

Prediction of Heart Disease

Team Name: - SNV

Team Members: -

Swetha Malladi;

University of New Haven; Mail Id: small13@unh.newhaven.edu

Nitisha Potluri;

University of New Haven; Mail Id: npotl1@unh.newhaven.edu

Vinuthna Vulligaddala;

University of New Haven; Mail Id: vvull1@unh.newhaven.edu

ABSTRACT: -

The diagnosis of heart disease in most cases depends on a complex combination of clinical and pathological data. Because of this complication, clinical professionals and researchers are particularly interested in the efficient and precise prognosis of cardiac disease. In this study, we create a heart disease predict system to help medical practitioners anticipate heart disease status based on patient clinical data. There are three steps to our strategy. Age, sex, chest pain kind, trestbps, cholesterol, fasting blood sugar, resting ecg, max heart rate, exercise induced angina, old peak, slope, number of vessels colored, and thal are the first 13 clinical parameters we look at. Second, based on these data, we create an artificial neural network method for classifying heart disease. The Heart Disease Prediction system (HDPS) will be consisted of multiple features, including input clinical data section and prediction performance display section (accuracy, sensitivity, F2 score, specificity, and predict result). Our approaches are effective in predicting the heart disease of a patient. The HDPS system created in this work provides a novel approach to heart disease classification that can be employed in the future.

Keywords- Logistic Regression, Random Forest, Decision Tree, Naïve Bayes, Support Vector Machine, Accuracy, python, F2 score, MSE score, Cleveland dataset, Precision, Recall and Sensitivity.

INTRODUCTION: -

Heart disease is one of the biggest causes of morbidity and mortality among the population of the world. Prediction of cardiovascular disease is regarded as one of the most important subjects in the section of clinical data analysis. The amount of data in the healthcare industry is huge. Data mining turns the large collection of raw healthcare data into information that can help to make informed decisions and predictions.

Heart disease is the leading cause of death for both women and men, according to a news report. The following is taken from the article: Every year, over 610,000 people in the United States die of heart disease, accounting for one out of every four deaths. For both men and women, heart disease is the leading cause of death. In 2009, men accounted for more than half of all heart disease deaths. CHD is the most common type of heart disease, claiming the lives of about 370,000 people each year. Approximately 735,000 Americans experience a heart attack each year. 525,000 of them are first-time heart attacks, while 210,000 occur in patients who have already experienced a heart attack.

As a result, heart disease is a significant issue that must be addressed. However, due to a number of contributing risk factors such as diabetes, high blood pressure, high cholesterol, an abnormal pulse rate, and a variety of other conditions, it can be difficult to diagnose heart disease. Because of these limitations, scientists have resorted to new technologies for disease prediction, such as Data Mining and Machine Learning. Machine learning (ML) has shown to be useful in assisting in the decision-making and prediction of outcomes from enormous amounts of data generated by the healthcare industry. In this article, I'll utilize one of the most widely used datasets, the Cleveland Heart Disease dataset from the UCI Repository, to apply Machine Learning algorithms (and then compare them) for determining whether or not a person has heart disease.

Research Question: In today's society, academics are working their hearts out to improve the smart health care system. Our goal is to predict a person's risk of heart disease.

Dataset: - In Cleveland dataset, database contains 76 attributes, but mostly all the practical experiments which are published used a subset of 14 of them.

Attributes: -

1. Sex
2. Chest pain type
3. Age
4. Resting Blood Pressure
5. Serum Cholesterol
6. Fasting Blood Sugar
7. Resting ECG
8. Max heart rate achieved
9. Exercise induced agina
10. Oldpeak
11. Slope
12. Ca
13. Thal
14. Num

Related Work:**Review 1:**

Title of paper: A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method.

Author Name: Xiao Liu, Xiaoli Wang, Qiang Su, Mo Zhang, Yanhong Zhu, Qiugen Wang, and Qian Wang.

Affiliation: Xiao Liu, Xiaoli Wang, Qiang Su - School of Economics and Management, Tongji University, Shanghai, China

Mo Zhang - 2 School of Economics and Management, Shanghai Maritime University, Shanghai, China

Yanhong Zhu - Department of Scientific Research, Shanghai General Hospital, School of Medicine, Shanghai Jiao tong University, Shanghai, China

Qiugen Wang and Qian Wang - Trauma Center, Shanghai General Hospital, School of Medicine, Shanghai Jiao tong University, Shanghai, China

Publication Date: 3 January 2017

Publisher Name: Hindawi

Review 2:-

Title of paper: Heart Disease prediction using data mining techniques.

