IJIRCCE

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

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 7, July 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.542

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907064 |

Prediction of Risk-Based Non-Contagious Diseases like Heart and Liver Using Supervised Machine Learning Algorithms

K.Deepthi Haritha[1], K.Kavya[2], M.Jagadeeswari[3], K.Sai Bhavana[4], Md.Shakeel Ahmad[5]

UG Students, Department of Information Technology, Vasireddy Venkatadri Institute of Technology, Nambur, Guntur, Andhra Pradesh, India [1,2,3,4]

Associate Professor, Department of Information Technology, Vasireddy Venkatadri Institute of Technology, Nambur, Guntur, Andhra Pradesh, India [5]

ABSTRACT: Our project "PREDICTION OF RISK-BASED NON-CONTAGIOUS DISEASES LIKE HEARTAND LIVERUSINGSUPERVISED MACHINE LEARNINGALGORITHMS"can be used by doctors or medical professionals to detect diseases in patients using the JUPITER Notebook. It is developed to predict the diseases such as Liver and Heart disease and also whether a person is suffering from that disease or not.Each of these diseases has different signs and symptoms among the patients. Different datasets are used in this project such as liver and heart datasets which are applied to the machine learning algorithms by this,we can get each algorithm accuracy which leads to decide where the person is suffering from liver or heart disease. For the classification calculation, Supervised Algorithmswe used in projectsuch as Decision Tree Algorithm, Random Forest Algorithm and Navie Bayes Algorithm.

KEYWORDS: Machine Learning, Decision Tree Algorithm, Random Forest Algorithm, Naïve Bayes Algorithm, Heart and Liver Diseases.

I. INTRODUCTION

The purpose of making our project is to predict the accurate disease of the patient using all their general information's and also the symptoms. If this Prediction is done at the early stages of the disease with the help of this project and all other necessary measures the disease can be cured and in general this prediction system can also be very useful in health industry. The work of the doctors can be reduced and they can easily predict the disease of the patient. This system will predict the most possible disease based on the symptoms

II. LITERATURE SURVEY

In the above study we will see different data mining techniquesthat were used to classify the heart diseases. In year 2000, research conducted by ShusakuTsumoto [5] says that as we human beings are unable to arrange data if it is huge in size
we should use the data mining techniques that are available for finding different patterns from the available huge database and
can be used again for clinical research and perform various operations on it. Y. Alp Aslan Dogan, et. al. (2004), worked on three different classifiers called K-nearest Neighbour (KNN), Decision Tree, Naïve Bayesian and used Dempster's' rule for thisthreeviewpoint to appear as one concluding decision. This classification based on the combined idea show increased accuracy [6]. Carlos Ordonez (2004), Assessed the problematic to recognize and forecast the rule of relationship for the heart disease. Adataset involving medical history of the patients having heart disease with the aspects of risk factors was accessed by him, measurements of narrowed artery and heart perfusion. All these restrictions were announced to shrink the digit of designs, these are as follows: 1) The features should seem on a single side of the rule. 2) The rule should distinct variousfeatures into the different groups. 3) The count of featuresavailable from the rule is organized by medical history of people having heart disease only. The occurrence or the nonappearance of heart disease was predicted by the author in four heart veins with the two clusters of rules [7]. Franck Le Duff (2004), worked on creating Decision tree quickly with clinical data of the physician or service. He suggested few data mining techniques which can help cardiologists in the predication survival of patients. The main

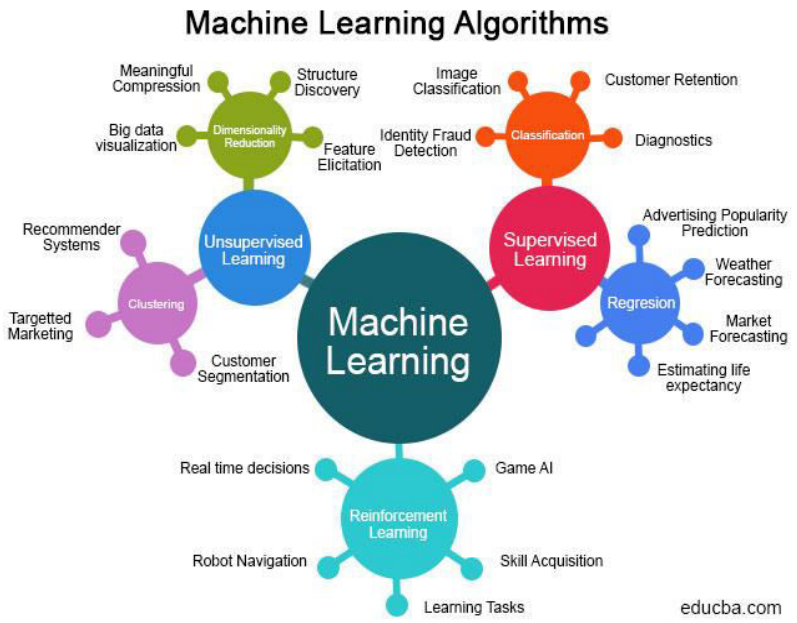
| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907064 |

