Final Paper

Health care: Heart attack possibility

Data Squad

Team Members:
Chanpreet Kaur
University of New Haven; Mail Id:ckaur1@unh.newhaven.edu

Hemant Rajpal
University of New Haven; Mail Id:hrajp2@unh.newhaven.edu

VishnuVardhan Rao Sidduri
University of New Haven; Mail Id:vsidd2@unh.newhaven.edu

ABSTRACT: -

The diagnosis of heart disease with most cases refers to a complex pairing of clinical and pathological data. Clinical experts and academics are especially interested in the accurate and efficient prognosis of heart disease because of this problem. In this study, we develop a method to forecast heart disease state to assist medical practitioners using patient clinical data. There really are 3 parts to our strategy. The very first 13 clinical parameters we consider are age, sex, chest pain type, trestbps, fasting blood sugar, cholesterol, resting ecg, maximum heart rate, exercise-induced chest pain, old peak, slope, amount of vessels colored, and thal. Next, based on these data, we construct an artificial neural network approach for identifying heart illness.The Heart Disease Prediction system (HDPS) would be consisted of various attributes, along with input clinical research section and prediction performance showcase section (sensitivity, accuracy, specificity, F2 score, and predict result). Our methodologies are

excellent in forecasting the cardiac illness of a patient. The HDPS system developed in this work offers a new strategy to heart disease categorization that can be utilised in the future.

INTRODUCTION: -

Heart disease is among the largest causes of death and morbidity as among world population. Forecasting of coronary heart disease is considered as among the most important topics in the portion of diagnostic data analysis. The number of data in healthcare industry is vast. Data mining converts the enormous accumulation of raw health records in information that can assist to make informed choices and forecasts.

Heart disease remains the major cause of death both for women and men, as per a news report. One in every four deaths in the United States occurs as a result of heart disease, according to the article, which states the following. Heart disease is the most common cause of death for both men and women. In 2009, men made up the majority of all heart disease deaths. CHD is by far the most common disease, taking the lives to around 370,000 people every year. Every year, about 735,000 Americans have a heart attack. 525,000 of them seem to be first-time heart attacks, whereas 210,000 occur in patients that have already encountered a heart attack.

Heart disease is therefore a serious problem that needs to be treated. However, because of several of making a contribution health risks such as diabetes, blood pressure, high blood cholesterol, an unusual pulse rate, and a diversity of many other situations, it may be difficult to identify heart disease. Due to these restrictions, scientists have needed to resort to technological advances for prediction and diagnosis, such as Machine Learning and Data Mining. Machine learning (ML) is demonstrated to be helpful in aiding in the judgment and forecasting of outcomes using large amounts of data produced by the healthcare sector.

Research Question:

At present, scholars are putting their efforts to upgrade the health care system. Our objective is prediction and estimate the risk of heart attack to a person.

Dataset:

https://www.kaggle.com/datasets/cherngs/heart-disease-cleveland-uci

Attributes:

- **Age**
- **Max heart rate achieved**
- **Sex**
- **Resting Blood Pressure**
- **Serum Cholesterol**
- **Fasting Blood Sugar**
- **Chest pain type**
- **Exercise induced agina**
- **Ca**
- **Oldpeak**
- **Num**
- **Thal**
- **Slope**

Related Reviews

Review 1: 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 2: 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 3: 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 Jiaotong University, Shanghai, China

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

Publication Date: 3 January 2017

Publisher Name: Hindawi

Review 4: Title of paper: Heart disease prediction by using novel optimization algorithm: A supervised learning prospective

Author Name: Sibo Prasad Patro, Gouri Sankar Nayak, Neelamadhab Padhy

Affiliation:

Sibo Prasad 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.

Review 5: 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

Proposed Methods:

Data Exploration Techniques:

1) Decision Tree:

One of the most prevalent approaches for building 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 weighing all factors, a choice 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. After that, the entire set is divided into 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 evaluate the system's precision.

2)Random Forest:

It is a classification technique that creates a multiple decision tree during training and creates a class by voting on individual branches. A forest of decision trees is built by the algorithm using attribute locations that are randomly selected. It has the advantage of improving prediction accuracy without significantly increasing computational costs.

