# Thyroid Disease Classification Using Machine Learning Algorithms

# Thyroid Disease Classification Using Machine Learning Algorithms

**Khalid salman**[1*], **Emrullah Sonuç**[2]

[1] Karabuk University, Computer Engineering 1.
[2] Karabuk, Turkey, Computer Engineering 2.

Email: [1]Khalidabdalstar20@gmail.com, [2]esonuc@karabuk.edu.tr

**Abstract.** With the vast amount of data and information difficult to deal with, especially in the health system, machine learning algorithms and data mining techniques have an important role in dealing with data. In our study, we used machine learning algorithms with thyroid disease. The goal of this study is to categorize thyroid disease into three categories: hyperthyroidism, hypothyroidism, and normal, so we worked on this study using data from Iraqi people, some of whom have an overactive thyroid gland and others who have hypothyroidism, so we used all of the algorithms. Support vector machines, random forest, decision tree, naïve bayes, logistic regression, k-nearest neighbors, multi-layer perceptron (MLP), linear discriminant analysis. To classification of thyroid disease.
**Keywords:** Machine learning, classification model, Thyroid diseases, Support vector machines, Random forest, Decision tree, Naïve bayes, logistic regression, K-nearest neighbors, Multi-layer perceptron (MLP), Linear discriminant analysis.
**Keywords:** Machine learning, classification model, Thyroid diseases, Support vector machines, Random forest, Decision tree, Naïve bayes, logistic regression, K-nearest neighbors, Multi-layer perceptron (MLP), Linear discriminant analysis.

## 1. Introduction

Thyroid disease is a subset of endocrinology which is one of the most misunderstood and undiagnosed diseases [1] [2].

Thyroid gland diseases are among the most prevalent endocrine disorders in the world, second only to diabetes, according to the World Health Organization. Hyper function hyperthyroidism and hypothyroidism affect about 2% and 1% of individuals, respectively. Men have about a tenth of the prevalence of women. Hyper-and hypothyroidism may be caused by thyroid gland dysfunction, secondary to pituitary gland failure, or tertiary to hypothalamic malfunction. Due to dietary iodine deficiency, goiter or active thyroid nodules may become prevalent in some regions, with a prevalence of up to 15%. The thyroid gland can also be the location of different kinds of tumors and can be a dangerous place where endogenous antibodies wreak havoc (autoantibodies) [3].

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 raises the odds of good treatment. Despite multiple trials, clinical diagnosis is often thought to be a difficult task [4].

The thyroid gland is a butterfly-shaped gland situated at the base of the throat. It comprises two active thyroid hormones, levothyroxine (T4) and triiodothyronine (T3), which are involved in brain functions such as body temperature control, blood pressure management, and heart rate regulation. Likewise, thyroid

disease is one of the most prevalent diseases worldwide, and it is mostly caused by a deficiency of iodine, but it may also be caused by other factors. The thyroid gland is an endocrine gland that secretes hormones and passes them through the bloodstream. It is situated in the middle of the front of the body. Thyroid gland hormones are responsible for aiding in digestion as well as maintaining the body moist, balanced, and so on. Thyroid gland treatments such as T3 (triiodothyronine), T4 (thyroid hormone), and TSH (thyroid stimulating hormone) are used to assess thyroid activity (thyroid stimulating hormone). Thyroid disorder is classified into two types: hypothyroidism and hyperthyroidism. Data mining [5] is a semi-automated method of looking for correlations in massive datasets.

Machine learning algorithms are one of the best solutions to many problems that are difficult to solve [6] Classification is a data extraction technique (machine learning) used to predict and identify many diseases, such as thyroid disease, which 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 [7]. Although the application of computer learning and artificial intelligence in medicine dates back to the early days of the field [8], 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 in the near future [9].

Hyperthyroidism is a disorder in which the thyroid gland releases so many thyroid hormones. Hyperthyroidism is caused by an increase in thyroid hormone levels [10]. Dry skin, elevated temperature sensitivity, hair thinning, weight loss, increased heart rate, high blood pressure, heavy sweating, neck enlargement, nervousness, menstrual cycles shortening, irregular stomach movements, and hands shaking are some of the signs [11]. Hypothyroidism is a condition in which the thyroid gland is underactive

Hypothyroidism is caused by a decline in thyroid hormone production. 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 are some of the symptoms. If not treated, these symptoms can escalate over time [12].

