Supervised Learning Algorithms for Classifying Thyroid Disease

Ashish Kumar Sen[1], Dr.Prabhat Pandey[2]

[1]*Research Scholar,AwadheshPratap Singh University,Rewa (M.P)-India*
[2]*Prof. Physics & O.S.D., Add. Directorate Office,Higher Education Rewa (M.P.)-India*

Abstract:*With the enormous amount of data and information that must be processed, particularly in the health system, machine learning algorithms and data mining techniques play a critical role in data management. We employed machine learning techniques to examine thyroid disorders in our study. The purpose of this study is to classify thyroid disease into three categories: hyperthyroidism, hypothyroidism, and normal. To accomplish this, we used data from Iraqi citizens, some of whom have hyperthyroidism and others who have hypothyroidism. Support vector machines, random forest, decision tree, naive bayes, logistic regression, k nearest neighbours, multilayer perceptron (MLP), and linear discriminant analysis are all terms that refer to support vector machines.*

Keywords: ***Classification model, machine learning Thyroid conditions, Automated support vector machines Forest of random numbers, Decision tree, Bayesian inference, logistic regression, K- next-door neighbours, Linear discriminant analysis (LDA), Multilayer perceptron (MLP)***

1. INTRODUCTION:

Thyroid disease is a subgroup of endocrinology that is widely misunderstood and underdiagnosed [1] [2].According to the World Health Organization, thyroid gland problems are the second most common endocrine disorder in the world, after diabetes. Hyper functioning hyperthyroidism and hypothyroidism affect approximately 2% and 1% of people, respectively. Men account for roughly a tenth of women's prevalence. Hyperthyroidism and hypothyroidism can be caused by thyroid gland dysfunction, pituitary gland dysfunction, or hypothalamus dysfunction. Goiter or active thyroid nodules may become prominent in some places due to dietary iodine shortage, with a prevalence of up to 15%. Additionally, the thyroid gland can be the site of several types of malignancies and can be a dangerous region for endogenous antibodies to wreak havoc (autoantibodies) [3].

According to experts, early disease detection, diagnosis, and treatment are critical for preventing disease progression and possibly death. Early detection and differential diagnosis improve the chances of successful treatment for a variety of various types of abnormalities. Despite extensive research, clinical diagnosis is frequently regarded as a challenging task [4].

At the base of the throat, the thyroid gland is a butterfly-shaped gland. It is composed of two active thyroid hormones, levothyroxine (T4) and triiodothyronine (T3), which regulates brain activities such as body temperature regulation, blood pressure control, and heart rate regulation. Similarly, thyroid disease is one of the most widespread disorders worldwide, and it is typically caused by an iodine deficit, but can also be caused by other reasons. The thyroid gland is an endocrine gland that produces and secretes hormones into the

bloodstream. It is located in the centre of the body's front. Thyroid gland hormones aid in digestion and help keep the body wet, balanced, and so on. Thyroid hormones such as T3 (triiodothyronine), T4 (thyroid hormone), and TSH (thyroid stimulating hormone) are used to monitor thyroid activity (thyroid stimulating hormone). Hypothyroidism and hyperthyroidism are the two forms of thyroid disorder. Data mining [5] is a semi-automated technique for discovering relationships within large datasets.

Machine learning algorithms are one of the most effective solutions to numerous difficult-to-solve situations [6]. Classification is a data extraction technique (machine learning) that is used to predict and identify a variety of diseases, including 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 fast and efficient [7]. Although computer learning and artificial intelligence have been used in medicine since the field's inception [8], there has been a recent push to consider the need for machine learning-driven healthcare solutions. As a result, analysts believe that machine learning will become increasingly prevalent in healthcare in the coming years [9].

Hyperthyroidism is a condition in which the thyroid gland produces an abnormally high amount of thyroid hormones. Increased thyroid hormone levels are the cause of hyperthyroidism [10]. Dry skin, increased temperature sensitivity, hair thinning, weight loss, increased heart rate, hypertension, excessive perspiration, neck enlargement, anxiousness, menstrual cycles becoming shorter, irregular stomach movements, and hands shaking are some of the symptoms [11]. Hypothyroidism is a medical term that refers to an underactive thyroid gland.