Author Name: S Anitha, N Sridevi

Affiliation: S Anitha - Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India

N Sridevi - College of Administrative and Financial Sciences, AMA International University, Salmabad, Kingdom of Bahrain

Publication Date: February 2019

Publisher Name: Hal Open Science

Review 3:-

Title of paper: Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques

Author Name: C. Beulah Christalin Latha, S. Carolin Jeeva

Affiliation: C. Beulah Christalin Latha, S. Carolin Jeeva - Karunya Institute of Technology and Sciences, India.

Publication Date: 02 July 2019

Publisher Name: Elsevier Ltd

Review 4:

Title of paper: Performance analysis of some selected machine learning algorithms on heart disease prediction using the noble UCI dataset

Author Name: Lamido Yahaya, Nathaniel David Oye, Abubakar Adamu Affiliation:

Lamido Yahaya: Department of Computer Science Gombe State University, Gombe, Gombe State, Nigeria

Nathaniel David Oye: Department of Computer Science, Modibbo Adama University of Technology, Yola, Adamawa State, Nigeria

Abubakar Adamu: Department of Mathematics, Gombe State University, Gombe, Gombe State, Nigeria

Publication Date: May 2020

Publisher Name: IJEAST

Review 5:

Title of paper: Early detection of coronary heart disease using ensemble techniques Author

Name: Vardhan Shorewala

Affiliation: Dhirubhai Ambani International School, Mumbai, India Publication Date: 11 July 2021

Publisher Name: Elsevier Ltd.

Review 6:

Title of paper: Heart disease prediction by using novel optimization algorithm: A supervised learning prospective

Author Name: Siboprasad Patro, Gouri Sankar Nayak, Neelamadhab Padhy Affiliation:

Siboprasad Patro : School of Engineering and Technology, Department of Computer Science and Engineering, GIET University, Gunupur, Odisha, India.

Gouri Sankar Nayak: School of Engineering and Technology, Department of Computer Science and Engineering, GIET University, Gunupur, Odisha, India.

Neelamadhab Padhy: School of Engineering and Technology, Department of Computer Science and Engineering, GIET University, Gunupur, Odisha, India.

Publication Date: 11 August 2021

Publisher Name: Elsevier Ltd.

Proposed Methods

Data Exploration Techniques:

- 1) Naïve Bayes: Bayes theorem allows us to compute a piece of data belonging to a given class. Bayes theorem is expressed as follows

$$P(\text{class}|\text{data}) = (P(\text{data}|\text{class}) * P(\text{class})) / P(\text{data})$$

Probability of given class of the data is measured by $P(\text{class}|\text{data})$.

```
def separate_by_class(dataset):
    separated = dict()
    for i in range(len(dataset)):
        vector = dataset[i]
        class_value = vector[-1]
        if (class_value not in separated):
            separated[class_value] = list()
        separated[class_value].append(vector)
    return separated
```

2) Decision Tree:

One of the most frequent methods for creating classifiers is to use a decision tree. It's similar to the flowchart structure, in which each internal node represents a condition on an attribute, each branch reflects the condition's conclusion, and each leaf node represents the class label. After computing all qualities, a decision is made. Classification rules are represented via a path from a root to a leaf.

In the medical industry, decision trees are used to decide the order of qualities. It first generates a set of solved problems. The entire set is then separated into two parts: a training set and a testing set. Where a training set is used for the induction of a decision tree. The testing set is used to determine the accuracy of the system.

3) Support Vector Machine:

SVM: Statistical learning models such as SVMs are becoming more popular. SVMs are supervised learning models that are applied mainly for classification, but they can also be used to solve regression problems. An SVM is a binary classifier that divides training data into two categories.

The SVM algorithm maps features into a higher-dimensional vector space, where in this space, a maximum margin hyperplane is established. On each side, the distance between the hyperplane and the closest data point is maximized. The method of maximizing the margin, and thus producing the largest possible distance between the separating hyperplane and the instances on either side of it, has proven to significantly reduce the expected generalization error.

4) Random Forest:

It is a method for classifying that is built by constructing a multiple decision tree at training time, and it produces a class by voting individual branches. The algorithm constructs a forest of decision trees based on attribute locations chosen at random. It has the advantage of improving prediction accuracy without significantly increasing computational costs.