## 2. literature Review

Chandel, Khushboo [13] Thyroid disorder is classified using different classification models based on parameters such as TSH, T4U, and goiter in this study. Several grouping methods, such as K-nearest neighbor, are used to justify this argument. The Naive Bayes and support vector machines algorithms are employed. The experiment was carried out using the Rapid miner instrument, 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 diagnosing a disease. KNN and Naive Bayes classifiers were used in this study. The Rapidminer tool is used to compare these two classifiers. The findings revealed that the K-nearest neighbor classifier is the most reliable, with a 93.44 percent accuracy, while the Naive Bayes classifier has a 22.56 percent accuracy. The proposed KNN technique improves classification accuracy, which contributes to improved results. As a result, Naive Bayes can only have a linear, elliptic, or parabolic decision boundary, so the decision boundary consistency of KNN is a huge plus. KNN outperforms most methods since the factors are interdependent.

Banu, G. Rasitha [14] Thyroid disease is one of the most common illnesses that humans suffer from. The hypothyroid data used in this study came from the data repository at the University of California, Irvine (UCI). The platform Waikato Environment of Information Analysis will be used for the whole research project (WEKA). The J48 technique was found 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, a number of data mining methods are used. In this analysis, we used dimensionality reduction to pick a subset of attributes from the original results, and we used J48 and decision stump data mining classification techniques to define hypothyroidism. The uncertainty matrix is used to assess classifier output in terms of

precision and error rate. The J48 Algorithm has 99.58 percent accuracy, which is higher than decision stump tree accuracy, and it also has a smaller error rate than Decision stump.

Umar Sidiq, Dr, Syed Mutahar Aaqib, and Rafi Ahmad Khan [15] Classification, which is used to characterize predefined data sets, is one of the most popular supervised learning data mining techniques. In the healthcare sector, the classification is commonly used 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 Nave bayes may be used. The Judgment Tree has the greatest accuracy of the other classes, at 98.89 percent.

Sindhya, Mrs K [16] 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 prediction takes very little time. In this study, I used classification algorithms to analyze thyroid data and came up with a result. A model's efficacy is primarily determined by two factors. 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. So we can assume that J48 is the best model for hypothyroid prediction since its accuracy is 99 percent, which is among the highest, and it takes 0.2 seconds to run, which is significantly less time than the Random Forest model.

AKGÜL, Göksu, et al [17] The aim of this study is to propose 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. The logical conclusion It was determined if the new samples were hypothyroid using data from the UCI machine learning database, which included 3163 samples, 151 of which were hypothyroid and the others were hypothyroid. Different sampling techniques were used in the data collection to eradicate the unbalanced distribution, and 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 percent, the region under the curve was 93.2 percent, and the Matthews correlation coefficient was 81.8 percent for this analysis, which was trained on the data set using over-sampling techniques.

VijiyaKumar, K., et al [18] The aim of this paper is to create a method that can 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 that is 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.

Chaurasia, Vikas, Saurabh Pal, and B. B. Tiwari [19] After all other cancers, breast cancer is the second most common cancer in women. The aim of this research paper is to provide a breast cancer study that incorporates cutting-edge techniques. Improving breast cancer survivability modeling models by incorporating recent research advances. We used a broad dataset and three common data mining algorithms (Nave Bayes, RBF Network, and J48) to construct prediction models (683 breast cancer cases). For accuracy comparison, we used 10-fold cross-validation approaches to measure the unbiased estimation of the three prediction models. The findings suggest that the Bay is a safe place to visit (based on an average precision Breast Cancer dataset). The RBF Network is the second-best predictor, with 93.41 percent accuracy on the holdout sample (better than any other prediction accuracy reported in the literature), and Nave Bayes is the third-best predictor, with 97.36 percent accuracy on the holdout sample (better than any other prediction accuracy reported in the literature) (better than any other prediction accuracy published in

the literature). In this study, we evaluated three breast cancer survivability prediction models using two criteria: benign and malignant cancer cases.

Begum, Amina, and A. Parkavi [20] The most recent research focuses on thyroid disease classification of two of the most frequent thyroid dysfunctions in the general population (hyperthyroidism and hypothyroidism). The researchers looked at and compared four different classification models: Naive Bayes, Decision Trees, Multilayer Perceptrons, and Radial Basis Function Networks. The findings reveal that all of the classification models listed above have a high degree of accuracy, with the Decision Tree model having the highest classification score. The classifier was built and validated using data from a Romanian data website and the UCI machine learning repository. KNIME Analytics Platform and Weka are two data sets. Data mining techniques were used as the foundation for developing and testing the classification models. A variety of studies in the field of thyroid classification use various data mining techniques to construct robust classifiers, according to the literature. The authors of this research explored the use of four classification models on thyroid data (Nave Bayes, Decision Tree, MLP, and RBF Network) to help classify thyroid dysfunctions such as hyperthyroidism and hypothyroidism. In all of the cases that were tested, the decision tree model was the correct classification model.