Hypothyroidism is a condition that results from a decrease in thyroid hormone production. In medical words, hypo denotes insufficient or less. Hypothyroidism is mostly caused by inflammation and thyroid gland damage. Obesity, a slow heart rate, increased temperature sensitivity, neck swelling, dry skin, numb hands, hair loss, irregular menstrual cycles, and intestinal difficulties are just a few of the symptoms. These symptoms can worsen over time if left untreated [12].

2. Review of Literature:

KhushbooChandel [13] in this study, thyroid disorders are identified using various classification models based on characteristics such as TSH, T4U, and goitre. This reasoning is supported by a variety of grouping techniques, including K-nearest neighbour. The algorithms Naive Bayes and support vector machines are used. The experiment was conducted using the Rapid miner instrument, and the results reveal that K-nearest neighbour is more effective in detecting thyroid disease than Naive Bayes. The researchers employed data mining classifiers to diagnose thyroid disorders. Thyroid dysfunction is critical to consider while diagnosing an illness. In this work, KNN and Naive Bayes classifiers were utilised. Rapid miner is used to compare the performance of these two classifiers. The K-nearest neighbour classifier was found to be the most accurate, with 93.44 percent accuracy, while the Naive Bayes classifier had 22.56 percent accuracy. The proposed KNN algorithm enhances classification accuracy, which results in enhanced performance.

As a result, because Naive Bayes can only have a linear, elliptic, or parabolic decision boundary, KNN's decision boundary consistency is a significant advantage. KNN outperforms the majority of approaches due to the interdependence of the components.

RasithaBanu, G. Thyroid disease is one of the most prevalent ailments in humans. The hypothyroid data utilised in this investigation were obtained from the University of California, Irvine's data repository (UCI). The entire study effort will be conducted using the platform Waikato Environment of Information Analysis (WEKA). It was discovered that the J48 technique is more effective than the decision stump tree strategy. Disease diagnosis is a difficult task in the field of health care. Numerous data mining techniques are employed in the decision-making process. We applied dimensionality reduction to extract a selection of characteristics from the original results, and we defined hypothyroidism using J48 and decision stump data mining classification approaches. The uncertainty matrix is used to evaluate the precision and error rate of the classifier output. The J48 Algorithm has a precision of 99.58 percent, which is higher than the decision stump tree's precision, and a lower error rate than the decision stump.

Dr. Syed MutaharAaqib, Umar Sidiq, and Rafi Ahmad Khan [15] Classification is one of the most widely used supervised learning data mining techniques. It is used to characterise preset data sets. The classification is frequently used in the healthcare sector to aid in medical decision-making, diagnosis, and administration. The data for this study were obtained from a reputable Kashmiri laboratory. The research will be conducted entirely on the ANACONDA3-5.2.0 platform. Classification methods such as k nearest neighbours, support vector machine, decision tree, and Nave bayes may be employed in an experimental investigation. At 98.89 percent accuracy, the Judgment Tree is the most accurate of the other classes.

Mrs K Sindhya [16] Thyroid disease is a chronic ailment that affects millions of individuals worldwide. Data mining in healthcare is achieving outstanding outcomes in terms of disease prediction. The predictive accuracy of data mining techniques is high, while the cost of prediction is low. Another big advantage is that prediction is a rather quick process. In this work, I analysed thyroid data using classification algorithms and arrived to a conclusion. The efficacy of a model is essentially governed by two things. The first is the precision of the prediction, and the second is the time required to make the prediction. According to our research, Nave Bayes forecasts in just 0.04 seconds. It is, however, less precise than J48 and Random Forest. When we looked at prediction accuracy, we found that the Random Forest model was 99.3 percent accurate. However, the model takes longer to develop than the other two iterations. As a result, we may claim that J48 is the best model for hypothyroid prediction because its accuracy is 99 percent, which is among the greatest, and it runs in 0.2 seconds, much less time than the Random Forest model.