5) Logistic Regression:

The Logistic regression is typically used to classify low dimensional data with nonlinear boundaries. also, it provides the difference in the percentages of the dependent variables and the rank of each variable. The main purpose of Logistic Regression is to determine the correct result of each variable Logistic regression is also known as Logistic Model,

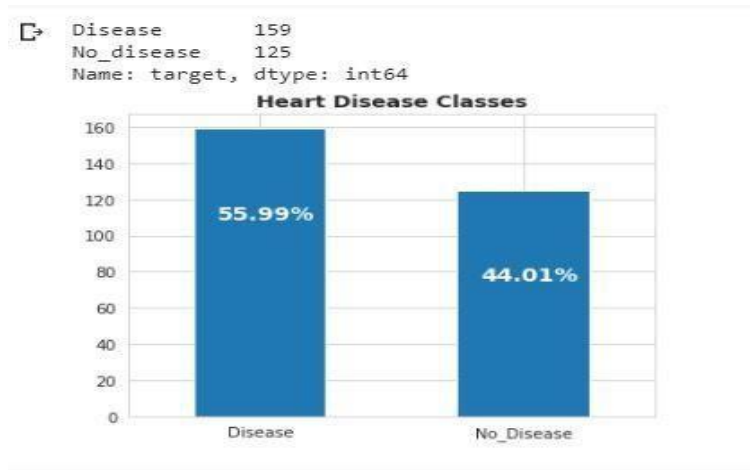
which is a categorical variable with two categories, for instance light or dark, slim/healthy.

EXPERIMENTAL RESULTS: -

Visualization Techniques:

1) Target Variable Distribution

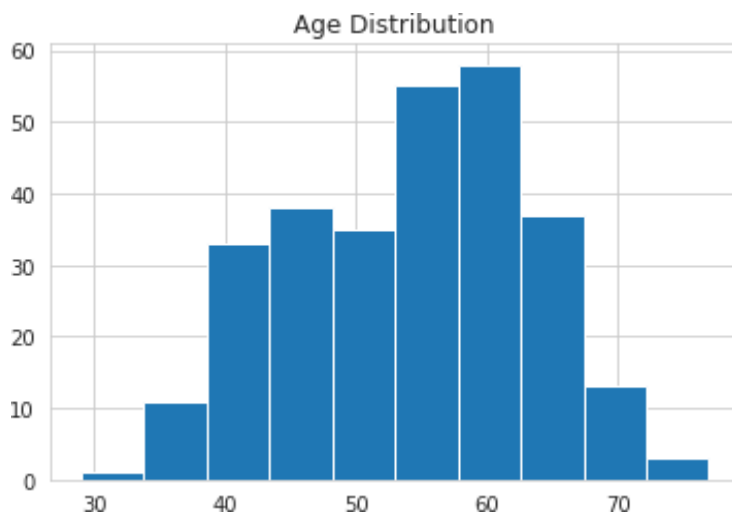
In this scenario, we're talking about group of people who have heart disease and



others who don't have heart disease. Most persons have cardiac disease (about 56 percent) and no disease (approximately 44 percent).

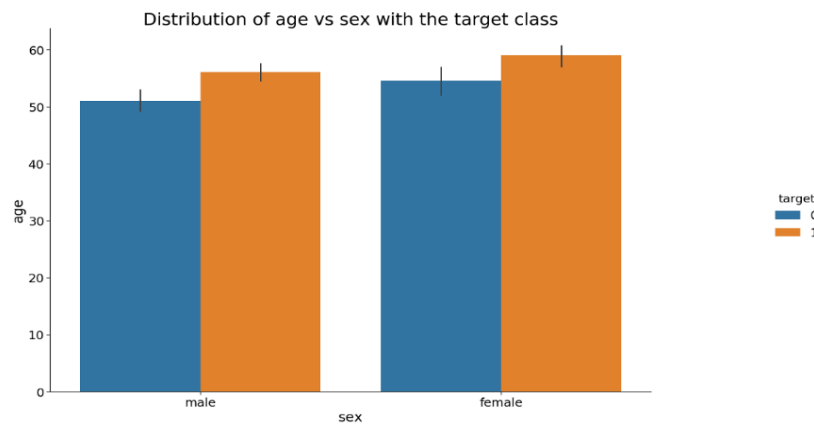
2) Age Variable Distribution

```
# print(df.age.value_counts())  
df['age'].hist().plot(kind='bar')  
plt.title('Age Distribution')
```



