Thyroid Disease Classification

Arsha Walia, Bhavesh Uppaluri, Vedic Tak and Astha Agrawal

***Abstract*—We have glands in our entire body, and they are distributed according to the requirement to release any substances that are required by our body. Thyroid is a vital gland for our body which helps in many functions. Men, women, teenagers, elderly, and infants alike can be affected by the thyroid disease. A person would be more likely to come across with this disease if they happen to have a history of thyroid disease in the family. In terms of ease of diagnosis, relative visibility and the ease of diagnosis, thyroid diseases are very much different. Early diagnosis and treatment remain the key factors. Different types of thyroid states are: Hyperthyroid, Hypothyroid, Euthyroid- sick, and Euthyroid (negative). In our journey through the project, we collected and worked on the data and then we used normalization followed by data splitting. Then, we trained several classification models: Logistic Regression,Naive Bayes,Decision Tree,Random Forest,SVM (Support Vector Machine),Gradient Boost,Adaboost and Bagging. Later we did hyperparameter tuning to evaluate models. We then ended the project with visual metrics. The first time we got the accuracy of 74.54% and later on after modifying the algorithm, we got an improved accuracy of 87.34%. This final result was obtained using the Decision Tree Classifier after Hyperparameter tuning.**

After training the data and testing it with all the models mentioned above, the final result we obtained was from DecisionTreeClassifier(maxfeatures='auto', maxleafnodes=2, randomstate=1). This model gave us the highest accuracy i.e. of 87.34491315136476%.

I. Introduction

THE thyroid is a gland. The thyroid gland is located in the front of the neck and is wrapped around the trachea(windpipe). This is a rather small gland, shaped like a butterfly. We have glands in our entire body, and they are distributed according to the requirement to release any substances that are required by our body. Thyroid is a vital gland for our body which helps in many functions of the body. Men, women, teenagers, elderly, and infants alike can be affected by the thyroid disease. This disease can be found in the body at the time or birth (typically hyperthyroidism), and it can even go forward and develop at a later stage of life generally after the menopause stage for women. A person would be more likely to come across with this disease if they have a family history of thyroid disease. The presence of some medical conditions like type 1 diabetes, pernicious anaemia, rheumatoid arthritis, Sjogren's syndrome and Turner syndrome can also be the cause for thyroid related problems. Intake of excess Iodine or being above 60 years of age (especially in women) can also be seen as some factors that affect in the turnout of thyroid related diseases. In terms of ease of diagnosis, relative visibility and the ease of diagnosis, thyroid diseases are very much different. Early diagnosis and treatment remain the key factors. Different types of thyroid states are: Hyperthyroid, Hypothyroid, Euthyroid- sick, and Euthyroid (negative).

Hyperthyroidism: Also known as overactive thyroid. This is a condition where; the thyroid gland starts releasing high amounts of thyroid hormone into the body and results in an increased metabolism. Symptoms include weight loss, increased appetite, anxiety, rapid heartbeat, etc. Here are some people who are more likely to have hyperthyroidism: people who have a family history of thyroid issues and problems, people having previous medical records that include problems like type 1 diabetes, anaemia, and Addison's disease; people who have an excess of mineral iodine (mineral used for production of thyroid hormones).

Hypothyroidism: This is a common condition in which the thyroid gland doesn't secrete the required amount of thyroid hormone into the blood. This results in slow metabolism. Symptoms include weight gain, tiredness, and inability to tolerate low temperatures. One of the most common primary causes of this type of disease is Hashimoto's disease which is a type of autoimmune condition. Thyroiditis is the inflammation of the thyroid gland; this is one of the other primary causes of Hypothyroidism. Treatment of Hyperthyroidism may also lead to the condition of Hypothyroidism if not done very carefully and under careful observation. Iodine deficiency and family hereditary conditions also play a big role in this disease.

Euthyroid-sick: Euthyroid sickness is a condition where the serum levels of hormones are low in patients with no systemic illness. This condition is usually treated by avoiding endocrine replacement.