3)Naïve Bayes:

Bayes theorem allows us to compute a piece of data belonging to a given class. The Bayes theorem is stated as follows:

P(class|data) = (P(data|class) * P(class)) / P(data)
Probability of given class of the data is measured by P(class|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
```

4) 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 creates two categories from training data.

A maximum margin hyperplane is established in the higher-dimensional vector space into which the SVM algorithm maps features. 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 hyper-plane and the instances on either side of it, has proven to significantly reduce the expected generalization error.

5) Logistic Regression:

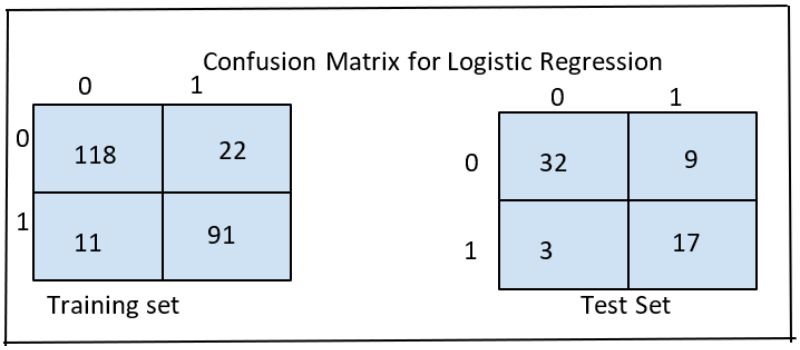
The Logistic regression is often used to categorize low dimensional data with nonlinear bounds. also, it provides the difference in the percentages of the dependent variables and the rank of each variable. The main purpose of Logistic Regression is to determine the correct result of each variable Logistic regression is also known as Logistic Model,
