# THYROID DISEASES CLASSIFICTION USING ML

# 1.INTRODUCTION:

* Diagnostics and prediction of diseases are among the most important

applications of machine learning techniques.

* Recently, machine learning algorithms have had an essential and convincing role in diagnosing and classifying diseases.
* Among these diseases, thyroid disease is a concern for human health, as the thyroid gland plays a critical role in regulating human health because it regulates human metabolism.
* This study used eight machine learning techniques (Support Vector Machines, Random Forest, Decision Tree, Naive Bayes, Logistic Regression, K-Nearest Neighbors, Multi- layer Perceptron (MLP), Linear Discriminant Analysis) to diagnose thyroid disease.
* Thyroid disease is classified in this study into three categories: hyperthyroidism, hypothyroidism, and normal.
* Machine learning algorithms have achieved promising results in diagnosing thyroid diseases, helping clinicians and health workers to diagnose the disease early, and increase the chance of treatment.
* The Random Forest algorithm got the highest accuracy of 98.93% with all the features, and the MLP algorithm got the highest accuracy of 95.73% with the deletion of three properties which are (query\_thyroxine, query\_hypothyorid and query\_hyperthyroid).
* The thyroid gland can also be the location of different kinds of tumors and dangerous where endogenous antibodies wreak havoc.
* Thyroid gland hormones are responsible for aiding in digestion and maintaining the body moist, balanced, and so on hormone). Thyroid disorder is classified into two types: hypothyroidism and hyperthyroidism.

1. Normal thyroid function test.

# Test Normal Value

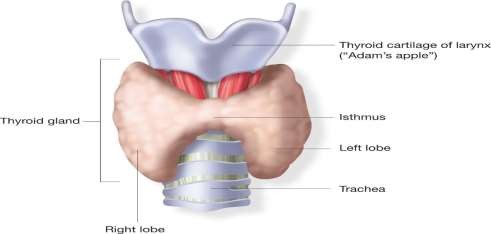
|  |  |
| --- | --- |
| TSH | (0.2 - 6.0) ulU/L |
| T3 | (0.9 - 2.35) nmol/L |
| T4 | (60 – 120) nmol/L |

* Hyperthyroidism is a disorder in which the thyroid gland releases so many thyroid hormones. Hyperthyroidism is caused by an increase in thyroid hormone level.
* Dry skin, elevated temperature sensitivity, hair thinning, Weight loss, increased heart rate, high blood pressure, heavy sweating, swollen neck, nervousness, menstruation cycle shortening, irregular stomach movements, and handshaking are some signs.
* A decline in thyroid hormone production causes hypothyroidism. Hypo means deficient or less in medical terms. Inflammation and thyroid gland injury are the two primary causes of hypothyroidism. Obesity, low heart rate, increased temperature sensitivity, neck swelling, dry skin, hand numbness, hair issues, heavy menstrual cycles, and intestinal problems. If not treated, these symptoms can escalate over time.
* Early disease detection, diagnosis, and care, according to doctors, are vital in preventing disease progression and even death. For several different forms of anomalies, early identification and differential diagnosis raise the odds of good treatment. Despite multiple trials, clinical diagnosis is often thought to be a difficult task.
* Data mining is a semi-automated method of looking for correlations in massive datasets.
* Machine learning algorithms are one of the best solutions for most complex problems.
* Classification is a data extraction technique (machine learning) used to predict and identify many diseases, such as thyroid disease. We researched and classified here because machine learning algorithms play a significant role in classifying thyroid disease and because these algorithms are high performing and efficient and aid in classification.
* Although the application of computer learning and artificial intelligence in medicine dates back to the early days of the field, there has been a new movement to consider the need for machine learning-driven healthcare solutions.
* As a result, analysts predict that machine learning will become commonplace in healthcare soon.

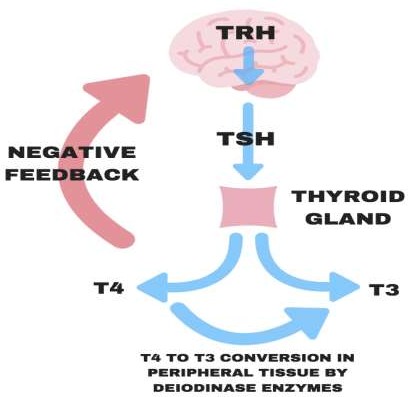
# 1.1Overview

* The thyroid gland is an endocrine gland divided into the right and left lobes, which are situated on opposite sides of the trachea in the throat, with an isthmus connecting them .
* It has the appearance of a butterfly gland and weighs around 25 grams in adults.
* When seen from the front, it is positioned under the chin,

It is found just below the cartilage protrusion, which is more pronounced in males and is referred to as Adam's Apple by the public, and it helps with swallowing. While it is small in size, the hormones it secretes play a critical role in human growth and development, and it is also known as the "regulator of all body functions.



Location of thyroid gland.

* The thyroid gland. It has two types of secretions within its functioning system, triiodothyronine (T3) and another is thyroxine (T4).
* And the element used for these two hormones is iodine in the blood because iodine is the main component to building these two hormones, T3 and T4. As for (T3), this label came because it consists of three atoms of iodine, while (T4) this label came because it consists of four atoms of iodine, and the critical role of these hormones is to control the metabolism process.
* TRH (TSH Releasing Hormone) is generated in the hypothalamus,located in the upper part of the brain and secreted by the thyroid gland. Thyroid hormone levels in the blood affect the volume of TRH hormone secreted by the hypothalamus.
* This hormone stimulates the pituitary gland, located in the lower part of the brain, to manufacture thyroid-stimulating hormone (TSH). The pituitary gland secretes the thyroid-stimulating hormone.
* This hormone is the central part of the synthesis of T4 (thyroxine). And also enters the formation of a small amount of the hormone T3 (triiodothyronine) and then excreted into the bloodstream.

**Thyroid hormones**

* T3 and T4 development occur until the body's requirements decide the optimum blood level .

Thyroid disease is classified into three important categories:

* + 1. Hyperthyroidism.
    2. Hypothyroidism.
    3. Normal.
* Hyperthyroidism (Thyrotoxicosis) refers to an abnormally high level of thyroid hormones in the blood.
* There are several subtypes of hyperthyroidism. Sublink suppressed TSH (TSH 0.5 mIU / L) with regular T3, T4 levels, and overt hyperthyroidism (high T3, T4 levels with suppressed TSH).
* Any signs Include dry skin, elevated temperature response, hair thinning, weight loss, increased heart rate, hypertension, heavy sweating, neck enlargement, nervousness, shortened menstrual cycles, irregular stomach movements, and hands shaking. Hypothyroidism is a typical concept that applies to an underactive thyroid gland. Hypothyroidism occurs when there is a shortage of tissue-level thyroid hormones or when these hormones are sometimes inactive.
* Hypothyroidism is mainly affected by inflammation and injury to the thyroid gland. Obesity, a sluggish heart rate, prolonged exposure to high temperatures, throat swelling, dry eyes, numb palms, hair loss, erratic menstrual cycles, and intestinal problems are only a few symptoms. If left untreated, these signs can deteriorate with time.
* As for the normal, that person who does not suffer from disorders in the thyroid gland, the level of secretion of thyroid hormones is moderate in the blood, so it does not suffer from any disease and be healthy.

