A Critical Analysis of Various Machine Learning Algorithm to Extract Disease Patterns

Ashish Tiwari, Pramod Kumar Maurya

School of Computer Science and Engineering
Vellore Institute of Technology ((Deemed to be University under section 3 of UGC Act, 1956)
Vellore (T.N), India
-632014

ABSTRACT

In this fast-pacing world, people desire to live a luxurious life, and to attain the same, they work tirelessly. This not only changes their food habits but also brings along a drastic change in their lifestyle. They are tenser and tend to develop many modern-day chronic diseases at an early age. Despite all this, many people don't bother to change their living or take proper medication, which in turn later leads to major threats. In this paper, we will deploy various machine learning algorithms like Regression, Decision tree classification, Random Forest, etc. to study the behavior of various chronic diseases and do a comparative analysis of the obtained results for accurate prediction. For this, we have used 3 disease datasets namely the PIMA Indian Diabetes dataset, Indian Liver Patient dataset, heart disease dataset for analysis. The final objective of the paper is to find which ML algorithm is apt for each of the datasets and what features were taken into consideration to achieve the best possible accuracy. This will not only help the patients to take the medications accordingly but will also help other people in day-to-day life to take proper care of themselves.

Keyword: *Machine Learning, Diabetes, Heart Disease, Liver, classification, prediction*

1. INTRODUCTION

Nowadays, it has become difficult to identify disease in the early stages as the symptoms are very minuscule. But as the stage advance, the disease deteriorates one's health and can prove to be fatal in the longer run. Therefore, it is very necessary for people to focus on their health beforehand and take necessary actions[6]. It will not only benefit an individual but also reduce the burden on the medical infrastructure. Every year, millions of deaths take place because of cardiovascular disorder, lung-related disease, liver infections, and other diseases but still, no necessary precautions are taken. It is very necessary to make people aware of the discourse & the parameters that are impacting their health and encourage them to take relevant actions in advance.

Currently, the medical industry relies on the traditional approach of treating the patient. i.e., an individualistic approach. Each and every patient is treated on case-to-case basis. As the fast and unhealthy livelihood of people are impacting their health, it's also little too late till any disease is diagnosed in them. This not only increase the burden on the healthcare industry but also in many countries people don't get easy and convenient access to healthcare which makes it difficult for treatment and diagnosis. Another reason such approach is disadvantageous is patient has to be constantly monitored by medical professionals and in countries with poor healthcare infrastructure it's arduous.

In this research paper, we have taken 3 disease datasets namely the PIMA Indian Diabetes dataset, Indian Liver Patient dataset, Cleveland Heart Disease dataset. Various machine learning algorithms will be applied on all three datasets and final accuracy will be marked. The final objective of the paper is to find which ML algorithm suits best each of the datasets and what features or attributes were taken into consideration to achieve the best accuracy. The different dataset will have different feature variable taken into the account and this will help us to predict that what are the parameters that affect or impacts the disease more. This will not only help the patients to take the medications accordingly but will also help other people in day-to-day life to take care of themselves.

2. LITERATURE SURVEY

Numerous risk factors related to diabetes mellitus was predicted in [1] using the machine learning techniques. The major supervised learning model that's considered are support vector machine, decision tree, naïve bayes and in the final prediction it was concluded that the decision tree is the superior model compared to rest of the ml model. One of the major drawbacks of the research was it didn't consider many other ml models that could have been more efficient. Plus, it didn't take many other features and attributes into consideration.

In [2], the major advantage of the research is Deep learning and machine learning technique were applied on 2 sets of datasets to cross check the validity of the prediction. The only drawback it has is, only Random Forest and Decision Tree technique was deployed for Machine Learning which isn't enough to compare the efficiency of the algorithm to the rest of the algorithms.

The [7] gave an insight into the dataset and the machine learning model that could be deployed for information mining and which info. mining technique could be efficient for medical findings. The only con of the paper is very limited machine learning supervised models were used. Additionally, [8] illustrated a decent flow of how to understand the underlying patterns in the diabetes dataset. Additionally, it also portrayed the correlation among the features and importance of each feature in the dataset for prediction. The only disadvantage being not having enough attributes to derive a solid conclusion.

In the earlier researches, limited algorithms were taken into account under a fixed training testing ratio and a final conclusion was made. But, in this research paper, we will deploy various supervised machine learning algorithms under different training and testing ratio and also check if the accuracy of the model changes with the change in the ratio.

3. PROPOSED SYSTEM

The flow diagram of the model can be seen below. It starts with importing all of the essential libraries and the dataset. It then moves on to the second stage, which is to show the relationship between the parameters that are most successful in doing the task and obtaining more accurate results. The next stage is data pre-processing, which involves cleaning the dataset and preparing it for use with our model. After pre-processing we perform exploratory data analysis in order to get the illustrated insights and statistical inferences. Our dataset is now ready to be used in the algorithms after pre-processing.

