Heart Disease Prediction Using Machine Learning Algorithms

Malar .K[1], Arjun .N[2], Hemamageswari .V[3], Jeevananthan .S[4]

[1]Assistant Professor, Dept of Computer Science Engineering

[2, 3, 4]Dept of Computer Science Engineering

[1, 2, 3, 4] Adhiyamaan College of Engineering, Krishnagiri District, Tamilnadu, India.

Abstract- *Heart plays significant role in living organisms. Diagnosis and prediction of heart related diseases requires more precision, perfection and correctness because a little mistake can cause fatigue problem or death of the person, there are numerous death cases related to heart and their counting is increasing exponentially day by day. 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 health care industry is huge. To deal with the problem there is essential need of prediction system for awareness about diseases. Machine learning is the branch of Artificial Intelligence (AI), it provides prestigious support in predicting any kind of event which take training from natural events. To calculate accuracy, machine learning algorithms are used for predicting heart disease. The Algorithms are k-nearest neighbor, decision tree, linear regression and support vector machine (SVM) by using UCI repository dataset for training and testing.*

Keywords- Heart Disease Prediction, Data Mining Techniaues, Machine Learning Supervised Algorithms.

I. INTRODUCTION

Heart plays significant role in living organisms. Diagnosis and prediction of heart related diseases requires more precision, perfection and correctness because a little mistake can cause fatigue problem or death of the person, there are numerous death cases related to heart and their counting is increasing exponentially day by day. 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 health care industry is huge. To deal with the problem there is essential need of prediction system for awareness about diseases. Machine learning is the branch of Artificial Intelligence (AI), it provides prestigious support in predicting any kind of event which take training from natural events. To calculate accuracy, machine learning algorithms are used for predicting heart disease. The Algorithms are k-nearest neighbor, decision tree, linear regression and support vector machine (SVM) by using UCI repository dataset for training and testing.

II. LITERATURE SURVEY

[1] Online Prediction of Exacerbation in Patients with Chronic Obstructive Pulmonary Disease Using Linear Discriminant Pattern Classification

Morten H. Jensena, Simon L. Cichosza, Birthe Dinesena, Ole K. Hejlesena Chronic obstructive pulmonary disease (COPD) is a burden on health care because of general care costs. Exacerbations alone cause an additional 23,000 hospitalizations each year in Denmark. Patients seeking treatment for exacerbations often delay consultation for several days after the onset of symptoms. Poor outcomes are often related to failure to seek appropriate treatment therapy. The aim of this study was to investigate whether the use of physiological data is suited for online prediction of COPD exacerbations. Home measurements from 57 patients were analysed and 273 different features were evaluated for their discrimination abilities between periods with and without exacerbations. Results show that if a sensitivity level of 70 % is assessed acceptable the specificity is 95 %, and AUC = 0.73, of the best classifier. Our findings indicate that it is possible to discriminate between periods of exacerbation and periods without. We suggest, that more research in this area should be conducted.

[2] LIBSVM: A Library for Support Vector Machines

Chih-Chung Chang and Chih-Jen Lin LIBSVM is a library for Support Vector Machines (SVMs). We have been actively developing this package since the year 2000. The goal is to help users to easily apply SVM to their applications. LIBSVM has gained wide popularity in machine learning and many other areas. In this article, we present all implementation details of LIBSVM. Issues such as solving SVM optimization problems, theoretical convergence, multi-class classification, probability estimates, and parameter selection are discussed in detail.

III. EXISTING SYSTEM

In the Existing System, Although Electronic Health Records (EHRs) have pulled in expanding research consideration in the information mining and AI people group. The methodology is restricted to a twofold arrangement issue (utilizing alive/perished marks) and therefore it isn't instructive about the particular illness territory in which an individual is in danger. Unlabelled information grouping are regularly taken care of through Semi-Supervised Learning (SSL) that gains from both named and unlabelled information, and Positive and Unlabelled (PU)learning, an exceptional instance of SSL that gains from positive and unlabelled information alone.

Disadvantages:

1. Prediction of cardiovascular disease results is not accurate.
2. Most existing arrangement techniques on medical care information don't think about the issue of unlabelled information.
3. Data mining techniques does not help to provide effective decision making.