# 1.2 Purpose

* The purpose of this article is to demonstrate how a prototype that uses text mining and machine learning approaches to detect strokes may be employed. Machine learning may be a significant tracker when correctly trained machine learning algorithms are used in surveillance, nursing, and data processing.
* The semantic and syntactic analysis of information monitoring is provided by the data mining methods utilized in this work. At Sugam Multispecialty Hospital in Kumbakonam, Tamil Nadu, India, 507 patient case sheets were gathered using the data collecting approach.
* Machine learning methods utilized to analyze the data included artificial neural networks, support vector machines, boosting and bagging, and random forests. Using a classification accuracy of 95% and a standard deviation of 14.69, artificial neural networks trained with a stochastic gradient descent approach outperformed the other techniques.
* Thyroid diseases are becoming increasingly common around the world. Hypothyroidism, hyperthyroidism, or thyroid cancer affect one out of every eight women.
* Thyroid categorization is important for medical researchers because medical reports show major thyroid dysfunctions among the population, with women being the most affected.
* The literature mentions several studies in thyroid classification that use various machine learning techniques to develop robust classifiers.
* The literature mentions several studies in thyroid classification that use various machine learning techniques to develop robust classifiers. The goal we want to reach or the primary goal:
* Comparison of the performance of the eight machine learning algorithms in predicting thyroid disease.
* Extract useful patterns from large and complex clinical data.
* Make the study work to show the following results.
  + - * hyperthyroidism.
      * hypothyroidism.
      * normal.

# 2. PROBLEM DEFINITION & DESIGN THINKING

* At present, diseases have become dangerous and rapidly spread, and their exploration and diagnosis require a great deal of time and effort.
* The correct and accurate diagnosis of the disease early has become one of the problems that the health system suffers from it.
* The critical role of early and correct diagnosis of the disease, including thyroid disease, is vital because it increases patient treatment opportunities and reduces mortality.
* Among the vast amount of clinical data, early diagnosis is a challenging task.
* Today the machine learning has had impressive and good results in many sciences.
* Hence, it had a prominent and valuable role in diseases, so this study used machine learning algorithms with thyroid disease.
* To detect and classify thyroid disease into three types hyperthyroidisms, hypothyroidism, and normal.
* In this study, researchers created a decision support system using machine learning techniques to classify Thyroid disorder using classification models based on TSH, T4U, and goiter.
* Several grouping methods, such as K-nearest neighbor, are used to justify this argument.
* The Naive Bayes and support vector machine algorithms are employed. And the findings indicate that K-nearest neighbor is more effective than Naive Bayes in detecting thyroid disease. To diagnose thyroid disorder, the researchers used data mining classifiers.
* Thyroid disorder is a vital factor to consider when analyzing a condition. KNN and Naive Bayes classifiers it used in this study.
* The findings revealed that the K-nearest neighbor classifier is the most reliable,
* 93.44 percent accuracy.
* While the accuracy obtained by the Naive Bayes classifier is

22.56 %, the proposed KNN technology improves classification accuracy, which contributes to better performance. Given that, Naive Bayes can only have linear, parabolic, or elliptical decision limits, so the consistency of KNN's decision limits is a good plus.

* KNN outperforms most methods because the factors are interrelate.
* In this study, researchers used machine learning techniques to diagnose and classify thyroid disease because Thyroid disease is one of the most common diseases in the world that humans suffer.
* The hypothyroid data used in this study came from the University of California, Irvine (UCI) data repository.
* This study will use the platform Waikato Environment of Information Analysis for the whole research project (WEKA).
* Where found the J48 technique to be more effective than the decision stump tree technique.
* In the world of health care, disease diagnosis is a difficult challenge. In the decision-making method, Used J48 and decision stump data mining classification techniques to define hypothyroidism. The J48 algorithm has 99.58 percent accuracy,which is higher than the decision stump tree accuracy, and it also has a lower error rate than the Decision stump.
* The researchers resorted to relying on machine learning techniques in this study to classify thyroid disease.
* Classification is used to characterize pre-defined data sets, is one of the most popular supervised learning data mining techniques.
* The classification is commonly used in the healthcare sector to aid in medical decision-making, diagnosis, and administration. The information for this study was gathered from a well- known Kashmiri laboratory.
* The entire research project will be conducted on the ANACONDA3-5.2.0 platform.
* In an experimental analysis, classification methods such as k nearest neighbors, Support vector machine, Decision tree, and Naive Bayes may be used.
* The Judgment Tree has the highest accuracy of the other classes, at 98.89 percent.
* In this study, the researchers studied the classification of the thyroid gland using machine learning because Thyroid disorder is a chronic illness that affects people all over the world.
* Data mining in healthcare is producing excellent results in the prediction of different diseases.
* The accuracy of data mining techniques for prediction is high, and the cost of prediction is low.
* Another significant benefit is that forecast takes very little time. In this study, I used classification algorithms to analyze thyroid data and came up with a result.
* Two factors primarily determine a model's efficacy. The first is prediction precision, and the second is prediction time.
* According to our findings, Nave Bayes took just 0.04 seconds to forecast. However, it is less accurate than J48 and Random Forest.
* When we looked at prediction accuracy, the Random Forest model came in at 99.3 percent. However, the model's construction time is longer than the other two iterations.
* Son can assume that J48 is the best model for hypothyroid prediction since its accuracy is 99 percent, which is among the highest. It takes 0.2 seconds to run, significantly less time than the Random Forest model [24].
* This study proposes a data mining-based method for enhancing the precision of hypothyroidism diagnosis by integrating patient questions with test results during the diagnosis process.
* Another goal is to reduce the risks that come with dialysis interventional trials.
* Using data from the UCI machine learning database, which included 3163 pieces, 151 were hypothyroid. And the others were hyperthyroid.
* Models were developed to diagnose hypothyroidism using Logistic Regression, K Nearest Neighbor, and Support Vector Machine classifiers. The thesis demonstrated the impact of sampling techniques on the diagnosis of hypothyroidism in this regard.
* The Logistic Regression classifier produced the best results of all the models created.
* The precision was 97.8%, the F-Score was 82.26 %, the region under the curve was 93.2 %, and the Matthews correlation coefficient was 81.8 percent for this analysis, which a trained on the data set using over-sampling techniques.
* This paper aims to create a method to predict diabetes in a patient early and accurately using the Random Forest algorithm in a machine learning technique.
* Random Forest algorithms are a type of ensemble learning system commonly used for classification and regression tasks.
* As compared to other algorithms, the performance ratio is higher. The suggested model gives the best outcomes for diabetic prediction, and the results revealed that the prediction system is capable of correctly, effectively, and most importantly, immediately forecasting diabetes disease.
* This study is based on building a decision using machine learning breast cancer classification.
* Breast cancer is the second most frequent cancer in women after all other malignancies.
* This research report intends to present a breast cancer study combining cutting-edge methodologies for enhancing breast cancer survival modeling models by embracing current research breakthroughs.
* According to the results, the Bay is a safe area to visit (based on an average precision Breast Cancer dataset).
* The RBF Network is the second-best predictor, with a holdout sample accuracy of 93.41 percent (better than any other prediction accuracy reported in the literature). With 97.36 percent accuracy on the holdout sample (greater than any other prediction accuracy recorded in the literature), Naive Bayes is the third-best predictor (better than any other prediction accuracy published in the literature).
* The researchers used two criteria to assess three breast cancer survivorship prediction models: benign and malignant cancer cases.
* The most recent study categorizes thyroid illness in two of the most common thyroid dysfunctions in the general population (hyperthyroidism and hypothyroidism).
* The researchers compared Naive Bayes, Decision Trees, Multilayer Perceptron.
* The results show that all of the classification mentioned above models are very accurate, with the Decision Tree model having the highest classification score.
* She used a Romanian data website and the UCI machine learning repository to build and evaluate the classifier. Weka and the KNIME Analytics Platform are two data sets.
* The categorization models were developed and tested using data mining methods as the basis.
* According to the literature, several researches in thyroid classification employs diverse data mining approaches to build robust classifiers.
* This study examined how four classification models (Nave Bayes, Decision Tree, MLP, and RBF Network) can be used on thyroid data to categorize thyroid dysfunctions, including hyperthyroidism and hypothyroidism.
* The decision tree model was the proper categorization model in every example that an examined.
* The machine learning application WEKA (Waikato Environment for Information Analysis) identifies the thyroid class (normal, hypothyroid, or hyperthyroid).
* In the UCI (University of California – Irvine) machine learning database, the thyroid dataset is represented by 215 instances.
* The MLP (Multilayer Perceptron) system, according to the results, delivers the highest accuracy, up to 96.74 percent.
* The BPA (Back Propagation Algorithm) approach, on the other hand, has the lowest accuracy among the six WEKA techniques, at 69.77 percent. MLP and RBF have the highest accuracy (96.74 and 95.35 percent, respectively), while BPA has the lowest (96.74 and 95.35 percent, respectively) (69.77 percent).
* A difficulty with the dataset, which comprises local minima or extreme values, causes BPA's poor accuracy. MLP, on the other hand, is the most accurate because of its excellent fault tolerance.
* The purpose of this article is to demonstrate how a prototype that uses text mining and machine learning approaches to detect strokes may be employed. Machine learning may be a significant tracker when correctly trained machine learning algorithms are used in surveillance, nursing, and data processing.
* The semantic and syntactic analysis of information monitoring is provided by the data mining methods utilized in this work.
* At Sugam Multispecialty Hospital in Kumbakonam, Tamil Nadu, India, 507 patient case sheets were gathered using the data collecting approach.
* Machine learning methods utilized to analyze the data included artificial neural networks, support vector machines, boosting and bagging, and random forests. Using a classification accuracy of 95% and a standard deviation of 14.69, artificial neural networks trained with a stochastic gradient descent approach outperformed the other techniques.