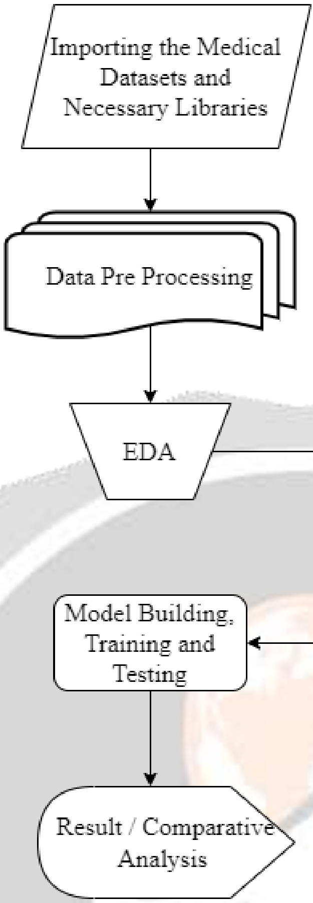

Importing the Medical Datasets and Necessary Libraries
Data Pre Processing
EDA
Data Visualization
Model Building, Training and Testing
Result / Comparative Analysis

Fig-1: Proposed Architecture

4. METHODOLOGY

4.1 Dataset Description

This dataset contains patient records with and without liver problem that is collected form North East of Andhra Pradesh region. The dataset contains around 11 attributes and 583 instance of data value. Following attributes were taken into consideration for experimentation:

1. Age of the patient
2. Gender of the patient
3. Total Bilirubin: An orange-yellow pigment that is produced after the break down of red blood cells.
4. Direct Bilirubin: Bilirubin that binds glucuronic acid, a glucose-derived acid to the liver to is called direct bilirubin.
5. Alkaline Phosphatase: An enzyme in a person's blood that aids in breaking down proteins.
6. Alamine Aminotransferase (ALT): An enzyme that is found in the kidney and liver. ALT increases with liver damage and is used to monitor liver disease.
7. Aspartate Aminotransferase (AST): An enzyme that is found mostly in the liver. AST is released into your bloodstream when the liver is damaged.
8. Total Proteins:
Albumin and globulin are two types of protein in the body.
9. Albumin: protein that stops the fluid in the blood to leak out in the tissues.
10. Albumin and Globulin Ratio: Indicator of the state of the liver

11. Dataset: patient with liver disease, or no disease

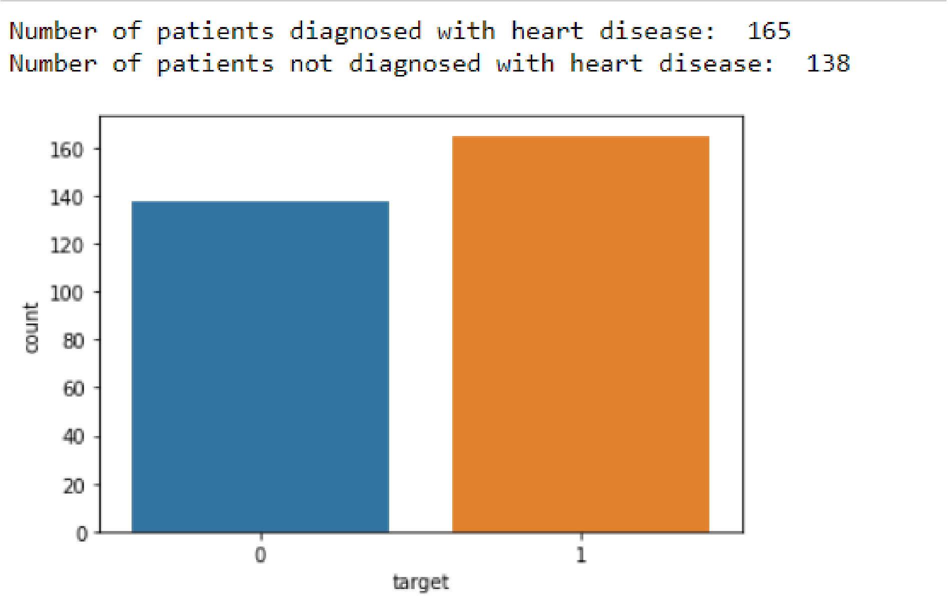
Another dataset that we considered consists of patient’s medical records based on diabetes. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset mentioned below. The dataset consisted 10 attributes and 768 instances of data.