AKGÜL, Göksu, and coauthors [17] The purpose of this work is to propose a strategy based on data mining for increasing the precision of hypothyroidism diagnosis by merging patient questions and test findings into the diagnosis process. A secondary objective is to mitigate the hazards associated with dialysis interventional trials. The conclusive conclusion the new samples' hypothyroid status was assessed using data from the UCI machine learning database, which contained 3163 samples, 151 of which were hypothyroid and the remainder were

normal. Different sample strategies were employed to eliminate the imbalanced distribution, and Logistic Regression, K Nearest Neighbor, and Support Vector Machine classifiers were utilised to create models to diagnose hypothyroidism. In this context, the thesis established the effect of sample methodologies on the diagnosis of hypothyroidism. Of all the models created, the Logistic Regression classifier delivered the best results. For this study, which was trained on the data set using over-sampling approaches, the precision was 97.8 percent, 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.

K. VijiyaKumar et al. The purpose of this study is to develop a method for reliably and early detection of diabetes in a patient by utilising the Random Forest algorithm in a machine learning technology. Random Forest algorithms are a subclass of ensemble learning algorithms that are frequently used for classification and regression applications. The performance ratio is superior to that of other methods. The proposed model produces the best outcomes for diabetic prediction, and the results indicate that the prediction system is capable of projecting diabetes disease accurately, effectively, and most importantly, quickly.

VikasChaurasia, Saurabh Pal, and B. B. Tiwari Breast cancer is the second most frequent type of cancer in women, after all other types. This research paper's objective is to present a breast cancer study that utilises cutting-edge approaches. By adding recent scientific developments, we can improve breast cancer survivability modelling models. We constructed prediction models for 68 3 breast cancer cases using a large dataset and three widely used data mining algorithms (Nave Bayes, RBF Network, and J48). To compare the accuracy of the three prediction models, we employed 10-fold cross-validation procedures to determine the unbiased estimation of the three models. The findings indicate that visiting the Bay is quite safe (based on an average precision Breast Cancer dataset). The RBF Network is the second-best predictor, achieving 93.41 percent accuracy on the holdout sample (better than any other predictor reported in the literature), and Nave Bayes is the third-best predictor, achieving 97.36 percent accuracy on the holdout sample (better than any other predictor reported in the literature) (better than any other prediction accuracy published in the literature). We investigated three breast cancer survivorship prediction models in this study using two different criteria: benign and malignant cancer cases.

Amina Begum and A. Parkavi [20] Recent study has concentrated on the classification of thyroid illness in two of the most common thyroid dysfunctions in the general population (hyperthyroidism and hypothyroidism). The authors examined and analysed four distinct categorization models: Naive Bayes, Decision Trees, Multilayer Perceptions, and Radial Basis Function Networks. The data indicate that all of the categorization models indicated above are highly accurate, with the Decision Tree model scoring the highest. The classifier was constructed and verified using data from a Romanian data website and the University of California, Irvine's machine learning repository. Two data sets are available: KNIME Analytics Platform and Weka. The categorization models were developed and tested using data mining techniques. According to the literature, numerous studies in the subject of thyroid classification employ various data mining techniques to develop robust classifiers. The authors of this study investigated the use of four classification models (Nave Bayes, Decision Tree, MLP, and RBF Network) on thyroid data to aid in the classification of thyroid dysfunctions such

as hyperthyroidism and hypothyroidism. The decision tree model was the proper classification model in all of the cases examined.

Table1. Demonstrates the literature review and the algorithms that were employed, as well as 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 Collecting Data:

Machine learning algorithms are utilised in the speedy and early diagnosis of thyroid and other disorders, since they have established a substantial presence in the health area and assist us in diagnosing and classifying diseases. As a result, we were able to collect a substantial amount of data on thyroid illnesses and are currently working on classifying diseases using this data in our study. The data that I used in our study is a collection of data from external hospitals and laboratories that specialise in analysing and diagnosing diseases, and the sample drawn from the data is the Iraqi population, with the type of data collected relating to thyroid disease. Data were collected on 1250 people, males and females, ranging in age from one year to one year. 90 years, as these samples include persons with thyroid disease who have hyperthyroidism or hypothyroidism, as well as healthy individuals who do not have thyroid disease. The data were collected over a one- to four-month period

with the primary purpose of classifying thyroid illnesses using machine learning techniques. Gender, age, T3 (triiodothyronine), T4 (thyroid hormone), TSH (thyroid stimulating hormone), and a variety of other variables are included in this data. As the data received contains 17 variables or features, all of which were included in our study (id, age, gender, query thyroxin, on anti-thyroid medication, sick, pregnant, thyroid surgery, query hypothyroid, query hyperthyroid, TSH M, TSH, T3 M, T3, T3, T4, Category).

Table2.Demonstrates the dataset's features

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

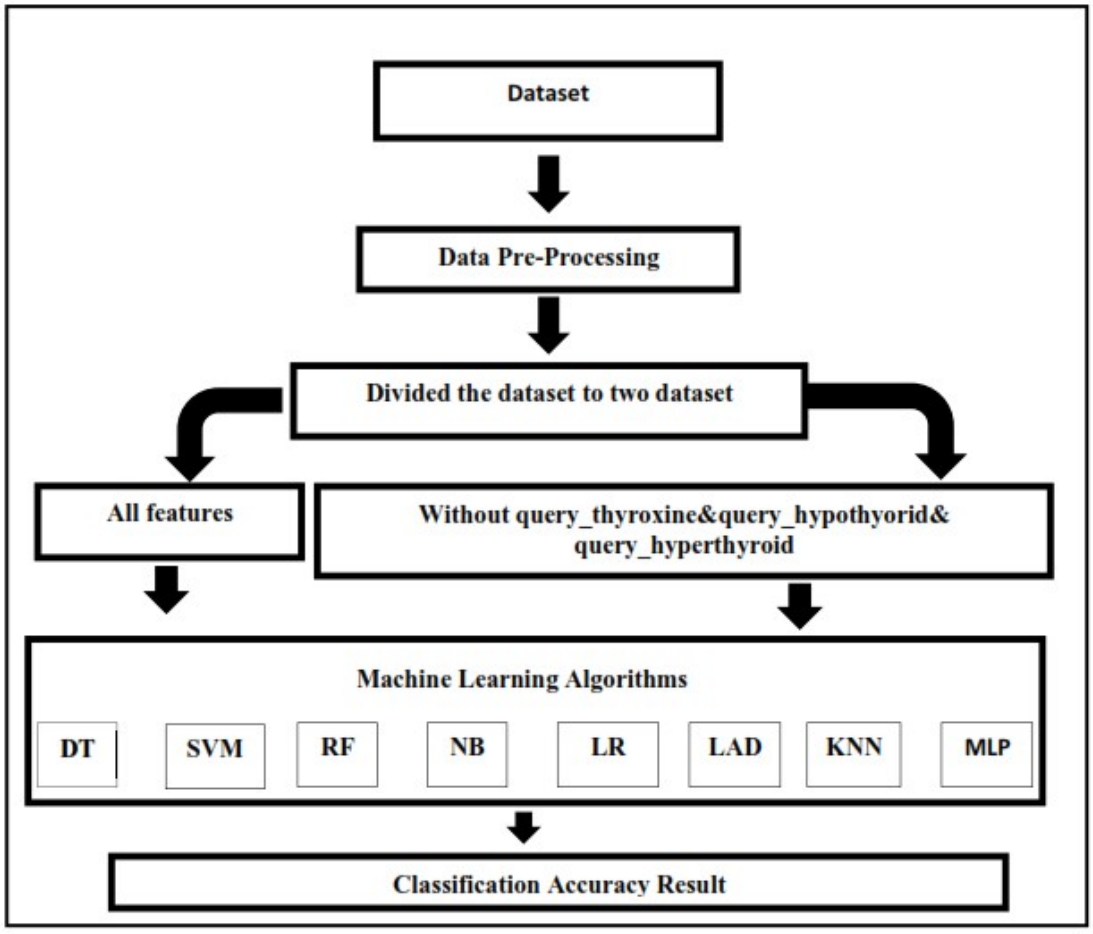

Figure1.Demonstrates how data is entered and the activities that occur

3.2 Pre-processing of Data:

Pre-processing data is critical and is a significant stage in data mining because it has a beneficial influence on the data, as the pre-processing process is used to disclose the data through analysis and discovery of lost data, as it carefully examines the data. The pre-processing stage entails cleaning and preparing the data. This stage or step involved cleaning and organising the data that we were able to obtain, where we identified a set of missing data in this data, specifically T4 by number 151 and T3 by number 112, where we were able to process this lost data by replacing it with the mediator's value, and after working in this manner, we were able to obtain the data in a good and better way that was free of errors. Additionally, we applied normalising techniques in conjunction with the MLP method.