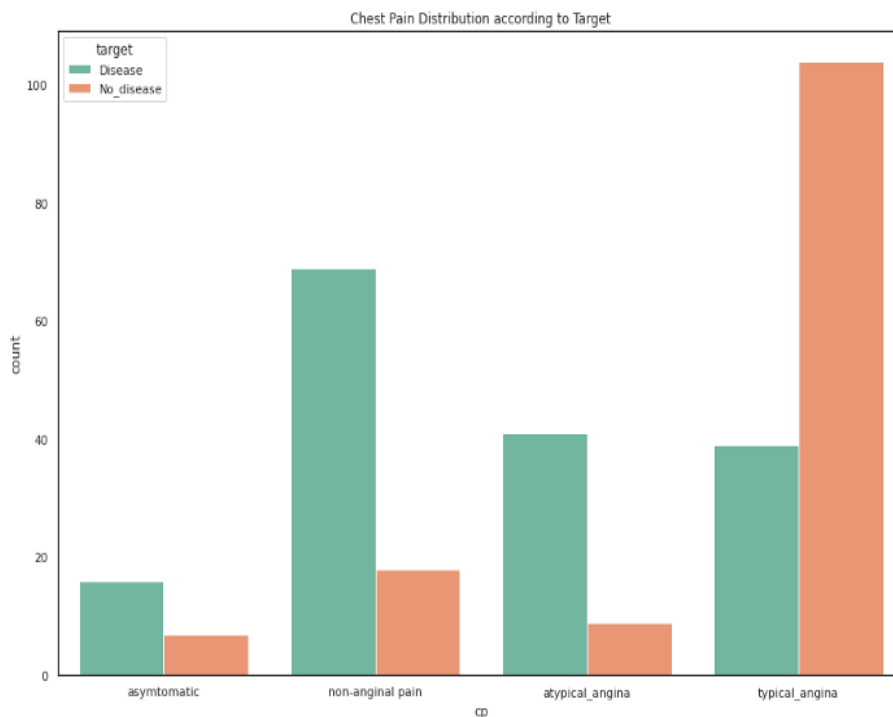
Using Cleveland dataset Heart disease affects the majority of people between the ages of 40 and 60. The youngest person diagnosed with heart disease is 29 years old, while the oldest is 77 years old. The average age where the people get heart disease is 54 years.

3) Gender distribution according to target variable



The graph describes the heart disease by comparing age and sex factor. We can see that females of upper age are more likely to suffer from heart disease than males.

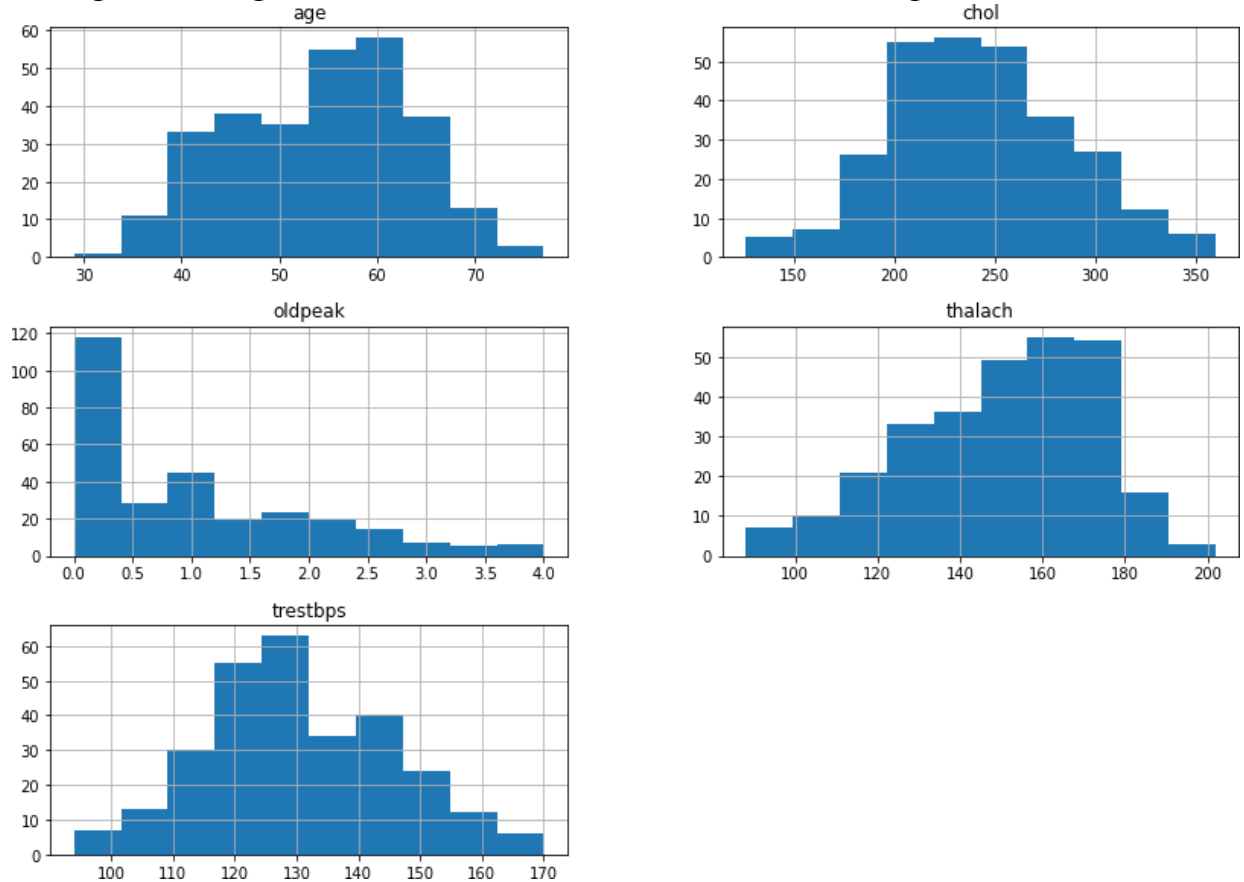
4) Chest Pain distribution according to target variable



The majority of individuals with heart disease experience non-anginal chest discomfort and people with typical angina have less chances of occurring the disease.

