An Efficient Early Prediction of Hypothyroid Disease with Ensemble Machine Learning Technique

BY

Kartick Nandi
ID: 191-15-2615
AND

Md. Ashraful Alam Ashik
ID: 191-15-2425

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering.

Supervised By

Tania Khatun
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Mohammad Jahangir Alam
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2023

APPROVAL

This Project/internship titled **"An Efficient Early Prediction of Hypothyroid Disease with Ensemble Machine Learning Technique"**, submitted by Kartick Nandi, ID No : 191-15-2615 and Md. Ashraful Alam ashik, ID No: 191-15-2425 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfilment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on *30/01/2023*.

BOARD OF EXAMINERS

Chairman

4/02/23

Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Mohammad Shamsul Arefin
Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Ms. Sharmin Akter
Lecturer (Senior Scale)
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Dr. Mohammad Shorif Uddin
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Tania Khatun (Assistant Professor), Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Tania Khatun

Tania Khatun
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Jahangir 31.01.23

Mohammad Jahangir Alam
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

Submitted by:

Kartiek

Kartick Nandi
ID: -191-15-2615
Department of CSE
Daffodil International University

Ashik

Md. Ashraful Alam Ashik
ID: -191-15-2425
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to Almighty God for His divine blessing in making us possible to complete the final year project/internship successfully.

We are grateful and wish our profound indebtedness to **Tania Khatun (Associate Professor)**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Machine Learning*" to carry out this project. His endless patience, scholarly guidance continual encouragement, constant and energetic supervision, constructive criticism valuable advice r and reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to **Professor Dr. Touhid Bhuiyan**, Head of**,** the Department of CSE, for his kind help to finish our project and also to other faculty members and the staff of the CSE department of Daffodil International University.

We would like to thank our entire course mate at Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

Abstract

Physical illnesses, such as hypothyroidism, have recently become more prevalent. The topic is well-known in today's society. Hypothyroidism is a problem for the majority of people. The differences between the ratios of the normal and affected diagnostic reports act as a barometer for the illness. The condition of hypothyroid disease has already been the focus of various studies. We have found a few fantastic opportunities to further the technique. We advocate using effective algorithmic models to identify threats and spread early awareness. Our proposed method is uncomplicated to implement in the real world and suitable for simple hypothyroid illness forecasts. The dataset was housed on the Kaggle website. Different classifiers, including the Random Forest (RF), Logistic Regression (LR), Gradient Boosting (GB), K-Nearest Classifier (KNN), Adaboost Classifier (ABC), and Decision Tree (DT) methods, have been implemented in our model. Before feature selection, Random Forest (RF) given an accuracy of 97.09%, Logistic Regression (LR) given accuracy of 94.7%, Gradient Boosting (GB) given accuracy of 94.04%, Decision Tree given accuracy of 97.62%, Adaboost Classifier (ABC) given the accuracy 97.62%, K-Nearest Classifier (KNN) given the accuracy 95.1%. We have used ensemble techniques to get the best accuracy. Our voting classifier RDAGLK gave the best accuracy of 98.01%. After feature selection, Random Forest (RF) given an accuracy of 98.26%, Logistic Regression (LR) given accuracy of 95%, Gradient Boosting (GB) given accuracy of 94.7%, Decision Tree given accuracy of 97.27%, Adaboost Classifier (ABC) given the accuracy 94.55%, K-Nearest Classifier (KNN) given the accuracy 95%. We have used ensemble techniques to get the best accuracy. Our voting classifier RDAGLK gave the best accuracy of 99.09%.

Keywords: Hypothyroid disease, Prediction, Machine Learning, Algorithms, Ensemble Model.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Introduction

The toughest part of our everyday lives is living with hypothyroidism, which refers to issues with creating adequate hormones, including failure and performance deterioration. This patient is appearing more frequently lately. The main problem, though, is identifying or locating the damage caused just at the time of diagnosis. Machine learning could be the most useful component of a significant part in determining the presence of hypothyroid illness from responsive health information by looking at numerous factors and patient diagnostic records. We have examined the patient's diagnostic records as part of our research and have identified many significant disease signs. The information was used to identify hypothyroid conditions in human bodies and diagnose them. Other academics have collaborated to create machine learning algorithms to identify the illness in the body. Their approach and accuracy, meanwhile, were neither suitable nor smooth for forecasting hypothyroidism. We offer our approach as a way to improve the human body's ability to forecast illness. There are two distinct categories of machine learning methods. While the other is unsupervised, one of them is. Based on examples of input-output pairs, supervised learning generates outputs from inputs using labeled data. The working data is the training data of the dataset. Building models with unlabeled data allows for the exploitation of previously unrecognized patterns and information.

1.2 Motivation

Other academics have collaborated to create machine learning algorithms to identify the illness in the body. Their approach and accuracy, meanwhile, were neither suitable nor smooth for forecasting hypothyroidism. We offer our approach as a way to improve the human body's ability to forecast illness. There are two distinct categories of machine learning methods. While the other is unsupervised, one of them is. Based on examples of input-output pairs, supervised learning generates outputs from inputs using labeled data.

The working data is the training data of the dataset. Building models with unlabeled data allows for the exploitation of previously unrecognized patterns and information. In order to predict the occurrence of hypothyroid illness in suspect or continuing patients, we have devised a procedure.

1.3 The rationale of the study

We developed a model in our research to predict hypothyroid illness in humans. We have recently seen that this illness is starting to negatively impact our society. But we also saw that there is a lack of information and diagnostic equipment. In our impoverished country, analyzing a patient's symptoms and diagnosing hypothyroidism are costly procedures. We are using machine learning in our study to try to solve the problem.

1.4 Research Questions

How well do the algorithms in this proposed model work?

What is the likelihood that a person will succeed whether they have hypothyroidism or not?

How may one forecast the early detection of hypothyroidism?

What benefits does our proposed model offer?

What real-world scenarios may this work be used to?

What is the anticipated course of the project?

Which safety precautions are necessary for this work?

How can we evaluate our model for predicting hypothyroidism?

How difficult is this task?

What credentials are required for this position?

1.5 Expected outcome

People are growing increasingly likely to have hypothyroidism. Nobody is also clear if she is affected or not. By examining the diagnosis report, we are advising the best course of action for predicting or recognizing the problem. Our method can find people with hypothyroidism, improve decision-making, and properly assess the outcome. It might measure life satisfaction and examine related problems. It may increase public knowledge of the hypothyroid condition. In the shortest period of time, the provided model can evaluate the sickness.

1.6 Project Management and Finance

Our proposed design is practical and cost-effective in daily living. Analyzing hypothyroidism may be a huge benefit for our nation. Common tools are needed in order to use the prediction process in real life. The use of high-conFigureuration tools will produce the best results and the smoothest functioning of our model. However, if we use basic tools, it is still feasible.

1.7 Report layout

Chapter 2 discusses the pertinent study conducted by the preceding scholars. We must look at the introduction and motive before we start the inquiry. Therefore, we discuss the introduction, which may thoroughly describe the proposed technique, and the motivational section, which may fully explicate the prognosis. We focused on pertinent research and acquired internal data for our work after concluding the Introduction part. We have selected machine learning algorithms, used them on our dataset, and then identified the best one in our methods section. We checked the data after the pre-processing stage, and eventually, we got the desired result—what we'll call the comparison result. That was covered in the conclusion, which is the last piece of our essay.

Chapter 2

BACKGROUND

2.1 Preliminaries

Machine learning methods are used to pinpoint the specific architecture of hypothyroid disease. The investigations related to the evaluative inspection of the patient's diagnosis report are what we aim to look at in this area. The techniques used by these models include Random Forest (RF), Logistic Regression (LR), Gradient Boosting (GB), K-Nearest Classifier (KNN), Adaboost Classifier (ABC), and Decision Tree (DT). In this part, the exploration is performed using machine learning models. In their research, a number of researchers utilized a variety of models; these researchers are described in the segment.

2.2 Related work

A few machine learning classifiers that we have used to categorize hypothyroid disease are suitable for the task we are proposing. Machine learning algorithms built on decision tree models are referred to as "tree structures" and are used to execute decision models [1] [2]. Researcher H. M. Almahshi, E. A. Almashri, H. Alquaran have proposed a machine learning model with DT 97.6%, NB with 96.7%, SVM with 75.1%, Ensemble with 97.3% accuracy [3]. Researchers Maheshwari, Afzal have proposed a machine learning model with RF 98.50%, SVM with 97.02% and KNN with 95.81% accuracy [4]. Researcher AKGUL, Goksu and others have proposed a model of machine learning with KNN 92%, SVM with 97.8% accuracy [5]. Chandel and Khushboo have proposed a model of machine learning with KNN 93.44% and NB with 22.56% accuracy [6].

2.3 Comparative Analysis and Summary

These days, a machine learning model is one that is often employed. We had to do a difficult task in order to find our separate jobs. All related works have subpar accuracy and model outcomes. We have to employ a number of machine learning models in order to determine the dataset's highest prediction accuracy. The models had to be run on specialized hardware, which was a challenge. We applied a number of separate algorithms to get the categorization rates. Complicated models could cause long runtimes by adding expensive GPUs.

2.4 Scope of the Problem

The challenge was to familiarize and simplify the diagnosis of hypothyroidism. Due to the large number of machine learning-related studies that are connected to our proposed model, we made an effort to offer the highest accuracy possible. Although there was minimal space for improvement in the procedure, the idea may be put into effect by applying straightforward technology to decrease the frequency of diagnosis for hypothyroidism.

2.5 Challenges

The dataset was kept on the Kaggle website. The material was very easy to utilize and highly practical. When the data collection is finished, we must manually check the dataset for any missing information. No one has ever been as exact as we are with this dataset.

Chapter 3

RESEARCH METHODOLOGY

3.1 Research Subject and Instrument

We employed a number of algorithms and hybrid models to get the best accuracy out of the dataset. Some tools, such as effective conFigureuration tools with the finest GPUs, were necessary. Python programming language and related tools, such as Jupiter Notebook, Google Collaboratory, and Anaconda, have been used. It enables the creation and execution of any Python code through the browser. On a machine running Windows 10 Pro 64-bit, an AMD Ryzen 5 3600 6-core CPU clocked at 3.59 GHz, and 8 GB of RAM, all tests were done.

3.2 Data Collection Procedure

When the dataset was obtained from Kaggle [8], it was essentially ready for use. There are 3772 rows and 30 columns total. The diagnostic column includes a category for the prevalence of hypothyroidism. Each trait was essential for identifying hypothyroidism. Patients are split into two groups based on conditions 0 and 1. These two situations are thought to happen often. There were 3481 persons at that stage, while the other 291 people did not have hypothyroidism. The ratio is shown in Figureure 3.1 below. There are two portions to the dataset. They go through testing and instruction. For the training section, we have selected 80% of the applicants, and for the exam portion, 20%.

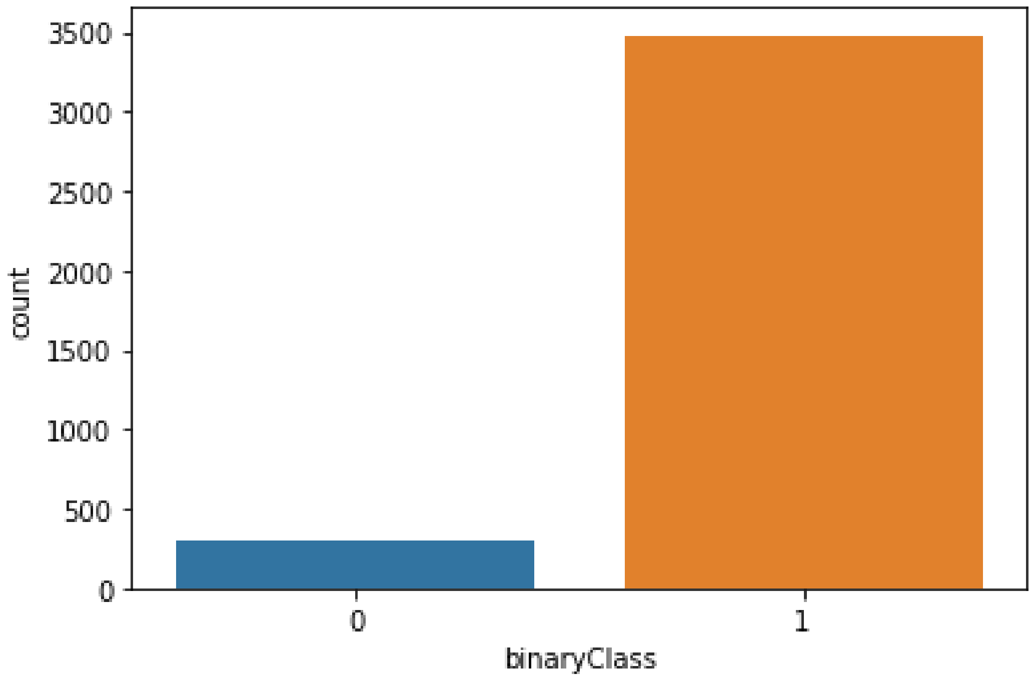


Figure 3.1: Number of target values

The dataset, which solely contains nominal values, included no missing or incorrect values. A detailed explanation of the dataset and its range is provided in Table 3.1.