which is a categorical variable with two categories, for instance light or dark, slim/healthy.

EXPERIMENTAL RESULTS: -

Results of data mining approaches employing various performance indicators and perspectives

Techniques for data mining applied to the performance metrics:

Logistic Regression: -



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: -

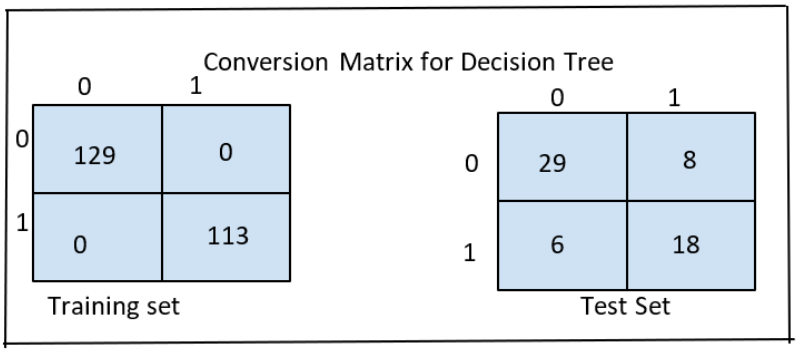
Accuracy: (32 + 17) / (3 + 9 + 32+ 17) * 100 = 80.32

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: -



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: -

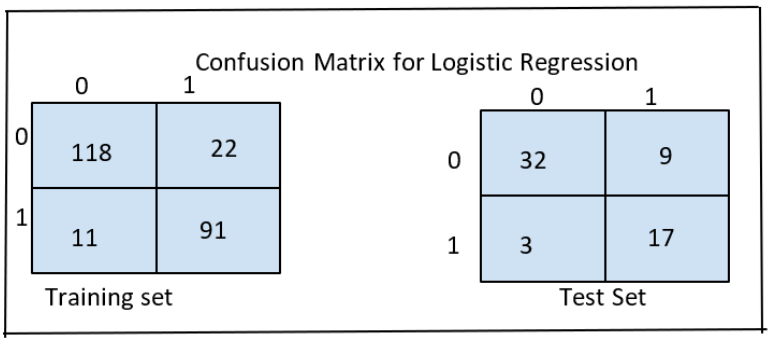
Accuracy: (30+18)/ (5+8+30+18) * 100 = 78.68

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: -



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: -

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: -

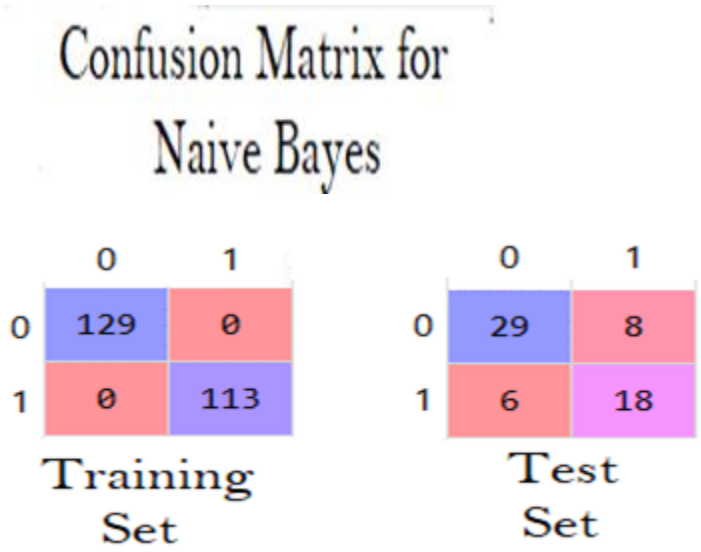
Accuracy: (32 + 16) / (3+10+32+16) * 100 = 75.40