**Table** 1: shows the literature review and the algorithms used and their accuracy.

| Study number | Authors | Reference | year | Algorithms | Accuracy |
|---|---|---|---|---|---|
| 1 | Chandel, Khushboo | [13] | 2016 | KNN, Naive Bayes | KNN 93.44, Naive Bayes 22.56 |
| 2 | Banu, G. Rasitha | [14] | 2016 | J48 | J48 99.85 |
| 3 | Umar Sidiq, Dr, Syed Mutahar Aaqib, and Rafi Ahmad Khan | [15] | 2019 | k nearest neighbors, Support vector machine, Decision tree, and Nave bayes | Nave bayes 98.89, SVM 96.30, KNN 98.89 |
| 4 | Sindhya, Mrs K | [16] | 2020 | Nave bayes,J48 and Random Forest | J48 99 , Random Forest 99.3, Nave bayes 95 |
| 5 | AKGÜL, Göksu, et al | [17] | 2020 | k nearest neighbors and SVM | k nearest neighbors 92, SVM 97.8 |
| 6 | VijiyaKumar, K., et al | [18] | 2019 | Random Forest | the results revealed that the prediction system is capable of correctly |
| 7 | Chaurasia, Vikas, Saurabh Pal, and B. B. Tiwari | [19] | 2018 | Nave Bayes, RBF Network, and J48 | J48 93.41, Nave Bayes 97.36, RBF Network 96.77 |
| 8 | Begum, Amina, and A. Parkavi | [20] | 2019 | Nave Bayes, Decision Tree, MLP, and RBF Network | Nave Bayes 91.63, Decision Tree 96.91, MLP 95.15, and RBF Network 96.03 |

## 3. Methodology

### *3.1. Data Collection*

Machine learning algorithms are used in the rapid and early diagnosis of thyroid diseases and other diseases, as they now in a significant position in the health field and help us in diagnosing and classifying diseases for this reason we were able to collect a good amount of data on thyroid diseases and we are working in our study on the classification of diseases using this data The data that I used in our study is a set of data taken from external hospitals and laboratories specialized in analyzing and diagnosing diseases, and the sample

taken from the data is the data of the Iraqi people and the type of data taken related to thyroid disease, where data were taken on 1250 people between males and females, and their ages range from 1 year to 1 year. 90 years as these samples contain people with thyroid disease who suffer from hyperthyroidism and hypothyroidism and normal people who do not suffer from thyroid disease. The data were collected over a period of one to four months, and the main goal of collecting the data was to classify thyroid diseases using machine learning algorithms. These data include gender, age, analysis of T3 (triiodothyronine), T4 (thyroid hormone), TSH (thyroid stimulating hormone), and a host of other characteristics. As the data obtained consist of 17 variables or attributes where all the attributes were taken in our study which consist of (id, age, gender, query thyroxine, on_antithyroid_medication, sick, pregnant,thyroid_surgery,query_hypothyorid,query_hyperthyroid, TSH_M, TSH, T3_M, T3, T3, T4, Category).

**Table** 2: shows the features contained in the dataset.

| No | Attribute Name | Value Type | Clarification |
|---|---|---|---|
| 1 | id | number | 1,2,3……..., |
| 12 | age | number | 1,10,20,50,……., |
| 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=hypothyroid,2=hyperthyroid |