IV. PROPOSED SYSTEM

The primary objective of this framework is to foresee coronary illness utilizing information mining method, for example, Naive Bayesian Algorithm. Crude clinic informational index is utilized and afterward pre-processed and changed the informational collection. At that point apply the information mining method, for example, Naïve Bayes calculation on the changed informational index. In the wake of applying the information mining calculation, coronary illness is anticipated and afterward exactness is determined.

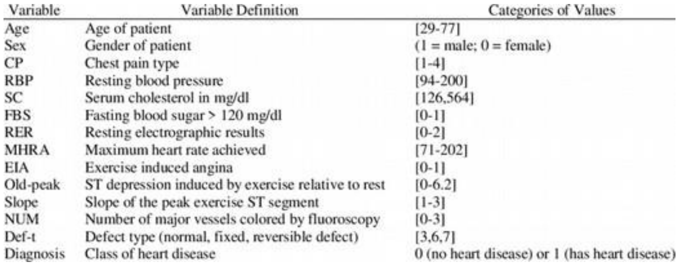
Advantages:

1. Increased accuracy for effective heart disease diagnosis.
2. Handles roughest(enormous) amount of data using random forest algorithm and feature selection.
3. Reduce the time complexity of doctors.
4. Cost effective for patients.

V. APPROACH

Data Pre-Processing:

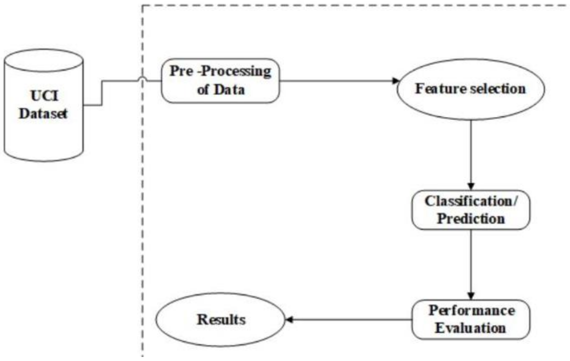
Heart disease data is pre-processed by using various collection of records. The dataset contains a total of 303 patient records.



Variable	Variable Definition	Categories of Values
Age	Age of patient	[29-77]
Sex	Gender of patient	(1 = male; 0 = female)
CP	Chest pain type	[1-4]
RBP	Resting blood pressure	[94-200]
SC	Serum cholesterol in mg/dl	[126,564]
FBS	Fasting blood sugar > 120 mg/dl	[0-1]
RER	Resting electrographic results	[0-2]
MHRA	Maximum heart rate achieved	[71-202]
EIA	Exercise induced angina	[0-1]
Old-peak	ST depression induced by exercise relative to rest	[0-6.2]
Slope	Slope of the peak exercise ST segment	[1-3]
NUM	Number of major vessels colored by fluoroscopy	[0-3]
Def-t	Defect type (normal, fixed, reversible defect)	[3,6,7]
Diagnosis	Class of heart disease	0 (no heart disease) or 1 (has heart disease)

Future selection and Reduction:

In this module is utilized to choose the highlights of the given dataset. Characteristic choice was performed to decide the subset of highlights that were exceptionally related with the class while having low bury relationship.

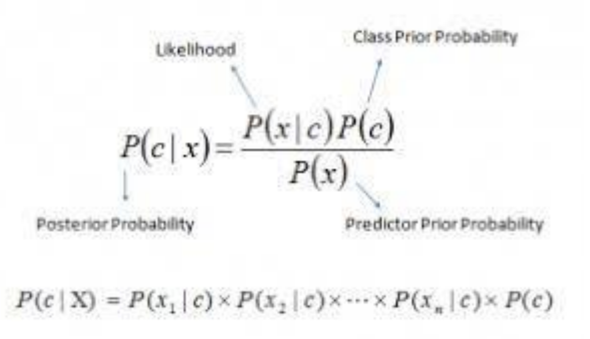


Classification Modeling:

The Naïve Bayesian Classification Algorithm speaks to a factual technique just as managed learning strategy for arrangement. Expects a probabilistic model which permits us to take care of the indicative and prescient issues. Bayes arrangement has been proposed which depends on Bayes rule of contingent likelihood. Innocent Bayesian guideline is a method used to assess the probability of a property from the given informational collection.

Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.

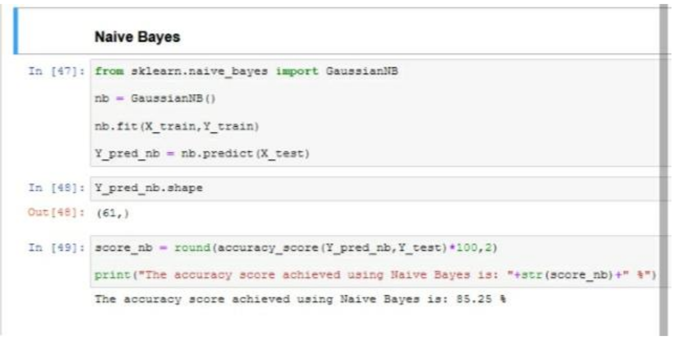
Equations:



$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Accuracy for Naive Bayes Algorithm:

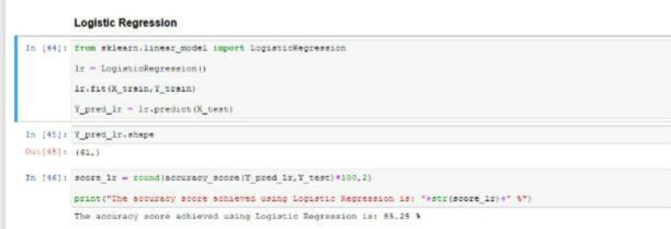


Naive Bayes

```
In [47]: from sklearn.naive_bayes import GaussianNB

         nb = GaussianNB()

         nb.fit(X_train,Y_train)

         Y_pred_nb = nb.predict(X_test)

In [48]: Y_pred_nb.shape

Out[48]: (61,)

In [49]: 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 %
```

Language Model:

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes. Mathematically, a logistic regression model predicts P(Y=1) as a function of X.



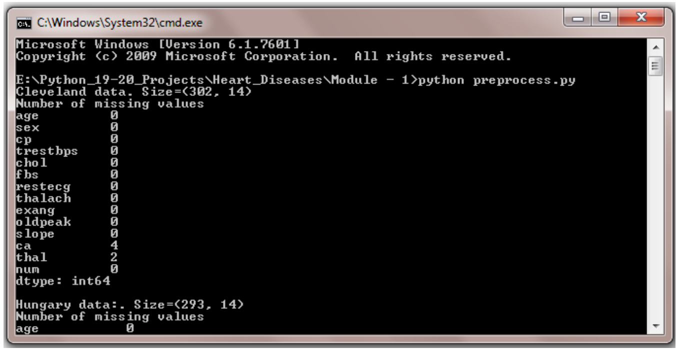
Logistic Regression

```
In [44]: from sklearn.linear_model import LogisticRegression

         lr = LogisticRegression()

         lr.fit(X_train,Y_train)

         Y_pred_lr = lr.predict(X_test)

In [45]: Y_pred_lr.shape

Out[45]: (61,)

In [46]: 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 %
```

Random Forest:

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. This ensemble classifier builds several decision trees and incorporates them to get the best result. For tree learning, it mainly applies bootstrap aggregating or bagging. For a given data, X = {x1, x2, x3, ..., xn} with responses Y = {x1, x2, x3, ..., xn} which repeats the bagging from b = 1 to B.

Generating the input using python and Random Forest classifier:



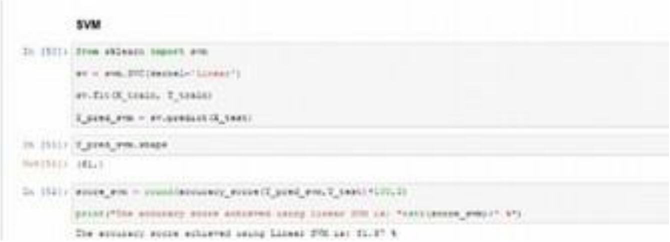
```
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation.  All rights reserved.

E:\Python_19-20_Projects\Heart_Diseases\Module - 1>python preprocess.py
Cleveland data. Size=(302, 14)
Number of missing values
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          4
thal        2
num         0
dtype: int64

Hungary data:. Size=(293, 14)
Number of missing values
age           0
```

Support Vector Machine:

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.