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.

Naïve Bayes: -



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

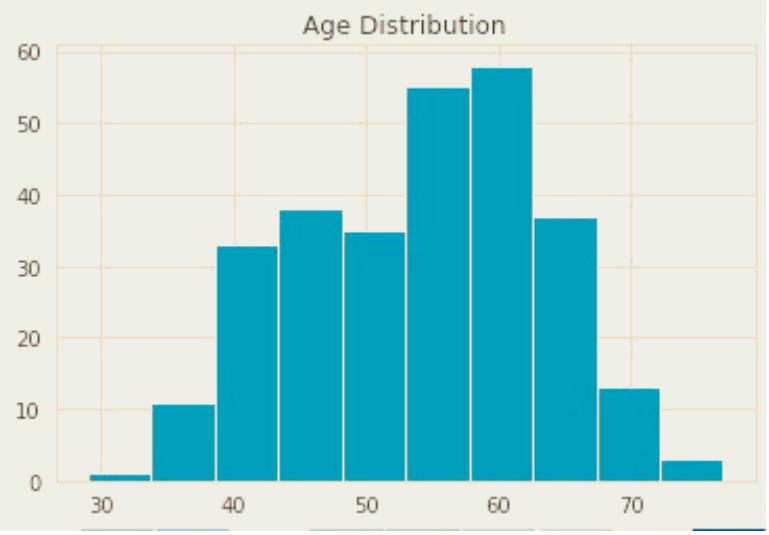
F1 score: 2 * (78.37*82.85) / (78.37+82.85) = 80.54

Visualization Techniques:

1)Age Variable Distribution

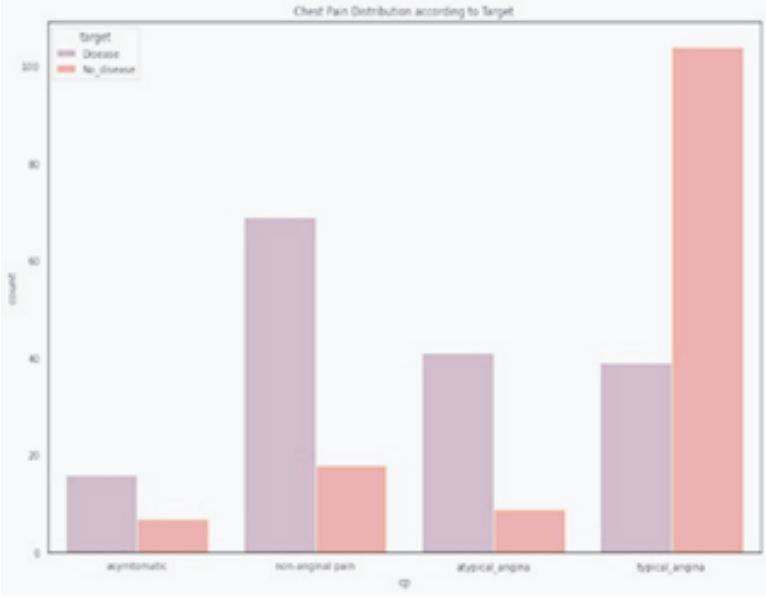
Utilizing the Cleveland dataset Most persons with heart disease are between the ages of 40 and 60. The oldest patient with a heart disease diagnosis is 77 years old, while the youngest patient is 29. Heart disease affects persons on average at the age of 54.

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



2) Distribution of chest pain based on the goal variable

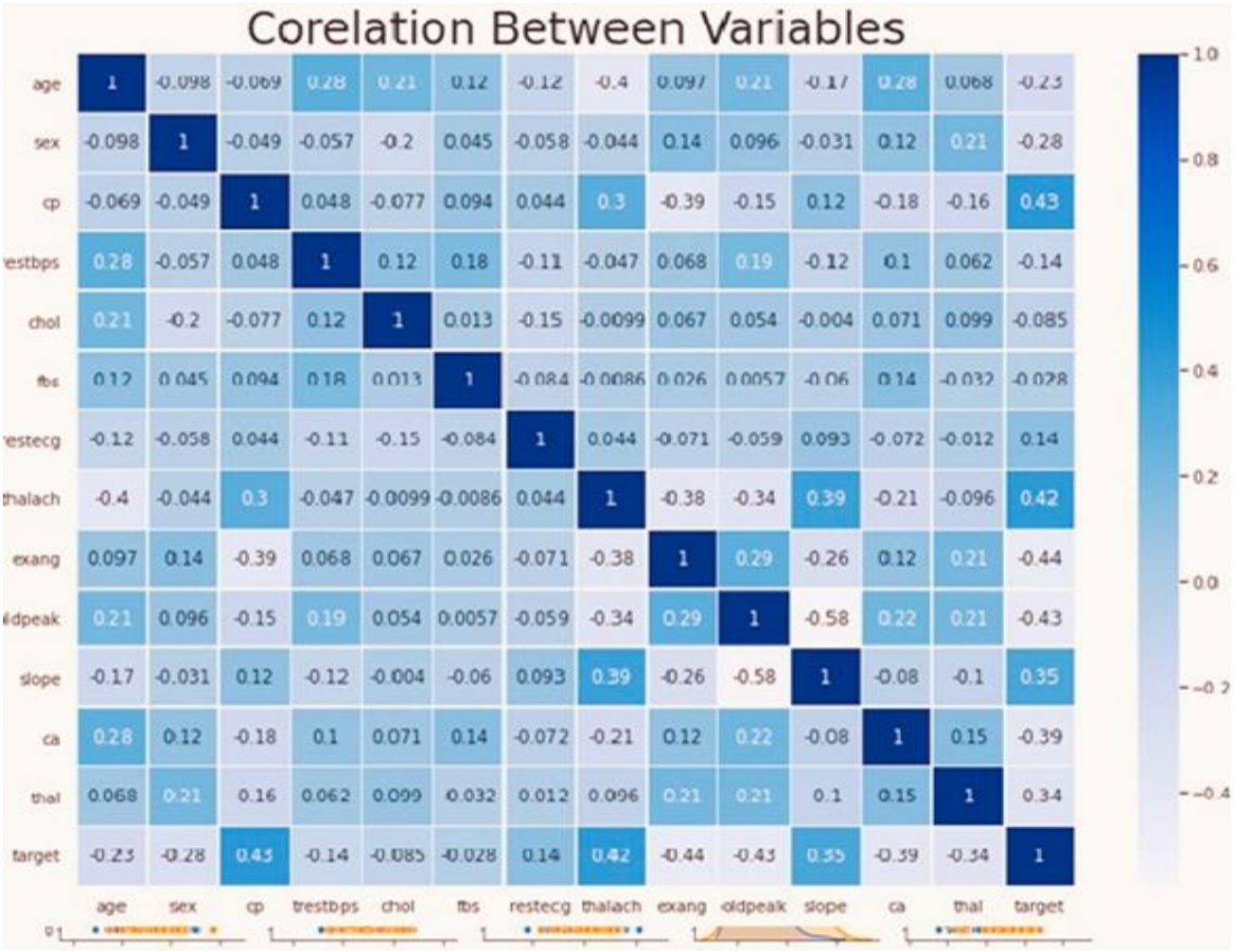
Persons with normal angina have a lower risk of developing heart disease, while most people with heart disease report non-anginal chest pain.



3) Cleveland dataset's 14 variables are correlated with one another.

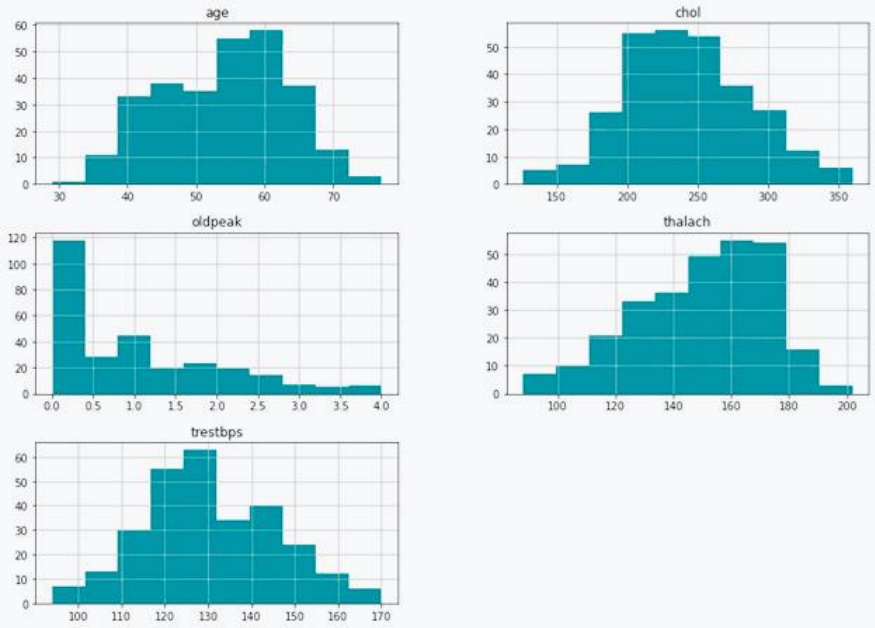
```
sns.set(style="white")