Euthyroid (negative): This is a state of not having normal functioning of thyroid gland. Having symptoms like Hyperthyroidism or Hypothyroidism.

A. Motivation

According to a study, in India alone, a population of 42 million suffers from thyroid related diseases. This is a worldwide problem, and it is claimed that among all the endocrine problems, thyroid disease is the most common. Even after all these years, there are still problems like misjudging of thyroid diseases or them being undiagnosed. In a time where technology is advancing beyond imagination it is hard to believe that this problem is still being neglected and undermined. There are major statistical proofs that show that this disease affects people worldwide and still hasn't been conquered by man. Here are some of the major reasons that have come forward after the research.

Around 12

Study shows that women are seven times more likely to get affected by this disease than man. In death, it is observed that half of the women had the gland inflamed whereas about quarter of men had inflamed thyroid. Women above the age of 60 years are shown to have an increased probability of coming across this condition. This is seen often in women after their menopause.

The symptoms of this disease vary from person-to- person. This leads to difficulty in diagnosing thyroid related problems. Sometimes, the symptoms can be misjudged for other irrelevant reasons and the main problem remains undiagnosed or gets misdiagnosed. As we already know, thyroid related problems are quite common, and they show symptoms that might resemble some other diseases. We need to eradicate this error in judgement and hence this project.

B. Background Knowledge

In recent years, much work has been done to diagnose benign thyroid disease. Many authors havedifferent types of data mining techniques. The authors have shown that they get adequate approach and assurance to reveal thyroid-like disease from work that includes various datasets and algorithms related to work to be done in the future for effective and better results. The intention of the work interprets various data mining techniques, mechanisms and statistical attributes that have become common in the interpretation of thyroid disorders in recent years, with the assurance of different authors of perspectives and different approaches. There are several Random Forest Count, Decision Tree, Naive Bayes, SVM and ANN machine learning algorithms that are widely used in common disease and prognosis problems.

Few features include diseases associated with heart disease [3], diabetes, Parkinson's disease, hypertension, Ebola virus (EV), diagnosis and prediction, analysis sequenced R-NA data, and biomedical image mapping. Despite Advances of Illness Positioned on Machine Learning, the mechanism of prediction and medical determination is a non-trivial task. There are core issues, such as capturing, compiling, and aggregating data, that are used to train machine learning installations., and are practically non- existent.

II. Related Works

1. Interactive Thyroid Disease Prediction System Using Machine Learning Technique

The intent of this work done is in order to cater the idiosyncratic techniques of Machine Learning in order to find the most appropriate medical results, and this project is specifically for Thyroid Disease Prediction. On the basis of the dataset available on UCI Machine Learning Repository author has created Analysis and Classification Models. In this work, they are mainly using Machine Learning Algorithms, support vector machine (SVM) SVM helps the researchers in performing the analysis in a precise way. SVM in the high

dimensional space can create a single hyper plane or multiple planes. Second algorithm which they are using is K-NN. The third algorithm is the decision trees in which there are 3 nodes as root node, internal node and leaf node. The internal node tests for a particular attribute, based on testing that the classes are assigned to the leaf node. The root node remains at the top of the C4.5 and ID3 decision trees. it is used to predict the risk of developing thyroid disease.

December 2018

2. Prediction of Thyroid Disorders Using Advanced Machine Learning Techniques

The main motive of the scientists was to diagnose the diseases as soon as possible or at least on time , so that there can be chances to treat the diseases at very early stage . This research was mainly focused on the techniques related to Data Mining which can predict the Thyroid Diseases along with various different classification technique. Techniques like KNN, Classification Model , Decision Tree, Native Bayes and SVM. One of the aim of this project was to find correlation of T3, T4 , THC towards different type of Diseases related to Thyroid . Native Bayes algorithm which can be also stated as earger learner since they can also build a model just after a training set is given . This Classifier is based on bayes theorem on conditional probability but here new problem have been seen which is in medical research . March 2019