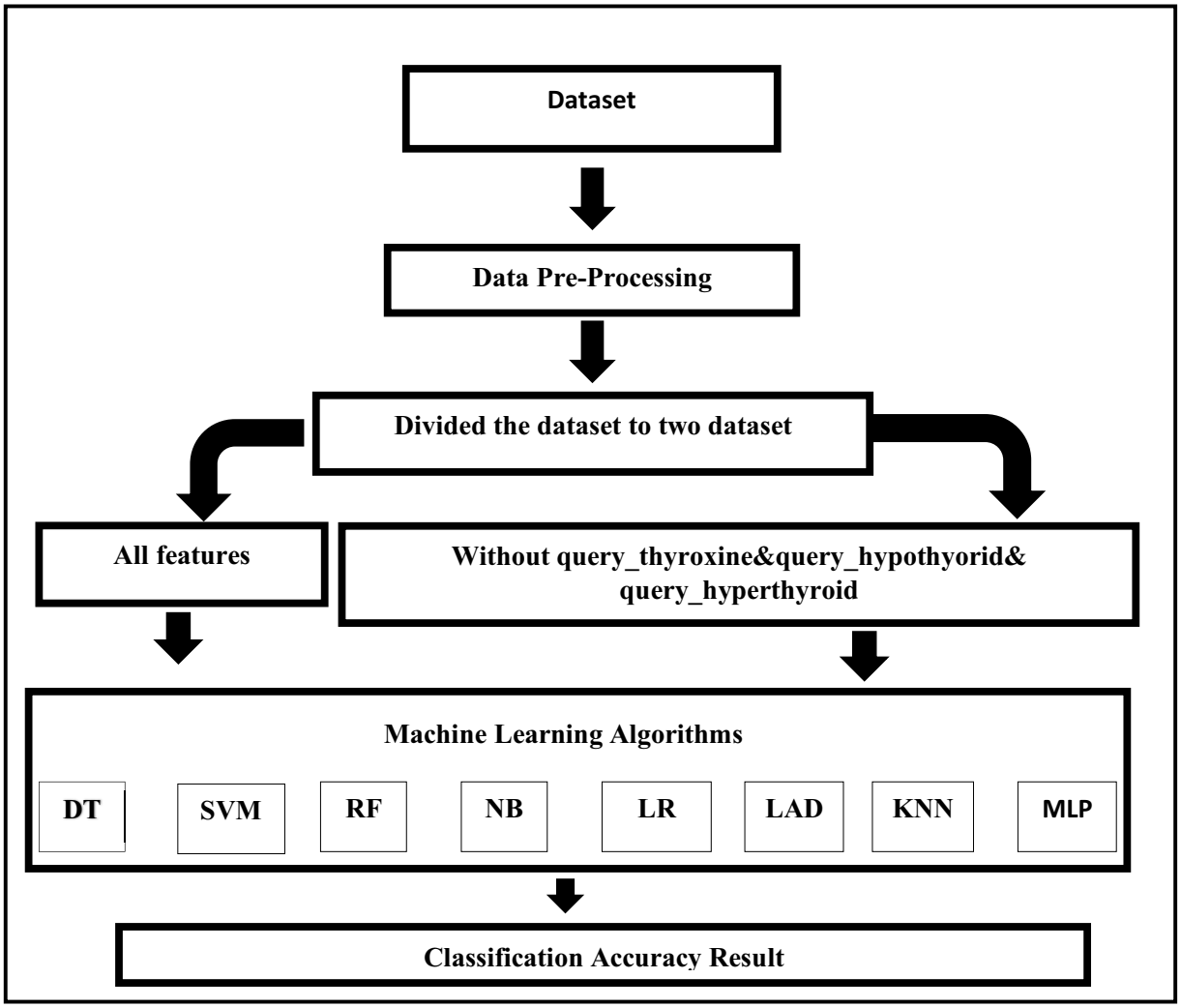


**Figure 1**. Shows how data is entered and the operations that take place .

### *3.2. Data Preprocessing*

The process of pre-processing the data is very important and it is a major step in data mining, as it has a good effect on the data, as the pre-processing process is used to reveal the data through analyzing the data and discovering the lost data, as it examines the data with great care. The pre-processing process includes cleaning the data, preparing the data, etc. In this stage or step we did is to clean and arrange the data that we were able to obtain, where we identified a set of missing data in this data where the missing features were identified, and among these properties that were missing T4 by number 151 and T3 by number 112, where we were able to Processing 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 good and free from any defect or problem so that we can work on it smoothly and well. We also used normalization technical with the MLP algorithm.

### *3.3. Data Machine Learning Techniques*

The key aim of using machine learning algorithms is to differentiate between three forms of thyroid disease. The first is hyperthyroidism, the second is hypothyroidism, and the third is stable patients who do not have any thyroid issues.

#### *3.3.1.Support Vector Machines*

The support vector machine (SVM) is a machine learning and data mining algorithm to determine the strongest predictors of this variable for energy consumption. The research used popular classification methods to answer our question: best subset selection, boosting trees, and generalized additive models. Our first approach was to use forward, backward and best subset selection to obtain a subset of predictors that most strongly predicted consumption with a linear relationship. The SVM provided an approach that was to use a tree-based method to stratify the predictor space into sample regions using recursive binary splitting. The research decided to use the boosting tree method as this is known to be one of the most powerful tree based models. SVM also has a good ability to deal with high dimensionality data.