plt.rcParams['figure.figsize'] = (15, 10)

sns.heatmap(df.corr(), annot = True, linewidths=.5, cmap="Blues")

plt.title('Corelation Between Variables', fontsize = 30)

plt.show()
```

The variables "cp," "thalach," and "slope" have a strong positive link with the target when the variables from the Cleveland dataset are correlated, whereas "oldpeak," "exang," "ca," "thal," "sex," and "age" have a negative connection with the target. Less association exists between "fbs," "chol," "trestbps," and "restecg" and the target.



4) Plot of the distribution for continuous variables

The distribution plot for age, trestbps, and almost for cholesterol has a normal distribution when taking into account continuous variables like age, cholesterol, oldpeak, thalach, and trestbps. Thalach is titled to the right, while Oldpeak is titled to the left.

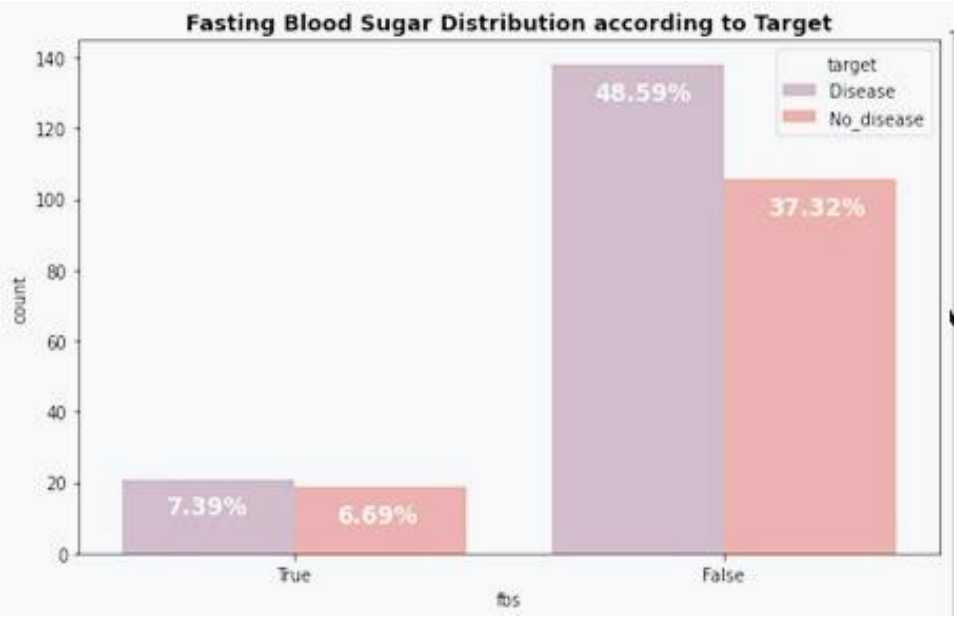


5) Target Variable Distribution

In this case, a group of people with heart illness and an other group of persons without heart disease are being discussed. About 56 percent of people have heart illness and no other condition (approximately 44 percent).

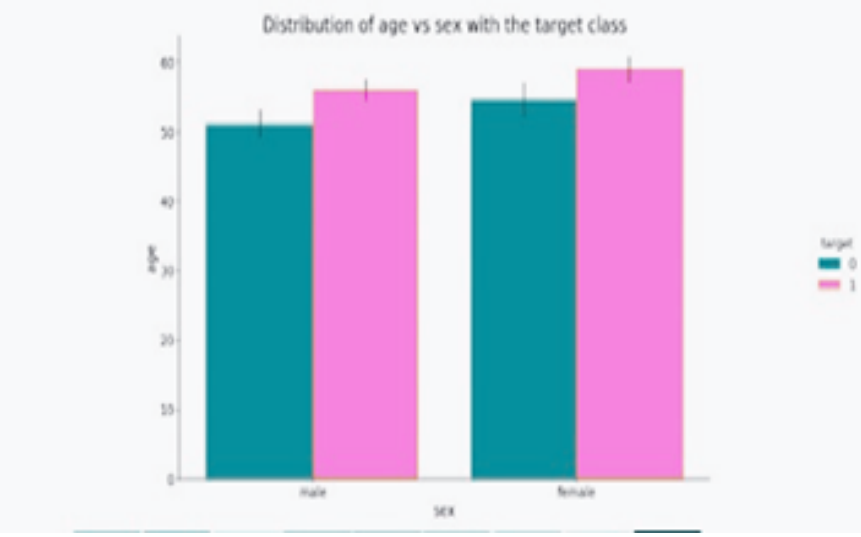
6) Fasting blood sugar distribution according to target variable

Fasting blood sugar, or fbs, is a sign of diabetes; if it is higher than 120 mg/d, you have the condition (True class). We can observe that the class true number is less than the class false number. If we look attentively, there are more patients with heart disease who do not have diabetes, as opposed to the other way around.

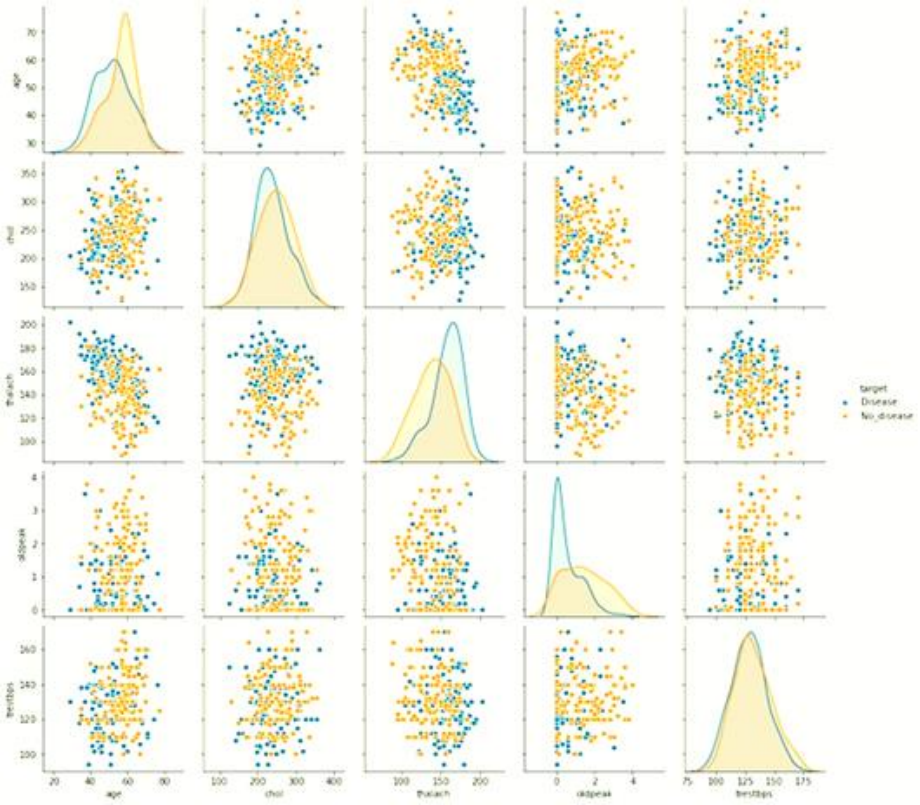


7) Gender distribution according to target variable

The graph illustrates cardiac disease by comparing the effects of age and sex. We can observe that older females have a higher risk of developing heart disease than older males do.



8**) Visualizing the distribution with the SNS pairplot**



DISCUSSION: -

