International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 SJIF 2023: 8.176 ISSN: 2582-3930

THYROID DISEASE CLASSIFICATION USING MACHINE LEARNING

AUTHORS AND EMAIL ID:

BINGUMALLA SUDHEESH 201910101095@presidencyuniversity.in

BUSAGANI GURUDEEPAK gurudeepak2001@gmail.com

DARURU PREM KUMAR CHOWDARY premchowdary1754@gmail.com

DERANGULA NARESH nareshnareshderangula7462@gmail.com

KASIREDDY SUDARSHAN REDDY sudarshanreddy855@gmail.com

Dr. RAGAVENTHIRAN ragaventhiran@presidencyuniversity.in

Abstract:

Thyroid disease is a common endocrine disorder that affects millions of people worldwide. Nowadays most of the women suffering from thyroid disease than male. There are two types in thyroid disease like hypothyroid and hyperthyroid disease. These diseases giving many side effects such as weight gain, weight loss, stress and so on to our human body. If this disease is detected in earlier stage, then physician can give proper treatment to the patients. Existing methods for thyroid detection, such as manual palpation and ultrasonography, have limitations in terms of sensitivity, specificity, and accuracy. Accurate diagnosis and classification of thyroid disease are critical for appropriate treatment and management of the disease. Machine learning techniques have emerged as promising tools for the automatic classification of thyroid diseases based on clinical, laboratory, and imaging features.

In this paper, we review recent studies on thyroid disease classification using machine learning techniques. We summarize the common machine learning algorithms used for thyroid disease classification decision tree, support vector machines, XG Boost, K-nearest neighbours and AdaBoost.

Overall, the studies suggest that machine learning techniques can achieve high accuracy in the classification of thyroid diseases, with some models achieving performance levels that are comparable to or better than those of human experts. However, further research is needed to validate the accuracy and generalizability of these models in large and diverse patient populations.

Keywords:

Machine Learning, decision tree, support vector machines, XG Boost, K-nearest neighbours, AdaBoost.

INTRODUCTION:

One of the most diagnosed and neglected ailments belongs to the endocrinology subgroup known as thyroid disease. Thyroid and other gland issues are second only to diabetes in terms of occurrence among endocrine disorders worldwide, according to the United Nations Medical Agency. The prevalence of hyperfunctioning hyperthyroidism and hypothyroidism is about 2% and 1%, respectively. The overall number of men is roughly one eighth that of women. Hyper- and hypothyroidism can be set on by thyroid cells dysfunction, either directly or indirectly because of pituitary or hypothalamic dysfunction. In some regions, a lack of dietary iodine can lead to goitre or active thyroid nodules, with a prevalence of up to 15%. A potentially risky location is the thyroid gland, which can also host a variety of cancers.

An endocrinologist gland, the thyroid gland is what creates hormones and delivers them to various parts of the body. It is situated in the upper midline of the body. Hormones produced by the thyroid gland are in charge of digesting and preserving the waste products of the body equilibrium and moisture content. Triiodothyronine (T3), thyroid hormone (T4), and thyroid stimulating hormone (TSH) are all treatments for the thyroid gland. The two types of thyroid diseases are hypothyroidism and hyperthyroidism. The data mining process is a partially computerised technique for looking for relationships in enormous amounts of data.

When a person has hyperthyroidism, their thyroid gland produces an excessive number of hormones that regulate thyroid function. Thyroid levels of hormones rising results in hyperthyroidism. Some of the symptoms include chapped lips, higher temperature sensitivity, lost hair, dropping weight, a rise in cardiac activity and blood sugar levels, excessive perspiration, neck expanding, brief period intervals, unexplained tummy actions and hand trembling are all symptoms of anxiety. In the illness known as a thyroid condition, the thyroid gland is underactive. Hypothyroidism results from a drop in production of thyroid hormones. Hypo means insufficient or lack of on medical jargon. Fever and injury to the know that the thyroid are both main roots of synthroid. Being obese, poor cardiac activity, raised temperature awareness, cervical edoema, cracked lips, tingling palms, difficulty with hair, heavy periods, and intestine difficulties are a few of the indicators. If ignored, these aches and pains may get greater gradually.