TABLE 3.1: DETAILS OF THE DATASET

Attributes	**Description**	**Value Range**	**Types of Values**
Age	Age	1 to 83	Integer
Sex	Gender	M and F	Char
On thyroxine	On thyroxine	True and False	Boolean
Query on thyroxine	Query on thyroxine	True and False	Boolean
On antithyroid medication	On antithyroid medication	True and False	Boolean

sick	sick	True and False	Boolean
pregnant	Pregnancy report	True and False	Boolean
Thyroid surgery	Thyroid surgery	True and False	Boolean
I131 treatment	I131 treatment	True and False	Boolean
Query hypothyroid	Query hypothyroid	True and False	Boolean
Query hyperthyroid	Query hyperthyroid	True and False	Boolean
lithium	lithium	True and False	Boolean
goitre	goitre	True and False	Boolean
tumor	tumor	True and False	Boolean
hypopituitary	hypopituitary	True and False	Boolean
psych	psych	True and False	Boolean
TSH measured	TSH measured	True and False	Boolean
TSH	TSH	0.03 to 45	float
T3 measured	T3 measured	True and False	Boolean
T3	T3	0.6 to 5.5	float
TT4 measured	T3 measured	True and False	Boolean

TT4	T3	11 to 175	Integer
T4U measured	T4U measured	True and False	Boolean
T4U	T4U	0.68 to 1.82	float
FTI measured	FTI measured	True and False	Boolean
FTI	FTI	51 to 312	Integer
TBG measured	TBG measured	True and False	Boolean
Binary Class	Target column	P and N	Char

3.2.1 Categorical Data Encoding

The categorical data encoding technique is the procedure for transforming category data into a numerical value. Since machine learning only takes and produces numeric input, the categorical encoding technique was essential to our research. We required the columns to alter in order to employ the categorical data encryption technique.

3.2.2 Missing Value Imputation

It entails substituting imputed values for missing or incomplete data that have been identified via analysis of data from other datasets. But it is encouraging that our dataset had no missing values.

3.2.3 Handling Imbalanced Data

It involves altering the class distribution of a dataset. By methodically including new instances in the collection, it controls the data. Minority data is boosted while utilizing the complete dataset as input.

3.2.4 Feature Scaling

It is a method for adjusting the many independent data variables. There is data with negative values. Modifications to the scale have been made. Two columns, TBG and referral source, have to be removed.

3.3 Statistical Analysis

An analysis section is necessary for all research projects of any kind. The development and evaluation of the algorithms I've used are necessary for this part. We have chosen to utilize a comma-separated value (CSV) file, thus we need to do a few steps to prepare the dataset for usage. We have conducted a variety of actions, such as pre-processing and data collection. We have implemented some different classifiers named Random Forest (RF), Logistic Regression (LR), Gradient Boosting (GB), K-Nearest Classifier (KNN), Adaboost Classifier (ABC), and Decision Tree (DT) algorithms. Before feature selection, Random Forest (RF) given an accuracy of 97.09%, Logistic Regression (LR) given accuracy of 94.7%, Gradient Boosting (GB) given accuracy of 94.04%, Decision Tree given accuracy of 97.62%, Adaboost Classifier (ABC) given the accuracy 97.62%, K-Nearest Classifier (KNN) given the accuracy 95.1%. We have used ensemble techniques to get the best accuracy. Our voting classifier RDAGLK gave the best accuracy of 98.01%. After feature selection, Random Forest (RF) given an accuracy of 98.26%, Logistic Regression (LR) given accuracy of 95%, Gradient Boosting (GB) given accuracy of 94.7%, Decision Tree given accuracy of 97.27%, Adaboost Classifier (ABC) given the accuracy 94.55%, K-Nearest Classifier (KNN) given the accuracy 95%. We have used ensemble techniques to get the best accuracy. Our voting classifier RDAGLK gave the best accuracy of 99.09%.

3.4 Proposed Methodology

Flow chart:

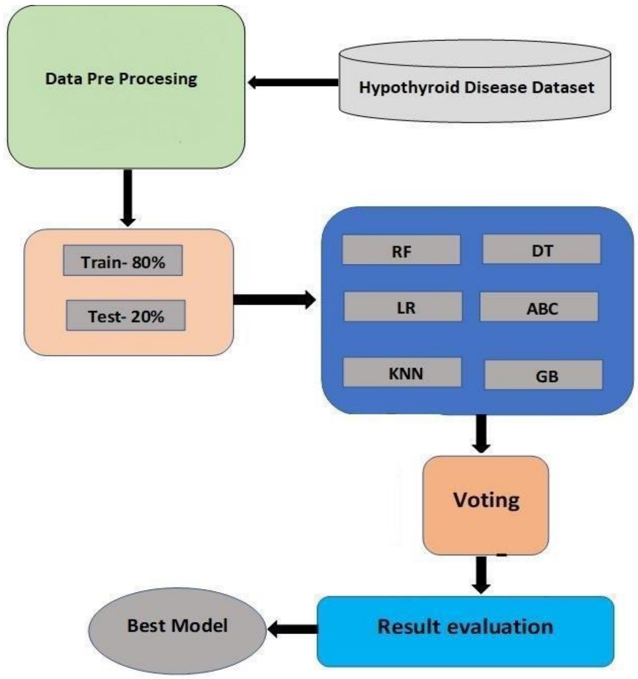


Figure 3.2: Methodology of Hypothyroid

Using a process diagram, we have predicted hypothyroidism in this part. The training and testing dataset for the system was originally presented. Then, methods for data pre-processing such Standard Scaler Transform were applied. translation of categorical data to numerical data. We used 80% of the resources for training and 20% for testing. Then, we put algorithms into practice and evaluated the results. Then, we used ensemble algorithms to get the maximum forecast accuracy. Voting is the name of a class of algorithms. The results of the used ensemble algorithms were then evaluated. The models that had been used were then evaluated using outcome analysis. The preferred model method is shown in Figure. 3.2.

A correlation subplot shows the underlying relationships between two variables or how one variable changes as a result of a change in another. The greater the dependency between variables, the more likely it is that one variable may be successfully predicted from another. It suggests a greater understanding of the dataset and facilitates our capacity to pinpoint the important variables [9]. In Figure. 3.3, all of the traits related to the predicted property "Hypothyroid Disease" were shown.

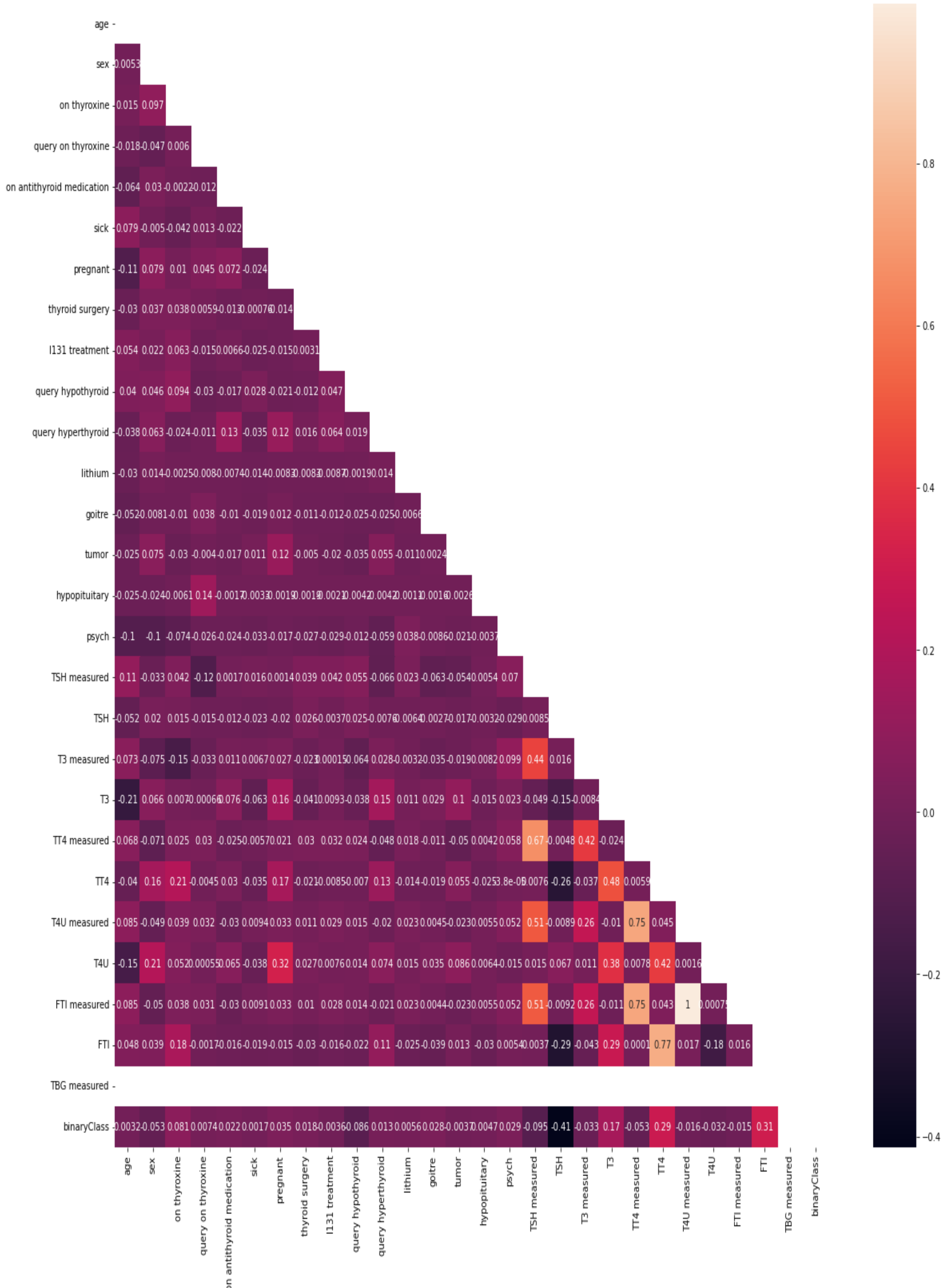


Figure 3.3: Correlated Features

3.5 Implementation Requirements

In order to test or train our proposed model, we need data sources. We must clean the dataset in order for things to run smoothly. The dataset will be cleaned using a variety of filtering methods. Then, methods for data pre-processing such as Standard Scaler Transform were applied. translation of categorical data to numerical data. We used 80% of the resources for training and 20% for testing. Then, we put algorithms into practice and evaluated the results. Then, we used ensemble algorithms to get the maximum forecast accuracy. One of the ensemble algorithms is voting. The results of the used ensemble algorithms were then evaluated. The result was then verified through hyperparameter tuning. The models that had been used were then evaluated using outcome analysis. The data analysis stage must next be completed in order to start the learning process. We must next use model learning and fit the predictions technique. Then, the models must be voted on to see which has the most accuracy. Then, depending on the model's accuracy, precision, recall, and F-1 score, the best one may be selected for use.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Experimental Setup

In this study, supervised learning—which runs on the principles of training and testing—was used. The training dataset is used to create the classification model. The produced model is used with the testing dataset to produce the result. The next parts will briefly demonstrate the machine-learning algorithm.