5) Fasting blood sugar distribution according to target variable

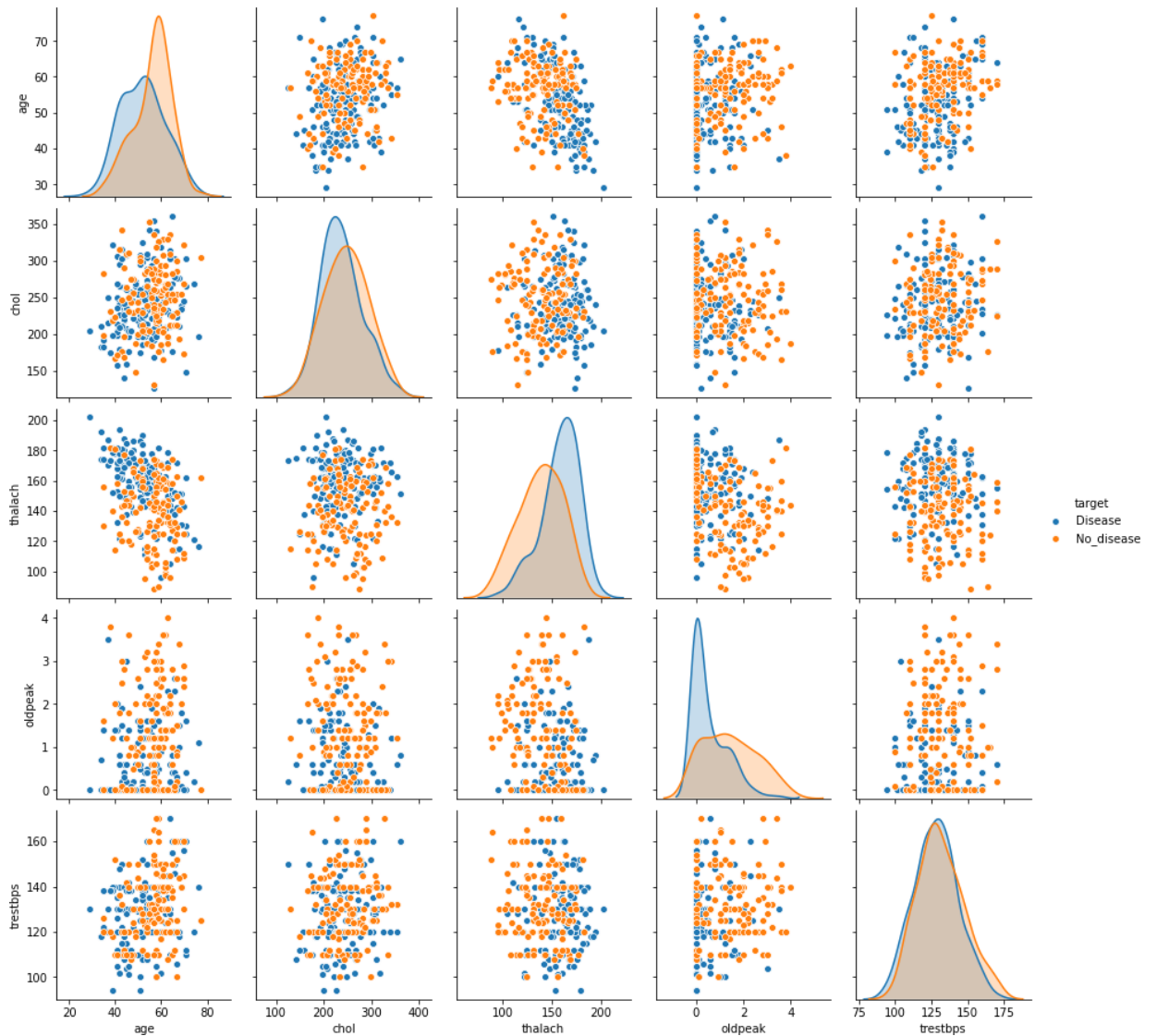
Fasting blood sugar, or fbs, is a diabetes indicator; if fbs is greater than 120



mg/d, you're diabetic (True class). We can see that the number for class true is lower than the number for class false. However, if we examine closely, we can see that there are more patients with heart disease who do not have diabetes. Distribution plot on continuous variables

Considering the continuous variables like age, cholesterol, oldpeak, thalach, trestbps the distribution plot for age, trestbps and nearly for cholesterol have normal distribution. Oldpeak is tilted to the left whereas thalach is tilted to the right.