1. Pregnancies: Number of times pregnant
2. Glucose: Plasma glucose concentration
3. BloodPressure: Diastolic blood pressure (in mm Hg)
4. SkinThickness: Triceps skin fold thickness (in mm)
5. Insulin: 2-Hour serum insulin (in mu U/ml)
6. BMI: Body mass index (in weight in kg/(height in m)^2)
7. DiabetesPedigreeFunction
8. Age: Age (in years)
10. Outcome: Class variable (0 or 1)

The third disease dataset for classification and prediction considered was Cleveland heart dataset. In this, different health attributes of the patients are recorded which could be useful to predict certain cardiovascular events or find any clear indications of heart health. It has around 14 attributes with 303 instances of data values.

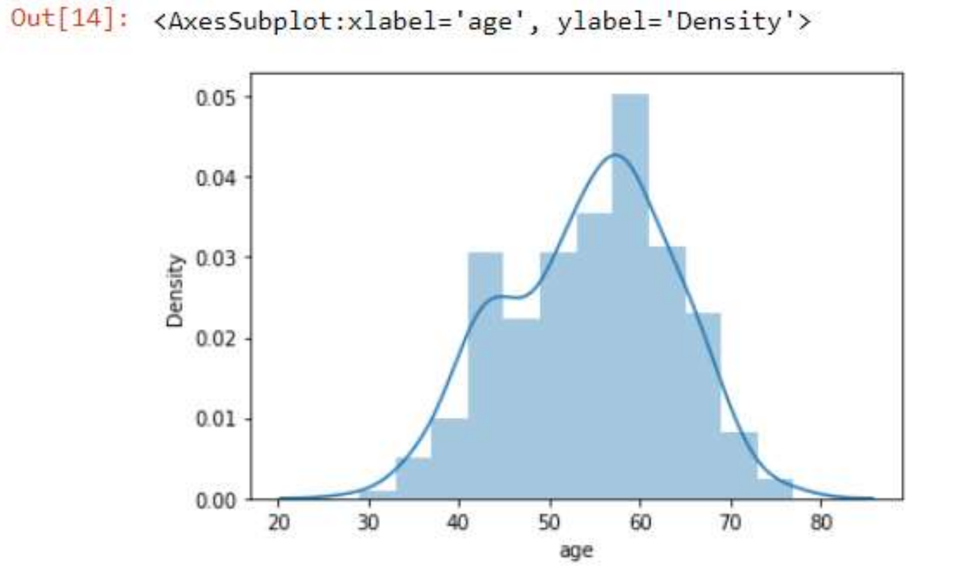
1. Age: displays the person's age
2. Sex: displays the person's gender
3. cp: Type of chest-pain experienced
4. trestbps: Resting blood pressure value of a person in mmHg (unit)
5. chol: Serum cholesterol in mg/dl (unit)
6. fbs: fasting blood sugar value of a person with 120mg/dl.
7. restecg : resting electrocardiographic results
8. thalach: max heart rate achieved by a person.
9. exang: Exercise induced angina
10. oldpeak: ST depression induced by exercise relative to rest
11. slope: Peak exercise ST segment
12. ca: Number of major vessels
13. Thal: displays the thalassemia
14. Diagnosis of heart disease: Displays whether the individual is suffering from heart disease or not

4.2 Exploratory Data Analysis



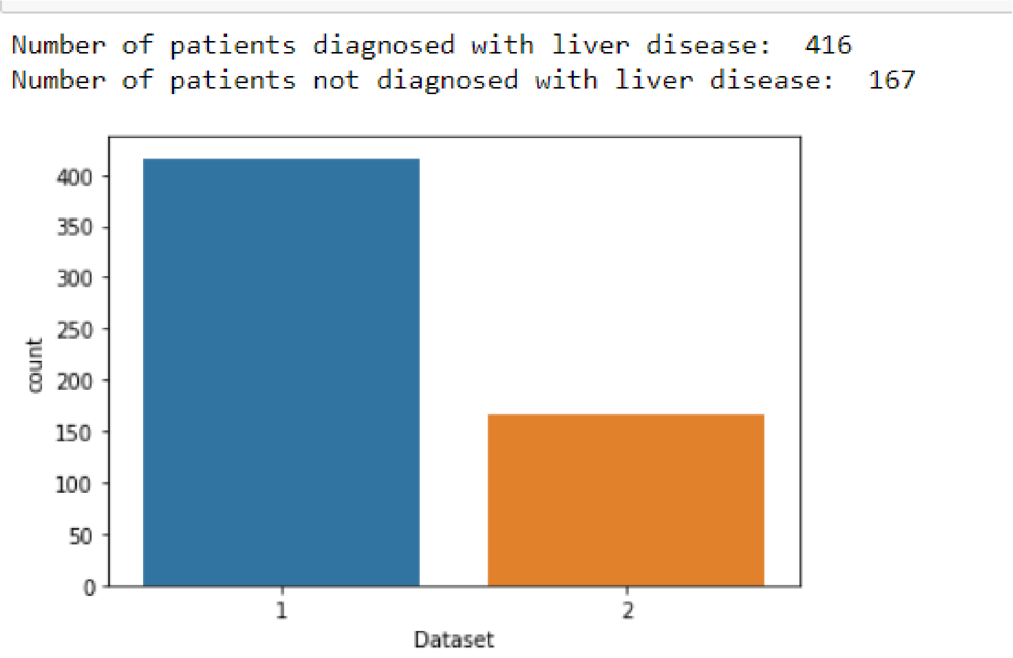
```
Number of patients diagnosed with heart disease:  165
Number of patients not diagnosed with heart disease:  138
```