One of the greatest answers for many challenging situations is machine learning algorithms. We investigated along with categorised thyroid disease here because machine learning techniques have a major impact in dividing related to thyroid gland illness along with those techniques are extremely successful and fast and help in grouping. The categorization is a gathering of data approach (machine education) used to forecast and recognise several illnesses such as thyroid illness. Despite the fact that automated instruction and intelligent technology have been used in medicine since its inception, there has recently been a push to address the demand for medical facilities powered by algorithms for learning. As a result, according to experts, artificial intelligence might soon be widely used in the healthcare industry.

Literature survey:

Dr. Syed Mutahar Aaqib, Umar Sidiq, and Rafi Ahmad Khan [15] One of the most well-liked automated methodologies for data mining is grouping, that serves to describe preset sets of information. A classification is frequently employed in the field of health care to support medical oversight, diagnosis, and making choices. A famous Kashmiri institute provided the data for this investigation. The ANACONDA3-5.2.0 framework will be used to carry out the full scientific study. Methods for categorization like k closest neighbours, vector support machines, decision trees, and naive bayes may be employed in an experiment. The Judgement The tree has a 94.89 percent success rate, which is the highest of all the other groups.

Mrs. K Sindhya [16] A long-term disorder that impacts patients all around globe is thyroid dysfunction. Fantastic results are being achieved in the field of healthcare via the analysis of data in terms of illness forecasting. The cost of forecasting is low and the reliability of data mining approaches is strong. The fact that forecasting requires a short period of time is an additional significant benefit. In this work, I analysed thyroid data using algorithmic classification and produced a conclusion. Two criteria are the main determinants of a model's effectiveness. Forecast precision comes first, followed by forecasting time. Our results show that Nave Bayes forecasting took only 0.04 seconds to complete. It is, however, less precise than J48 and Random Forest. The Random Forest model achieved forecasting accuracy of 99.3 percent. The design's development took a bit longer than the prior two incarnations, though. Therefore, given that J48's precision is 99 percent, which is amongst the greatest, and that it runs in just under two seconds, which is a lot less time than the model using Random Forests, we can guess that it is the most suitable model for hypothyroidism forecasting.

Rasitha G. Banu [14] The one of the greatest widespread ailments that affect people is thyroid disease. The campus of California, Irvine (UCI) data repository provided the undernourished information used in this investigation. The entire study endeavour will be conducted on the WEKA platform, which stands for the Waikato Valley of Data Investigation. The decision root tree methodology was proven to be less efficient than the J48 methodology. The recognition of diseases is a challenging task in the field of health care. Many kinds of methods based on data mining are applied in the course of making choices. In this investigation, we used J48 and decision root data mining software for classification to identify hypothyroidism and diminished dimensionality to choose a smaller number of variables from the initial results. The final result of the predictor is evaluated in terms of sensitivity and degree of error using the unpredictability array. 98.56 percent efficiency, that is greater than the decision stumps tree efficiency and offers a lower mistake rate than Decisions the stump, is achieved by the J48 methodology.

Parkavi, Amina, and Begum [20] The majority of recent work deals with defining the categories of hyperthyroidism and hypothyroidism, a pair from the greatest common thyroid disorders in the society at huge at large. Four independent categorization models were examined and contrasted by the researchers: Naive Bayes, Decision Trees, Multilayer Perceptrons, and Radial Basis Function Networks. All of the methods of classification discussed earlier have a high level of accuracy, according to the results, given that the Option Tower framework an excellent rating for classification. Taking information the Romanian database source and the machine learning repository at UCI, the prediction model was created and evaluated. Weka for the first and KNIME Intelligence System are two distinct info packages. The development and validation of the categorization models was based on data mining techniques. Based to the literary works, numerous investigations in the subject of thyroid grouping build reliable learners using a range

of data mining methods. In order to divide up thyroid disorders such hyperthyroidism and hypothyroidism, the writers of this study investigated the usage of four classification models on thyroid data. The choice matrix technique is the appropriate classification model in each of the examined scenarios.