3.3 Techniques for Machine Learning using Data:

The primary goal of machine learning algorithms is to distinguish between three distinct types of thyroid illness. The first category is hyperthyroidism, the second category is hypothyroidism, and the third category is stable patients with no thyroid problems.

3.3.1 Sustaining Vector Machines:

The support vector machine (SVM) is a machine learning and data mining approach used to identify the most powerful predictors of this energy consumption variable. To address our challenge, we used popular categorization methods: best subset selection, boosting trees, and generalised additive models. Our first step was to employ forward, backward and best subset selection to identify a subset of predictors that most strongly predicted consumption in a linear fashion. The SVM proposed a method for stratifying the predictor space into sample regions using recursive binary splitting. The researchers chose the boosting tree method since it is one of the most powerful tree-based models available. Additionally, SVM is well-suited for dealing with data with a high degree of dimension.

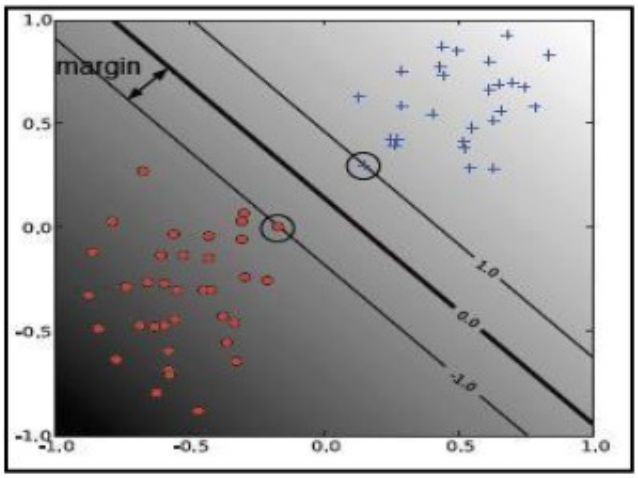

Figure2. A two-class linear SVM classifier is an illustration [21].

3.3.2 Forest of unknown origin:

The random forest algorithm determines the mean response of each predictor to the energy consumption variable. Then, for each sample, a random forest calculates the absolute distance between each response and the mean of each predictor, resulting in a cumulative sum of the distances between each answer and the data averages. A high distance score indicates that people in each sample were consistently far from the mean response. Detecting rates that classify samples frequently was straightforward—a function that determined the mode of each answer was utilised. If a response method accounted for more than 90% of the total number of questions, the research classified the response as possibly high in terms of energy consumption. Numerous responses are highlighted. Visual assessment of these responses revealed that the subjects had all sampled the same response.

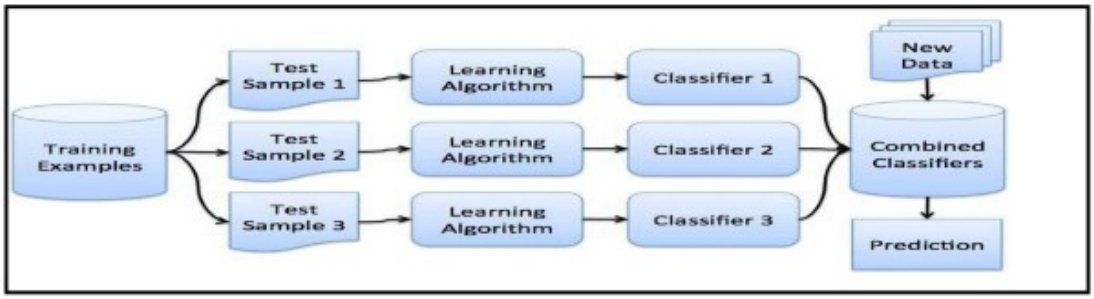

Figure3. Ensemble random forest techniques to classification issues in phases [22]