Fig-2: Number of patients with & without heart disease



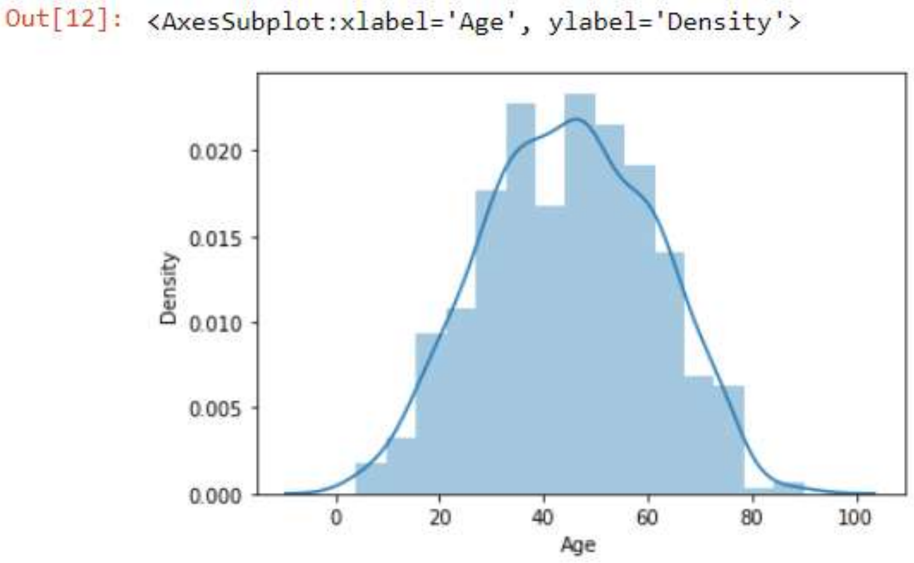
```
Out[14]: <AxesSubplot:xlabel='age', ylabel='Density'>
```

Fig-3: Age distribution of patients having heart problems



```
Number of patients diagnosed with liver disease:  416
Number of patients not diagnosed with liver disease:  167
```

Fig-4: Number of patients with & without liver disease



```
Out[12]:  <AxesSubplot:xlabel='Age', ylabel='Density'>
```

Fig-5: Age distribution of patients having liver problems

```
Number of patients diagnosed with diabetes disease:  500
Number of patients not diagnosed with diabetes disease:  268
```

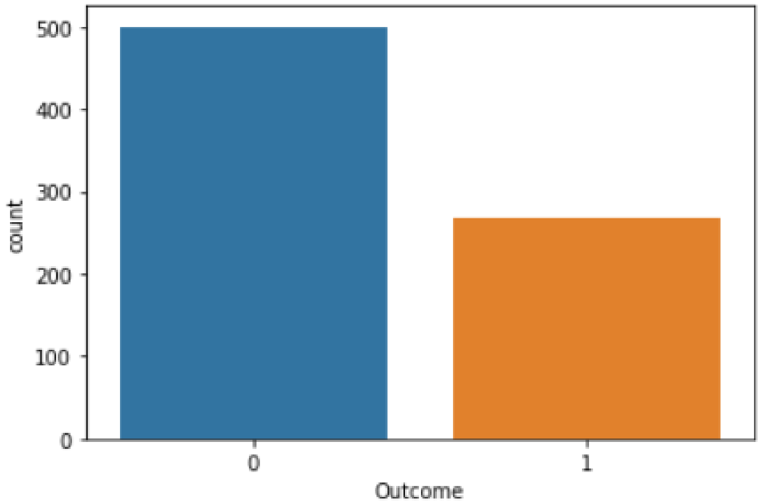


Fig-6: Number of patients with & without diabetes

```
Out[19]: <AxesSubplot:xlabel='Age', ylabel='Density'>
```