4.1.1 Classifier Algorithms

Random Forest (RF), Logistic Regression (LR), Gradient Boosting (GB), K-Nearest Classifier (KNN), Adaboost Classifier (ABC), and Decision Tree (DT) methods are some of the classifiers we've created.

Logistic Regression

Logistic regression (LR), a machine learning-based classifier technique, uses a binary (0/1) scale with two options for the class label: yes or no. Logistic regression is suited for discrete variables even if it allows for the combined value of continuous variables and discrete predictors [10] [11]. The concept is shown in Figure. 4.1 below. The method of supervised machine learning is used in logistic regression. The basic equation is shown below [12].

$$h\Theta(x) = \frac{1}{1} + e - (\beta o + \beta 1X) \text{ ……….(1)}$$

'hΘ(x)' is the output of the logistic function, where $0 \leq h\Theta(x) \geq 1$

'β1' is the slope

'βo' is the y-intercept

'X' is the independent variable

(βo + β1X) – derived from the equation of a line Y (predicted) = (βo + β1X) + Error.

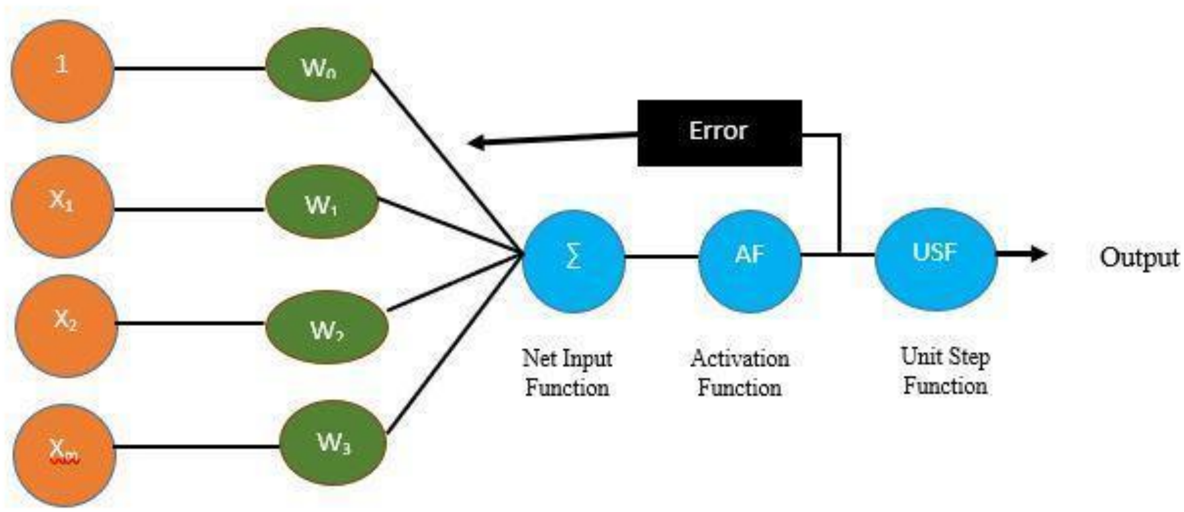


Figure 4.1: Logistic Regression

Random Forest

Random Forest (RF), a Machine Learning (ML) based classifier ensemble technique, is made up of many Decision Tree algorithms [13]. While the algorithm is being trained, RF constructs many decision trees in order to give an optimal decision model with more accuracy than the single decision tree model. This idea is seen in Figure. 4.2 below. It can, however, be used to large datasets. The Random Forest technique is used to obtain the average of all decision tree approaches [14] [15] [19] [20]. The average of two decision tree algorithms was determined using the Random Forest technique.

$$j = \frac{1}{B} + \sum_{b=1}^{B} fb(X') \ldots\ldots\ldots\ldots(2)$$

Concerning X = {x1,x2,x3,.................. xn} with respect to Y = {y1,y2,y3,.................. yn} with the lower to upper limit is 1 to B.

Sample x′ = mean of the sum of the prediction $\sum^{B}{}_{b=1} fb(X')$ for every summation.

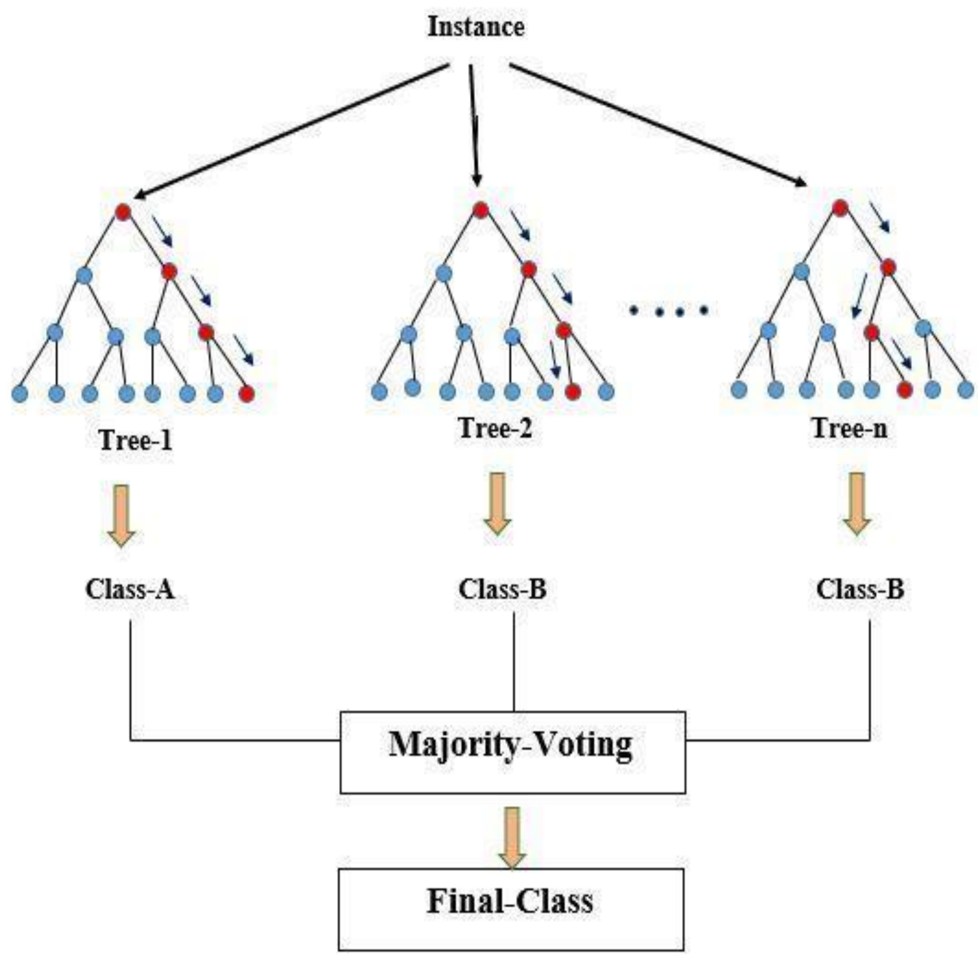


Figure 4.2: Random Forest

Gradient Boosting

The Gradient Boosting (GB) technique of boosting, which is based on Machine Learning, is primarily composed of the loss function (ML). The idea is shown in the following Figure. 4.3. It reduces a model's loss function by merging and optimizing weak learners. Overfitting is reduced to enhance algorithm performance [25] [26] [27][29].

Here fi(x) = loss function with correlated negative gradients (−ρi x gm(X)), m = number of iterations.

Feature increment (i) = 1,2,3, ... , m. Therefore, the optimal function F (X) after m−th iteration is shown below [16].

$$F\ (X) = \sum_{i=0}^{m} \quad fi(x) \ldots\ldots\ldots (3)$$

Here, gm = the path of loss function's fast decreasing F(X) = Fn − 1(X) the decision tree's target is to solve the mistakes by previous learners [17][18].

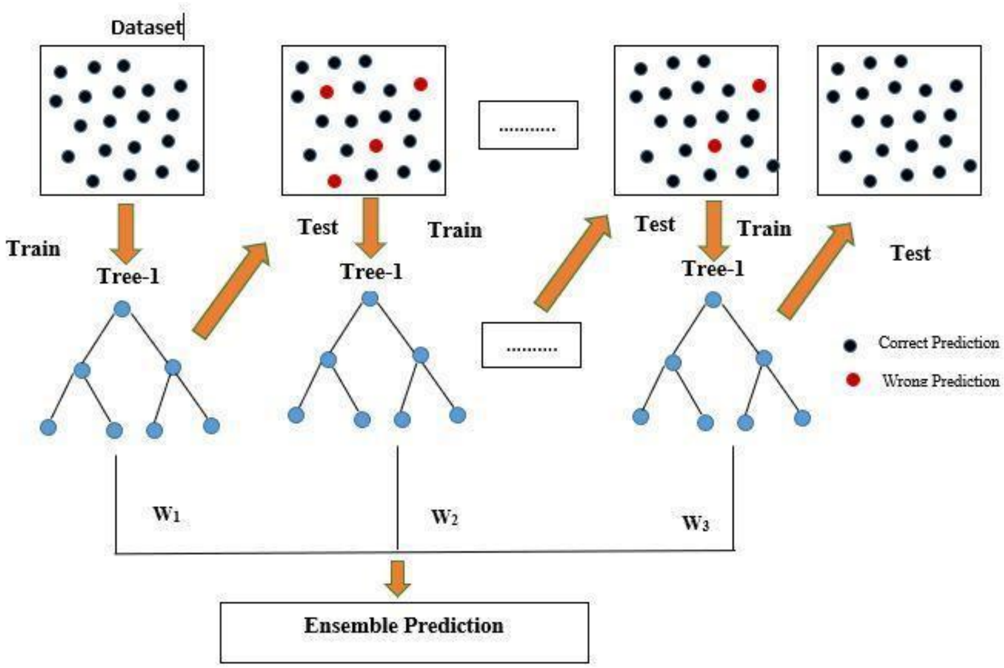


Figure 4.3: Gradient Boosting

Adaboost Classifier

AdaBoost is a boosting classifier that combines many weak classifiers to produce a strong classifier. Using 1000 samples, ABC predicts TA. ABC fixes the weights that vary across classifiers and samples. This makes it difficult for classifiers to focus on the result [28]. The final formula to achieve TA is,

$$H_k(P) = l - \left(\sum_{k=1}^{k} \quad a_k h_k(P) \right) \ldots\ldots (4)$$

Here, N=frequency of training data, k = total number of weak classifiers combined to use, hk = output of weak classifier (lowerlimit 1 to upper limit k), ak = weight of classifier. ABC combines sample trainers, fixes the weights of samples and classifiers to geta more accurate and efficient TA. The notion is depicted in Figure. 4.4 below.