Khushboo and Chandel [13] In the present research, numerous models of classification for thyroid-related illnesses are employed and are dependent on characteristics. including TSH, T4U, and goitre. To bolster this claim, a variety of categorization approaches, such as K-nearest neighbour, are applied. The algorithms Naive Bayes and Support Vector Machines are used. K-nearest neighbour was found to be more accurate than Naive Bayes in diagnosing thyroid disease in the test, which was conducted employing the Fast a miner tool. Data-driven A system for classification was used by scientists to identify thyroid disorders. When evaluating a condition, thyroid dysfunction is an important element to take into account. In this work, KNN and Naive Bayes classifiers were employed. These two algorithms are compared using the Rapidminer programme. The outcomes showed that the classifier based on Naive Bayes has an accuracy of 84.57 percentage form, while the K-nearest neighbours classifier has a consistency by 93.44 percent. Outcomes are boosted by the recommended KNN technique's increased accuracy in classification. Naive Bayes may only have a straightforward, elliptical or parabola selection margin as a result; hence, KNN's selection margin constancy is a major advantage. As the variables are interrelated, KNN performs better than most approaches.

Objectives:

The major goal is to create an approach that can identify the sort of thyroid illness a patient is suffering from.

to use the greatest possible number of factors to forecast thyroid illness.

To forecast every potential thyroid disease kind.

EXISTING METHOD

Sensitivity, specificity, and accuracy are limits of currently used thyroid methods of identification including tactile examination and ultrasound.

The gland that produces thyroid hormone is physically examined by a health care expert using hand manipulation. However, the success of this procedure depends heavily on the knowledge and expertise of the examiner, and it might not be able to find tumours that are small or deeply buried. Further, if the examiner is unqualified, the patient may feel painful during the examination, which increases the possibility of thyroid damage.

Sonic waves with a high frequency are used in an ultrasound to produce pictures of the gland that produces thyroid hormone. Despite the fact that this procedure is safe and can give a detailed picture of the organ and any tumours that may be in existence, it is additionally operator-dependent and might miss tiny nodules that are situated in regions that are challenging to see. In addition, contingent upon the physical features of the nodule, ultrasonic efficiency in differentiating harmless from malignant tumours will differ greatly, and additional screening or excision may be needed for an accurate identification.

Proposed Method:

Several machine learning algorithms have been put out to categorise thyroid disease, but none have satisfactorily resolved the issue of incorrect diagnoses. Additionally, identical studies that presented models for evaluating this illness categorization frequently ignore the varied nature and quantity of the data. As a result, we suggest Support Vector Machines, Decision Trees, XGBoost, AdaBoost, and KNN. Performing classifiers testing based on machine learning.