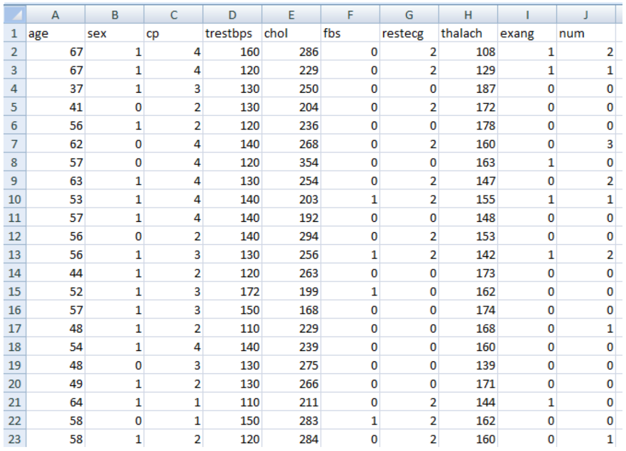
Let the training samples having dataset Data = {yi, xi}; i = 1, 2, . . . , n where xi ∈ R n represent the i th vector and yi ∈ R n represent the target item.

Accuracy for Support Vector Machine:



SVM

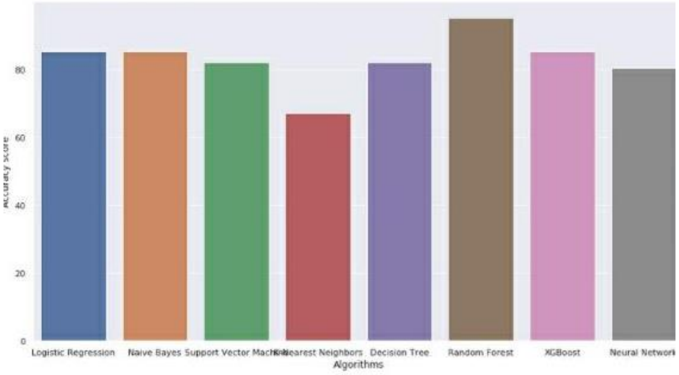
```
In [50]: from sklearn import svm

         sv = svm.SVC(kernel='linear')

         sv.fit(X_train, Y_train)

         Y_pred_svm = sv.predict(X_test)

In [51]: Y_pred_svm.shape

Out[51]: (61,)

In [52]: 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 %
```



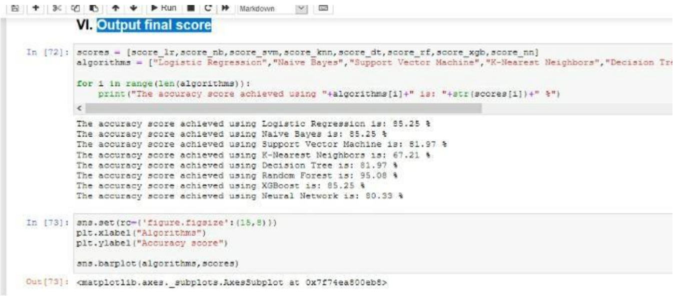
	A	B	C	D	E	F	G	H	I	J
1	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	num
2	67	1	4	160	286	0	2	108	1	2
3	67	1	4	120	229	0	2	129	1	1
4	37	1	3	130	250	0	0	187	0	0
5	41	0	2	130	204	0	2	172	0	0
6	56	1	2	120	236	0	0	178	0	0
7	62	0	4	140	268	0	2	160	0	3
8	57	0	4	120	354	0	0	163	1	0
9	63	1	4	130	254	0	2	147	0	2
10	53	1	4	140	203	1	2	155	1	1
11	57	1	4	140	192	0	0	148	0	0
12	56	0	2	140	294	0	2	153	0	0
13	56	1	3	130	256	1	2	142	1	2
14	44	1	2	120	263	0	0	173	0	0
15	52	1	3	172	199	1	0	162	0	0
16	57	1	3	150	168	0	0	174	0	0
17	48	1	2	110	229	0	0	168	0	1
18	54	1	4	140	239	0	0	160	0	0
19	48	0	3	130	275	0	0	139	0	0
20	49	1	2	130	266	0	0	171	0	0
21	64	1	1	110	211	0	2	144	1	0
22	58	0	1	150	283	1	2	162	0	0
23	58	1	2	120	284	0	2	160	0	1