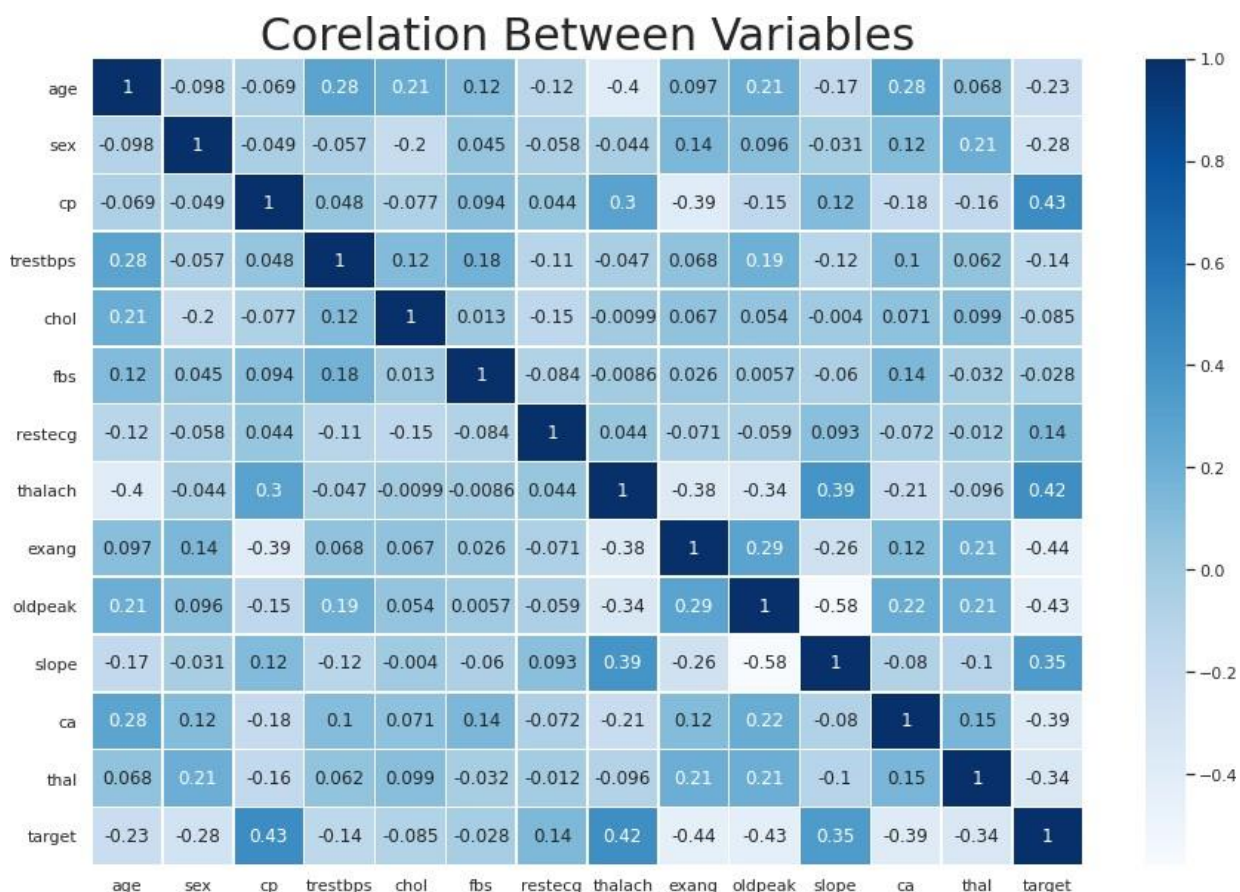
6) Sns pairplot to visualize the distribution



By drawing the seaborn pairplot for variables age, cholesterol, thalach, oldpeak, trestbps to visualize the distribution there is a linear difference between disease and non-disease in oldpeak and for thalach there is a slight distinction between disease and non-disease. Other characteristics don't make a clear distinction.

7) Correlation between 14 variables of Cleveland dataset

```
sns.set(style="white")
plt.rcParams['figure.figsize'] =
(15, 10)
sns.heatmap(df.corr(), annot = True, linewidths=.5, cmap="Blues")
plt.title('Correlation Between Variables', fontsize = 30)
plt.show()
```



By correlating the variables of Cleveland dataset the variables 'cp', 'thalach', and 'slope' have a strong positive connection with the target whereas 'oldpeak', 'exang', 'ca', 'thal', 'sex', 'age' have a negative correlation with the target. 'fbs', 'chol', 'trestbps', 'restecg' has less correlation with the target.