Related literature review.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No** | **Authors** | **Reference** | **year** | **Algorithms** |
| **1** | Chandel, Khushboo | [22] | 2016 | KNN, Naive Bayes |
| **2** | Banu, G. Rasitha | [10] | 2016 | J48 |
| **3** | Umar Sidiq, Dr, Syed Mutahar Aaqib, and Rafi  Ahmad Khan | [23] | 2019 | k nearest neighbors, SVM, DT, Nave bayes |
| **4** | Sindhya, Mrs K | [24] | 2020 | Nave bayes, J48 and  Random Forest |
| **5** | AKGÜL, Göksu, et al | [25] | 2020 | k nearest neighbors and  SVM |
| **6** | VijiyaKumar, K., et al | [26] | 2019 | Random Forest |
| **7** | Chaurasia, Vikas, Saurabh  Pal, and B. B. Tiwari | [27] | 2018 | Nave Bayes, RBF  Network, and J48 |
| **8** | Begum, Amina, and A. Parkavi | [4] | 2019 | Nave Bayes, Decision Tree, MLP, and RBF |
| **9** | Maysanjaya, I. Md Dendi, Hanung Adi Nugroho, and Noor Akhmad Setiawan | [28] | 2015 | RBF, MLP, BPA |
| **10** | Govindarajan, Priya, et al. | [29] | 2020 | neural networks |

# DECISION TREE

* A decision tree is a sort of tree system that builds classification and regression models.
* It cuts down a dataset into smaller and smaller sections while creating a decision tree in the background.
* The outcome is a tree having both decision and leaf nodes. In a decision node (for example, Outlook), there are two or three divisions (e.g., Sunny, Overcast, and Rainy).
* The leaf node represents a categorization or judgment (for example, Play).
* The root node in a tree corresponds to the most decisive indication. Both category and numerical data may be handled using decision trees. The decision tree system is based on a decision-boosting computer tested for forecasting energy use and finding the most significant predictors of service using a tree-based methodology.
* I utilized the decision tree methodology that was provided to me to do this.
* The model must fit hundreds of trees to develop, each grown use information from the preceding three. Among the tuning parameters are the number of trees, the shrinkage parameter, and the splits in each tree.
* This is one of the numerous types of CARTS that we employed in our research using the Decision Trees algorithm. CART (Classification and Regression Trees) is similar to C4.5, except it cannot generate rule sets and does not allow numerical target variables (regression). CART creates binary trees depending on the function and threshold that deliver the most benefit to each node in terms of knowledge. Used CART. This is a formula.

(𝑄𝑚) = ∑𝑘 𝑝𝑚𝑘 (1 − 𝑝𝑚K)

* Class k is the percentage of its views in node m. If m is a leaf node, then the prediction for this region is set to pmk. Common impurity measures are the following.

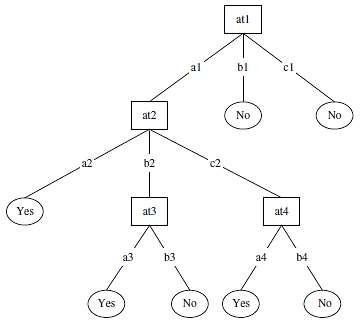


Figure 3.1 Decision tree algorithm structure

* From top to bottom a decision tree is created, starting with the root node, which involves dividing the data in the form of subsets of instances with identical (homogeneous) values.
* The homogeneity of the sample in the ID3 algorithm is calculated using entropy. The entropy of a wholly homogeneous model is negative, whereas the entropy of an evenly distributed example is one. To build a decision tree, we must use frequency tables to quantify two forms of entropy:
* Entropy utilizing one attribute's frequency table:

E= (S) =∑𝐶 log2 𝑃𝑖

𝑖=1

* Entropy using the frequency table of two attributes:
* E= (T,S) =∑𝐶
* 𝑃(𝐶) 𝐸(𝐶)

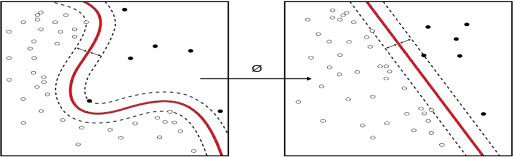
E= (T,S) =∑𝐶

𝑐∈𝑋

𝑃(𝐶) 𝐸(𝐶)

# NAIVE BAYES

* Naive Bayes can compare multiple generalized additive models featuring the output variables of subset selection, variables with the most potent relative influence in the classifying, and a mix of variables from both.
* It compared the prediction accuracy of each best model to compare them directly.
* It narrowed down the relationships between single predictors and the response by fitting naïve Bayes with various splines, 2nd- degree polynomials, and linear predictor variables.
* More polynomials and splines were applied to predictors proved to have nonlinear relationships with our response variable. Among the many types of Naive Bayes algorithm, we dealt with the quality of Gaussian Naive Bayes in this study.
* A normal distribution is also known as a Gaussian distribution.
* We should use it while all of the functions are constant.
* The benefit of this distribution is that you need to measure the mean and standard deviation when working with Gaussian results



Naïve Bayes algorithm on left with respect to support vector machine.