The purpose of this study is to predict whether a patient has heart disease or not using clinical data that will help in the diagnosis process because early detection of heart problems improves survival rates. The Cleveland dataset on cardiac illnesses is used to examine five supervised machine learning algorithms: Logistic Regression, Decision Tree, Support Vector Machine, Nave Bayes, and Random Forest. We utilized the Python programming language and the Weka data analysis tool. We divided the dataset into training and testing sets and computed the Accuracy, Precision, Recall, and F1 scores. The experimental results reveal that the Naive Bayes algorithm predicts cardiac disease with 100 percent accuracy.

Four different categorization algorithms were employed to compare and see who predicted heart disease concerns with greater accuracy and fewer error.
The precision, accuracy, and sensitivity of the methods utilized are displayed in each confusion matrix.
An SSA-optimized Neural Network outperforms a Neural Network alone in terms of accuracy. The support vector machine is also optimized via Bayesian optimization.

The KNN and Naive Bayes classifications are used for comparative analysis.
The KNN and Naive Bayes techniques are used individually for classification, while the Salp Swarm Algorithm optimizes the bias and weight values for Neural Networks.
SVM's weight and kernel function are also optimized via Bayesian Optimization.

The optimization techniques are very effective in predicting cardiac disease.
The Bayesian Optimized SVM-based strategy outperforms other methods, as evidenced by confusion matrix charts, with a maximum accuracy of 93.3 percent.

CONCLUSION: -

Heart disease is the most frequent sickness in India. The purpose of this study is to predict whether or not a patient has heart disease using clinical data that will help in the diagnostic procedure. The Cleveland dataset on cardiac illnesses is used to examine five supervised machine learning algorithms: Logistic Regression, Decision Tree, Support Vector Machine, Nave Bayes, and Random Forest. We utilized the Python programming language and the Weka data analysis tool. We divided the dataset into training and testing sets and computed the Accuracy, Precision, Recall, and F1 scores. The experimental results reveal that the Naive Bayes algorithm predicts cardiac disease with 100 percent accuracy.

FUTURE WORK: -