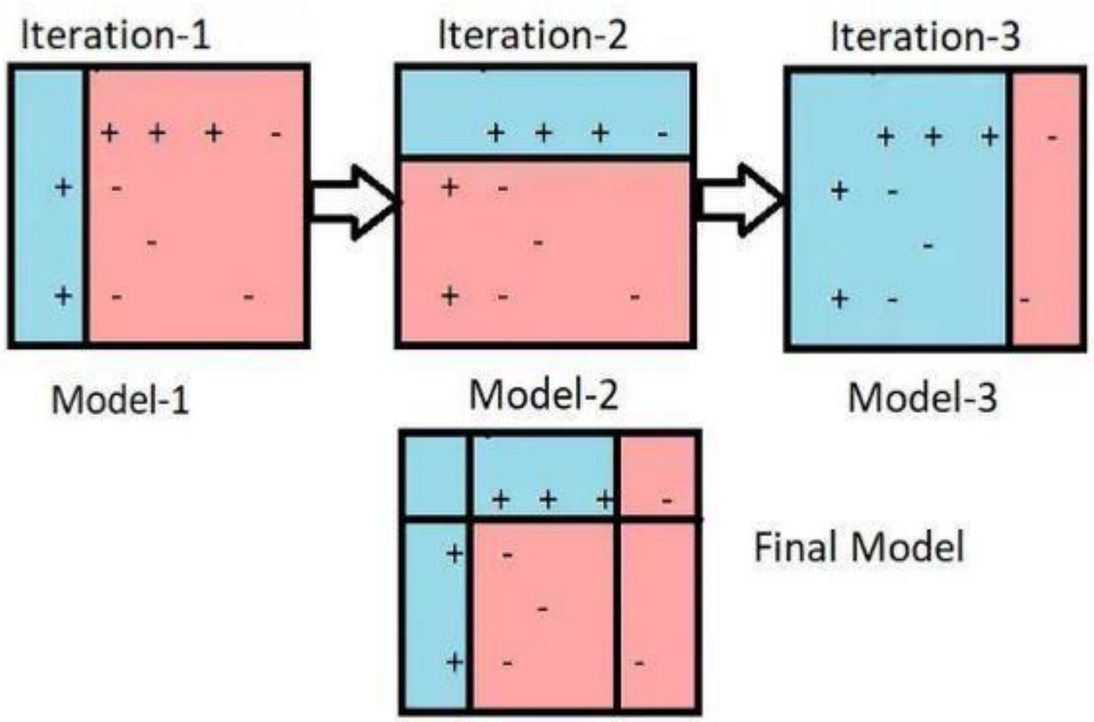


Figure 4.4: Adaboost Classifier

4.1.2 Ensemble Methods of Machine Learning

The phrase "ensemble method" describes the application of a number of classifiers to transform weak classifiers into strong classifiers by achieving the highest accuracy and efficiency. Because it reduces variances, combines predictions from several models, and reduces prediction spread, it was employed in our inquiry owing to variable handling, bias, and uncertainty [7] [21] [22]. A single ensemble technique was used in our research. Ensemble modeling using voting and boosting was used.

Voting

Voting classifiers are a collection of classifiers that predict which class will receive the most number of votes. It suggests that by combining the vote results, the model develops several models to predict outcomes. The idea is shown in the following Figure. 4.5. The calculation we used is displayed below [24]..

Here, w_j = weight that can be assigned to the j^{th} classifier.

$$y' = argmax \sum_{j=1}^{m} w_j p_{ij} \text{ ………(5)}$$

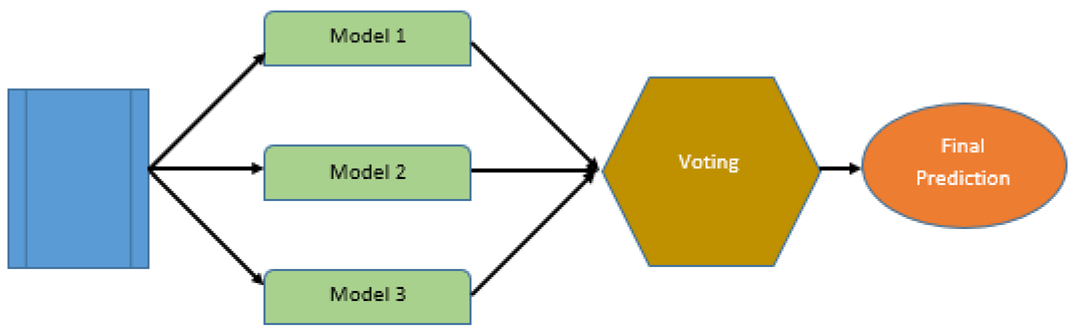
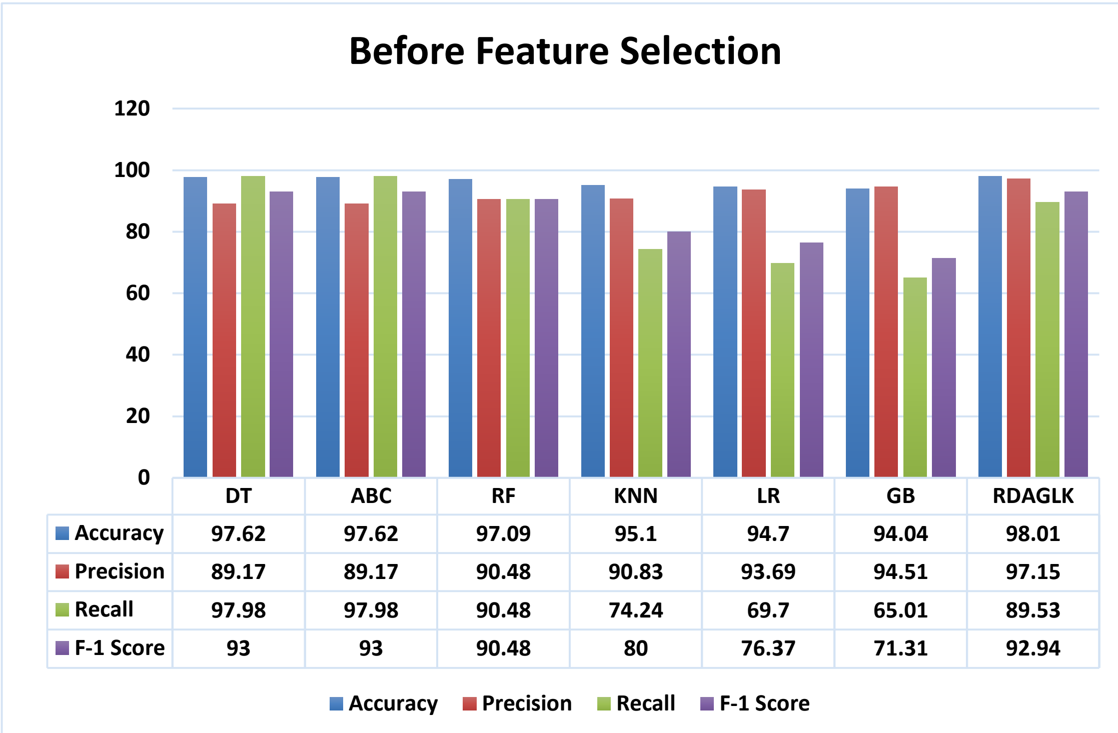


Figure 4.5: Voting

4.2 Experimental Result & Analysis

We now had to evaluate how well the present models worked. We may employ a number of performance evaluation metrics and techniques to confirm the efficient functioning of our proposed model. These methods use fictitious data to calculate the overall performance. The analysis report from our machine-learning experiments on the hypothyroid disease target dataset must be presented in this part. We first use our selected dataset in practice. Any missing or incorrect numbers have been removed from our dataset by filtering. We tested a range of algorithms and judged how well they performed. We evaluated Confusion matrices' Accuracy, Precision, Recall, and F-1 Score for our proposed methods. For traditional approaches, these confusion matrices have been measured. Decision Tree (DT), K-Nearest Classifier (KNN), Logistic Regression (LR), Gradient Boosting (GB), Adaboost

Classifier (ABC), and Random Forest (RF) techniques have all been explored. We have seen numerous ensemble techniques in action with confusion matrices. We evaluated ensemble voting techniques. We have calculated the results before and after feature selection.



	DT	ABC	RF	KNN	LR	GB	RDAGLK
Accuracy	97.62	97.62	97.09	95.1	94.7	94.04	98.01
Precision	89.17	89.17	90.48	90.83	93.69	94.51	97.15
Recall	97.98	97.98	90.48	74.24	69.7	65.01	89.53
F-1 Score	93	93	90.48	80	76.37	71.31	92.94

Figure 4.6: Experimental Results of Classifiers before feature selection

Firstly, we considered the performances of algorithmic classifiers, the best accuracy had obtained before feature selection, Random Forest (RF) given an accuracy of 97.09%, Logistic Regression (LR) given accuracy of 94.7%, Gradient Boosting (GB) given accuracy of 94.04%, Decision Tree given accuracy of 97.62%, Adaboost Classifier (ABC) given the accuracy 97.62%, K-Nearest Classifier (KNN) given the accuracy 95.1%. The output is shown in Figure 4.6. The precision score was Decision Tree (DT), K-Nearest Classifier (KNN), Logistic Regression (LR), Gradient Boosting (GB), Adaboost Classifier (ABC), and Random Forest (RF) achieved respectively 89.17%, 90.83%, 93.69%, 94.51%, 89.17% and 90.48%. The Recall score was Decision Tree (DT), K-Nearest Classifier (KNN), Logistic Regression (LR), Gradient Boosting (GB), Adaboost Classifier (ABC), and Random Forest (RF) achieved respectively 97.98%, 74.24%, 69.7%, 65.01%, 97.98%

and 90.48%. The F-1 score was Decision Tree (DT), K-Nearest Classifier (KNN), Logistic Regression (LR), Gradient Boosting (GB), Adaboost Classifier (ABC), and Random Forest (RF) achieved respectively 93%, 80%, 76.37%, 71.31%, 93% and 90.48%. Ensemble classifier RDAGLK gives the accuracy 98.01%, precision 97.15%, recall 89.53%, F-1 Score 92.94%.

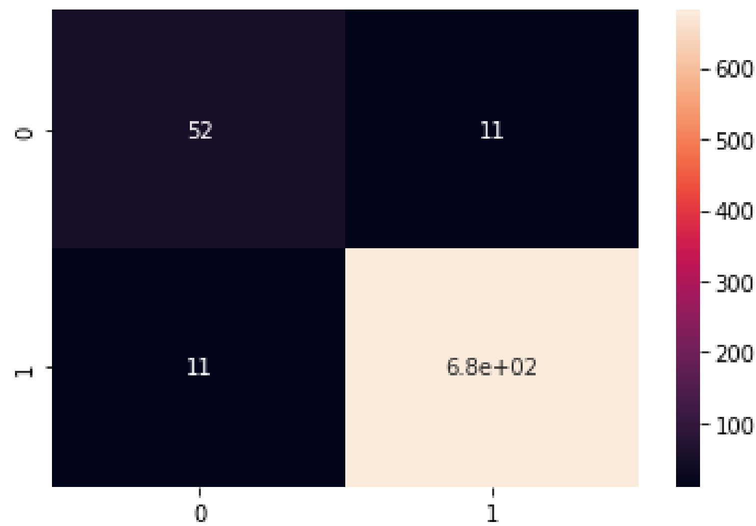


Figure 4.7: Random Forest Confusion Matrix

For Random Forest (RF) the True positive value was 52, False Positive value was 11, False Negative value was 11 and True Negative value was 9. The output is shown in Figure 4.7.

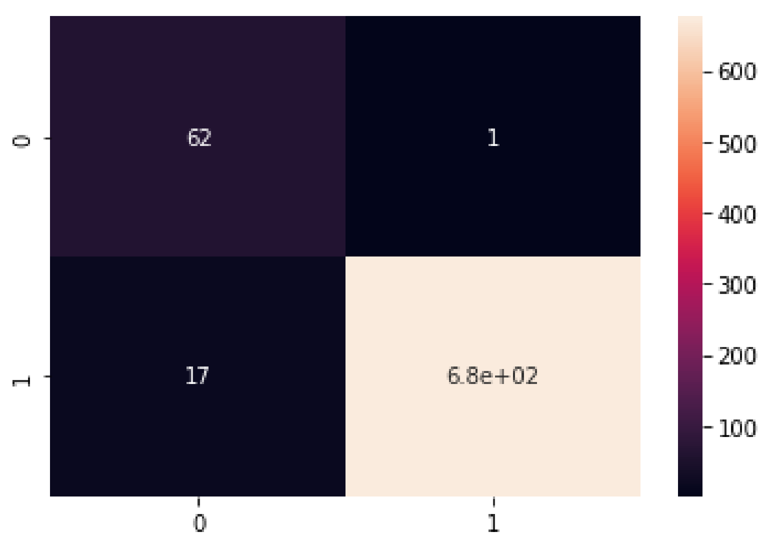


Figure 4.8: Decision Tree Confusion Matrix

For Decision Tree (DT) the True positive value was 62, False Positive value was 1, False Negative value was 17 and True Negative value was 9. The output is shown in Figure 4.8.