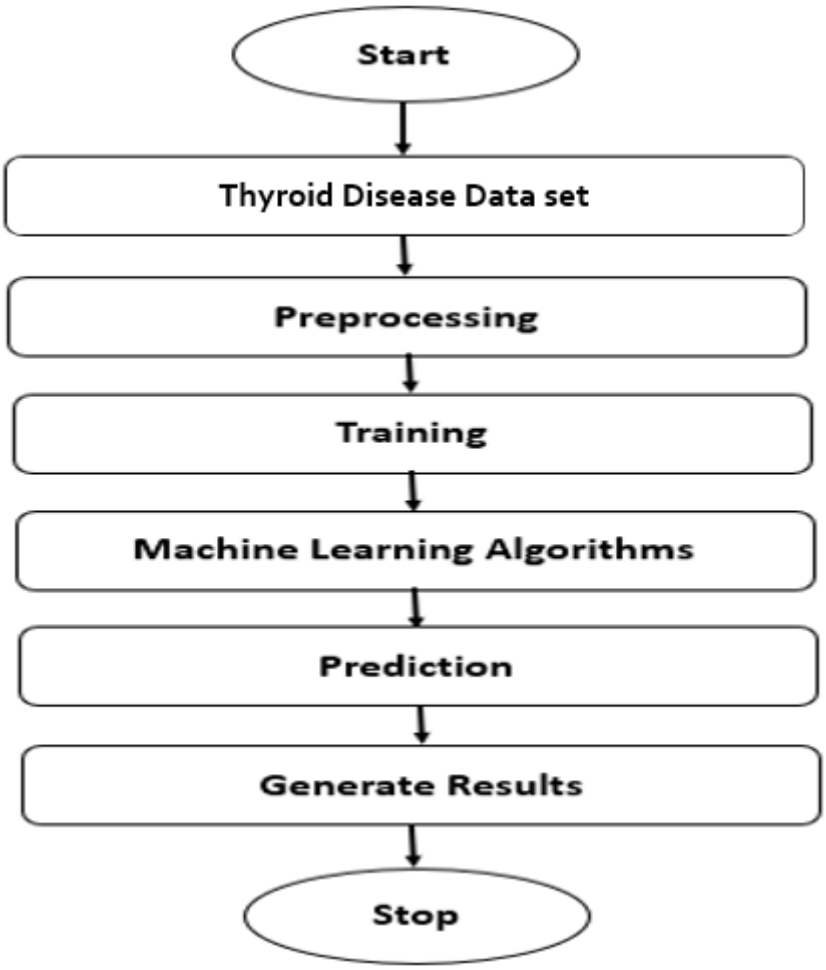
Advantages:

- Highest accuracy
- Reduces time complexity

Methodology

Data Collection:

We were able to gather an enormous amount of information on thyroid conditions as well as are now employing the information in our research on the categorization of illness. Machine learning strategies are utilised in the swift and prompt detection of thyroid function illness along with various illness because they now hold an important place in health care and aid us in recognising and categorising illness. Information from outside healthcare facilities and labs that specialise in analysing and treating illnesses were collected and utilised by me for our investigation. The specimen was used Information on the Iraqi population and their thyroid illness kind was extracted from the records, and information on 1250 males and females with ages that varied from one year to One year was collected. Ninety years of age, as these specimens include both healthy individuals and those without thyroid conditions who have hyperthyroidism and hypothyroidism as well as thyroid disorders. The information was gathered over a between one and four months timeframe with the primary objective of applying learning methods to categorise gland illnesses. . As the information collected comprise 31 factors or qualities, all of them were used in our research to figure out(['age', 'sex', 'on_thyroxine', 'query_on_thyroxine', 'on_antithyroid_medication', 'sick', 'pregnant', 'thyroid_surgery', 'I131_treatment', 'query_hypothyroid', 'query_hyperthyroid', 'lithium', 'goitre', 'tumor', 'hypopituitary', 'psych', 'TSH_measured', 'TSH', 'T3_measured', 'T3', 'TT4_measured', 'TT4', 'T4U_measured', 'T4U', 'FTI_measured', 'FTI', 'TBG_measured', 'TBG', 'referral_source', 'Target', 'ID'])



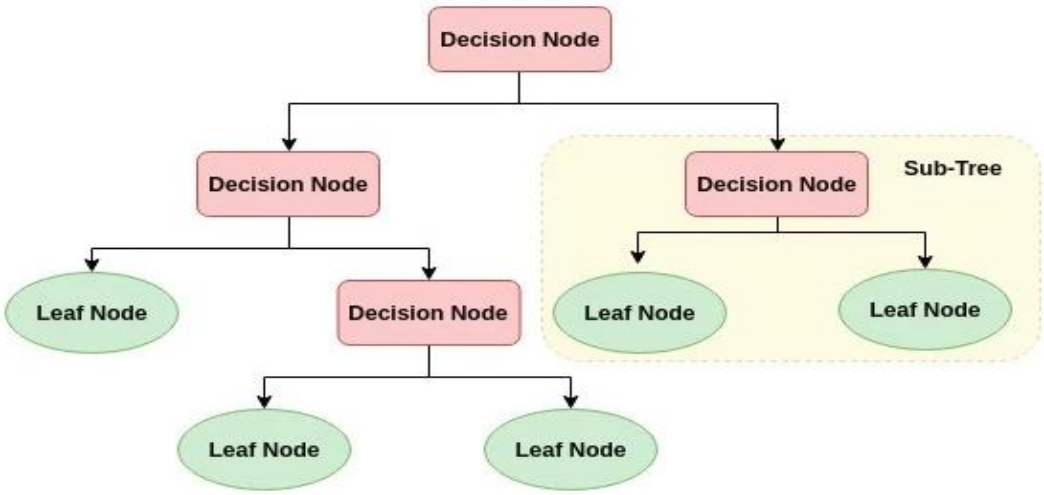
Data Preprocessing:

Pre-processing the information is a crucial stage in data analysis because it makes a positive impact on the information. The pre-processing method is applied to disclose the material by analysing it as well as finding what was deleted because it carefully investigates the information. Data preparation and cleansing are all part of the pre-processing phase. We cleaned & organised the information that they had the opportunity to get in that phase or gradually and we also discovered a set of missing information where the absent characteristics are. We were capable of to analyse the data that was missing through changing it with the current value of the intermediary after identifying the attributes that were the ones absent it, including T4 by the amount 151 and T3 by the amount 112. As a result, our team was able to get the information in a satisfactory and superior way and recover about missing information, as the information grew organised and satisfactory and was free via any flaw or issue so that that we were able to operate on it easily and effectively. Additionally, we combined the MLP algorithm with normalisation techniques.

Machine Learning Techniques

DECISION TREE:

An interior graph indicates a characteristic (or property), an extension indicates a choice rule of thumb, and every leaf node indicates the result in a tree of decisions, which resembles a schematic. The starting point in the decision tree is the first branch from the highest pointIt acquires the capacity to partition the information into subsets based on the data points of the characteristic. Repetitive The act of continuously separating branches is known as partitioning. This building apparently looks like a diagram, aids in making decisions. It is a flowchart-like visualisation that exact replication of human thought. Decision-making models are simple to as a result of it, understand along with evaluate.



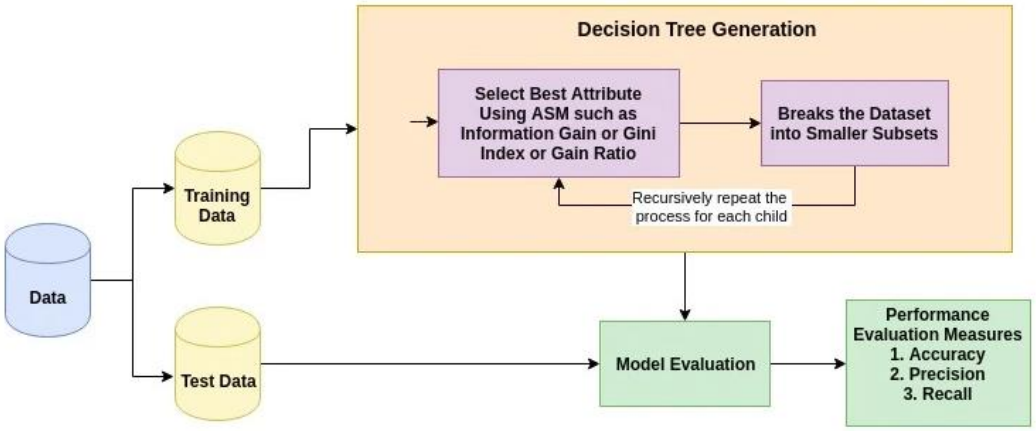
Any decision tree procedure's fundamental principle is as listed below:

1. To divide each record, choose the most suitable parameter using Attribute Selection Measures (ASM).

2. Divide the information set up into smaller groups to create that characteristic the choice branch.

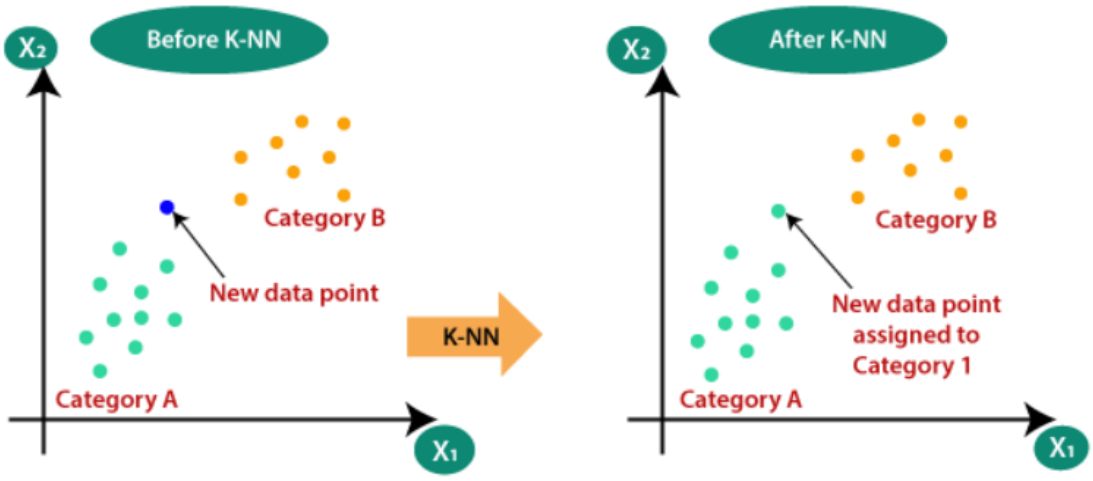
3. Repetitively repeats the procedure for every kid to begin growing the tree until one of the required is met:

- A single Each one and every pair of tuples has the same property worth.
- There were no longer any traits to be found.
- No further occurrences exist.



K-Nearest neighbours:

In contrast to the other techniques, the k-nearest neighbour algorithm classifies information immediately without initially creating a model. As a result, no specific modelling is required, and k, the number of nearest neighbours to utilise in class a membership estimate, is only one factor in the design: The ratio of the number of excellent member across x's k closest neighbours is what determines the value of p(y/x). The framework will become less or more stable depending on whether the value of k is little or large (respectively). The simplicity of k-nearest neighbours' technique makes it superior than other ones. Neighbours can offer an explanation for the categorization outcome; such method of reasoning is useful when black-box models aren't comprehensive. The process of determining the radius scope of the scenario, which calls for identifying a measurement that estimates the distance between data objects, is the predominant drawback of k-nearest neighbour.



The K-NN is in operation.

Step 1: Decide which neighbour has K as their K-number.

Step 2: Find the Euclidean separation amongst K partners.

Step 3: Pick the K closest peers dependent off of the Euclidean radius that was computed.

Step 4: measure the number information points there are in every group amongst those k peers.

Step 5: Assign the fresh information to the group of data points where the neighbour count is highest.

Step 6: The design is complete.

ADABOOST CLASSIFIER:

Yoav Freund and Robert Schapire proposed the Ada-boost or Adaptive Boosting ensembles boosting algorithm in 1996. To enhance the classification technique efficiency, It integrates a number of classifications. An continuous AdaBoost's aggregation methodology. AdaBoost classification integrates several unsuccessful plans classifications to develop an efficient classification algorithm with an elevated level of performance. The primary premise of Adaboost is to provide instruction users information in the samples and set the classification weights in every round in a manner offering reliable forecasts of uncommon data. An machine learning methodology that takes predefined parameters of training data can be used as the basis predictor.