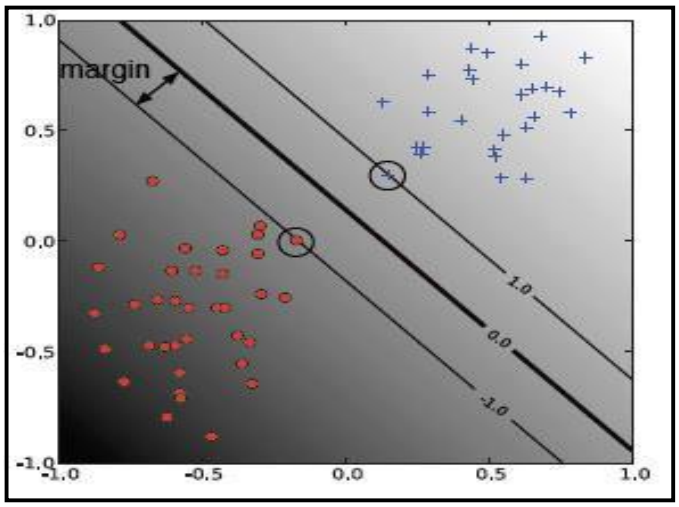


**Figure 2**. Example a two-class linear SVM classifier [21].

*3.3.2. Random forest*

The random forest computes the mean response of every predictor for energy consumption. Then, for each sample, a random forest adds the absolute distance each response was from the mean of each predictor for a total sum of the distance that each answer was from the means of the data. A high distance value will signify individuals who were consistently far away from the mean response in each sample. Detecting rates who repeatedly classify the samples was simple--a function that calculated the mode of each response was used. If the mode of a response was over 90% of the total number of questions, the research marked the response as potentially high in energy consumption. There are many responses marked. It was clear from a visual examination of these responses that the individuals had sampled with the same response.

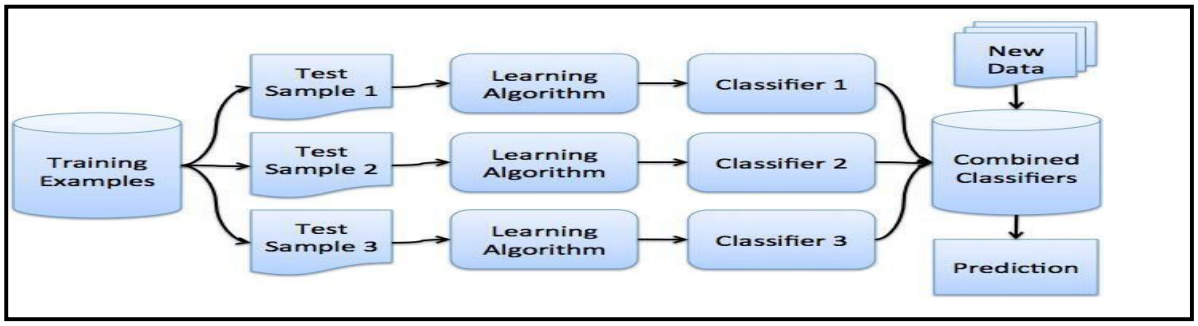


**Figure 3.** Phases of ensemble random forest approaches to solve classification problems [22].

*3.3.3.Decision Tree*

The decision tree method is based on a decision-boosting machine which is analyzed for predicting the energy consumption factor to try a tree based approach to determine the most significant predictors of consumption. To do so, used the decision tree approach as provided. The decision entails fitting thousands of trees, each of which is grown using information from the previous tree, in order for the model to improve over time. There are a few tuning parameters, including: number of trees, shrinkage parameter, number of splits in each tree.

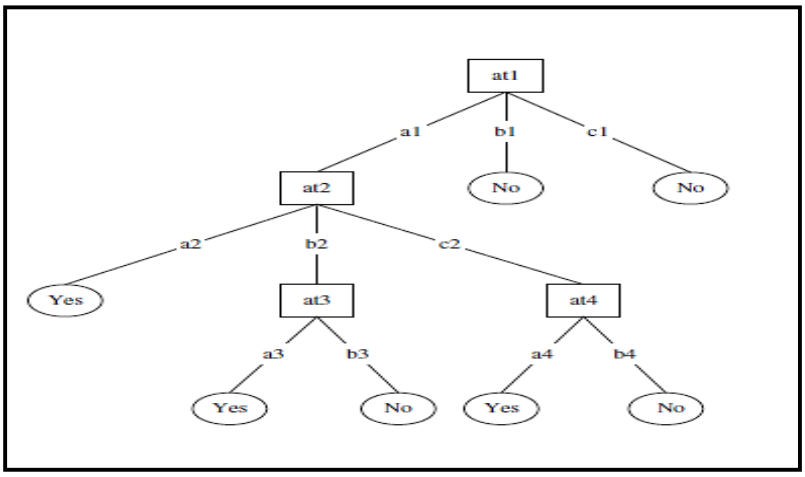


**Figure 4.** Decision tree algorithm structure [23].

*3.3.4.Naïve Bayes*

Naïve Bayesian are able to compare multiple generalized additive models featuring the output variables of subset selection, variables with the highest relative influence in the classifying and a mix of variables from both. It compared the prediction accuracy of each best model to directly compare them. By fitting naïve Bayes with various combinations of splines, 2nd degree polynomials and linear predictor variables, it narrowed down the relationships between single predictors and the response. More polynomials and splines were applied to predictors that were proved to have nonlinear relationships with our response variable.