Outcomes of data mining techniques

Naïve Bayes: -

Confusion Matrix for Naive Bayes			
		0	1
	0	129	0
	1	0	113
Training Set			
		0	1
	0	29	8
	1	6	18
Test Set			

Training Set:

Accuracy: $(129 + 113) / (129 + 113) * 100 = 100$

Precision: $(129) / (129 + 0) * 100 = 100$

Recall: $(129) / (129 + 0) * 100 = 100$

F1 score: $2 * (100 * 100) / (100 + 100) = 100$

Test Set: -

Accuracy: $(29 + 18) / (6 + 8 + 29 + 18) * 100 = 77.04$

Precision: $(29) / (29 + 8) * 100 = 78.37$

Recall: $(29) / (29 + 6) * 100 = 82.85$

F1 score: $2 * (78.37 * 82.85) / (78.37 + 82.85) = 80.54$

Logistic Regression: -

Confusion Matrix for Logistic Regression			
		0	1
	0	118	22
	1	11	91
Training Set			
		0	1
	0	32	9
	1	3	17
Test Set			

Training Set: -

Accuracy: $(118 + 91) / (11 + 22 + 118 + 91) * 100 = 86.36$

Precision: $118 / (118 + 22) * 100 = 84.28$

Recall: $-118 / (118 + 11) * 100 = 91.47$

F1 score: $- 2 * (91.47 * 84.28) / (84.28 + 91.47) = 87.87$

Test Set: -

Accuracy: $(32 + 17) / (3 + 9 + 32 + 17) * 100 = 80.32$

Precision: $(32) / (32 + 9) * 100 = 65.3$

Recall: $(32) / (32 + 3) * 100 = 91.42$

F1 score: $2 * (65.3 * 91.42) / (65.3 + 91.42) * 100 = 76.32$

Decision Tree: -

Confusion Matrix for Decision Tree

	0	1		0	1
0	117	20	0	30	8
1	12	93	1	5	18
	Training Set			Test Set	

Training Set: -

Accuracy: $(117 + 93) / (12 + 20 + 117 + 93) * 100 = 86.77$

Precision: $(117) / (117 + 20) * 100 = 85.40$

Recall: $(117) / (117 + 12) * 100 = 90.69$

F1 score: $2 * (90.69 * 85.40) / (90.69 + 85.40) = 87.96$

Test Set: -

Accuracy: $(30 + 18) / (5 + 8 + 30 + 18) * 100 = 78.68$

Precision: $(30) / (30+8) * 100 = 78.94$

Recall: $(30) / (30+5) * 100 = 85.71$

F1 score: $2 * (78.94 * 85.71) / (78.94+85.71) = 82.18$

Support Vector Machine: -

Confusion Matrix for SVM

	0	1
0	124	13
1	5	100

Training Set

	0	1
0	32	9
1	3	17

Test Set

Training Set: -

Accuracy: $(124+100) / (5+13+124+100) * 100 = 92.51$

Precision: $(124) / (124+13) * 100 = 90.51$

Recall: $(124) / (124+5) * 100 = 96.12$

F1 score: $2 * (90.51 * 96.12) / (90.51+96.12) = 93.23$

Test Set: -

Accuracy: $(32+17) / (9+3+32+17) * 100 = 80.32$

Precision: $(32) / (32+9) * 100 = 78.04$

Recall: $(32) / (32+3) * 100 = 91.42$

F1 score: $2 * (78.04 * 91.42) / (78.04+91.42) = 84.20$

Random Forest: -

Confusion Matrix for
Random Forest

	0	1
0	129	2
1	0	111

Training Set

	0	1
0	32	10
1	3	16

Test Set

Training Set: -

Accuracy: $(129 + 111) / (0+2+129+111) * 100 = 98.76$

Precision: $(129) / (129+2) * 100 = 98.47$

Recall: $(129) / (129+0) * 100 = 100$

F1 score: $2 * (98.47 * 100) / (98.47+100) = 99.22$

Test Set: -

Accuracy: $(32 + 16) / (3+10+32+16) * 100 = 75.40$

Precision: $(32) / (32+10) * 100 = 76.19$

Recall: $(32) / (32+3) * 100 = 91.42$

F1 score: $2 * (76.19*91.42) / (76.19 + 91.42) = 83.11$

DISCUSSION: -

Early detection of heart disorders improves survival rates, hence the goal of this study is to predict whether a patient has heart disease or not using clinical data that will aid in the diagnosis process. Five supervised machine learning algorithms namely Logistic Regression, Decision tree, Support Vector Machine, Naïve Bayes, Random Forest are compared utilizing the Cleveland dataset on heart disorders. We used the python programming language and performed data analysis using Weka tool. We have segregated the dataset into training set and test set and calculated the Accuracy, Precision, Recall and F1 score. The experimental findings show that the Naive Bayes algorithm is effective with an accuracy of 100 percent, it predicts heart disease.