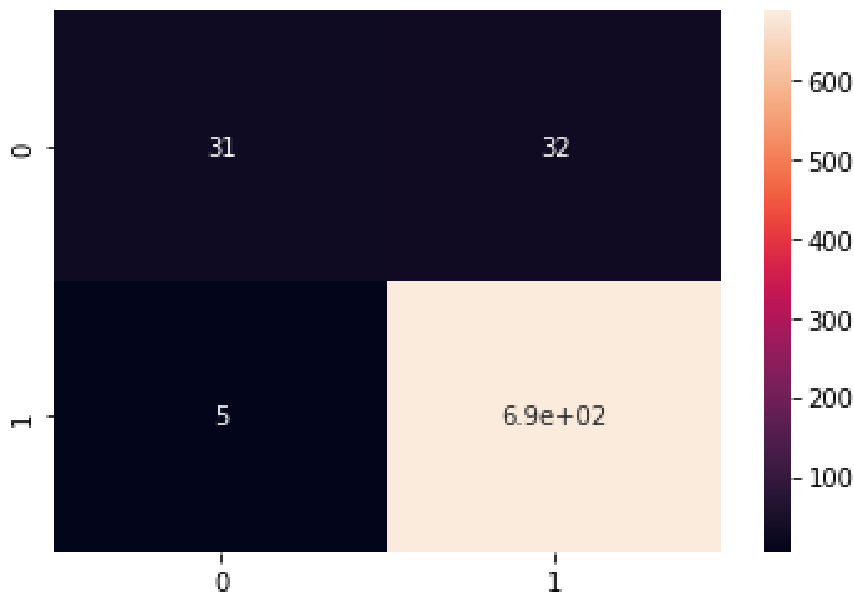


Figure 4.9: K-Nearest Confusion Matrix

For K-Nearest the True positive value was 31, False Positive value was 32, False Negative value was 5 and True Negative value was 9. The output is shown in Figure 4.9.

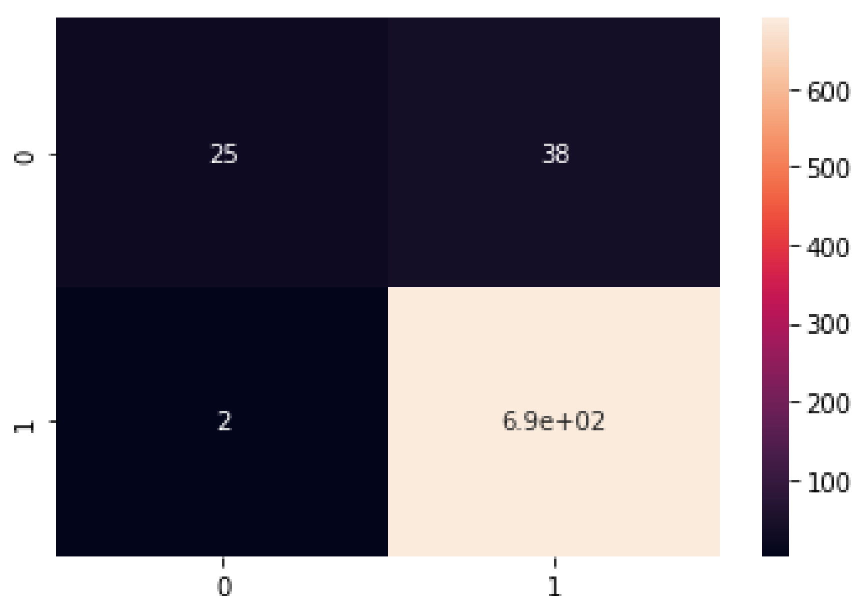


Figure 4.10: Logistic Regression Confusion Matrix

For Logistic Regression (LR) the True positive value was 25, False Positive value was 38, False Negative value was 2 and True Negative value was 9. The output is shown in Figure 4.10.

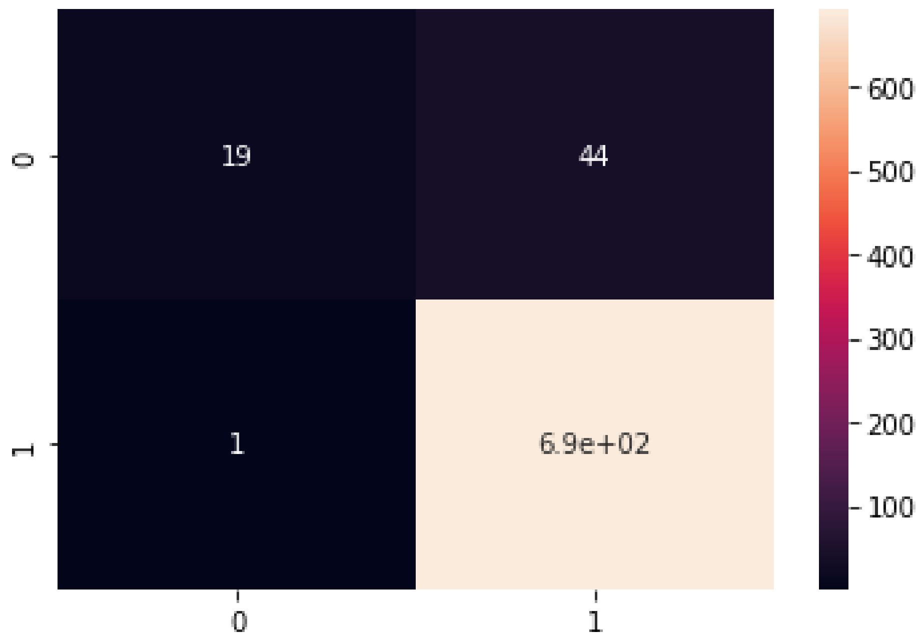


Figure 4.11: Gradient Boosting Confusion Matrix

For Gradient Boosting (GB) the True positive value was 19, False Positive value was 44, False Negative value was 1 and True Negative value was 9. The output is shown in Figure 4.11.

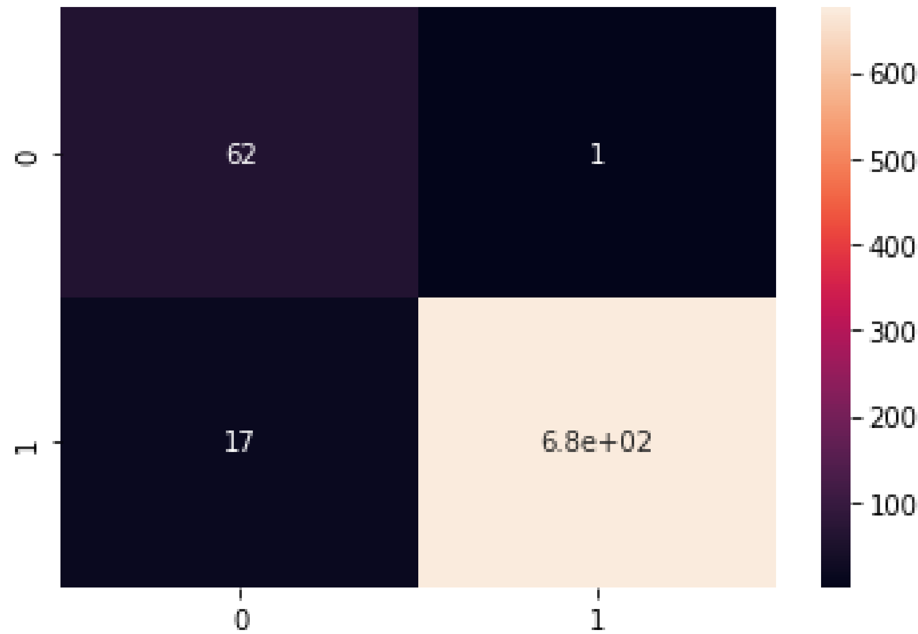


Figure 4.12: Adaboost Classifier Confusion Matrix

For Adaboost Classifier (ABC) the True positive value was 62, False Positive value was 1, False Negative value was 17 and True Negative value was 9. The output is shown in Figure 4.12.

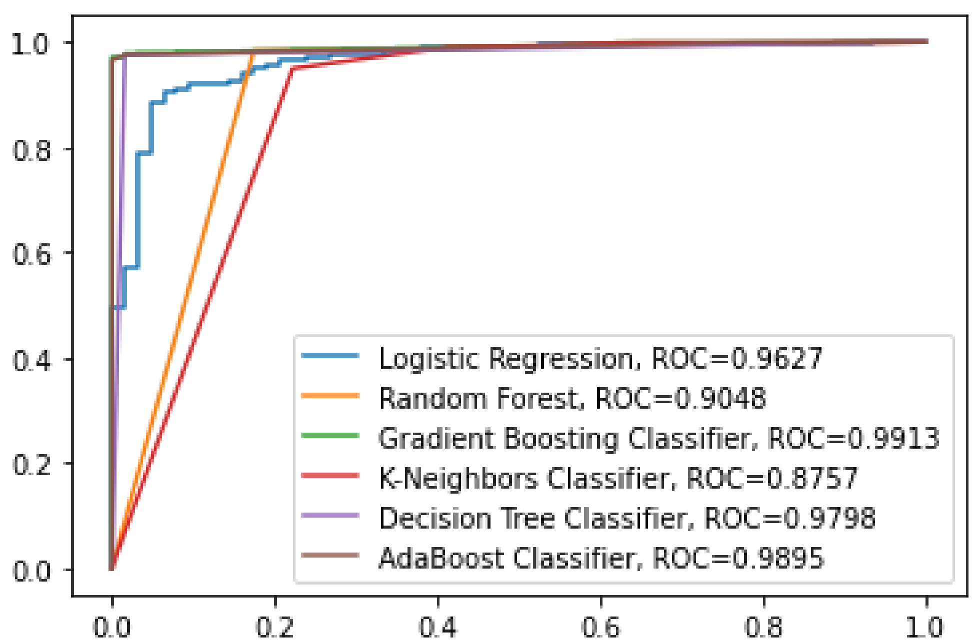


Figure 4.13: AUC Curve of Classifiers

The AUC Curve of Decision Tree (DT), K-Nearest Classifier (KNN), Logistic Regression (LR), Gradient Boosting (GB), Adaboost Classifier (ABC), and Random Forest (RF) is shown in Figure 4.13.

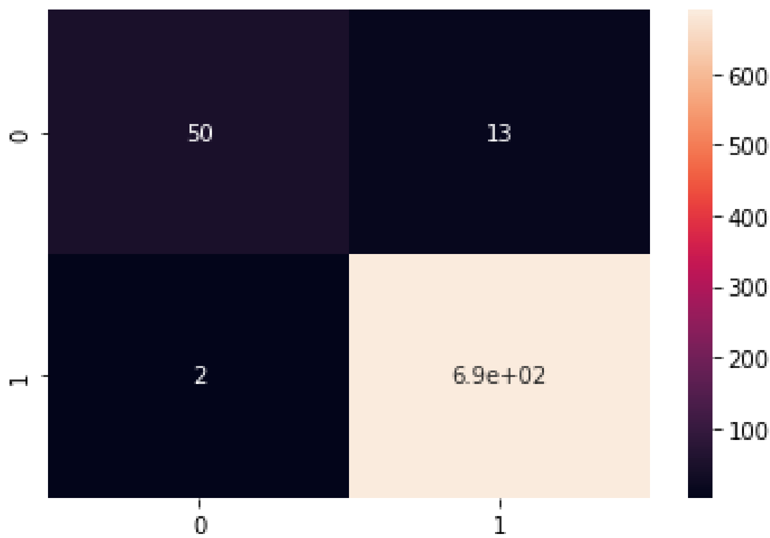


Figure 4.14: Voting Classifier Confusion Matrix

For Voting Classifier, the True positive value was 50, False Positive value was 13, False Negative value was 2 and True Negative value was 9. The output is shown in Figure 4.14.