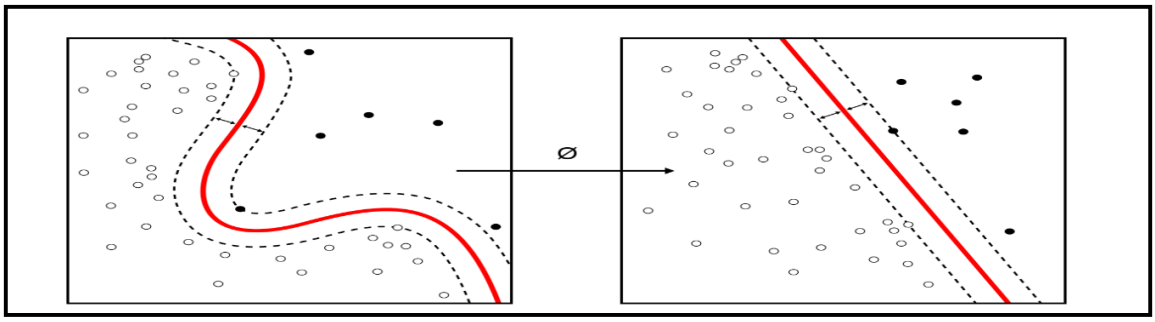


**Figure 5**. Naïve Bayes algorithm on left with respect to support vector machine on right side for classification structure [24].

3.3.5.*Logistic* Regression

Under the Supervised Learning technique, one of the most common Machine Learning algorithms is logistic regression. It's a method for estimating a categorical dependent variable from a number of independent variables. A categorical dependent variable's contribution is predicted using logistic regression. As a result, the result must be a singular or categorical value. It may be Yes or No, 0 or 1, true or false, and so on, but instead of providing exact values like 0 and 1, it gives probabilistic values that are somewhere between 0 and 1. Except for how they are used, Logistic Regression is somewhat similar to Linear Regression. Regression problems are solved using Linear Regression, although the classification problems are solved using logistic regression [25].

*3.3.6.k-Nearest neighbors*

The k-nearest neighbor algorithm differs from the other methods in that it uses the data directly for classification rather than first building a model [26]. As a result, no special model construction is necessary, and the only variable in the model is k, the number of nearest neighbors to use in class membership estimation: the value of p(y/x) is simply the ratio of members of classy among the k nearest neighbors of x. Changing the value of k will make the model more or less stable (small or big values of k, respectively). The advantage of k-nearest neighbors over other algorithms is its ease of use. Neighbors can provide a rationale for the classification result; in cases where black-box models are incomplete, this case-based reasoning can be helpful. The key disadvantage of k-nearest neighbors [27] is the calculation of the case neighborhood, which requires defining a metric that measures the distance between data objects.

*3.3.7.Multi-Layer Perceptron's (MLP)*

A multilayer perceptron is a feedforward artificial neural network that generates a series of outputs from a set of inputs (MLP). In an MLP, several layers of input nodes form a directed graph between the input and output layers. MLP employs backpropogation to train the network. MLP is a deep learning technique. A multilayer perceptron is a neural network that connects several layers in a guided graph, meaning that the signal only travels in one direction between nodes. Each node, with the exception of the input nodes, has a nonlinear activation function. An MLP uses backpropagation as a supervised learning process. MLP is a deep learning method that employs several layers of neurons. MLP is a commonly used system in supervised learning problems, computational biology, and parallel distributed processing analysis. Applications include speech recognition, image recognition, and automatic translation [28].

*3.3.8.Linear Discriminant Analysis*

It is one of the most commonly used dimensionality reduction techniques. It's used in pattern recognition systems like machine learning and other systems. The aim of LDA is to project elements from a high-dimensional space into a lower-dimensional space. This is achieved in order to avoid common dimensionality issues while still lowering spatial costs and capital. Linear discriminant analysis, a supervised classification method, is used to construct machine learning models. These dimensionality reduction models are used in a number of applications, such as ad prediction and image recognition [29].