Adaboost is limited by a pair of rules:

1. The predictor needs to be continuously trained using a variety of weighed samples for training.

2. By reducing learning error, it seeks to offer a superb fit for these samples in every iteration of training.

The process is as follows:

1. Adaboost firstly decides on an exercise sample at arbitrarily.

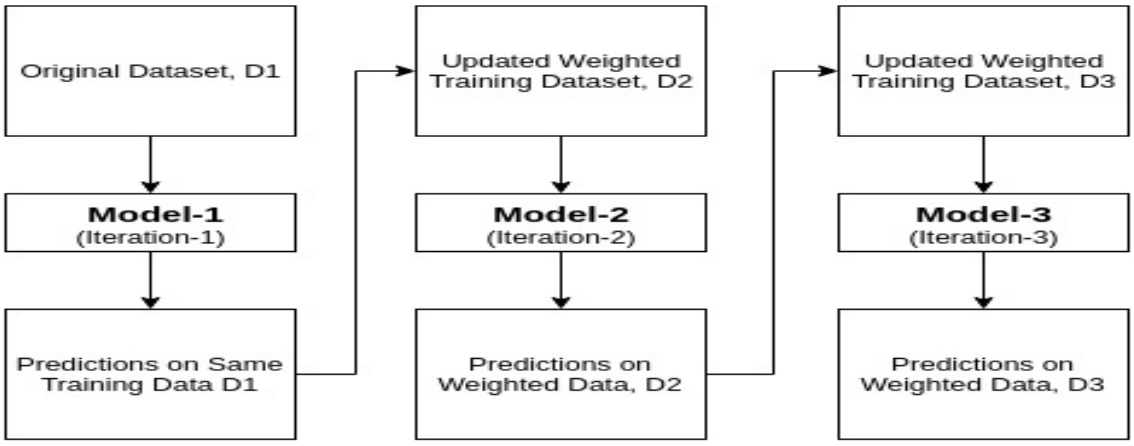
2. by picking the instruction sample according to the accurate forecast Without any prior instruction, it repeatedly learns the AdaBoost learning model.

3. It In error delivers categorised observation a heavier weight to increase their likelihood of getting detected correctly categorised in the following round.

4. In every iteration, the amount of weight is assigned to the trained predictor based on how accurate it is. The classifier that is There can be greater clarity. Additional size.

5. This approach up to the full workout set works, repeat flawlessly or until the given ultimate amount of estimation models successfully attained.

6. Cast a "vote" among all of the prediction methods you created in order to categorise.



XGBOOST:

"Extreme Gradient Boosting" is the abbreviation for XGBoost. XGBoost is a networked gradient boosting toolkit to that extent optimised should be extremely efficient, flexible, and adaptable. It uses the Gradient Boosting technology for developing machine learning techniques. It offers the use of parallel tree strengthening to swiftly and precisely handle a range of issues related to data science.

Boosting

Boosting is a group discovering methodology which generates a powerful classification created from a series of weaker classifiers. To contend with the bias-variance compromise, raising techniques are essential. Boosting methods regulate all the components of a model—bias and variance—and are believed to have been better with bagging procedures, they only provide additional variability.

Just a few forms of boosting algorithms are listed below:

- AdaBoost (Adaptive Boosting)
- Gradient Boosting

- XGBoost
- CatBoost
- Light GBM

Extreme Gradient Boosting is known by the abbreviation XGBoost. Due to its ability to scale lately increased in fame and is currently achieving Kaggle competitions for structural data and applicable machine learning.

Gradient-boosted decision trees (GBM) have an extension called XGBoost, which was developed particularly to boost effectiveness as well as speed.

Support Vector Machine:

Support vector machines are typically thought of as a categorization strategy, and their yet are often applied on both types of problems categorization- and regression-related issues. Numerous categorical as well as Continuous variables are straightforward to manage. To separate different groups, SVM creates a plane of division across three dimensions. SVM progressively creates the perfect grid to minimise loss. identifying a best-classifying maximum marginal hyperplane (MMH) to determine the information in question is the central goal of SVM.

A model created with SVM is just a plane in space with multiple dimensions that represents a number of classes. SVM will produce a hyper plane in a sequence in order to reduce loss. SVM seeks to find the largest margin hyperplane (MMH) by classifying the available data sets.