# 

# LOGISTIC REGRESSION

* Logistic regression, employed in the Supervised Learning technique, is one of the most widely used Machine Learning algorithms.
* It's a method for estimating a categorical dependent variable using a group of independent factors.
* Logistic regression is used to predict the contribution of a categorical dependent variable.
* As a result, the final value must be absolute or singleton.
* It might be yes or no, 0 or 1, true or false, and so on, but instead of even integers like 0 and 1, it returns probabilistic values in the middle.
* In terms of application, Logistic Regression is similar to Linear Regression. To solve regression issues, linear regression is employed, whereas logistic regression is utilized to solve classification issues.

# K-NEAREST NEIGHBOR

* For both regression and classification, the supervised learning technique K-nearest neighbors (KNN) is utilized.
* KNN tries to predict the proper class for the test data by evaluating the difference between the test data and all training points.
* Then choose the K number of points that most closely resemble the test findings. The KNN method calculates the chance of test data belonging to any of the 'K' training data groups.
* Then chooses the class with the highest probability.
* The average of the 'K' training points selected determines the significance in regression.
* Unlike the previous approaches, the k-nearest neighbor approach utilizes the data directly for classification rather than first developing a model.
* Consequently, no additional model building is required, and the model's only variable is k, the number of closest neighbors to employ in estimating class membership: the value of p(y/x) is just the ratio of members of classy among the k nearest neighbors
* The simplicity of usage of k-nearest neighbors over other algorithms is one of its main advantages.
* Neighbors can justify the categorization result; this case- based reasoning might be helpful in circumstances when black-box models are insufficient.
* The main drawback of k-nearest neighbors is that [53] Is the case neighborhood computation, which necessitates the definition of a metric that estimates the distance between data items.
* Assume we have two kinds, Category A and Category B, with a new data point x1. Would this data point fit into one of these categories? To address this kind of issue, a K-NN method is required. With the use of K-NN, we can rapidly categorize the type or class of a dataset.

The following algorithm can be used to illustrate how K-NN works:

* + Initially the number of neighbors (K) is determined.
  + Hence the Euclidean distance between the neighbors of K is determined.
  + Then the nearest neighbors K is searched by using the measured Euclidean distance.
  + Then count the number of data points for each group among these k- neighbours.
  + The most recent data points are then assigned to the group that has the most neighbours.finished the model.

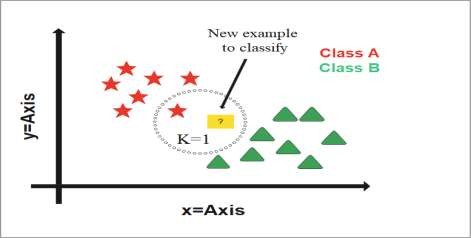


Figure 3.2. K-nearest neighbors (KNN) algorithm.

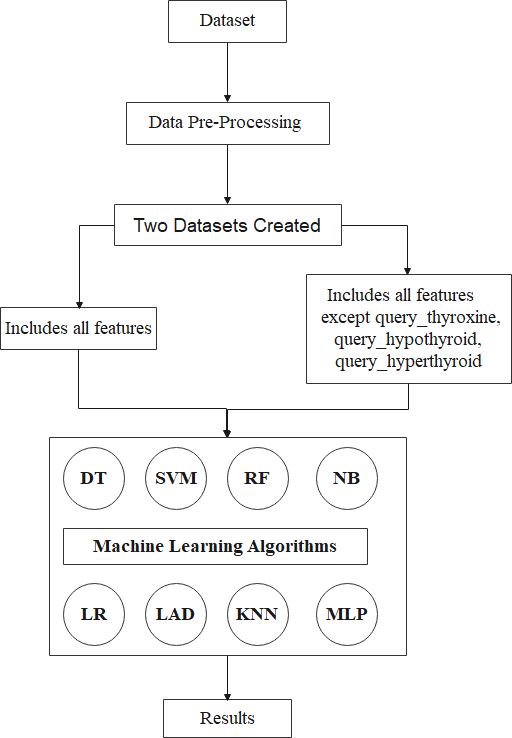
* There are no pre-defined mathematical methods for determining the most advantageous K value to choose the best K. Set a random K value as the starting point and begin computation.
* As K is set to a low value, the judgment boundaries become unpredictable. The higher the K value, the smoother the judgment boundaries get, which is better for classification.

# METHODOLOGY

* In this chapter, the process of discussing the proposed method of work in this study is illustrated in Figure.
* In the first stage, the data set is collected, and after that, the data is prepared by pre-processing as shown in figure. Second, two datasets are created, the first containing all the attributes and the second containing all but three of the attributes, which are query\_thyroxine & query\_hypothyorid & query\_hyperthyroid.
* In the final stage, classification algorithms are used to predict thyroid disease and compare the performance of different classifiers by calculating the accuracy of each classifier.
* In this study, we use eight machine learning techniques represented by (Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), Naive Bayes (NB), Logistic Regression (LR), Linear Discriminant Analysis (LDA), K-Nearest Neighbor (K-NN), and multi-layer perceptron (MLP).

Platform Used

* The development of machine learning algorithms used in this study with the Python programming language (version 3.8.3), data pre-processing was done with panda’s libraries (version 1.0.5), and the developed machine learning algorithms with scikit libraries (version 0.23 .1). The primary platform used for this work is Spyder (Anaconda) on a computer with a 64-bit operating system, x64-based processor, 4.00 GB of RAM, on an Intel (R) Core (TM) i7-7230M CPU.



Flowchart of the study.

# DATA COLLECTION

* We collected a good amount of data on thyroid diseases, and we are working in our study to classify diseases using this data.
* The data used in our study is a data set taken from hospitals and private laboratories specialized in analyzing and diagnosing diseases in Iraq, Baghdad governorate, and the type of data taken related to thyroid diseases. Data were taken on 1,250 male and female individuals whose ages range from one year to Ninety years as these samples contain thyroid disease subjects with hyperthyroidism and hypothyroidism and normal subjects without thyroid disease.
* We collected data over one to four months, and the main objective of data collection was to classify thyroid diseases using machine learning algorithms.

shows the features contained in the dataset.