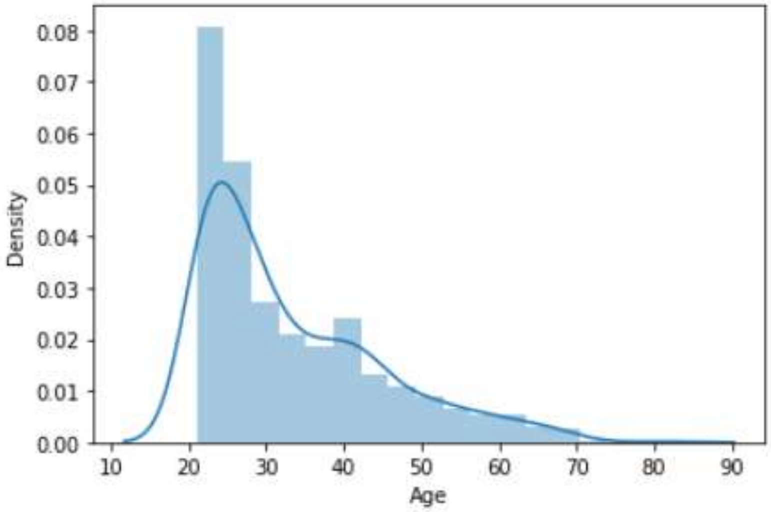


Fig-7: Age distribution of patient having diabetes

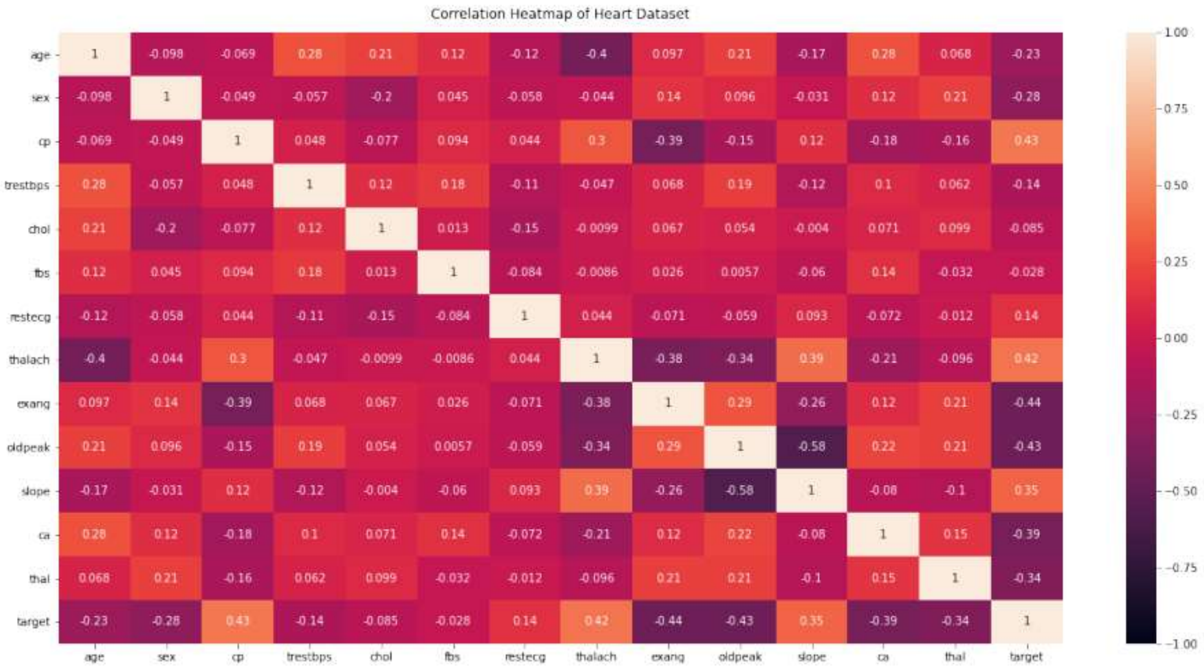


Fig-8: Correlation matrix of heart disease (depicts the influence of the attribute in diagnosing the heart disease). Higher value means more correlation