drawback of the system was that the user needs to have knowledge of the techniques and we should collect sufficient data for creatingan suitable model [8]. Boleslaw Szymanski, et. al. (2006), operated on a novel experiential to check the aptitude of calculation of scarce kernel in SUPANOVA. The author used this technique on a standard boston housing market dataset for discovering heart diseases, measurement of heart activities and prediction of heart diseases were found 83.7% correct which were measured with the help of support vector machine and kernel equivalent to it. A quality result is gained by spline kernel with the help of standard boston housing market database [9]. Kiyong Noh, et. al. (2006) made use of a classification technique for removal of multi-parametric structures by accessing HRV and ECG signals. Kiyong used the FPgrowth algorithm as the foundation of this technique that is associative. A rule consistency degree was gained which allows a robust press on trimming designs in the method of producing designs[10].

III. DESCRIPTION OF MACHINE LEARNING ALGORITHMS



Machine learning is a method of data analysis that automates analytical model building. It is a branch of ARTIFICIAL INTELLIGENCE based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.

Once Machine Learning Algorithm scan pinpointon certain correlations, the model can either use these relationships to predict future observations or generalize the data to reveal interesting patterns. In Machine Learning there are various types of algorithms such as Regression, Linear Regression, Logistic Regression, Naive Bayes Classifier, Bayes theorem, KNN (K-Nearest Neighbor Classifier),DecisionTress,Entropy,ID3,SVM(SupportVectorMachines),K-meansAlgorithm,RandomForestandetc.,

Supervised learning:

Supervised learning algorithms are trained using labeled examples, such as an input where the desired output is known. For example, a piece of equipment could have data points labeled either “F” (failed) or “R” (runs). The learning algorithm receives a set of inputs along with the corresponding correct outputs, and the algorithm learns by comparing its actual output with correct outputs to find errors. It then modifies the model accordingly. Through methods like classification, regression, prediction and gradient boosting, supervised learning uses patterns to predict the values of the label on additional unlabeled data. Supervised learning is commonly used in applications where historical data predicts likely future events. For example, it can anticipate when credit card transactions are likely to be fraudulent or which insurance customer is likely to file a claim.

In this project there are mainly 3 **Supervised Machine Learning Algorithms** we used. They are:

1. Decision Tree Algorithm.
2. Random Forest Algorithm.

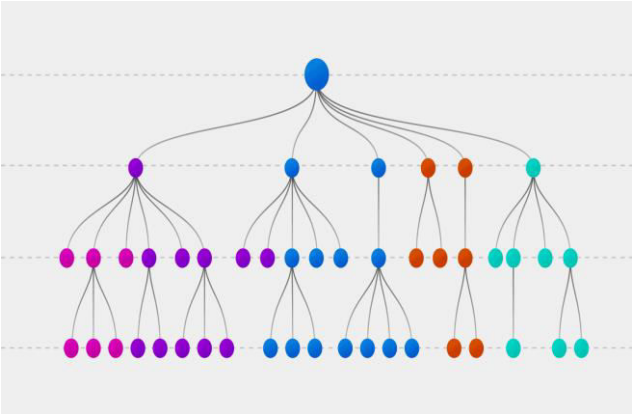
| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907064 |

3. Navie Bayes Algorithm.

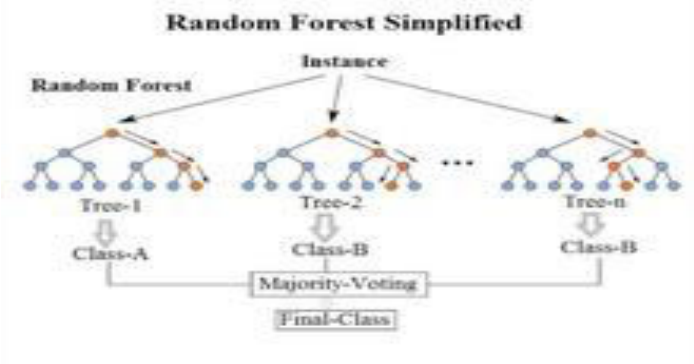
1. Decision Tree Algorithm



Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represent the outcome.