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Attribute Name** | **Value Type** | **Clarification** |
| 1 | Id | number | 1,2,3…etc. |
| 2 | Age | number | 1,10,20, 50…etc. |
| 3 | Gender | 1,0 | 1=m,0=f |
| 4 | query\_thyroxine | 1,0 | 1=yes,0=no |
| 5 | on\_antithyroid\_medication | 1,0 | 1=yes,0=no |
| 6 | Sick | 1,0 | 1=yes,0=no |
| 7 | Pregnant | 1,0 | 1=yes,0=no |
| 8 | thyroid\_surgery | 1,0 | 1=yes,0=no |
| 9 | query\_hypothyorid | 1,0 | 1=yes,0=no |
| 10 | query\_hyperthyroid | 1,0 | 1=yes,0=no |
| 11 | TSH measured | 1,0 | 1=yes,0=no |
| 12 | TSH | Analysis ratio | Numeric value |
| 13 | T3 measured | 1,0 | 1=yes,0=no |
| 14 | T3 | Analysis ratio | Numeric value |
| 15 | T4 measured | 1,0 | 1=yes,0=no |
| 16 | T4 | Analysis ratio | Numeric value |
| 17 | Category | 0,1,2 | 0=normal,1=hypothyroi d,  2=hyperthyroid |

# DATA PRE-PROCESSING

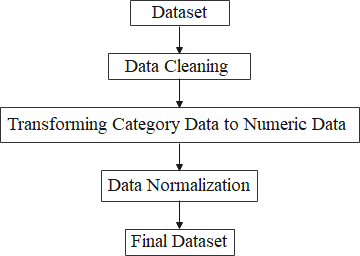
* The data pre-processing process is critical and is an essential step in data mining; it has a good effect on the data.
* The data pre-processing process is used to discover the data through data analysis and find the missing data; it scrutinizes the data.
* The pre- processing process involves data cleaning, preparation, etc., because real-world data tends to be inconsistent, noisy, and may contain missing, redundant, and irrelevant data. This may adversely affect the functioning of the algorithms.
* Pre-processing is used to clean the data, scale the data, and convert the data into a format compatible with the algorithms used.

The pre-treatment stage in this work goes through the following stage:

* + - Check for missing values and if found, process them.
    - Check for duplicate data if you find and remove it.
    - Convert all categorical features in the data set to a standardized numerical representation because of the presence of categorical and numerical features.

In the data processing stage, we identified missing data, and among these properties were missing T4 by number 151 and T3 by number 112. where we were able to Process this lost data by replacing it with the value of the mediator, and after working in this way, we were able to obtain the data in a good and better way and free from lost data, as the data became arranged and excellent and free from any defect or problem so that we can work on it smoothly and well.

* Data Normalization may refer to more extensive alterations aiming at harmonizing the whole probability distributions of transformed data in more complicated settings.
* Before averaging, normalization of ratings requires translating data from many scales to a single, apparently typical scale.
* In this study, data normalization was used with the MLP algorithm only using the Min-Max normalization method.
* After applying pre-processing data, the final data set after pre-processing consists of 1250 rows and 17 columns and does not contain any null or redundant values. All features are represented in numerical form.



Data Pre-processing.

# PERFORMANCE MEASUREMENT

In our study, we relied on evaluating the eight machine learning techniques using the following metrics:

* Accuracy
* specificity
* sensitivity
* precision
* F-score
* and recall as shown in table (4.2).
* Subjects without thyroid disease were considered false positive, and people with thyroid disease were both hypothyroid and hyperthyroid as true negative.

The formulae statistics to evaluate the predictive capability.

**Measure Formula Description**

**accuracy** 𝑡𝑝 + 𝑡𝑛

𝑡𝑝 + 𝑡𝑛 + 𝑓𝑝 + 𝑓𝑛

The percentage of successfully identified observations compared to the total number of classifications made by the classifier.

**precision** 𝑡𝑝

𝑓𝑝 + 𝑡𝑝

The percentage of successfully categorized bad loans compared to all bad loans categorized by the classifier.

**specificity**

𝑡𝑛

𝑓𝑝 + 𝑡𝑛

The proportion of successfully identified excellent loans in the test dataset compared to the total number of excellent

loans.

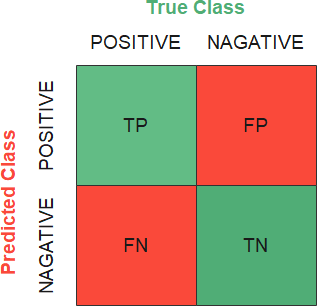
**sensitivity** 𝑡𝑝

𝑡𝑝 + 𝑓𝑛

The percentage of properly categorized bad loans in the test dataset compared to the total number of bad loan

# CONFUSION MATRIX

* It is the confusion matrix.
* The matrix is used to evaluate the effectiveness of the classification model and is N × N.
* The number of target groups, where N is equal to the total number of target groups.
* The matrix compares the actual target values with the predictions of the machine learning algorithm.
* Its use gives us a clear view of how well our rating model is performing as well as the errors it results in. True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) are the four types of false positive (FN). Figure 1 depicts the confusion matrix (4.4) [57].



Confusion Matrix.

# IMPLEMENTING MACHINE LEARNING ALGORITHMS

At this point, we will implement our dataset on the eight machine learning algorithms. Since we created the first dataset with all 17 attributes and then applied it to the algorithms, this is the first step.

After that, we created a second dataset containing all the attributes except for three points (query\_thyroxine&query\_hypothyorid& query\_hyperthyroid), and we applied it to the eight algorithms.

# IMPLEMENTING DECISION TREE (DT) ALGORITHM

In the first step we apply our data with all attributes to DT algorithm using the following parameters set as (criterion='gini', max\_features=” auto”, random\_state=0, max\_leaf\_nodes=9, max\_depth=6).

In the second step, we applied the data with the omission of three attributes that used the same parameters. The following parameters are:

criteria: The primary function or assigned to it is to measure the quality (disaggregation).

* Criterion: ("Geni," "Entropy") The function it performs is to measure the quality of the split.
* max\_features: The job assigned to it is to know and consider the advantages of searching for the best division.
* random\_state: The process set to control the randomness of the amount since the characteristics often change randomly for each division.
* max\_leaf\_nodes: In this parameter, a tree with (max\_leaf\_nodes) grows in the best way.
* max\_depth: This parameter sets the funnel for the tree.

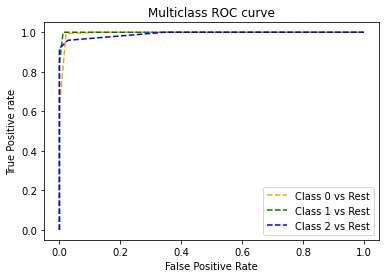


Figure 4.6.1. Roc curve for performing DT algorithm with all features.

# IMPLEMENTING SVM ALGORITHM

In the first stage, the following parameters are used with the SVM algorithm (kernel='poly',degree=4,gamma='scale',coef0=3,shrinking=False,probability=False,d ecision\_function\_shape='ovo',random\_state=0).

The following parameters are:

* the kernel: This parameter specifies the type of kernel used in this algorithm.
* degree: A chasm means the degree denoting the polynuclear ("poly")
* gamma: huh means kernel modulus, and they are all of "RBF," "poly," and "sigmoid."
* coef0: It is an independent term in the kernel function.
* shrinking: mean, will contraction inference be used or not
* probability: This means you will enable estimates (probabilities).
* decision\_function\_shape: Huh, will it return one decision function against the rest ('over) of the shape (n\_samples, n\_classes) or not
* random\_state: It controls the generation of pseudorandom numbers for data shuffling and for probability estimations.

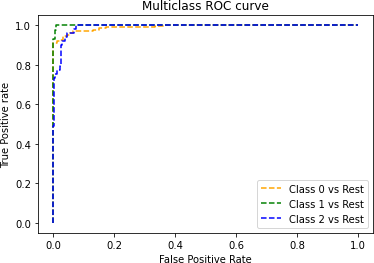


Figure 4.6.2. Roc curve for performing SVM algorithm with all features.

# IMPLEMENTING RANDOM FOREST ALGORITHM

A random forest is a meta estimator that employs averaging to increase predicted accuracy and control over-fitting by fitting multiple decision tree classifiers on different sub-samples of the dataset.