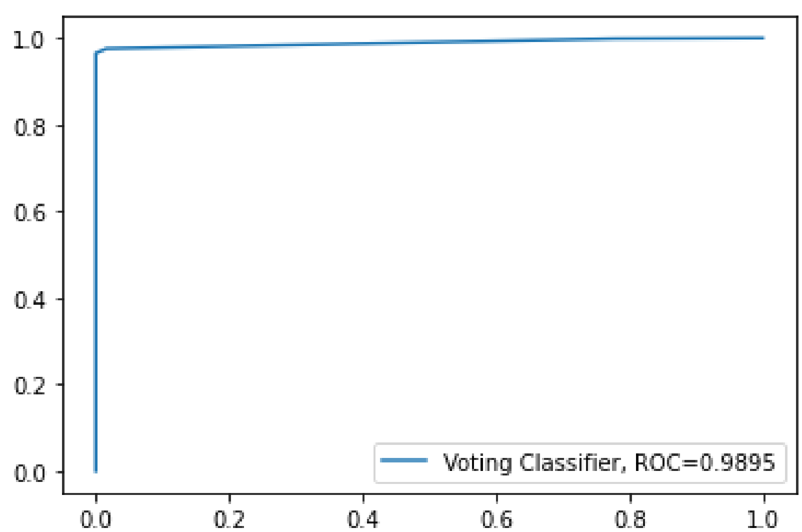
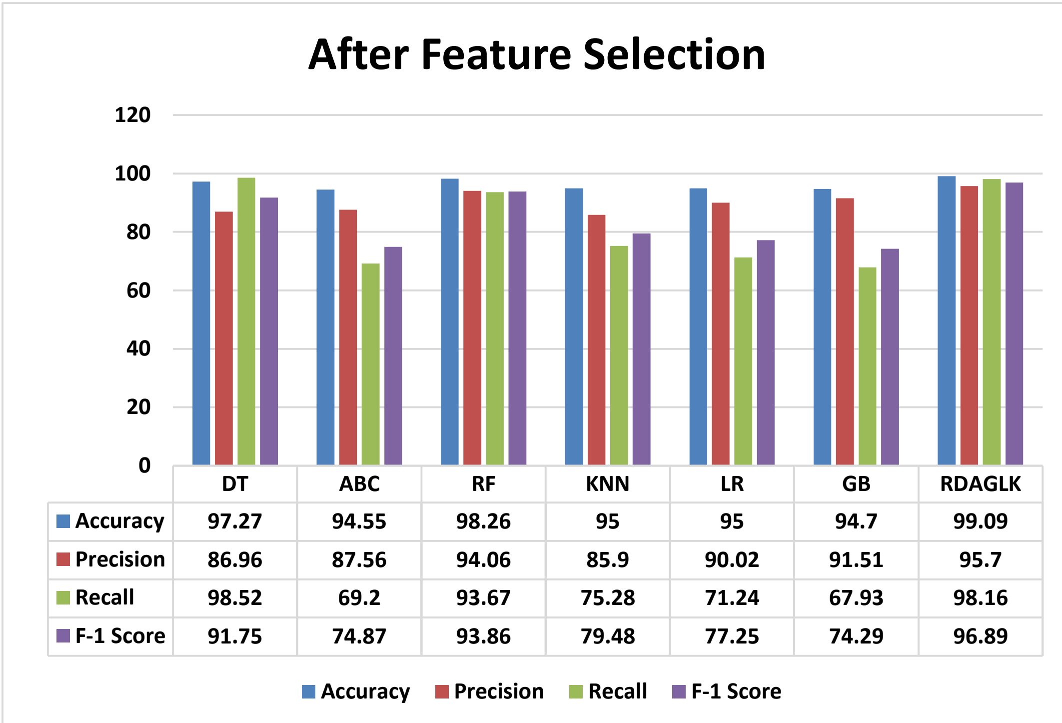


Figure 4.15: AUC Curve of Voting Classifiers

The AUC Curve of Voting Classifier is shown in Figure 4.15.



	DT	ABC	RF	KNN	LR	GB	RDAGLK
Accuracy	97.27	94.55	98.26	95	95	94.7	99.09
Precision	86.96	87.56	94.06	85.9	90.02	91.51	95.7
Recall	98.52	69.2	93.67	75.28	71.24	67.93	98.16
F-1 Score	91.75	74.87	93.86	79.48	77.25	74.29	96.89

Figure 4.16: Experimental Results of Classifiers after feature selection

We considered the performances of algorithmic classifiers, the best accuracy had obtained after feature selection, Random Forest (RF) given an accuracy of 98.26%, Logistic Regression (LR) given accuracy of 95%, Gradient Boosting (GB) given accuracy of 94.7%, Decision Tree given accuracy of 97.62%, Adaboost Classifier (ABC) given the accuracy 94.55%, K-Nearest Classifier (KNN) given the accuracy 95%. The output is shown in Figure 4.16. The precision score was Decision Tree (DT), K-Nearest Classifier (KNN), Logistic Regression (LR), Gradient Boosting (GB), Adaboost Classifier (ABC), and Random Forest (RF) achieved respectively 86.96%, 85.9%, 90.02%, 91.51%, 87.56% and 94.06%. The Recall score was Decision Tree (DT), K-Nearest Classifier (KNN), Logistic Regression (LR), Gradient Boosting (GB), Adaboost Classifier (ABC), and Random Forest (RF) achieved respectively 98.52%, 75.28%, 71.24%, 67.93%, 69.2% and 93.67%. The F-1 score was Decision Tree (DT), K-Nearest Classifier (KNN), Logistic Regression (LR), Gradient Boosting (GB), Adaboost Classifier (ABC), and Random Forest (RF) achieved respectively 91.75%, 79.48%, 77.25%, 74.29%, 74.87% and 93.48%.

Ensemble classifier RDAGLK gives the accuracy 99.09%, precision 95.7%, recall 98.16%, F-1 Score 96.89%.

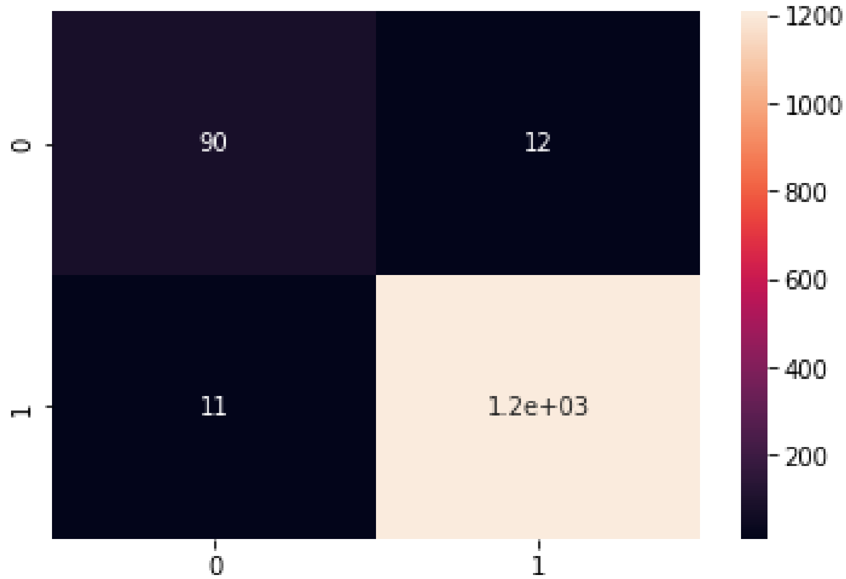


Figure 4.17: Random Forest Confusion Matrix AF

For Random Forest (RF) the True positive value was 90, False Positive value was 12, False Negative value was 11 and True Negative value was 3. The output is shown in Figure 4.17.

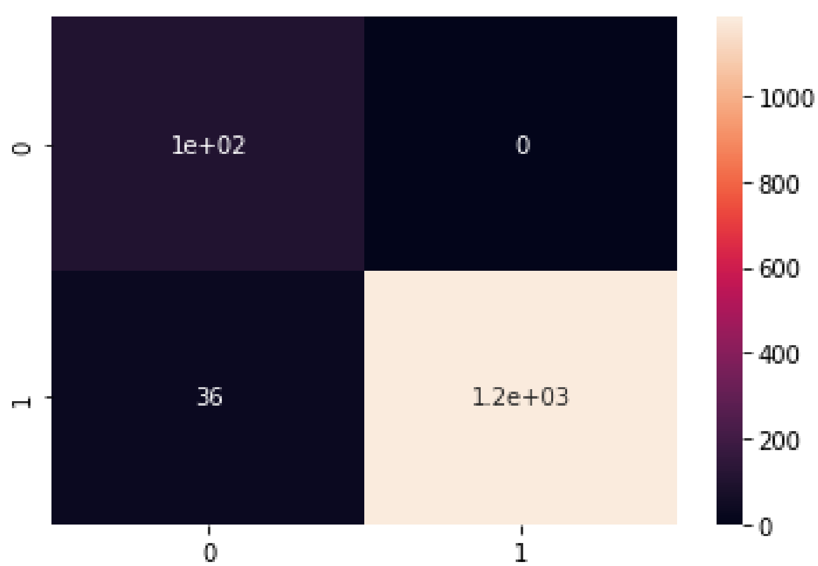


Figure 4.18: Decision Tree Confusion Matrix AF

For Decision Tree (DT) the True positive value was 62, False Positive value was 0, False Negative value was 36 and True Negative value was 4. The output is shown in Figure 4.18.

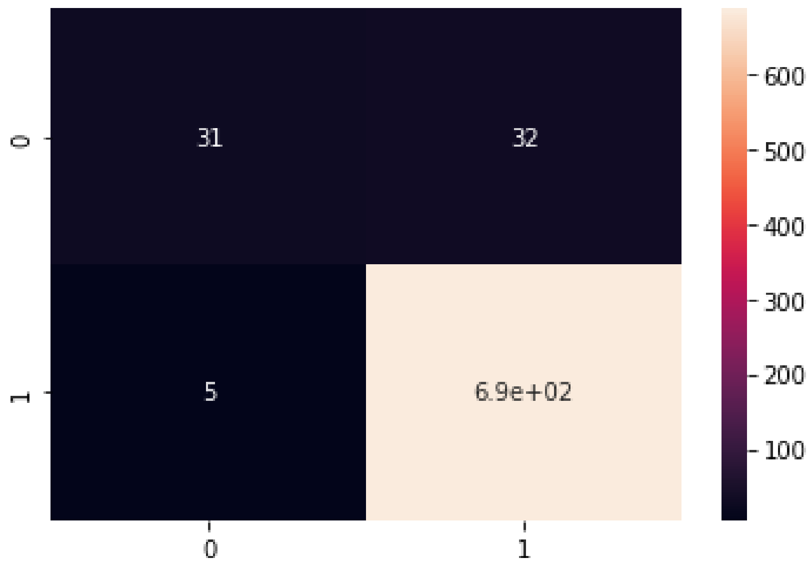


Figure 4.19: K-Nearest Confusion Matrix AF

For K-Nearest the True positive value was 31, False Positive value was 32, False Negative value was 5 and True Negative value was 9. The output is shown in Figure 4.19.

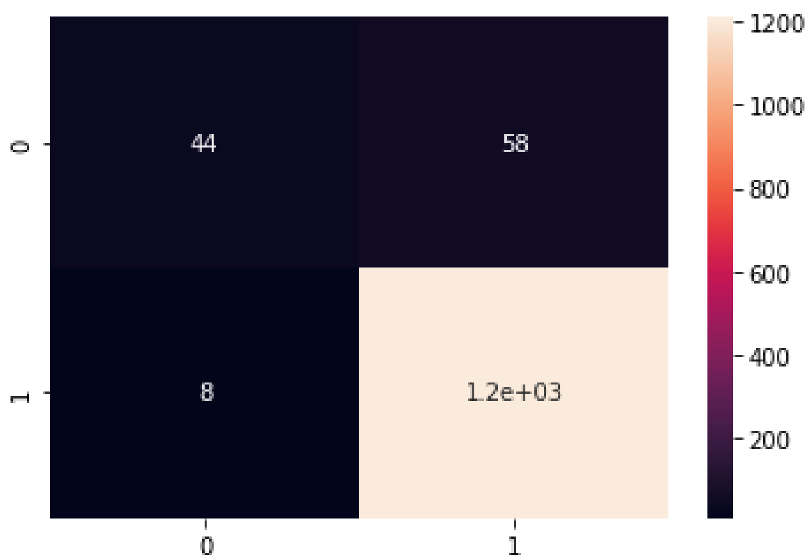


Figure 4.20: Logistic Regression Confusion Matrix AF

For Logistic Regression (LR) the True positive value was 44, False Positive value was 58, False Negative value was 8 and True Negative value was 4. The output is shown in Figure 4.20.