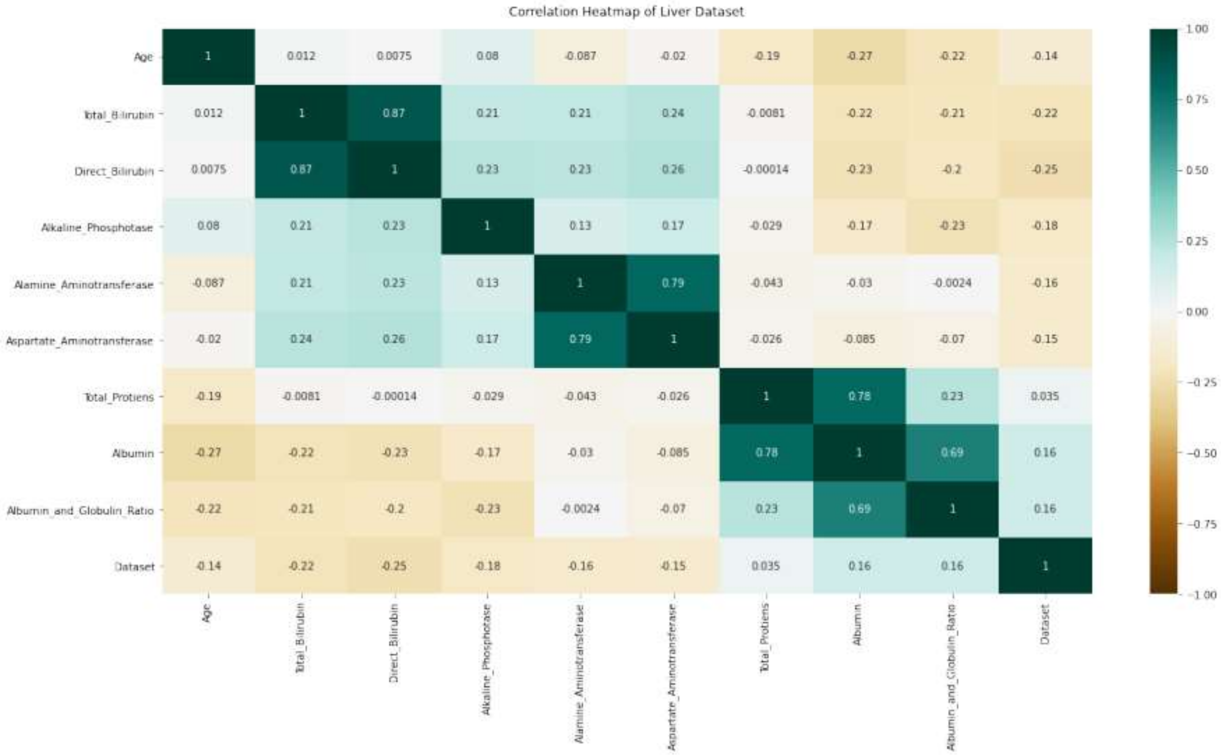


Fig-9: Correlation matrix of liver disease (portrays the impact of the attribute in diagnosing the liver disease)

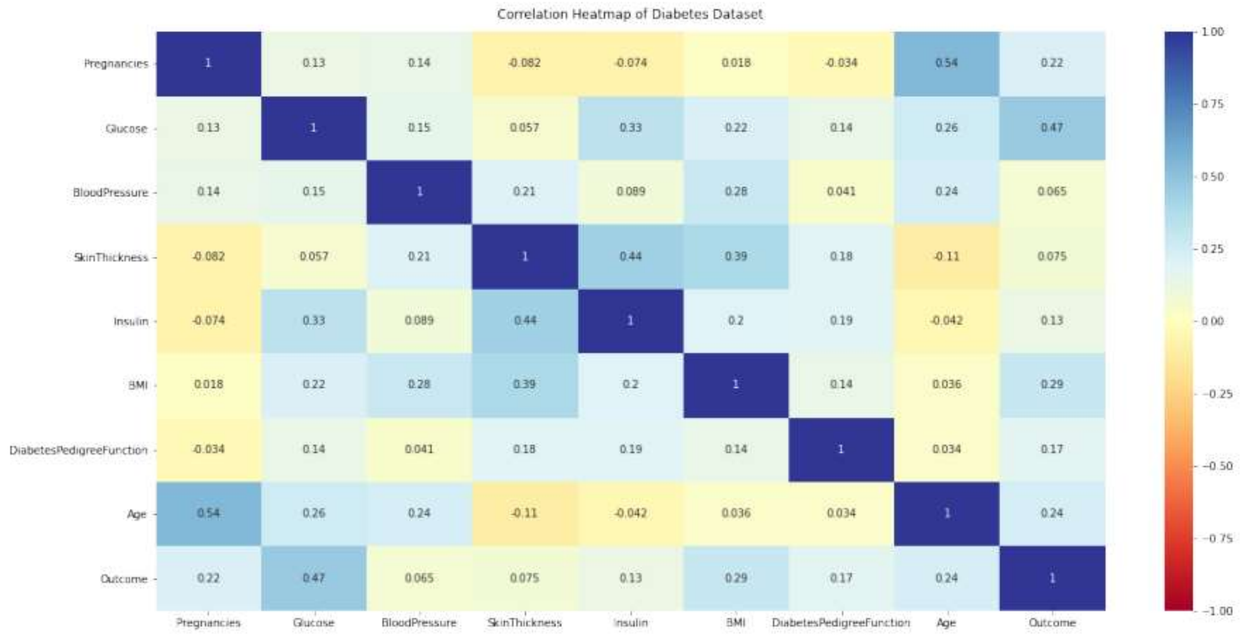


Fig-10: Correlation matrix of diabetes (illustrates the effect of different attribute in diagnosing the diabetes)

4.3 Algorithms Used

4.3.1 Regression

Regression is one of the supervised machine learning techniques that aids in the discovery of different variable correlations and enables us to forecast/predict a continuous output variable with the help of one or more predictor variables. Prediction, forecasting, time series modelling, and recognizing the causal-effect link among the variables are all its' common applications.