In the first stage, the following parameters are used with the Random Forest algorithm, and with the second stape, the same parameters are used (random\_state=0, max\_depth=5).

The following parameters are:

* random\_state: The primary cycle is to control the random take-off of the samples used at the beginning of the tree.
* max\_depth: It is the maximum depth that the tree can reach.

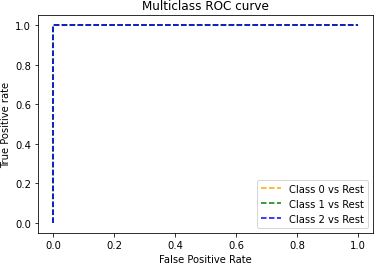


Figure 4.6.3. Roc curve for performing RF algorithm with all features.

# IMPLEMENTING (NB) ALGORITHM

In the first stage, are used the following parameters with the NB algorithm (priors, var\_smoothing).

In the second stage, the same parameters are used in the appropriate location. The following parameters are:

* Priors: Here, indicate the last possibilities for the categories.
* var\_smoothing: This means the enormous variance part of all the features added to the variances, and the aim is to stabilize the account.

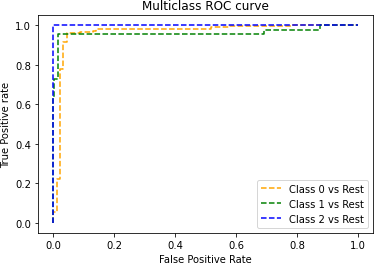


Figure 4.6.4. Roc curve for performing NB algorithm with all features.

# IMPLEMENTING LOGISTIC REGRESSION ALGORITHM

The data is applied to the LR algorithm with the following parameters in each of the two steps (solver='liblinear', multi\_class='ovr').

The following parameters are:

* Solver: It means that the algorithm will be used in the (optimization) problem.
* multi\_class: When it is 'over,' the classification is appropriate because the problem will be binary.

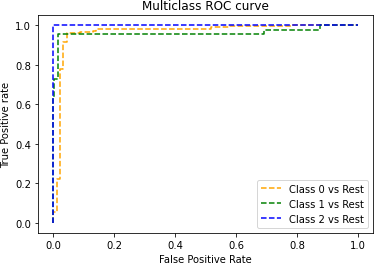


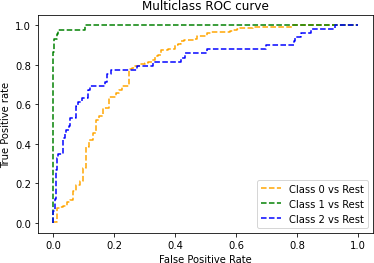
Figure 4.6.5. Roc curve for performing LG algorithm with all features.

# IMPLEMENTING (LDA) ALGORITHM

* A linear decision boundary classifier is created by fitting class conditional densities to data and using Bayes' rule.
* The model gives each class a Gaussian density, which assumes that all levels have the same covariance matrix.
* Using the transform approach, it can use the appropriate model to minimize the dimensionality of the input by projecting it in the most distinctive aspects.
* In first stage LDA parameters set as (solver='eigen',shrinkage='auto', store\_covariance=True). The second stage is using the same parameters in the first stage.

The following parameters are:

* solver: Its primary role is to analyze the intrinsic value
* shrinkage: spontaneous contraction is determined using (Ledoit-Wolf lemma).
* store\_covariance: If true, the covariance matrix is explicitly calculated.



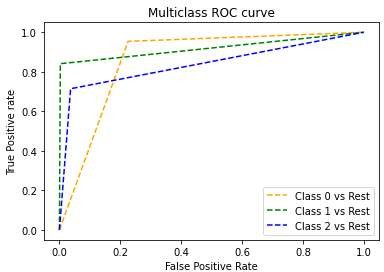
Roc curve for performing LDA algorithm with all features.

# IMPLEMENTING (KNN) ALGORITHM

* In the first stage, we apply the data to the KKN algorithm using parameters (n\_neighbors = 1, p = 1). The same parameters will also be used in the second step.

The following parameters are:

* n\_neighbors: where the number of neighbors per sample.
* p: It is the energy parameters of the Minkowski scale.



Roc curve for performing KNN algorithm with all features.

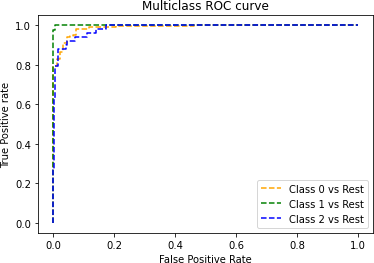
# IMPLEMENTING (MLP) ALGORITHM

In the first and second step of applying the data to the MLP algorithm, the following parameters are used (activation='identity', alpha=0.1,

batch\_size=min (200,600),max\_iter=100000,hidden\_layer\_sizes=(10,2)). Each parameter has its own function.

The following parameters are:

* Activation: A chasm indicates the activation function in the hidden layer.
* alpha: where denotes L2's penalty coefficient (regulation term).
* batch\_size: and the minibatch size gap (for random optimizers).
* max\_iter: It indicates the maximum number of (repetitions).
* hidden\_layer\_sizes: The I element represents the number of neurons in hidden layer i.
* In this parameter, (i) represents the number of neurons that make up the hidden layer in MLP.



Roc curve for performing MLP algorithm with all features.

# 3.RESULT

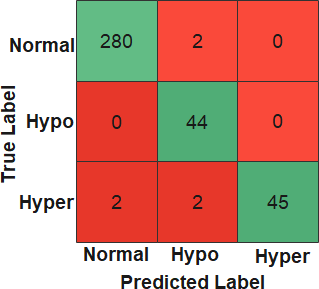
* Thyroid illness is estimated to affect more than 200 million individuals worldwide.
* It is also widespread among Iraqis, particularly women. This condition has the potential to have a significant impact on adult bodily processes as well as infant development.
* Thyroid diseases are generally treatable, but they may be dangerous if they progress to the advanced stage, and they may even result in death.
* The successful classification was performed using various techniques using the thyroid disease data set obtained from private laboratories and hospitals.
* We used eight algorithms in this study (Decision Tree (DT), Supporting Vector Machine (SVM), Random Forest (RF), Naive Bayes (NB), Logistic Regression (LR), Linear Discriminant Analysis (LDA), K- Nearest Neighbor (K-NN), Multilayer (MLP)).
* We divided the existing data into two parts, 30% for training and 70% for testing, as this training is the first training on this data.
* In the first step, we took all the properties in our data and applied them to the eight algorithms.
* We take all the features in our data in the second step except for query\_thyroxine, query\_hypothyorid, and query\_hyperthyroid and they ared used to predict using eight algorithms.
* Our data contains 16 attributes as one input and one output. That is, the sum of the attributes in this data is 17 attributes.