3.3.3 Tree of Decisions:

The decision tree method is based on a decision-boosting machine that is assessed for predicting energy consumption in order to discover the most significant predictors of consumption using a tree-based methodology. To accomplish this, I used the suggested decision tree approach. The decision requires fitting thousands of trees, each of which is generated using data from the preceding tree, in order to continuously improve the model. A few tuning parameters are available, including the number of trees, the shrinkage parameter, and the number of splits in each tree.

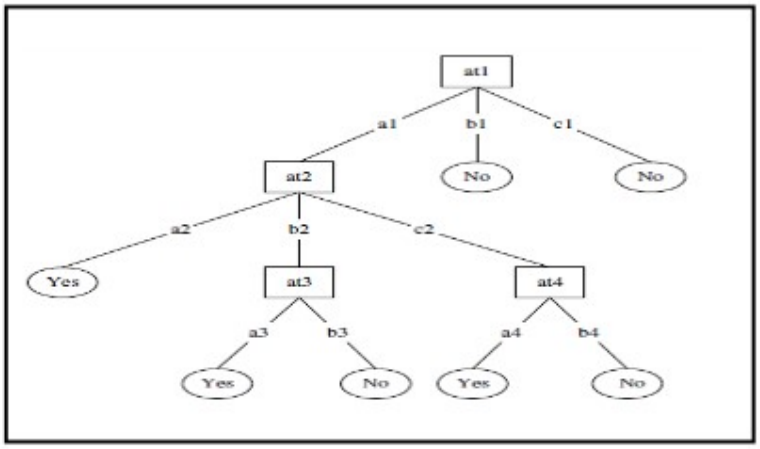

Figure4. Structure of a decision tree algorithm [23]

3.3.4 Naïve Bayes:

The naive Bayesian is capable of comparing different generalised additive models that include the output variables of subset selection, the variables having the greatest relative influence on classification, and a combination of the two. It directly compared the forecast accuracy of each best model. It restricted the correlations between single predictors and the response by fitting naive Bayes with various combinations of splines, second degree polynomials, and linear predictor variables. Additional polynomials and splines were applied to predictors whose nonlinear interactions with our response variable were established.

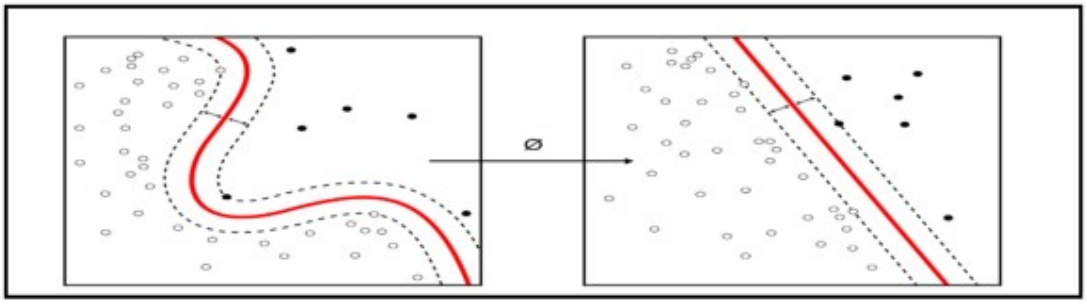

Figure5.On the left, the naive Bayes algorithm is compared to the support vector machine on the right for classification structure [24]

3.3.5 Regression Logistic:

Under the Supervised Learning approach, logistic regression is one of the most often used Machine Learning algorithms. It is a technique for deriving the value of a categorical dependent variable from a set of independent factors. The contribution of a categorical dependent variable is predicted using logistic regression. As a result, the outcome must be either solitary or categorical. It may be Yes or No, 0 or 1, true or false, and so on, but rather than offering exact values such as 0 or 1, it provides probability values between 0 and 1. Except for their application, Logistic Regression and Linear Regression are very similar. While linear regression is used to solve regression problems, logistic regression is used to address classification difficulties [25].