A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees

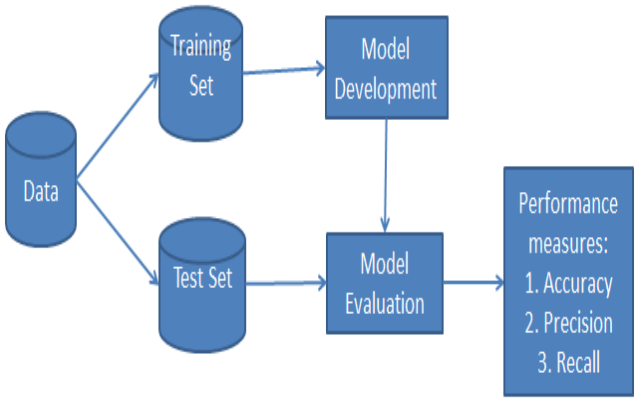
2. Random Forest Algorithm.



"Random Forest is a classifier that contains anumber of decision trees on various subsets of the given dataset and takesthe average to improve the predictiveaccuracy 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.

3. Navie Bayes Algorithm.

Naive Bayes is one of the fast and easy ML algorithms to predict a class of datasets. It can be used for Binary as well as Multi-class Classifications.



It performs well in multi-class predictions as compared to the other Algorithms. It is the most popular choice for text classification problems. This algorithm works quickly and can save a lot of time. Naive Bayes is suitable for solving multi-class prediction problems. If its assumption of the independence of features holds true, it can perform better than other models and requires much less training data

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 7, July 2021 ||

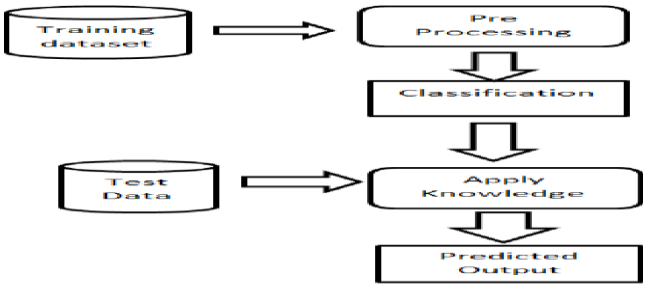
| DOI: 10.15680/IJIRCCE.2021.0907064 |

IV. EXISTING SYSTEM

Prediction using traditional methods and models involves various risk factors and it consists of various measures of algorithm such as datasets, programs and much more to add on.High-risk and Low-risk based patient classification is done on the basis of the tests that are done in group.But these models are only valuable in clinical situation and not in big industry
Sector.So, to include the disease predictions in various health related industries,we have used the concepts of machine learning and supervised leaning methods to build the prediction system.After doing the research and comparison of all the algorithms and theoremsofMachine learning we have come to conclusion that all those algorithms such as DecisionTree,KNN,Naïve Bayes ,Regression and Random Forest Algorithm all are important in building apredictionof heart and liver diseases which predicts the disease of the patients from which he/she is suffering from and to do this we have used some performance measures like ROC, KAPPA Statistics, RMSE, MEA and various other tools. After using various techniques such as neural networks to make predictions of the diseases and after doing that we come toconclusion that it can predicts up to less accuracy rate after doing the experimentation and verifying the results. The information of patient statistics, results, disease history in recorded in EHR, which enables to identify the potential data centric solution, which reduces the cost of medical case studies.
Existing system can predict the disease but notthe sub type of the disease and it fails to predict the correct presence of the disease, the predictions ofdisease have been in definite and non-specific.

V. PROPOSED SYSTEM

The proposed system of this project is that we have usedmany techniques and algorithms and all other various tools to build a system which predicts the liver and heart disease of the patient using the symptoms and by taking those symptoms we are comparing with the system's dataset that is previously available. By taking those datasets and comparing with the patient's disease we will predict the accurate percentage disease of the patient. The dataset and symptoms go to the prediction model of thesystem where the data is pre-processed for the future references and then the feature selection is done by the user where he will be inputting the various symptoms. Then the classification of those data is done with the help of various algorithms and techniques such as Decision Tree, Naïve Bayes, Random Forest. Then the data goes in there commendation model, there it shows accuracy rates of the system such that from their final result and also from their symptoms like it can show what to use and what not to use from the given datasets and the final results. Here we have combined the overall structure and unstructured form of data for the overall risk analysis that is required for doing the prediction of the disease. After inputting the results Data Cleaning and Data Transformation is done. After that we will train and test the data accordingly.This system takes symptomsfrom the user and predicts the liver or heart disease accordingly based on the symptoms that it takes and also from the previous datasets, then it predicts the appropriate and accurate disease.