3. Thyroid diagnosis based on the artificial immune recognition system (AIRS)

The most important concern diagnosis of thyroid disease is proper understanding of the functional data of thyroid gland. Major function of thyroid gland is to help control the metabolism of the body. This is provided by the thyroid hormone released by the thyroid gland. There can be two different thyroid disorders, it completely depends upon the thyroid gland, if it releases very little thyroid hormone then it leads to hypothyroidism whereas if glands produce more thyroid hormones it leads to hyperthyroidism. Artificial Immune System are recent yet powerful branch of artificial intelligence, Artificial Immune Recognition System (AIRS) as suggested has so far been one of the systems introduces in this area. Cross – validation process is used to check the strength of the samples.

4. Diagnosis Method of Thyroid Disease Combining Knowledge Graph and Deep Learning

This research proposes diagnosis of Disease based on Deep Learning and Knowledge graph to collect the scattered knowledge in various Medical Information System, taking Thyroid Disease as an Example. Initially it creates Biomedical Knowledge graph. The model used to diagnose disease is Bidirectional long-short term memory network which is trained by using Pathological Disease relationship Data.

August 2020

August 2020

A. Classification of Thyroid

In our project our main focus is that we will be classifying the different types of thyroid according to different symptoms, and for the classification we will use SVM (Support Vector Machine). SVM's are very accurate and shows good results for the small datasets which can be clearly divided and also SVM uses subsets of training points for attaining more accurate and efficient results. SVM will help in gaining good understanding of the thyroid data.

Support Vector Machines is a Machine learning Algorithm which is mostly used for the classification and regression purpose and is famous for its good and accurate results. Support Vector Machines uses geometric method for the classification