It's one of the most used algorithms that may be found in the Linear Regression class. One or more output variables are predicted using a single input variable (the significant one), provided that the input variables are uncorrelated. It's represented as y=b*x + c. where y = dependent variable, x = independent variable, b = slope of the best fit line that provides correct results, and c = intercept. There could be a loss in output except if there is an exact line linking the dependent & independent variables, which is usually calculated as the squared value of the difference between the expected output & actual output, that is, the loss function.

4.3.2 Decision Tree

A Decision Tree is another supervised learning approach which is used to solve both classification & regression tasks, however, it is routinely employed to work out on the classification issues. Internal nodes portray the dataset “attributes”, the branches illustrate the “decision rules”, and each one of the leaf nodes produces the conclusion in the tree-structured classifier format.

The Leaf Node and the Decision Node are the 2 nodes of a Decision tree. The Leaf nodes are the branches/output of those decisions and hence do not accommodate any further branches, while the Decision nodes are handed down to produce additional decisions and have various branches.

The decisions are put together on the basis of the features in the given dataset. It's a graphical presentation for acquiring all optimal solutions to a decision/problem based on certain parameters.

It's named a decision tree as, like a tree, it begins with the root node and further grows into a tree kind of structure with additional branches.

We employ the CART algorithm, which translates to the Classification & Regression Tree algorithm, to design a tree.
A decision tree unambiguously puts forth a question and splits up the tree into subtrees depending on the answer i.e., (Yes/No).

4.3.3 Support Vector Machine

The SVM or the Support Vector Machine is one more Supervised Learning methodology that could be used to answer both regression and classification tasks. However, it is time and again availed in machine learning for Classification problems.
The SVM algorithm's primary purpose is to determine the decision boundary or optimum line to categorize the n-dimensional space into separate classes, thereupon, that the additional data vector/points can be easily placed in the right category in future predictions. A hyperplane is a known word for the optimal choice boundary.
The extreme points/vectors that help to generate the hyperplane are selected via SVM. Support vectors are extreme instances/points, and the algorithm is known as a Support Vector Machine.

4.3.4 Ensemble Learning

Ensemble algorithms integrate various decision trees to generate stronger predicted results, which would then be merged into a single decision tree. The ensemble model's basic premise is that a number of weak learners join up to create an active learner.
Ensemble decision trees are constructed using one of two strategies listed below.

Bagging

Whenever the demand to minimize the variance of a decision tree is desired, we employ the bagging technique. The objective is to extract a few subsets of data from the training sample, which is picked at random and replaced. Now, each subset of data is used to create its decision trees, resulting in an ensemble of different models. It is more powerful than a single decision tree as the average of all the decision trees from multiple trees are combined.

Boosting

Another ensemble approach for creating a group of predictors is boosting. In other words, we fit a series of trees, usually random samples, with the goal of solving net error from the previous trees at each stage.
If a specific input is incorrectly classified by model, its weight is enhanced so that the next hypothesis has a better chance of correctly classifying it. Finally, integrating the entire collection transforms the weak learners into higher-performing models.

4.3.4.1 Random Forest

Bagging is extended to include Random Forest. To forecast a random subset of data, one has to need to do one more step. It also constructs trees using a random selection of traits/subsets rather than all of them. The Random Forest is created when there are several random trees.

4.3.4.2 XGBoost

XGBoost is an extended implementation of gradient boosting. It is a perfect combination of hardware and software optimization methods in order to yield superior results by utilizing less computational resources in the smallest amount of time.

4.3.4.3 ExtraTree Classifier

Extra Trees is similar to Random Forest since it creates numerous trees and splits nodes by utilizing random subsets of features, but there are two significant differences: it does not bootstrap observations (means it samples the observation without replacement) and nodes are partitioned on random splits rather than optimal splits.

4.4 Comparative Analysis of Result

Training and testing the data under different training and testing percentage (e.g.: 70% training & 30% testing i.e., 70:30) using different supervised machine learning models and comparing which machine learning model suits best for different dataset. Different dataset will have different accuracy based on the attributes and the data instances. Earlier experimentation on this was done under fixed training:testing proportions and a conclusion about the most adept machine learning model was drawn. But in this paper, we will see how the efficiency of model changes with change in the training & testing parameter.