The ones that follow are basic SVM ideas effectively:

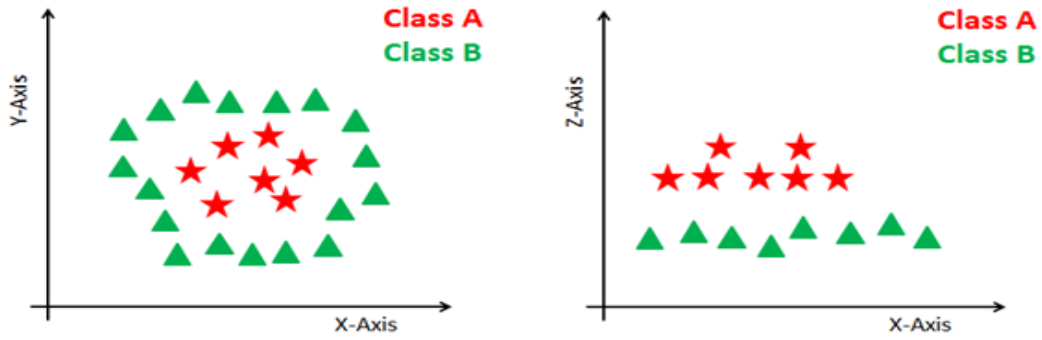
- Support vectors – These are the information that are most closely related to the hyperplane. These pieces of information will be used to define the dividing line.
- Hyperplane – As seen in the diagram previously, this is a choice planes or region that is partitioned amongst a group of items of various types.
- Margin – The distance between two lines on the closet data points of different classes can be referred to as the border. The angle that is perpendicular relationship the directions as well as the boundary line that support it can be used to compute it. A large margin is viewed favourably, whilst a small margin is unfavourably.

Separation is a crucial objective of the data set provided as effectively as feasible. The width of the margin is the separation among the two closest points. The goal is to choose that hyperplane in the data set in question that has the

greatest difference among the support vectors. The procedures that SVM executes to obtain the highest marginal plane are as follows:

1. Make separation-effective the hyperplanes the groups of people. Figure showing three separate intersections on the left in orange, blue, and black. despite the fact that the black in this particular case adequately distinguishes the two separate teams, the blue and orange exhibit more serious misclassifications.

2. As shown in select a suitable plane in the closest figure. has the greatest separation between the 2 closest a few informational points.



Result:

We employed a vari2 closestniques for machine learning with the information we provided. As this training is the initial training on this information, we split the current information into two sections: 30% for training and 70% for testing. In the initial phase, we submitted every characteristic in our information to the group of methods listed in the list further down, and the outcomes that emerged were those that were displayed there. This useful component is being put into practise using the Python programming platform, which is thought of as a full-fledged integrated framework. All of the qualities, which include 28 inputs and 1 output, have been taken.

NO	Algorithms	Accuracy
1	Decision Tree	96.26%
2	KNeighbors Classifier	97.52%
3	Ada Boost	98.28%
4	XGBoost	98.63%
5	Support Vector Machine(SVM)	98.99%

Conclusion

One of the illnesses that affects everyone worldwide and is becoming more prevalent is thyroid disease. Our study examines the categorization of thyroid disease between hyperthyroidism and hypothyroidism in light of medical reports that demonstrate substantial abnormalities in thyroid diseases. Algorithms were used to classify this illness. Using multiple methods, machine learning produced favourable outcomes. The accuracy of the SVM produced a value of 98.99% in the model, which is the highest accuracy of the other algorithms when considering all the features with 25 inputs and 1 output.

However, it is important to note that the success of machine learning models in thyroid disease classification heavily relies on the quality and quantity of the input data. Adequate and diverse datasets are essential to train robust models that can generalize well to new cases. Furthermore, the interpretability and explain ability of machine learning models remain important considerations, particularly in the medical field, where transparency and trust are crucial.

In summary, machine learning techniques have shown great potential in the classification of thyroid diseases. With further advancements in data collection, algorithm development, and collaboration between researchers and medical professionals, machine learning can continue to play a valuable role in improving thyroid disease diagnosis and management.

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/. 2nd International Conference on Physics and Applied Sciences (ICPAS 2021) Journal of Physics: Conference Series 1963 (2021) 012140 IOP Publishing doi:10.1088/1742-6596/1963/1/012140 11

[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.