VI. SYSTEM REQUIREMENTS

6.1 SOFTWARE REQUIREMNETS

- Operating System-Windows7/8
- ProgrammingLanguage-Python(python3.6.3)

6.2HARDWARE REQUIREMENTS

- Processor - Pentium –IV
- Speed - 1.1 Ghz
- RAM - 256 MB(min)
- Hard Disk - 20 GB

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907064 |

- Key Board - StandardWindows Keyboard
- Mouse - Two or Three Button Mouse
- Monitor - SVGA

VII. HEART AND LIVER DATASETS

7.1 Heart Dataset

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	age	sex	cp	trestbps	chol	fbs	resteceg	thalach	exang	oldpeak	slope	ca	thal	target
2	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
3	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
4	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
5	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
6	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
7	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
8	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
9	44	1	1	120	263	0	1	173	0	0	2	0	3	1
10	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
11	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
12	54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
13	48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
14	49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
15	64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
16	58	0	3	150	283	1	0	162	0	1	2	0	2	1
17	50	0	2	120	219	0	1	158	0	1.6	1	0	2	1
18	58	0	2	120	340	0	1	172	0	0	2	0	2	1
19	66	0	3	150	226	0	1	114	0	2.6	0	0	2	1
20	43	1	0	150	247	0	1	171	0	1.5	2	0	2	1
21	69	0	3	140	239	0	1	151	0	1.8	2	2	2	1

7.2 Liver Disease

	A	B	C	D	E	F	G	H	I	J	K
	Age	Gender	Total_Bilirubin	Direct_Bilirubin	Alkaline_Phosphotase	Alamine_Aminotransferase	Aspartate_Aminotransferase	Total_Protiens	Albumin	Albumin_and_Globulin_Ratio	target
	65	Female	0.7	0.1	187	16	18	6.8	3.3	0.9	1
	62	Male	10.9	5.5	699	64	100	7.5	3.2	0.74	1
	62	Male	7.3	4.1	490	60	68	7	3.3	0.89	1
	58	Male	1	0.4	182	14	20	6.8	3.4	1	1
	72	Male	3.9	2	195	27	59	7.3	2.4	0.4	1
	46	Male	1.8	0.7	208	19	14	7.6	4.4	1.3	1
	26	Female	0.9	0.2	154	16	12	7	3.5	1	1
	29	Female	0.9	0.3	202	14	11	6.7	3.6	1.1	1
	17	Male	0.9	0.3	202	22	19	7.4	4.1	1.2	2
	55	Male	0.7	0.2	290	53	58	6.8	3.4	1	1
	57	Male	0.6	0.1	210	51	59	5.9	2.7	0.8	1
	72	Male	2.7		260	31	56	7.4	3	0.6	1
	64	Male	0.9	0.3	310		58	7	3.4	0.9	2
	74	Female	1.1	0.4	214	22	30	8.1	4.1	1	1
	61	Male	0.7	0.2	145	53	41	5.8	2.7	0.87	1
	25	Male	0.6	0.1	183	91	53	5.5	2.3	0.7	2
	38	Male	1.8	0.8	342	168	441	7.6	4.4	1.3	1
	33	Male	1.6	0.5	165	15	23	7.3	3.5	0.92	2
	40	Female	0.9	0.3	293	232	245	6.8	3.1	0.8	1

VIII. DESIGN OVERVIEW

Steps involved in Design Overview:

1.DataSet Collection

In this step we will be collecting the datasets of both liver and heart diseases.The dataset is the matrix where the rows represent the patient details and the columns represents the factors or attributes to be tested.