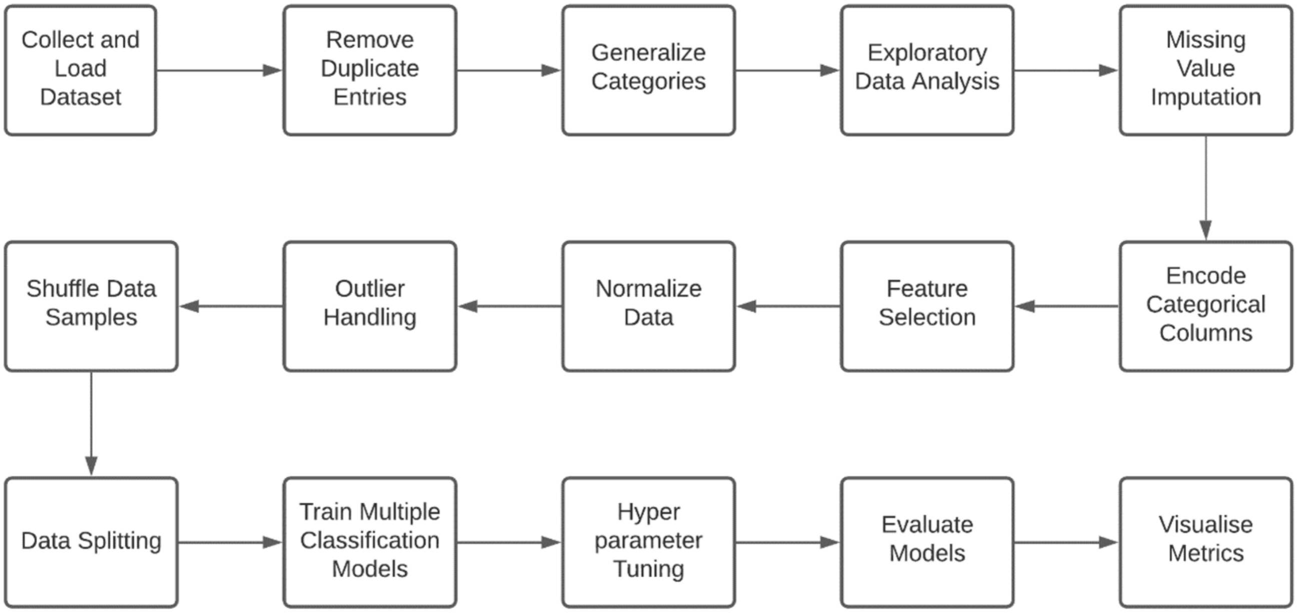


Fig. 1. Tentative approach model

and regression. In Support Vector machine the data is linearly divided into two parts, the linear line which best separates the set of data and classifies it is termed as Hyperplane. We will ideally target to find the clear Hyperplane in a single plane as if the data set classification is not defined than we would have to find a Hyperplane in 3D view. In 3D the dataset dots are to be separated by a plane rather than a linear line and this 3D separation mechanism is called kernelling. If the classification of dataset can also not be defined in this, then we will have to increase the dimensions until the distinction is proper. the linear line which best separates the set of data and classifies it is termed as Hyperplane. We will ideally target to find the clear Hyperplane in a single plane as if the data set classification is not defined than we would have to find a Hyperplane in 3D view. In 3D the dataset dots are to be separated by a plane rather than a linear line and this 3D separation mechanism is called kernelling. If the classification of dataset can also not be defined in this, then we will have to increase the dimensions until the distinction is proper.

The nearest data points to the hyperplane are called Support Vector. As clarified in the figure. Our target will be that the data lies as scattered and far as possible and also on the correct plane, as in the SVM the far the data points are the better is the result and we can also be sure that the classification is correct. The data will be divided into two classes and we will assign the classes on the basis of which side of the hyperplane the data falls on. The distance between the dataset point and the hyperplane is termed as Margin, there are many possibilities for the hyperplane but our target will be to choose the hyperplane so that the margin is greatest since the dataset points will then be far that will make the results more accurate.

One the drawback of Support Vector Machine technique is that it is less accurate and takes high training time for larger dataset and also for the datasets which are not having clear distinction with overlapping classes.

B. . DIAGRAM - WORKFLOW

As shown in the Fig. 1.,we first collect and load the thyroid dataset from kaggle repository found through the google datasetsearch into our local memory. We then pre-process the data, remove all duplicate samples from the dataset. We generalize the 8 classes of the target column (3 types of hypothyroid diseases [compensated hypothyroid,primary hypothyroid,secondary hypothyroid], 3 types of hyperthyroid diseases [goitre,hyperthyroid,t3 toxic], sick but not thyroid disease, negative) into 4 columns (hypothyroid, hyperthyroid, sick, negative). We visualize the dataset using uni-variate, bi-variate and multi-variate analysis methods to get a better understanding of the data and its distribution throughout. We plot histograms, boxplots and correlation matrix of the numeric data to find features with outliers and correlated features. After visualizing the data, we handle all the missing values by filling the empty values with the most frequent values of their columns. Then, if a given column is categorical and it has more than 5 unique values, we label encode that column otherwise if the categorical column has less than 5 unique values, we can one hot encode the column.After encoding categorical data, we select the 15 best features based on each feature's ANOVA f-value to determine the relevance of the feature with respect to the target we normalize the complete dataset into a range of 0 to 1 using min-max scaling method, other methods can also be decided manually depending on their accuracy metric. Although there should not be any more outliers left after normalising the data, but we can check it once by finding skewness from a normal distribution curve, and if we find any

