accuracy_score(Y_pred_dt,Y_test)*100,2)

   print("The accuracy score achieved using Decision Tree is: "+str(score_dt)+" %")

   The accuracy score achieved using Decision Tree is: 81.97 %
```

LOGISTIC REGRESSION:



```
score_lr = round(accuracy_score(Y_pred_lr,Y_test)*100,2)

print("The accuracy score achieved using Logistic Regression is: "+str(score_lr)+" %")

The accuracy score achieved using Logistic Regression is: 85.25 %
```

SUPPORT VECTOR MACHINE:



```
Y_pred_svm.shape
```

```
(61,)
```

```
score_svm = round(accuracy_score(Y_pred_svm,Y_test)*100,2)

print("The accuracy score achieved using Linear SVM is: "+str(score_svm)+" %")
```

```
The accuracy score achieved using Linear SVM is: 81.97 %
```

1.6 POTENTIAL APPLICATIONS

The use of supervised machine learning techniques for heart disease diagnosis, coupled with feature selection methods, can have various applications in the field of healthcare.

Early Diagnosis and Risk Prediction:

Detecting heart disease at an early stage is crucial for effective treatment. Machine learning models, such as k-Nearest Neighbors (KNN), Support Vector Machines (SVM), Naive Bayes, Logistic Regression, and Decision Trees, can be trained on patient data to identify patterns associated with heart disease risk.

Personalized Medicine:

Tailoring treatment plans based on individual patient characteristics is a growing area in healthcare. Machine learning models can analyze patient data to predict how individuals might respond to different treatments and interventions.

Remote Patient Monitoring:

Implementing machine learning models for heart disease diagnosis allows for continuous remote monitoring of patients. Wearable devices and sensors can collect real-time data, which can be analyzed by the models to provide timely alerts or recommendations.

Decision Support Systems:

Integrating machine learning into decision support systems for healthcare professionals can enhance their ability to make accurate and timely decisions. These systems can provide additional insights based on the analysis of patient data.

1.7 CONTIBUTIONS

The study you described contributes to the field of Medical Informatics or Health Informatics. Specifically, it falls within the domain of Clinical Decision Support Systems (CDSS), where machine learning techniques are applied to aid healthcare professionals in making accurate and timely decisions related to patient care. the study contributes to the intersection of healthcare, data science, and machine learning, with a specific emphasis on improving the accuracy and efficiency of heart disease diagnosis through the application of various supervised machine learning algorithms and feature selection techniques**.**

Interpretability and Explainability:

If the study addresses the interpretability and explainability of the chosen machine learning models, it contributes to the broader discussion on making AI-driven diagnostic tools more understandable and trustworthy for healthcare practitioners.

Validation and Generalization:

Demonstrating the robustness of the developed models through thorough validation on different datasets or patient populations contributes to the generalizability of the findings, making them applicable across diverse healthcare settings.

Potential for Real-World Implementation:

If the study considers practical challenges, such as data privacy, integration with existing healthcare systems, and scalability, it contributes to the feasibility of implementing the developed models in real-world clinical settings.

Contribution to Research community:

Sharing the results, methodologies, and datasets (while respecting privacy and ethical considerations) contributes to the broader research community, enabling further advancements in the field of medical informatics.

2. LITERATURE REVIEW

A literature review for a project on heart disease diagnosis using supervised machine learning and feature selection techniques involves examining existing studies, research articles, and publications related to the use of K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Logistic Regression, Decision Trees, and Naive Bayes in the context of heart disease prediction.

1. Introduction to Heart Disease Diagnosis:

- Provide a brief overview of the significance of heart disease as a leading cause of mortality worldwide.

- Highlight the importance of early and accurate diagnosis for effective treatment and prevention.

2. Machine Learning in Healthcare:

- Explore the general application of machine learning in healthcare, emphasizing its potential in disease prediction and diagnosis.

- Discuss the advantages of using machine learning techniques over traditional methods.

3. Supervised Machine Learning Algorithms:

- K-Nearest Neighbors (KNN):

- Summarize studies that have applied KNN to heart disease diagnosis.

- Discuss the strengths and limitations of KNN in this context.

- Support Vector Machines (SVM):

- Review research that utilizes SVM for heart disease prediction.

- Evaluate the performance of SVM in comparison to other algorithms.

- Logistic Regression:

- Examine studies employing logistic regression for heart disease diagnosis.

- Discuss the interpretability and simplicity of logistic regression models.

- Decision Trees:

- Survey literature on the use of decision trees in heart disease prediction.

- Explore how decision trees contribute to feature selection.

- Naive Bayes:

- Summarize research applying Naive Bayes to heart disease diagnosis.

- Discuss the assumptions and effectiveness of Naive Bayes in this domain.

4. Feature Selection Techniques:

- Explore studies that focus on feature selection in the context of heart disease diagnosis.

- Discuss how feature selection improves model interpretability and reduces overfitting.

5. Performance Evaluation Metrics:

- Describe common metrics used to evaluate the performance of machine learning models in healthcare, such as accuracy, sensitivity, specificity, ROC-AUC, and F1 score.

6. Challenges and Future Directions:

- Identify challenges faced in applying machine learning to heart disease diagnosis.

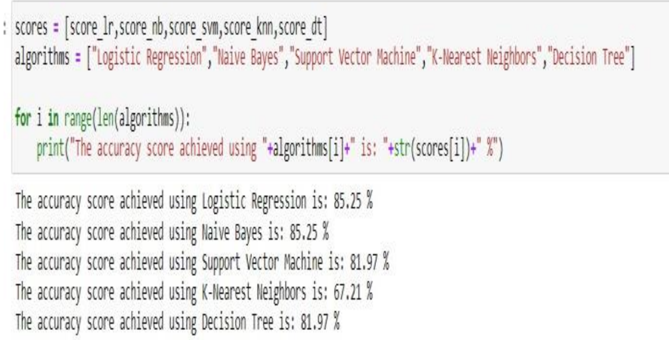
- Suggest potential areas for future research and improvement in model performance.

7. Comparative Analysis:

- Conduct a comparative analysis of studies that directly compare the performance of different algorithms for heart disease diagnosis.

- Highlight the strengths and weaknesses of each algorithm in different scenarios

3. EXPERIMENTAL RESULTS