2. Data Wrangling

```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        %matplotlib inline
        import seaborn as sns
```

3.Data Cleaning

In this Data Cleaning we will be removing the noisy data like Null Values which are seen when Datasets are gathered from Online or Hospitals.By Removing All the Null Values with the help of its median value. The Output will be shown as below:

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907064 |

```
In [15]: df.isnull().sum()

Out[15]: Age                         0
         Gender                      0
         Total_Bilirubin             0
         Direct_Bilirubin            0
         Alkaline_Phosphotase        0
         Alamine_Aminotransferase    0
         Aspartate_Aminotransferase  0
         Total_Protiens              0
         Albumin                     0
         Albumin_and_Globulin_Ratio  0
         target                      0
         dtype: int64
```

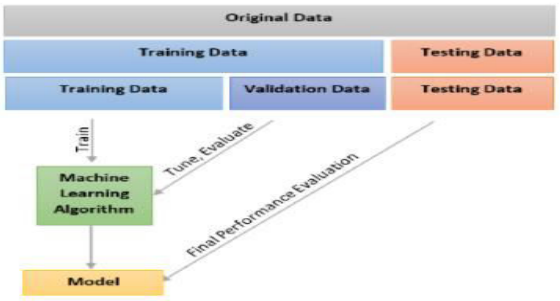
In this we can see all represents 0 that means all null values are removed by its median. Hence Data Cleaning is successful.

4.Data Transformation

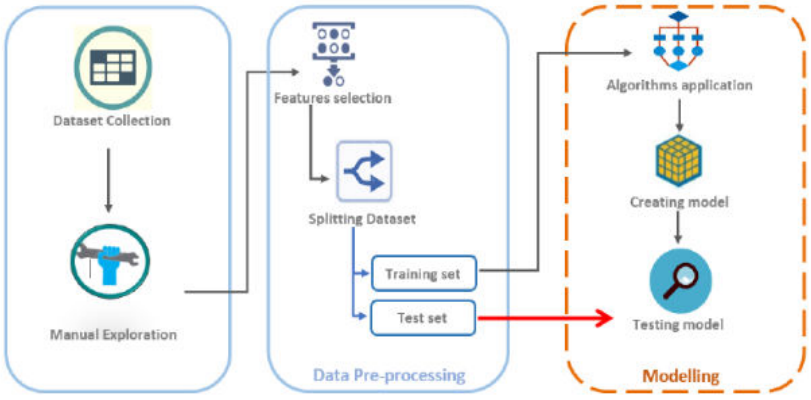
In this we will be transforming the data
Male→ 0 and Female → 1.

```
In [18]: df['Gender'].value_counts()

Out[18]: Male      820
         Female    257
         Name: Gender, dtype: int64

In [19]: df['Gender']=df['Gender'].replace(('Male','Female'),(0,1))

In [20]: df['Gender'].value_counts()

Out[20]: 0    820
         1    257
         Name: Gender, dtype: int64
```

5.Prepare Data for Modelling



6.Modelling/Training



| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907064 |

7. Choosing the best Algorithm

We finish our design overview by selecting the best Algorithm that gives the best accuracy,and next present the results obtained.

IX. TOOLS, METHODOLOGIES ANDTECHNOLOGIES

9.1 Methodologies Used

- Data Cleaning and Data Transformation
- Machine Learning - Supervised Learning Algorithms

Decision Tree
Random Forest
Navie Bayes

- Evaluation Metrics

Confusion Matrix
Classification Report

9.2 Tools Used

- Anaconda Application Tool
- Jupyter Notebook

9.3 Technologies Used

- Anaconda
- Jupyter Notebook
- Python
- Pandas
- NumPy
- Matplotlib
- Seaborn
- Sklearn

X. EXPERIMENTAL RESULTS ANDANALYSIS

For Heart and Liver Disease

Algorithm (Random Forest)Accuracy Score:

```
In [15]: a1

Out[15]: 0.8360655737704918
```

Output (Prediction):(Heart Disease)

1 indicates presence of heart disease
0 indicates absence of heart disease

Output (Predictions):(Liver Disease)

1 indicates Presence of liver Disease
2 indicates absence of liver Disease

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907064 |

'age', 'sex', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach','exang', 'oldpeak', 'slope', 'ca', 'thal'

```
In [65]: rf.predict([[56,1,1,120,236,0,1,178,0,0.8,2,0,2]])
Out[65]: array([1], dtype=int64)

In [26]: rf.predict([[51,0,0,130,305,0,1,142,1,1.2,1,0,3]])
Out[26]: array([0], dtype=int64)
```

Algorithm (Naïve Bayes) Accuracy Score:

```
In [33]: a2
Out[33]: 0.8360655737704918
```

Output (Prediction):

'age', 'sex', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach','exang', 'oldpeak', 'slope', 'ca', 'thal'

```
In [42]: nb.predict([[63,1,3,145,233,1,0,150,0,2.3,0,0,1]])
Out[42]: array([0], dtype=int64)
```