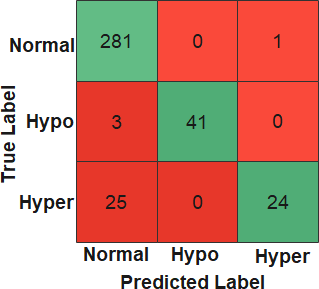
# RESULTS IN THE FIRST STEP WITH ALL FEATURES IN DATASET

* In this section, the predicted thyroid disease by taking the data set with all the features for thyroid disease classification.
* An implemented classification algorithm on this data set and the accuracy achieved by each algorithm was calculated and compared the achieved results.
* Shows us the accuracy of each algorithm, as it received an algorithm Decision Tree 98.4% accuracy SVM 92.27% accuracy Random Forest 98.93% accuracy Naive Bayes 81.33% accuracy Logistic Regression 91.47% accuracy Linear Discriminant Analysis 83.2% accuracy KNeighbors Classifier 90.93% accuracy and MLP (NN) 97.6% accuracy and through these results.
* The Random Forest algorithm is very accurate, and then the algorithm follows the decision tree and other algorithms.
* The Random Forest algorithm enables the rapid discovery of essential data from large databases.
* The main advantage of Random Forest is that it relies on a range of different decisions to arrive at the answer. Therefore, she was able to obtain the highest accuracy.

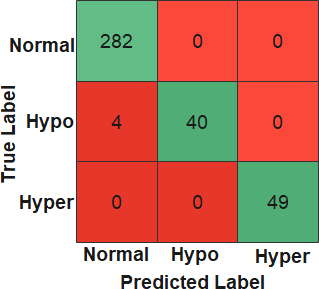
Evaluation measurements for classification models with all features**.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No** | **Algorithms** | **Accuracy (%)** | **Sensitivity (%)** | **Specifity (%)** |
| **1** | Decision Tree | 98.40 | 99.29 | 97.84 |
| **2** | SVM | 92.27 | 90.93 | 98.48 |
| **3** | Random Forest | 98.93 | 98.60 | 100 |
| **4** | Naive Bayes | 81.33 | 98.63 | 57.69 |
| **5** | Logistic Regression | 91.47 | 98.80 | 100 |
| **6** | Linear Discriminant  Analysis | 83.20 | 82.49 | 89.47 |
| **7** | KNeighbors Classifier | 90.93 | 92.75 | 84.70 |
| **8** | MLP | 96.00 | 94.93 | 98.75 |

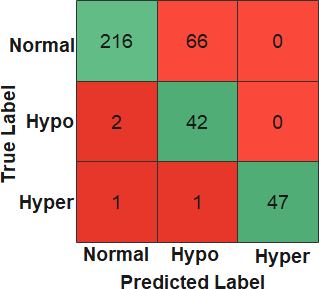


 Confusion Matrix for Decision Tree Classifier.

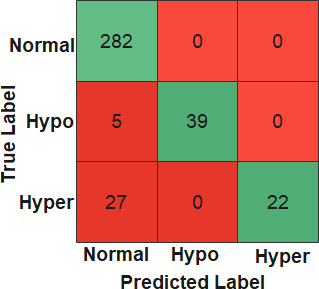
Confusion Matrix for SVM Classifier.



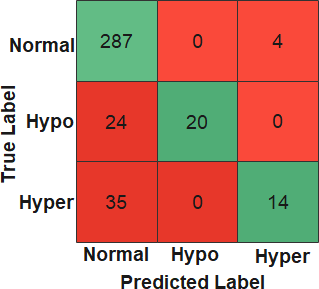
Confusion Matrix for Random Forest Classifier.



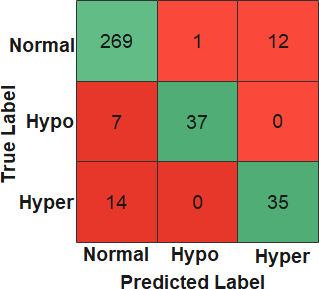
Confusion Matrix for Naive Bayes Classifier.



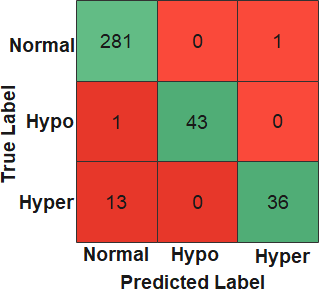
Confusion Matrix for Logistic Regression Classifier.



Confusion Matrix for Linear Discriminant Analysis Classifier.



Confusion Matrix for KNeighbors Classifier.



Confusion Matrix for MLP Classifier.

# RESULTS IN THE SECOND STEP WITH OUT THREE FEATURES

* In the second step, we removed three traits based on a previous study Ioniţă, Irina, and Liviu Ioniţă [58].
* The deleted attributes were both query\_thyroxine&query\_hypothyorid& query\_hyperthyroid. After deleting these attributes, we applied our data also to the algorithm group, and by using the Python script, we were able to obtain these [results

Algorithm Decision Tree 90.13%

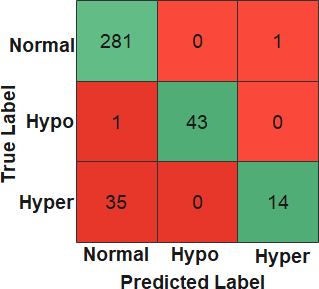
accuracy SVM 92.53% accuracy Random Forest 91.2% accuracy Naive Bayes 90.67% accuracy Logistic Regression 91.73% accuracy Linear Discriminant Analysis 83.20%accuracy KNeighbors Classifier 91.47% accuracy and MLP(NN) 95.73% accuracy and through these results.

As it seems to us that the Naive Bayes algorithm has a high accuracy of 90.67 after the three traits have been omitted, the SVM algorithm, the logistic regression algorithm, and the neighbor's Classifier algorithm have increased slightly and reduced the accuracy of the other algorithms.

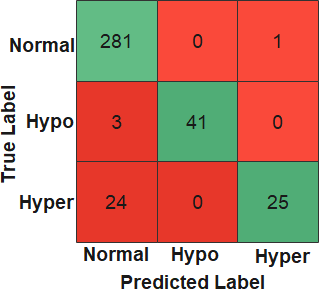
* We show here that the accuracy of algorithms used in our data changes with the change of the characteristics used in the data.
* Experience had demonstrated this evident change; algorithms were obtained when three of the traits were deleted, as the accuracy of some algorithms decreased, and some increased.
* The MLP algorithm received the highest accuracy after deleting three attributes, as its accuracy reached 95.73%.

Evaluation measures of the classification models without three features.

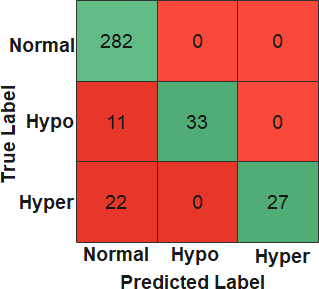
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **No** | **Algorithms** | **Accuracy (%)** | **Sensitivity (%)** | **Specifity (%)** |
| **1** | Decision Tree | 90.13 | 88.64 | 98.27 |
| **2** | SVM | 92.53 | 91.23 | 98.50 |
| **3** | Random Forest | 91.20 | 89.52 | 100 |
| **4** | Naive Bayes | 90.67 | 99.60 | 73.60 |
| **5** | Logistic Regression | 91.73 | 90.00 | 100 |
| **6** | Linear Discriminant  Analysis | 83.20 | 82.49 | 89.47 |
| **7** | KNeighbors Classifier | 91.47 | 93.10 | 85.88 |
| **8** | MLP | 95.73 | 94.93 | 98.73 |



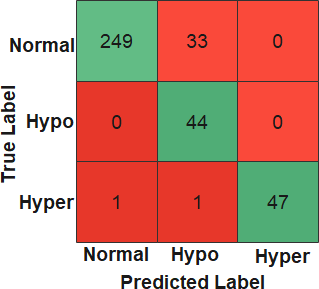
Confusion Matrix for Decision Tree Classifier without three attributes.