## 4. Results

We have applied our data to a range of machine learning algorithms (Decision Tree, SVM, Random Forest, Naive Bayes, Logistic Regression, Linear Discriminant Analysis, k-Nearest neighbors, Multi-Layer Perceptron) 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 a group of algorithms shown in the table below, and after the application process these results appeared to us. This practical part has been implemented on the python platform and is considered a complete and integrated platform. All attributes have been taken which are 16 inputs and one output.

**Table 3.** Evaluation measurements for classification models with all attribute of dataset

| **NO** | Algorithms | Accuracy |
|---|---|---|
| 1 | Decision Tree | 90.13 |
| 2 | SVM | 92.53 |
| 3 | Random Forest | 91.2 |
| 4 | Naive Bayes | 90.67 |
| 5 | Logistic Regression | 91.73 |
| 6 | Linear Discriminant Analysis | 83.2 |
| 7 | KNeighbors Classifier | 91.47 |
| 8 | MLP | 96.4 |

And as shown to us in this table, it 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, this logic Random Forest algorithm has obtained high accuracy Then an algorithm follows Decision Tree. Most of the algorithms that I used to classify thyroid disease have proven their worth in diagnosing the disease, and this will help us a lot in the health system, as it will be an aid to the health sectors. In the second step, we removed 3 traits, based on a previous study Ioniţă, Irina, and Liviu Ioniţă [30] 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 also by using the Python script, we were able to obtain these results listed below in Table (4).

**Table 4.** Evaluation measures of the classification models without three attributes of the data set.

| NO | Algorithms | Accuracy |
|---|---|---|
| 1 | Decision Tree | 98.4 |
| 2 | SVM | 92.27 |
| 3 | Random Forest | 98.93 |
| 4 | Naive Bayes | 81.33 |
| 5 | Logistic Regression | 91.47 |
| 6 | Linear Discriminant Analysis | 83.2 |
| 7 | KNeighbors Classifier | 90.93 |
| 8 | MLP | 97.6 |

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 KNeighbours Classifier algorithm have increased slightly and reduced the accuracy of the other algorithms. We show here that the accuracy of the algorithms used on our data changes with the change of the characteristics used in the data, as experience has demonstrated this clear change, which obtained the accuracy of the algorithms when three of the characteristics were deleted, as the accuracy of some algorithms decreased and some of them increased.

## 5. Conclusion

Thyroid disease is one of the diseases that afflict the world's population, and the number of cases of this disease is increasing. Because of medical reports that show serious imbalances in thyroid diseases, our study deals with the classification of thyroid disease between hyperthyroidism and hypothyroidism. This disease was classified using algorithms. Machine learning showed us good results using several algorithms and was built in the form of two models. In the first model, all the characteristics consisting of 16 inputs and one output were taken, and the result of the accuracy of the random forest algorithm was 98.93, which is the highest accuracy among the other algorithms. In the second embodiment, the following characteristics were omitted based on a previous study. The removed attributes were 1- query_thyroxine 2- query_hypothyorid 3-query_hyperthyroid. Here we have included the increased accuracy of some algorithms, as well as the retention of the accuracy of others. It was observed that the accuracy of Naive Bayes algorithm increased the accuracy by 90.67. The highest precision of the MLP algorithm was 96.4 accuracy.

## 6. References

[1] Azar, a.T, Hassanien, A.E. and Kim, T. Expert system based on neural fuzzy rules for thyroid diseases diagnosis, Computer Science, Artificial Intelligence, arXiv:1403.0522, Pp. 1-12,2012.
[2] Keles, A. ESTDD: Expert system for thyroid diseases diagnosis, Expert Syst Appl., Vol. 34, No.1, Pp.242–246,2008.
[3] a. c.c.Heuck, "World Health Organization," 2000. [Online]. Available: https://www.who.int/.