Four optimization techniques were used to compare and see who predicted heart disease problems with more accuracy and less error using four different classification algorithms. Each confusion matrix displays the precision, accuracy, and sensitivity of the methods used. An SSA-optimized Neural Network has a higher accuracy than a Neural Network alone. Similarly, Bayesian optimization is used to optimize the support vector machine. For comparative study, the KNN and Naive Bayes classifications are applied. For classification, the KNN and Naive Bayes methods are utilized separately, while the Salp Swarm Algorithm optimizes the bias and weight values for Neural Networks. Bayesian Optimization is also used to optimize the weight and kernel function of SVM. The optimization approaches are particularly useful in heart disease prediction. As we can see from confusion matrix charts, the Bayesian Optimized SVM-based strategy outperforms other methods with a maximum accuracy of 93.3 percent.

CONCLUSION: -

In India, heart disease is the most common ailment. Early detection of heart disorders improves survival rates; Hence the goal of this study is to predict whether a patient has heart disease or not using clinical data that will aid in the diagnosis process. Five supervised machine learning algorithms namely Logistic Regression, Decision tree, Support Vector Machine, Naïve Bayes, Random Forest are compared utilizing the Cleveland dataset on heart disorders. We used the python programming language and performed data analysis using Weka tool. We have segregated the dataset into training set and test set and calculated the Accuracy, Precision, Recall and F1 score. The experimental findings show that the Naive Bayes algorithm is effective with an accuracy of 100 percent, it predicts heart disease.

FUTURE WORK: -

The use of RF trees, SVM Nave Bayesian, neural networks, and logistic regression analysis-based classifiers has been proposed as part of an ensemble approach-based prototype smart heart disease prediction model. The suggested system, which has been developed on the WEKA platform, is GUI-based, user-friendly, scalable, stable, and extendable. By delivering initial diagnostics on time, the proposed working model can also help to reduce treatment expenses. The model can also be used as a teaching tool for medical students and as a soft diagnostic tool for doctors and cardiologists. This tool can be used by general practitioners to diagnose cardiac patients in the early stages. To improve the scalability and accuracy of this prediction system, a number of changes could be investigated. Handling several class labels in the prediction process can considerably improve the performance of health diagnosis, and this could be another promising study direction. Because the dimensionality of the heart database in DM warehouses is often high, identifying and selecting key variables for better detection of heart disease will be difficult challenges for future research.

APPENDIX FOR LINK TO THE GITHUB REPOSITORY: -



https://github.com/SwethaMalladi/Final_Report_Academic_Paper


REFERENCES: -

1. This work of predicting heart disease is evaluated using the dataset from the UCI machine learning repository and Weka tool.
2. H. B. F. David and S. A. Belcy, "Heart Disease Prediction Using Data Mining Techniques", ICTACT Journal On Soft Computing, vol. 09, no. 01, 2018.
3. X. Liu, X. Wang, Q. Su, M. Zhang, Y. Zhu, Q. Wang, et al., "A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method", Computational and Mathematical Methods in Medicine, vol. 2017.
4. C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques", Informatics in Medicine Unlocked, vol. 16, 2019.
5. S. Palaniappan and R. Awang, "Intelligent Heart Disease Prediction System using Data Mining Techniques", International Journal of Computer Science and Network Security, vol. 8, no. 8, pp. 1-6, 2008.
6. J. R. Quinlan, "Induction of Decision Trees", Machine Learning, vol. 1, no. 1, pp. 81-106, 1986.
7. H. Kahramanli and N. Allahverdi, "Design of a hybrid system for the diabetes and heart diseases", Expert Systems with Applications, vol. 35, no. 1-2, pp. 82-89, 2000.

PROODREADING WITH AN EMAIL FROM WRITING CENTER: -

Review the IEEE report








 1 





Malladi, Swetha

To: Writing Center

Cc: Khare, Shivanjali; Potluri, Nitisha; Vulligaddala, Vinuthna



Wed 5/4/2022 8:40 PM

 Prediction of Heart Disease_f...
551 KB 

Hi Team,

Please find the attachment for final report of our project of data mining(CSCI-6674).
Please do suggest any changes on the report for the final submission.

Thanks & Regards,
Swetha Malladi.

[Reply](#) | [Reply all](#) | [Forward](#)