3.3.6 k-Proximate neighbours:

The k-nearest neighbour technique is distinct from the others in that it does classification directly on the data, rather than first developing a model [26]. As a result, no additional model building is required, and the models only variable is k, the number of nearest neighbours to employ in estimating class membership: the value of p(y/x) is just the ratio of classy members to x's k nearest neighbours. By varying the value of k, the model becomes more or less stable (small or big values of k, respectively). The primary advantage of k-nearest neighbours over other algorithms is its simplicity of implementation. Neighbors can provide context for the categorization result; in situations where black-box models are insufficient, this case-based reasoning can be beneficial. The primary downside of k-nearest neighbours [27] is that it requires specifying a metric for measuring the distance between data elements.

3.3.7 Perceptron with Multiple Layers (MLP):

A multilayer perceptron is a type of feed forward artificial neural network that generates a sequence of outputs in response to a set of inputs (MLP). Between the input and output layers of an MLP, numerous layers of input nodes create a directed graph. MLP trains the network using back propagation. MLP is a technique for deep learning. A multilayer perceptron is a type of neural network that connects several layers in a directed graph, which means that the signal between nodes only travels in one way. Except for the input nodes, each node has a nonlinear activation function. A MLP is a supervised learning procedure that makes use of back propagation.

MLP is a method for deep learning that makes use of many layers of neurons. MLP is a system that is frequently used in supervised learning issues, computational biology, and research of parallel distributed processing. Speech recognition, picture recognition, and automatic translation are all examples of applications [28].

3.3.8 Discriminant Analysis Using Linear Discriminants:

It is one of the most frequently used approaches for dimensionality reduction. It is employed in systems for pattern recognition, such as machine learning and others. LDA is used to project items from a three-dimensional space to a two-dimensional space. This is accomplished by avoiding frequent dimensionality difficulties while maintaining minimal spatial costs and capital. Machine learning models are constructed using linear discriminant analysis, a supervised classification technique. These dimensionality reduction methods are applied in a variety of applications, including advertisement prediction and picture recognition [29].

4. RESULTS:

We used a variety of machine learning algorithms on our data (Decision Tree, SVM, Random Forest, Naive Bayes, Logistic Regression, Linear Discriminant Analysis, k-Nearest Neighbors, and Multi-Layer Perceptron). We divided the existing data into two parts: 30% for training and 70% for testing, as this is the first time we have used this data for training. In the first stage, we applied all the properties in our data to a collection of algorithms provided in the table below, and these results appeared to us as a result of the application process. This practical section was developed using the Python programming language and is regarded a complete and integrated platform. All qualities have been accounted for, which totals sixteen inputs and one output.

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

Table3. Metrics for evaluating classification models with all of the dataset's attributes

And, as demonstrated in this table, it demonstrates the accuracy of each algorithm in the order in which it received an algorithm. Decision Tree with an accuracy of 98.4 SVM accuracy of 92.27 Random Forest has an accuracy of 98.93 percent. 81.33 accuracy for Naive Bayes Accuracy of 91.47 in Logistic Regression 83.2 accuracy for Linear Discriminant Analysis With a KNeighbors Classifier accuracy of 90.93 and an MLP(NN) accuracy of 97.6, this logic Random Forest technique has achieved a high level of accuracy. Then, an algorithm is applied to the Decision Tree. The majority of the algorithms I used to classify thyroid disease have demonstrated their value in detecting the ailment, which will benefit the health system significantly by assisting the health sectors. In the second phase, we eliminated three features based on Ioniţă, Irina, and LiviuIoniţă'sprior

study [30]. Both query thyroxin &query hypothyroid &query hyperthyroid were eliminated. After eliminating these attributes, we applied our data to the algorithm group as well, and using the Python script, we obtained the following findings in Table (4).