Sample	70:30			60:40			80:20		
Data / ML Algo	Diabetes	Liver	Heart	Diabetes	Liver	Heart	Diabetes	Liver	Heart
Regression	32%	10%	48%	32%	10%	54%	32%	12%	49%
Decision Tree	67%	68%	74%	69%	65%	72%	70%	64%	75%
SVM	74%	69%	82%	75%	71%	79%	74%	74%	79%
Random Forest	72%	69%	84%	72%	69%	82%	71%	68%	87%
XGBoost	71%	65%	79%	72%	68%	77%	69%	72%	80%
ExtraTree Classifier	73%	68%	86%	73%	71%	84%	70%	75%	88%

Table 1: Comparative Analysis of ML Algorithm

5. CONCLUSIONS

It is observed that when different machine learning algorithms were applied on a disease dataset under different training and testing ratios, the effective machine learning model changes for the liver dataset while it remained the same for the heart and diabetes dataset. In the previous similar experimentation and papers, either of the datasets was taken into consideration and the machine learning models were applied and the concluding statement made it clear as to which is the perfect model for that dataset. But here in this paper, we have changed the training and testing parameters and concluded how the accuracy of the model changes with the change in training and testing specifications. Additionally, we have also used models like the ExtraTree classifier and XGBoost which were not there in the previous papers. It is discovered that with change in training and testing boundary SVM was best suited for diabetes dataset with around 75% accuracy, ExtraTree Classifier was most adapted for heart classifier with 84-88% accuracy while the accuracy of the liver dataset changed with training testing parameter. In liver dataset for 70:30 (training: testing), accuracy was observed to be 69% with SVM as the best model, for 60:40, accuracy is around 71% with SVM & ExtraTree as the most feasible model and for 80:20, the accuracy was noticed to be 88% with ExtraTree as the best match for the dataset. Therefore, it is necessary to perform extermination under different train-test parameters and then finally conclude as to which is the most befitting model based on the requirement.

6. FUTURE SCOPE

More data from different geographic regions/places can be collected and studied separately & combined with the existing dataset to know if the particular ML algorithm effectively holds the particular disease dataset. Also, more features/attributes could be added to the dataset in order to fetch more accurate results. Apart from this, various Deep learning techniques like Deep Neural Network (DNN), Multi Layer Perceptron (MLP), Back Propagation techniques, and other neural network methods could also be applied. we can also experiment with the hidden layers in neural networks in order to improve the accuracy.

7. REFERENCES

[1]. Faruque, M. F., & Sarker, I. H. (2019, February). Performance analysis of machine learning techniques to predict diabetes mellitus. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1-4). IEEE.

[2]. Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics, 9, 515.

[3]. Repaka, A. N., Ravikanti, S. D., & Franklin, R. G. (2019, April). Design and implementing heart disease prediction using naives Bayesian. In 2019 3rd International conference on trends in electronics and informatics (ICOEI) (pp. 292-297). IEEE.

[4]. Dangare, C. S., & Apte, S. S. (2012). Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10), 44-48.

[5]. Muthuselvan, S., Rajapraksh, S., Somasundaram, K., & Karthik, K. (2018). Classification of liver patient dataset using machine learning algorithms. Int. J. Eng. Technol, 7(3.34), 323.

[6]. Sontakke, S., Lohokare, J., & Dani, R. (2017, February). Diagnosis of liver diseases using machine learning. In 2017 International Conference on Emerging Trends & Innovation in ICT (ICEI) (pp. 129-133). IEEE.

[7]. Durai, V., Ramesh, S., & Kalthireddy, D. (2019). Liver disease prediction using machine learning. Int. J. Adv. Res. Ideas Innov. Technol, 5(2), 1584-1588.

[8]. Dutta, D., Paul, D., & Ghosh, P. (2018, November). Analysing feature importances for diabetes prediction using machine learning. In 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) (pp. 924-928). IEEE.

[9]. Rajdhan, A., Agarwal, A., Sai, M., Ravi, D., & Ghuli, P. (2020). Heart disease prediction using machine learning. International Journal of Research and Technology, 9(04), 659-662.

[10]. Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE access, 7, 81542-81554.

[11]. Mohanty, S., Gantayat, P. K., Dash, S., Mishra, B. P., & Barik, S. C. (2021). Liver Disease Prediction Using Machine Learning Algorithm. In Data Engineering and Intelligent Computing (pp. 589-596). Springer, Singapore.

[12]. Mujumdar, A., & Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. Procedia Computer Science, 165, 292-299.