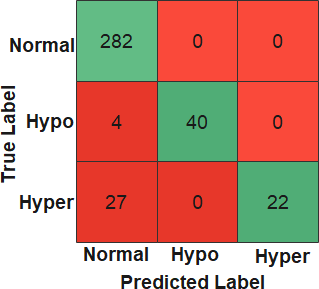
Confusion Matrix for SVM Classifier without three attributes.



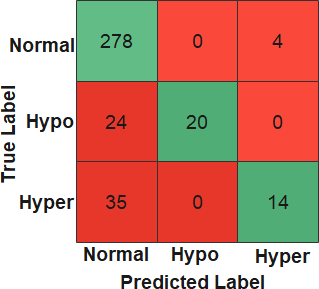
Confusion Matrix for Random Forest Classifier without three attributes.



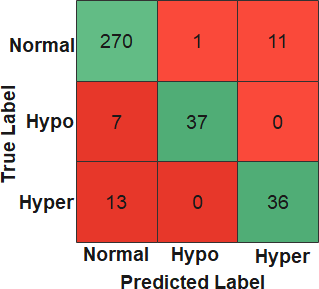
Confusion Matrix for Naive Bayes Classifier without three attributes.



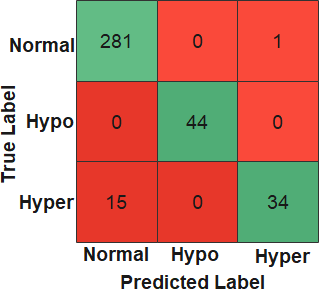
Confusion Matrix for LG Classifier without three attributes.



Confusion Matrix for LDA Classifier without three attributes.



Confusion Matrix for KNN Classifier without three attributes.



Confusion Matrix for MLP Classifier without three attributes.

**4.TRAILHEAD PROJECT PUBLIC URL**

# Team Lead - <https://trailblazer.me/id/hgahlot3>

# 

# Team member 1-

# 

# Team member 2-

# 

# Team member 3-

**5.1 ADVANTAGES**

* The ML algorithms can approach in one of two ways.
* The first is what It is known as directed learning.
* The field of supervised learning entails the use of specific examples to train algorithms.
* The computer is given a set of inputs and a set of correct outputs, and it learns by comparing empirical outcomes to the right results to classify errors [34]. When past events can be used to forecast future events, this method of learning is used.
* Since the aim is always to get the machine to understand a classification scheme that we have developed, supervised learning is popular in classification problems.
* Once again, digit identification is a typical example of classification learning. Classification learning is helpful for any issue where deducing a classification is usable, and the category is simple to evaluate.
* In other instances, assigning preset categories to each example of a dilemma might not be appropriate if the agent can figure out the classifications independently .

# 5.2 DISADVANTAGE

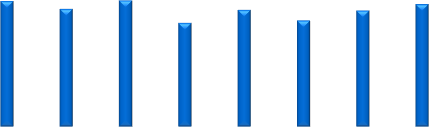
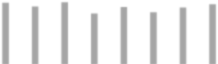
* It instruction is the other choice.
* The computer must explore the data and construct some pattern or structure under this approach; it must create models from scratch and not give any correct information.
* The results Outliers are often identified and distinguished using this approach [36]. Unsupervised learning seems much harder: the goal is to have the computer learn how to do something that we don't tell it how to do. Unsupervised learning has yielded a slew of milestones, including world-champion backgammon projects and even self-driving vehicles
* When assigning values to behavior is easy, it can be an effective technique

**6.APPLICATIONS**

* Today the health field needs solid support to help overcome many obstacles. Among the vast amount of clinical data, it becomes tough to work with.
* We note the excellent and effective progress of artificial intelligence techniques in the medical field, especially machine learning algorithms, and their valuable and prominent role in classifying, predicting, or diagnosing diseases

Our study deals with machine learning algorithms to classify thyroid disease. These algorithms are (Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), Naive Bayes (NB), Logistic Regression (LR), Linear Discriminant Analysis (LDA), K-Nearest Neighbor (K-NN), and multi-layer perceptron (MLP), as well as comparison of algorithm results.

* After arranging and cleaning the dataset, we created two datasets with all the attributes, and the second, we erased three of them, which are (query\_thyroxine&query\_hypothyorid& query\_hyperthyroid).
* With the first step, the Random Forest algorithm obtained 98.93% accuracy, followed by other algorithms.
* In the second step, the MLP algorithm received 95.73% accuracy, followed by the different algorithms.
* With this step also, the accuracy of the Naive Bayes algorithm increased and obtained 90.67% after its accuracy in the first step was 81.33%. The KNN algorithm increased to 91.47% from 90.93%, the accuracy of some other algorithms increased slightly, the accuracy of some of the algorithms decreased.



120

100

80

60

40

20

0

98.40

92.27

98.93

91.47

90.…

96.00

81.33

83.20

DT

SVM

RF

NB

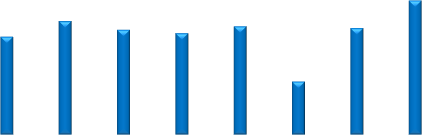
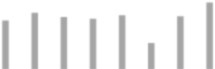
LR

LDA KNN MLP

Accuracy (%)

Evaluation measurements for classification models with all features.

* It indicates that the three characteristics had a negative performance on the results of algorithms whose accuracy increased after deletion, while they had a good effect with algorithms whose accuracy decreased after deletion; this indicates that these properties had an essential and valuable role with these algorithms for diagnosing disease.
* In each of the two steps, we reached excellent and effective results in analyzing and classifying thyroid disease to contribute to helping the health system and its workers.



100

95

90

85

80

75

95.73

92.53

90.13

91.20 90.67

91.73

91.47

83.2

DT SVM RF NB LR LDA KNN MLP

Accuracy (%)

Evaluation measures of the classification models without three features.

# 

# 7.CONCLUTION

* The study showed that the could diagnose thyroid functional disease with we use eight machine learning techniques.
* We were represented by (Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), Naive Bayes (NB), Logistic Regression (LR), Linear Discriminant Analysis (LDA), K-Nearest Neighbor (K-NN), and multi-layer perceptron (MLP)).
* This study assists to doctors and medical staff in the healthcare field. We were also able to compare the eight algorithms and which one could reach the best accuracy.

# 8. FUTURE SCOPE

* We predicted and classified thyroid disease by applying machine learning techniques to a data set consisting of 1250 actual samples. We divided the dataset as follows: 30% of the data were used for training, and 70% were used for testing.
* After applying these techniques to dataset one that consists of all the characteristics, the random forest algorithm obtained an accuracy rate of 98.93%. In the second step, and based on a previous study, we deleted a set of features which are 1- query\_thyroxine 2- query\_hypothyorid 3-query\_hyperthyroid.
* We applied machine learning techniques to this data, and the MLP algorithm got the highest accuracy of 95.73%. The results obtained in this study help us in the rapid prediction of thyroid disease. And the classification of the disease (Hyperthyroidism or Hypothyroidism).
* Future work should focus on improving the performance of classification algorithms and using different approaches from feature selection methods to obtain better results.