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

Table4.Evaluation metrics for classification models that do not include the data set's three features

As the Naive Bayes method appears to have a high accuracy of 90.67 when the three features are excluded, the SVM algorithm, the logistic regression algorithm, and the KNeighbours Classifier algorithm all gained somewhat in accuracy while the other algorithms decreased significantly in accuracy. We demonstrate here that the accuracy of the algorithms used on our data changes as the characteristics used in the data change, as experience has demonstrated this clear change in the accuracy of the algorithms obtained when three of the characteristics were deleted, as the accuracy of some algorithms decreased while the accuracy of others increased.

5. CONCLUSION:

As the Naive Bayes method appears to have a high accuracy of 90.67 when the three features are excluded, the SVM algorithm, the logistic regression algorithm, and the KNeighbours Classifier algorithm all gained somewhat in accuracy while the other algorithms decreased significantly in accuracy. We demonstrate here that the accuracy of the algorithms used on our data changes as the characteristics used in the data change, as experience has demonstrated this clear change in the accuracy of the algorithms obtained when three of the characteristics were deleted, as the accuracy of some algorithms decreased while the accuracy of others increased.

References:

[1] *Azar, a.T, Hassanien, A.E. and Kim, T. Expert system based on neural fuzzy rulesfor thyroid diseases diagnosis, Computer Science, Artificial Intelligence, arXiv:1403.0522, Pp. 1-12,2019.*

[2] *Keles, A. ESTDD: Expert system for thyroid diseases diagnosis, Expert Syst Appl., Vol. 34, No.1,Pp.242–246,2017.*

[3] *a. c.c.Heuck, "World Health Organization," 2010, [Online]. Available: https://www.who.int/.*

[4] *Kouroua, K., Exarchosa, T.P. Exarchosa, K.P., Karamouzisc, M.V. andFotiadisa, D.I. (2017) Machine learning applications in cancer prognosis and prediction, Computational and Structural Biotechnology Journal, Vol. 13, Pp.8–17.*

[5] *Shukla, A. &Kaur, P. (2018). 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 EmrullahSonuç. "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 (2017): 6470.*

[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, 2018.*

[9] *Travis B Murdoch and Allan S Detsky. The inevitable application of big data to health care,Jama, 309(13):1351–1352, 2017.*

[10] *Dr.Srinivasan B, Pavya K "Diagnosis of Thyroid Disease: A Study" International Research Journal of Engineering and Technology Volume: 03 Issue: 11 | Nov – 20.*

[11] *AytürkKeles and Keles, 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] *KhushbooTaneja, ParveenSehgal, 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 2018*

[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 (2017): 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 (2017): 64-70.*

[15] *Umar Sidiq, Dr.SyedMutaharAaqib, and Rafi Ahmad Khan, "Diagnosis of various thyroid ailments using data mining classification techniques", Int J Sci Res CoputSciInfTechnol 5 (2019): 131-6.*

[16] *Sindhya, Mrs K. "EFFECTIVE PREDICTION OF HYPOTHYROID USING VARIOUS DATA MINING TECHNIQUES."*

[17] *AKGÜL, Göksu, et al. "HipotiroidiHastaligiTeshisindeSiniflandirmaAlgoritmalarinin,Kullanimi", BilisimTeknolojileriDergisi 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 2016; 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 LucilaOhno-Machado. "Logistic regression and artificial neural network classification models: a methodology review", Journal of biomedical informatics 35.5-6 (2012): 352-359.*

[26] *Dasarathy B. Nearest neighbor pattern classification techniques. Silver Spring, MD: IEEE Computer Society Press; 2001.*

[27] *Ripley B. Pattern recognition and neural networks. Cambridge: Cambridge University Press; 2006.*

[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] *Ionita, Irina, and LiviuIonita, "Prediction of thyroid disease using data mining techniques", BRAIN. Broad Research in Artificial Intelligence and Neuroscience 7.3 (2016): 115-124.*