An ensemble approach-based prototype smart heart attack prediction model including RF trees, Nave Bayesian, neural networks, and logistic regression analysis-based classifiers has been proposed. The proposed system is graphical user interface (GUI)-based, user-friendly, scalable, dependable, and customizable. It was built on the WEKA platform. The recommended working method can also help to reduce treatment expenses by giving timely diagnosis. Doctors and cardiologists can use the model as a soft diagnostic tool as well as a teaching tool for medical students. This tool can help general practitioners establish an early diagnosis of cardiac patients. Modifications to this prediction system might improve its accuracy and scalability. Another intriguing topic for investigation is how using several class labels in the prediction process may considerably improve health diagnosis performance. Because of the vast dimensionality of the heart information in the DM database, future research will confront difficult difficulties in identifying and selecting critical elements for better heart detection.

APPENDIX FOR LINK TO THE GITHUB REPOSITORY: -

https://github.com/HemantRajpal-9018/Academic-Paper-IEEE-

REFERENCES: -

1. This work of predicting heart attack is evaluated using the dataset from the UCI machine learning repository and Weka tool.

2. H. B. F. David and S. A. Belcy, "Heart attack 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 attack 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 attack risk based on ensemble classification techniques", Informatics in Medicine Unlocked, vol. 16, 2019.

5. S. Palaniappan and R. Awang, "Intelligent Heart attack 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 attacks", Expert Systems with Applications, vol. 35, no. 1-2, pp. 82- 89, 2000

PROOFREADING WITH AN EMAIL FROM WRITING CENTER:

Review the IEEE report

1

SR

Sidduri, Vishnuvardhan Rao
To: Writing Center
Cc: Kaur, Chanpreet; Rajpal, Hemant

Thu 08/12/2022 10:01 PM

Academic_Paper.docx
1 MB

Hi Team,

Please find the attachment for final report of our project of data mining(CSCI-6401). Please do suggest any changes on the report for the final submission.

Thanks & regards,
Vishnuvardhan Sidduri

Reply | Reply all | Forward