[4] Kouroua, K., Exarchosa, T.P. Exarchosa, K.P., Karamouzisc, M.V. andFotiadisa, D.I. (2015) Machine learning applications in cancer prognosis and prediction, Computational and Structural Biotechnology Journal, Vol. 13, Pp.8–17.
[5] Shukla, A. & Kaur, P. (2009). Diagnosis of thyroid disorders using artificial neural networks, IEEE International Advance computing Conference (IACC 2009)– Patiala, India, pp 1016-1020.
[6] Aswad, Salma Abdullah, and Emrullah Sonuç. "Classification of VPN Network Traffic Flow Using Time Related Features on Apache Spark." 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). IEEE, 2020.
[7] Banu, G. Rasitha. "A Role of decision Tree classification data Mining Technique in Diagnosing Thyroid disease." International Journal of Computer Sciences and Engineering 4.11 (2016): 64-70.
[8] Chandio, Jamil Ahmed, et al. "TDV: Intelligent system for thyroid disease visualization." 2016 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube). IEEE, 2016.
[9] Travis B Murdoch and Allan S Detsky. The inevitable application of big data to health care. Jama, 309(13):1351–1352, 2013.
[10] Dr. Srinivasan B, Pavya K "Diagnosis of Thyroid Disease: A Study" International Research Journal of Engineering and Technology Volume: 03 Issue: 11 | Nov – 2016
[11] Aytürk Keleş and Keleş, Ali. "ESTDD: Expert system for thyroid diseases diagnosis." International Research Journal of Engineering and Technology (IRJET) Volume: 03 Issue: 11 | Nov -2017 34.1 (2017): 242- 246
[12] Khushboo Taneja, Parveen Sehgal, Prerana "Predictive Data Mining for Diagnosis of Thyroid Disease using Neural Network" International Journal of Research in Management, Science & Technology (E-ISSN: 2321- 3264) Vol. 3, No. 2, April 2016
[13] Chandel, Khushboo, et al. "A comparative study on thyroid disease detection using K-nearest neighbor and Naive Bayes classification techniques." CSI transactions on ICT 4.2-4 (2016): 313-319.
[14] Banu, G. Rasitha. "A Role of decision Tree classification data Mining Technique in Diagnosing Thyroid disease." International Journal of Computer Sciences and Engineering 4.11 (2016): 64-70.
[15] Umar Sidiq, Dr, Syed Mutahar Aaqib, and Rafi Ahmad Khan. "Diagnosis of various thyroid ailments using data mining classification techniques." Int J Sci Res Coput Sci Inf Technol 5 (2019): 131-6.
[16] Sindhya, Mrs K. "EFFECTIVE PREDICTION OF HYPOTHYROID USING VARIOUS DATA MINING TECHNIQUES."
[17] AKGÜL, Göksu, et al. "Hipotiroidi Hastalığı Teşhisinde Sınıflandırma Algoritmalarının Kullanımı." Bilişim Teknolojileri Dergisi 13.3 (2020): 255-268.
[18] VijiyaKumar, K., et al. "Random Forest Algorithm for the Prediction of Diabetes." 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN). IEEE, 2019.
[19] Chaurasia, Vikas, Saurabh Pal, and B. B. Tiwari. "Prediction of benign and malignant breast cancer using data mining techniques." Journal of Algorithms & Computational Technology 12.2 (2018): 119-126.
[20] Begum, Amina, and A. Parkavi. "Prediction of thyroid disease using data mining techniques." 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS). IEEE, 2019.
21. C. Fan, F. Xiao, Z. Li, J. Wang. Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review. Energy Build. 2018, 159, 296–308.
[22] W. Kleiminger, C. Beckel, T. Staake, S. Santini. Occupancy Detection from Electricity Consumption Data. In Proceedings of the 5th ACM Workshop on Embedded Systems for Energy-Efficient Buildings, Rome, Italy, 14–15 November 2013; pp. 1–8.
[23] D. Mora, G. Fajilla, M. Austin, D. Simone. Occupancy patterns obtained by heuristic approaches: Cluster analysis and logical flowcharts. A case study in a university office. Energy Build. 2019, 186, 147–168
[24] V. Cerqueira, L. Torgo, M. Mozetic. Evaluating time series forecasting models: An empirical study on performance estimation methods. Mach. Learn. 2020, 109, 1997–2028.
[25] Dreiseitl, Stephan, and Lucila Ohno-Machado. "Logistic regression and artificial neural network classification models: a methodology review." Journal of biomedical informatics 35.5-6 (2002): 352-359.
[26] Dasarathy B. Nearest neighbor pattern classification techniques. Silver Spring, MD: IEEE Computer Society Press; 1991.
[27] Ripley B. Pattern recognition and neural networks. Cambridge: Cambridge University Press; 1996.

[28] Pacheco, Wolffang D. Niño, and Fabián R. Jiménez López. "Tomato classification according to organoleptic maturity (coloration) using machine learning algorithms K-NN, MLP, and K-Means Clustering." 2019 XXII Symposium on Image, Signal Processing and Artificial Vision (STSIVA). IEEE, 2019.
[29] Ye, Jieping. "Least squares linear discriminant analysis." Proceedings of the 24th international conference on Machine learning. 2007.
[30] Ioniţă, Irina, and Liviu Ioniţă. "Prediction of thyroid disease using data mining techniques." BRAIN. Broad Research in Artificial Intelligence and Neuroscience 7.3 (2016): 115-124.