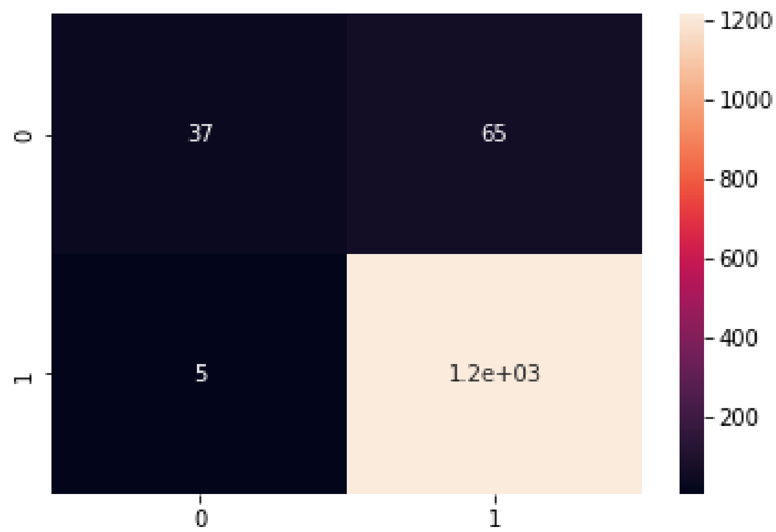


Figure 4.21: Gradient Boosting Confusion Matrix AF

For Gradient Boosting (GB) the True positive value was 37, False Positive value was 65, False Negative value was 5 and True Negative value was 4. The output is shown in Figure 4.21.

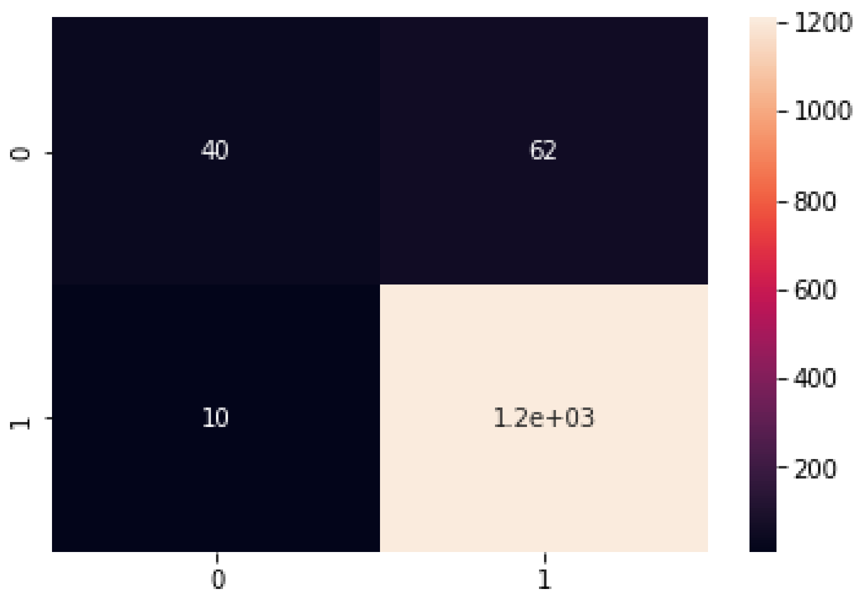


Figure 4.22: Adaboost Classifier Confusion Matrix AF

For Adaboost Classifier (ABC) the True positive value was 62, False Positive value was 1, False Negative value was 17 and True Negative value was 9. The output is shown in Figure 4.22.

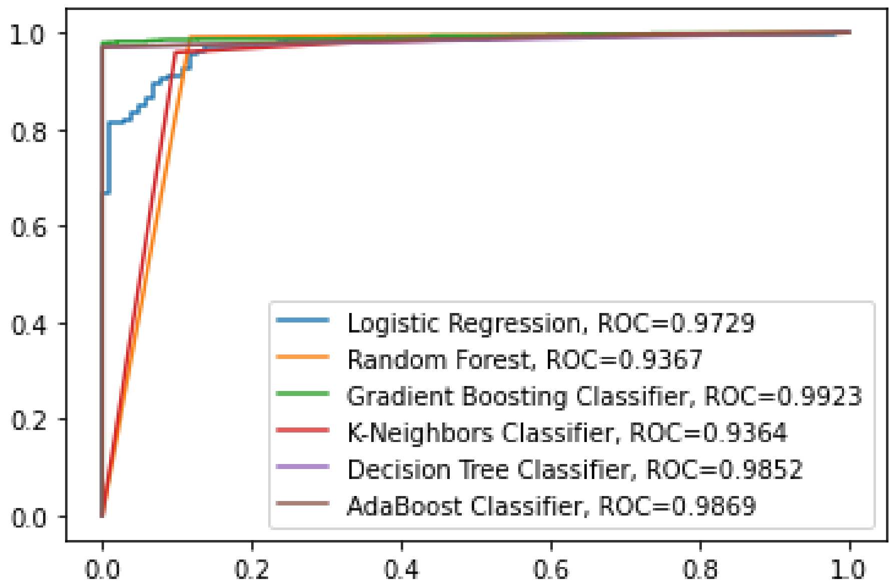


Figure 4.23: AUC Curve of Classifiers AF

The AUC Curve of Decision Tree (DT), K-Nearest Classifier (KNN), Logistic Regression (LR), Gradient Boosting (GB), Adaboost Classifier (ABC), and Random Forest (RF) is shown in Figure 4.23.

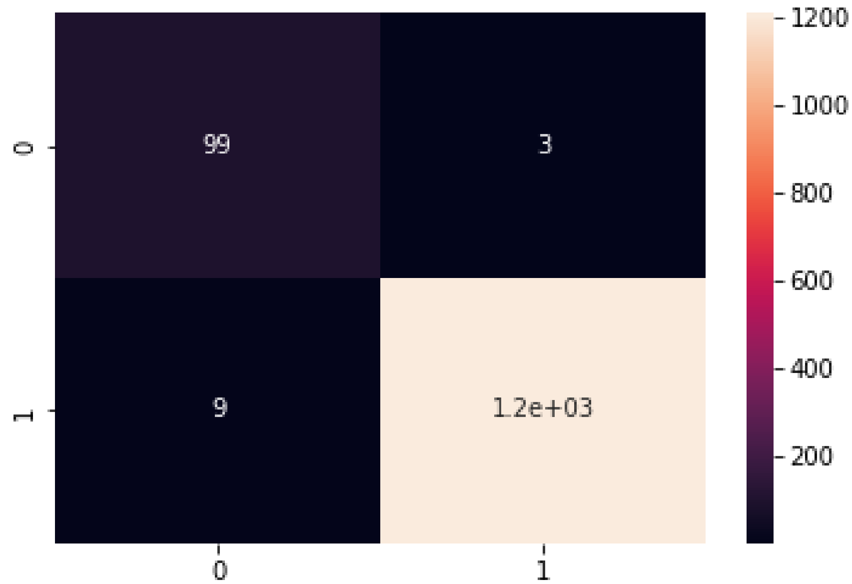


Figure 4.24: Voting Classifier Confusion Matrix AF

For Voting Classifier, the True positive value was 99, False Positive value was 3, False Negative value was 9 and True Negative value was 4. The output is shown in Figure 4.24.

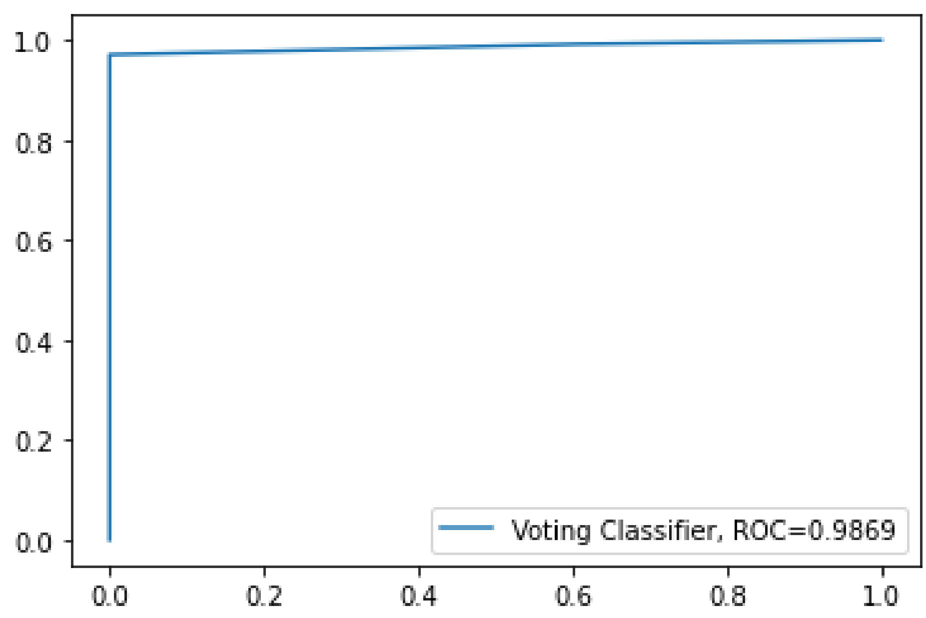


Figure 4.25: AUC Curve of Voting Classifiers AF

The AUC Curve of Voting Classifier is shown in Figure 4.25.

4.3 Discussion

The judicial system of our proposed paradigm will now be defined. The F-1 score, recall, accuracy, and precision have all been considered. Before feature selection, Random Forest (RF) given an accuracy of 97.09%, Logistic Regression (LR) given accuracy of 94.7%, Gradient Boosting (GB) given accuracy of 94.04%, Decision Tree given accuracy of 97.62%, Adaboost Classifier (ABC) given the accuracy 97.62%, K-Nearest Classifier (KNN) given the accuracy 95.1%. We have used ensemble techniques to get the best accuracy. Our voting classifier RDAGLK gave the best accuracy of 98.01%. After feature selection, Random Forest (RF) given an accuracy of 98.26%, Logistic Regression (LR) given accuracy of 95%, Gradient Boosting (GB) given accuracy of 94.7%, Decision Tree given accuracy of 97.27%, Adaboost Classifier (ABC) given the accuracy 94.55%, K-Nearest Classifier (KNN) given the accuracy 95%. We have used ensemble techniques to get the best accuracy. Our voting classifier RDAGLK gave the best accuracy of 99.09%.

4.3.1 Accuracy

It refers to the percentage of predictions made using testing data that were accurate. Accuracy is accomplished in contrast to accessibility of the measures using real measurements. It is based on just one variable. Accuracy solely handles intentional errors. It is among the simplest measuring techniques for any model. We must aim for maximal accuracy in our models.

Accuracy = (TruePositive + TrueNegative)/(TruePositive + FalsePositive + TrueNegative + FalseNegative)

4.3.2 Precision

It discusses the proportion of optimistically anticipated observations that really happened. Precision identifies the actual true portion of all the instances when they properly predicted true. A high recall for any sort of model might also be quite misleading.

Precision = (TruePositive)/(TruePositive + FalsePositive)

4.3.3 Recall

It refers to the proportion of data from a model that are expected to be positive. However, high accuracy might also be misleading. Normally recall determines the ratio of anticipated positives to all positive labels.

Recall = (TruePositive)/(TruePositive + FalseNegative)

4.3.4 F-1 Score

It mentions the accuracy and harmonic ways of recollection. Relevant metrics include the recall and accuracy ratios. If the harmonic mean is smaller, we infer the model is really bad.