Performance Measures:

The performanceof the algorithms was analyzed using parameters such as Accuracy, Precision, AUC and F1-score. From the experimental result, it is found that the Random Forest is more accurate for predicting the heart disease with accuracy of 95.08% compared with other supervised machine learning algorithms.



Output Final Scores:



VI. Output final score

```
In [72]: scores = [score_lr,score_nb,score_svm,score_knn,score_dt,score_rf,score_xgb,score_nn]
         algorithms = ["Logistic Regression","Naive Bayes","Support Vector Machine","K-Nearest Neighbors","Decision Tr

         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 accuracy score achieved using Random Forest is: 95.08 %
         The accuracy score achieved using XGBoost is: 85.25 %
         The accuracy score achieved using Neural Network is: 80.33 %

In [73]: sns.set(rc={'figure.figsize':(15,8)})
         plt.xlabel("Algorithms")
         plt.ylabel("Accuracy score")

         sns.barplot(algorithms,scores)

Out[73]: <matplotlib.axes._subplots.AxesSubplot at 0x7f74ea800eb8>
```

VI. CONCLUSION

Information mining applications are utilized immeasurably in the clinical field to recognize sicknesses and finding the heart tolerant dependent on the informational collection and the traits gave. Scientists have been exploring applying diverse information mining procedures to assist wellbeing with caring experts in the analysis of coronary illness. In the proposed work arbitrary woods calculation is utilized to order the informational collection since irregular woodland gives precise outcomes, with these outcomes heart infections among individuals is anticipated. In this manner heart illnesses expectation framework effectively analyse the clinical information and predicts the heart infections. The outcomes in this manner got shows that irregular woodland calculation gives 95.30% of precision least time.

REFERENCES

[1] B.Azhagusundari, Antony Selvadoss Thanamani: Feature Selection based on Information Gain :International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-2, January (2013).

[2] Beant Kaur, Williamjeet Singh: Review on Heart Disease Prediction System using Data Mining Techniques: International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 2 Issue: 10, October (2014).

[3] R.Chitra1 and V. Seenivasagam : Review Of Heart Disease Prediction System Using Data Mining And Hybrid Intelligent Techniques: Ictact Journal On Soft Computing, Volume: 03, Issue: 04, July (2013).

[4] T.Georgeena.S. Thomas, Siddhesh.S. Budhkar, Siddhesh.K. Cheulkar, Akshay.B.Choudhary, Rohan Singh: Heart Disease Diagnosis System Using Apriori Algorithm: International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 2, February (2015).

[5] Institute for health metrics and evaluation- Article Deaths from cardiovascular disease increase globally while mortality rates decrease.

[6] Muhammad Arif Mohammad, Haswadi Hassan, Dewi Nasien & Habibollah Haron,:A Review on Feature Extraction and Feature selection for Handwritten Character Recognition: International Journal of Advanced Computer Science and Applications,Vol 6, No 2, (2015).

[7] Ron Kohavi and Dan Sommerfield: Feature Subset Selection Using the Wrapper Method:Overfitting and Dynamic Search Space Topology: KDD-95 Proceedings, (1995).

[8] K.Sudhakar & Dr.M.Manimekalai : Study of Heart disease prodiction using data mining: International journal of Advanced Research in Computer Science and Software Engineering, Vol 4, Issues 1, ISSN: 2277 i28x, pp 1157-1160, January (2014).

[9] Swati A sonawale & Roshani Ade: Dimensionality Reduction: An Effective Technique for Feature Selection International Journal of Computer Applications, Volume 117-No 3, May (2015).

[10] Tarun Kumar Gupta, Chanchal Kumar, Shiv Prakash and Mukesh Prasad: Dimensionality Reduction Techniques and its Applications: Computer Science Systems Biology, (2015).