```
scores = [score_lr,score_nb,score_svm,score_knn,score_dt]
algorithms = ["Logistic Regression","Naive Bayes","Support Vector Machine","K-Nearest Neighbors","Decision Tree"]

for i in range(len(algorithms)):
    print("The accuracy score achieved using "+algorithms[i]+" is: "+str(scores[i])+" %")
```

```
The accuracy score achieved using Logistic Regression is: 85.25 %
The accuracy score achieved using Naive Bayes is: 85.25 %
The accuracy score achieved using Support Vector Machine is: 81.97 %
The accuracy score achieved using K-Nearest Neighbors is: 67.21 %
The accuracy score achieved using Decision Tree is: 81.97 %
```

The proposed Logistic regression and Naïve Bayes algorithms has been shown to be very effective in terms of heart disease prediction with maximum accuracy of 85.25%.

4. CONCLUSION

The identification of raw healthcare data of heart information processing would aid in the long-term saving of human lives and the early detection of defects in heart conditions. Based on the project's findings, it can be inferred that machine learning algorithms have a lot of potential in the medical field. In the medical world, predicting heart disease is both difficult and crucial. . However, the death rate can be drastically controlled if the disease is detected at the early stages and preventative measures can be adopted as soon as possible. It would be ideal if this research could be expanded to include real-world datasets rather than only theoretical methods and simulations. The proposed Logistic regression and Naïve Bayes algorithms has been shown to be very effective in terms of heart disease prediction.

5. FUTURE WORK

Hyperparameter Tuning:

Perform an extensive hyperparameter tuning process to optimize the performance of the chosen models. This can significantly impact the model's accuracy and generalization to new data.

Dynamic Model Updating:

Develop methods for dynamic model updating to adapt the model over time as new data becomes available. This is important in the healthcare domain where the distribution of data may change over time.

Integration of Multiple Data Modalities:

If available, consider integrating multiple data modalities such as genetic data, imaging data, or patient demographics to improve the overall predictive performance of the model.

External Validation:

Validate the model's performance on external datasets from different sources or populations to assess its generalizability and robustness.

Collaboration with Healthcare Professionals:

Collaborate closely with healthcare professionals to gather domain-specific insights, validate the model's predictions, and ensure that the developed system aligns with clinical needs.

Ethical Considerations:

Address ethical considerations, including privacy concerns, data security, and potential biases in the dataset or model predictions. Ensure that the deployment of the model complies with relevant healthcare regulations.

6. REFERENCES

[1] T.M. Mitchell, "Machine Learning", McGraw-Hill, 1997.

[2] Machine Learning, Saikat Dutt, Subramanian Chandramouli, Amit Kumar Das, Pearson, 2019.

[3] Ethern Alpaydin,"Introduction to Machine Learning", MIT Press, 2004

[4] Stephen Marsland, "Machine Learning -An Algorithmic Perspective", Second Edition, Chapman and Hall/CRC Machine Learning and Pattern Recognition Series, 2014.

[5] Andreas C. Müller and Sarah Guido "Introduction to Machine Learning with Python: A Guide for Data Scientists", Oreilly.