Algorithm (Decision Tree) Accuracy Score:

```
In [49]: a3
Out[49]: 0.7704918032786885
```

Output (Prediction):

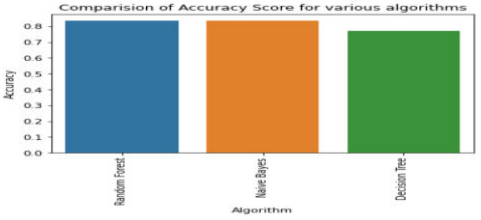
```
In [58]: tree.predict([[63,1,3,145,233,1,0,150,0,2.3,0,0,1]])
Out[58]: array([0], dtype=int64)
```

Comparison of accuracy and loss levels for various algorithms:

Out[60]:

	Algorithm	Accuracy	Loss
0	Random Forest	0.836066	0.163934
1	Naive Bayes	0.836066	0.163934
2	Decision Tree	0.770492	0.229508

```
Out[61]: Text(0.5, 1.0, 'Comparision of Accuracy Score for various algorithms')
```



| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542 |

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907064 |

XI. CONCLUSION

In this project we can predict whether the person is suffering with heart and liver disease or not by accuracy rates of each algorithm for heartand liver disease using different datasets.Our project can be used by doctors or medical professionals to detect diseases in patients.The work of the doctors can be reduced and they can easily predict the disease of the patient.Our machine learning algorithm can now classify patients with Heart and liver Disease. Now we can properly diagnose patients, & get them the help they need to recover. By diagnosing detecting these features early, we may prevent worse symptoms from arising later. Our Random Forest algorithm yields the highest accuracy, 80%. Any accuracy above 70% is considered good, but We have to be careful because if the accuracy is extremely high, it may be too good to be.Thus, 80% is the ideal accuracy.This is the final conclusion of our project.

XII. FUTURE ENHANCEMENTS

Future enhancements of this project is to predict the disease in a different area are hospital, Clinic, smartphone, smart wear, hospital/police emergency system and integrate with fitness mobile application. We will integrate this model in hospital and clinic system to predict heart and liver disease. We will implement this project into smart wears to detect essential attributes of Heart and Liver disease. we will also apply this model into a mobile app to easily test ourselves. we will integrate smart wear to the hospital and police emergency system to save the life of the patient at the emergency condition.

REFERENCES

[1] Deepika Verma AND Dr. Nidhi Mishra “Analysis and Prediction of Breast cancer and Diabetes disease datasets using Data mining classification Techniques” ©2017 IEEE.
[2] Saba Bashir, Zain Sikander Khan, Farhan Hassan Khan, Aitzaz Anjum, Khurram Bashir “Improving Heart Disease Prediction Using Feature Selection Approaches” Proceedings of 2019 16th International Bhurban Conference on Applied Sciences & Technology (IBCAST) Islamabad, Pakistan, 8th – 12th January, 2019.
[3] Prerna Jain, Amandeep Kaur “Big Data Analysis for Prediction of Coronary Artery Disease” ©2018 IEEE.
[4] Anjinkya Kunjir, Harshal Sawant, Nuzhat F.sheikh “Data Mining and Visualization for Prediction of Multiple Diseases in Healthcare” ©2017 IEEE.
[5] Ayman Mir, Sudhir N. Dhage “Diabetes Disease Prediction using Machine Learning on Big Data of Healthcare” ©2018 IEEE.
[6] Dr. S. N. Singh, Shivani Thakral “Using Data Mining Tools for Breast Cancer Prediction and Analysis” ©2018 IEEE.
[7] Thirunavukkarasu K., Ajay S. Singh, Md Irfan, Abhishek Chowdhury “Prediction of Liver Disease using Classification Algorithms” 2018 4th International Conference on Computing Communication and Automation (ICCCA) 978-1-5386-6947- 1/18/$31.00 ©2018 IEEE.
[8] Arvindkumar.s,Arun.P, Ajith.A,” Prediction of Chronic Disease by Machine Learning’ @IEEE 978-1-2-7281-1524-5.
[9] K. Vijiya Kumar,B.Lavanya,I.Nirmala,S.Sofia Carroline ,”Random forest Algorithm For The Prediction of Diabetes “IEEE 978-1-7281-1524-5.
[10] Divya Krishnani,Akash Dewangan, Aditya Singh, Nenavath srinivas Naik,” Prediction of coronary Heart Disease Using Superviseed Machine Learning Algorithms” 978-1-7281-1895-6/19/$31.00©2019IEEE.

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 7.542

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