$F - 1$ Score = $2 *$(Recall + Precision)/(Recall + Precision)

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT, AND SUSTAINABILITY

5.1 Impact on Society

Numerous benefits are provided by our recommended strategy, both economically and socially. Our model is based on a real-life dataset and was developed to investigate and determine the essential components or characteristics of a hypothyroid illness patient. Benefits to society include the capacity to educate individuals about the incidence of hypothyroidism and available prevention strategies. We are able to suggest early therapy because of precise diagnosis and regular checks. due to the fact that they are more likely to be knowledgeable about illnesses and able to foresee whether they would be affected or not. Our method takes less time and calls for fewer compilations. This makes sickness prediction easy and precise. We have examined the data in our model to identify the underlying cause of hypothyroidism using enhanced diagnosis techniques. On a social level, we hope that our advised course of action will be embraced and implemented.

5.2 Impact on Environment

Our proposed paradigm is particularly effective in distant areas due to the simplified diagnosis methods. We can reduce complexity and time using the device model. We can assure that the environment will benefit from our technique because it is straightforward and has no unfavorable effects. To find out if they have hypothyroidism or not, people don't need to travel to urban areas. The prediction model, which also predicts probable outcomes, may easily complement the patient's diagnostic report. Patients won't be concerned about the cost of local therapy or the cheap cost of identifying hypothyroidism. Anyone at any level can use it since it is less complicated.

Using our recommended model, it is possible to determine if a patient has hypothyroidism or not. Our recommended model will make the political and social environment better. We are convinced that if our proposed model is put into use, it will result in a considerable improvement in the state of medical scientific technology.

5.3 Ethical Aspects

We must take certain moral precautions before the system is put into use to stop the publication of private information, diagnostic results, or humor. The diagnosis and treatment of hypothyroidism in the actual world as well as upcoming research projects may make use of our recommended methodology. We've found that the problem impacts the entire world, not just a tiny area or region. Any victim or knowledgeable person may predict how quickly their hypothyroid condition will affect them using the proposed methodology.

5.4 Sustainability Plan

We can ensure that our proposed model can be accepted by the technology used in hypothyroid illness diagnosis and research throughout the globe. We are confident that the victim women who are aware of their propensity to acquire hypothyroid condition would benefit from our suggested course of action. If we are given the proper resources and room to operate, we could be motivated and ready to assist the rural areas. We believe that our proposed paradigm will be both practical and long-lasting.

CHAPTER 6

SUMMARY, CONCLUSION, RECOMMENDATION, AND IMPLICATION FOR FUTURE RESEARCH

6.1 Summary of the Study

In this engaging essay, we use algorithms to evaluate the influence rate of our people. We are able to accurately predict the future using our model. The diagnostic technique could be useful for the prediction system. Knowing whether or not they will make an influence may be beneficial for individuals. They can mistakenly think that they need to be aware of the hypothyroid condition. With the aid of our method, people may recognize the various stages of hypothyroidism with ease. Using our recommended methodology can be advantageous to diagnosis authority as well. We have used a range of well-established, quick-to-implement, low-training, and highly accurate algorithms.

6.2 Conclusion

We currently live in a contemporary world. Today's world is both basic and highly advanced. Anyone in the globe has access to the new technology. Technology has made what we've proposed incredibly rapid and easy. The method of anticipating hypothyroid disorder in individuals has been made as simple as possible. Our folks will benefit from our cutting-edge models. We must ensure that the idea is practical, and we pledge to add a ton more features and focus on more popular subjects in the future. This expectation is being set now.

6.3 Implication for Further Study

Because we are human, we are mortal. Numerous illnesses affect us on a daily basis. While the majority of us have hypothyroid disease, some of us also have the tools to heal. Since we reside in a developing nation, the technologies for therapy and diagnostics are more sophisticated and accurate. Thanks to improved technologies, the procedure for diagnosing hypothyroid sickness is now easier and takes less time. We have made an attempt to offer something new to our customers. We hope that other people will use our strategy. We have improved a few algorithms for better performance, and we want to add more in the future.

REFERENCE

[1] “hypothyroid”, Last Accessed: December 2018, Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6822815/#:~:text=Hypothyroidism%20affects%20up%20to%205,patients%20suffer%20from%20primary%20hypothyroidism.

[2] L. Yang and A. Shami, “On hyperparameter optimization of machine learning algorithms: theory and practice,” *Neurocomputing*, vol. 415, pp. 295–316, 2020.

[3] H. M. Almahshi, E. A. Almasri, H. Alquran, W. A. Mustafa and A. Alkhayyat, "Hypothyroidism Prediction and Detection Using Machine Learning," 2022 5th International Conference on Engineering Technology and its Applications (IICETA), 2022, pp. 159-163, doi: 10.1109/IICETA54559.2022.9888736.

[4] Shahid, A. H., Singh, M. P., Raj, R. K., Suman, R., Jawaid, D., &Alam, M. (2019, July). A Study on Label TSH, T3, T4U, TT4, FTIin Hyperthyroidism and Hypothyroidism using Machine LearningTechniques. In 2019 International Conference on Communicationand Electronics Systems (ICCES) (pp. 930-933). IEEE.

[5] Akgül, G., Çelik, A. A., AYDIN, Z. E., & ÖZTÜRK, Z. K. (2020).Hipotiroidi Hastalığı Teşhisinde Sınıflandırma AlgoritmalarınınKullanımı. Bilişim Teknolojileri Dergisi, 13(3), 255-268.

[6] Chandel, K., Kunwar, V., Sabitha, S., Choudhury, T., & Mukherjee,S. (2016). A comparative study on thyroid disease detection using K-nearest neighbor and Naive Bayes classification techniques. CSItransactions on ICT, 4(2), 313-319.

[7] V. Lahoura, H. Singh, A. Aggarwal et al., “Cloud computing-based framework for breast cancer diagnosis using extreme learning machine,” *Diagnostics*, vol. 11, no. 2, p. 241, 2021.

[8] “hypothyroid.csv”, Accessed: December 29, 2019, Available: https://www.kaggle.com/datasets/nguyenthilua/hypothyroidcsv

[9] “What is Correlation in Machine Learning?”, Accessed: August 6, 2020, Available: https://medium.com/analytics-vidhya/what-is-correlation-4fe0c6fbed47

[10] L. Mary Gladence, M. Karthi, V. Maria Anu. “A statistical Comparison of Logistic Regression and Different Bayes Classification Methods for Machine Learning” ARPN Journal of Engineering and Applied Sciences, ISSN 1819-6608, Vol -10, No-14, August 2015.

[11] “What is Correlation in Machine Learning?”, Accessed: November 8, 2021, Available: https://medium.com/analytics-vidhya/what-is-correlation-4fe0c6fbed47

[12] “Logistic Regression for Machine Learning”, Accessed: August 6, 2021, Available: https://www.capitalone.com/tech/machine-learning/what-is-logistic-regression/

[13] Ghosh, Pronab, Asif Karim, Syeda Tanjila Atik, Saima Afrin, and Mohd Saifuzzaman. "Expert cancer model using supervised algorithms with a LASSO selection approach." International Journal of Electrical and Computer Engineering (IJECE) 11, no. 3 (2021): 2631.

[14] Nahar, Nazmun, and Ferdous Ara. "Liver disease prediction by using different decision tree techniques." International Journal of DataMining & Knowledge Management Process 8, no. 2 (2018): 01-09

[15] Aljahdali, Sultan, and Syed Naimatullah Hussain. "Comparative prediction performance with support vector machine and random forest classification techniques." International journal of computer applications 69, no. 11 (2013).

[16] Bentéjac, Candice, Anna Csörgő, and Gonzalo Martínez-Muñoz. "A comparative analysis of gradient boosting algorithms." ArtificialIntelligence Review 54, no. 3 (2021): 1937-1967.

[17] Drucker, Harris, Corinna Cortes, Lawrence D. Jackel, Yann LeCun, and Vladimir Vapnik. "Boosting and other ensemble methods." Neural Computation 6, no. 6 (1994): 1289-1301.

[18] Pasha, Maruf, and Meherwar Fatima. "Comparative Analysis of Meta Learning Algorithms for Liver Disease Detection." J. Softw. 12, no.12 (2017): 923-933.

[19] Wang, Yizhen, Somesh Jha, and Kamalika Chaudhuri. "Analyzing the robustness of nearest neighbors to adversarial examples." In International Conference on Machine Learning, pp. 51335142. PMLR, 2018.

[20] Sharma, Ajay, and Anil Suryawanshi. "A novel method for detecting spam email using KNN classification with spearman correlation as distance measure." International Journal of Computer Applications 136, no. 6 (2016): 28-35.

[21] hou, Zhi-Hua. Ensemble methods: foundations and algorithms. CRC Press, 2012.

[22] Drucker, Harris, Corinna Cortes, Lawrence D. Jackel, Yann LeCun, and Vladimir Vapnik. "Boosting and other ensemble methods." Neural Computation 6, no. 6 (1994): 1289-1301.

[23] emmens, Aurélie, and Christophe Croux. "Bagging and boosting classification trees to predict churn." Journal of Marketing Research43, no. 2 (2006): 276-286.

[24] Islam, Rakibul, Abhijit Reddy Beeravolu, Md Al Habib Islam, Asif Karim, Sami Azam, and Sanzida Akter Mukti. "A Performance Based Study on Deep Learning Algorithms in the Efficient Prediction of Heart Disease." In 2021 2nd International Informatics and Software Engineering Conference (IISEC), pp. 1-6. IEEE, 2021.

[25] Shorove Tajmen, Asif Karim, Aunik Hasan Mridul, Sami Azam, Pronab Ghosh, Alamin Dhaly, Md Nour Hossain. “A Machine Learning based Proposition for Automated and Methodical Prediction of Liver Disease”. In April 2022 The 10th International Conference on Computer and Communications Management in Japan.

[26] Molla, S., Shamrat, F. J. M., Rafi, R. I., Umaima, U., Arif, M. A. I., Hossain, S., & Mahmud, I. (2022). A predictive analysis framework of heart disease using machine learning approaches. Bulletin of Electrical Engineering and Informatics, 11(5), 2705-2716.

[27] Afrin, S., Shamrat, F. J. M., Nibir, T. I., Muntasim, M. F., Moharram, M. S., Imran, M. M., & Abdulla, M. (2021). Supervised machine learning based liver disease prediction approach with LASSO feature selection. Bulletin of Electrical Engineering and Informatics, 10(6), 3369-337.

[28] Ghosh, P., Azam, S., Jonkman, M., Karim, A., Shamrat, F. J. M., Ignatious, E., ... & De Boer, F. (2021). Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. IEEE Access, 9, 19304-19326.

[29] Jubier Ali, M., Chandra Das, B., Saha, S., Biswas, A.A., Chakraborty, P. (2022). A Comparative Study of Machine Learning Algorithms to Detect Cardiovascular Disease with Feature Selection Method. In: Skala, V., Singh, T.P., Choudhury, T., Tomar, R., Abul Bashar, M. (eds) Machine.

Kartick

ORIGINALITY REPORT

25%	16%	17%	13%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Shorove Tajmen, Asif Karim, Aunik Hasan Mridul, Sami Azam, Pronab Ghosh, Al-Amin Dhaly, Md Nour Hossain. "A Machine Learning based Proposition for Automated and Methodical Prediction of Liver Disease", The 10th International Conference on Computer and Communications Management, 2022 Publication	7%
2	Charles A. Morgan, Yaron Rabinowitz, Robert Leidy, Vladimir Coric. "Efficacy of Combining Interview Techniques in Detecting Deception Related to Bio-threat Issues", Behavioral Sciences & the Law, 2014 Publication	3%
3	dspace.daffodilvarsity.edu.bd:8080 Internet Source	3%
4	Submitted to Daffodil International University Student Paper	3%
5	Hebatullah Mohammad Almahshi, Esraa Abdallah Almasri, Hiam Alquran, Wan Azani Mustafa, Ahmed Alkhayyat. "Hypothyroidism	1%