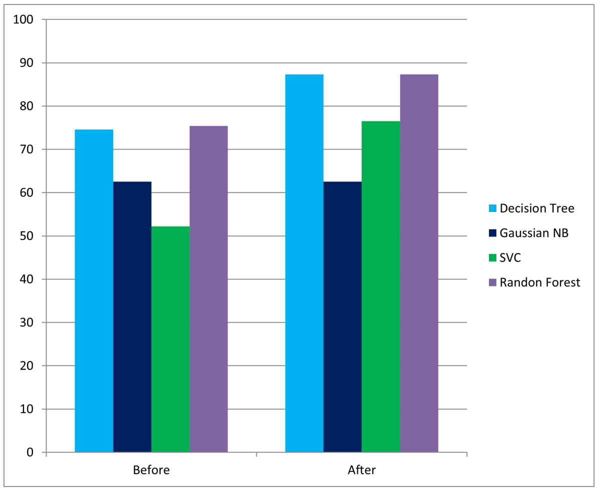


Fig. 2. Accuracy of different Models before and after Training

columns with outliers we can handle those columns through the log transformation method which replaces the values of the columns with their logs. After all these pre-processing methods we randomize the dataset entries to ensure the trained model stays generalised. We split the dataset into training and testing datasets and use the training dataset to fit the various multi-class classification models. the trained models are the evaluated using the testing dataset through various classification metrics like accuracy, precision and recall, f1-score. In the end, we display the metrics of the various models and the classifications of the generalised data and determine the model with the best metrics as the best model for this problem

III. PROPOSED METHODOLOGY

A. Data Manipulation

Target column has values in the format of ’category.—score’ where category is the classification of the particular data and score is a value calculated based on some unknown parameters defined by the publisher of the dataset. As we don’t know the use of this score we would only use the category data to train and validate our models. There are 8 categories in the original dataset - 3 types of hyperthyroid diseases, 3 types of hypothyroid disease, sick but not thyroid disease, negative - these categories are generalised to 4 classes (hyperthyroid, hypothyroid, sick, negative) for the sake of simplicity. The dataset comprises of 29 features and 1 target for all 3221 unique samples, 7 features are of integer data type and all the remaining features are nominal in nature. Other than the T4U column, all the integer type features (Age, TSH, T3, TT4, FTI, TBG) have a large number of outliers. TBG feature is completely empty while TBG Measured feature is constant. Missing Values are handled using Most_Frequent Strategy, filling empty values with the mode of the particular column. Non-numeric features which have more than 5 unique values are Label Encoded and those which have less than 5 are One-Hot Encoded and inserted back into the same index

B. Feature Selection

We have used ANOVA f-value of each of all the features using f_classif method to determine the top 15 most relevant features to the target.

Analysis of variance (ANOVA) is used to statistically assess the equality of means when you have four or more groups by using F-tests, and the target variable like classification predictive modelling , the two most common used feature selection methods are ANOVA F-test statistic and the mutual information statistic. A F-statistic is termed as the ratio of two variance and was names after Sir Ronald Fisher . Technically it can also be termed as two mean squares.

C. Algorithm and our approach

We have used several methods in our project to classify the thyroid disease. Here we have explained in detail all of them. Algorithms :-

1) Logistic Regression :- it is a classification model rather than a regression model, as its name suggests. Logistic regression is a more efficient and simple method for linear and binary classification problems. It is a very easy to build classification model with excellent performance with linearly separable classes. Logistic regression classification model algorithm is used widely in industry. The logistic regression model, like Adaline and perceptron, is a statistical method of binary classification that can be generalized to multi-class classification. Scikit-learn has a highly optimized version of the logistic regression implementation, which supports the multiclass classification activity (Raschka, 2015).

Logistic Regression and its types: 1. Binary Logistic Regression 2. Ordinal Logistic Regression 3. Multinomial Logistic Regression

2) Naïve Bayes:- The Naive Bayes model is a very easy-to-build model and it is mostly preferred because of its usefulness for very large data sets. This is based on Bayes’ theorem. Although it is simple, Naive Bayes is also well known for leaving behind the most complex and advanced methods of classification.

3) Decision Tree:- This Machine learning algorithm splits the data into subsets. The prime motive of this algorithm is to compute the training data in the more small possible tree. This Partitioning begins with a binary split and continues till no further can be made. In this model various branches of different length are formed. Steps which are included in this model are: Splitting, Pruning, Tree Selection. This model is mainly used among non-statisticians as their model is very easy to explicate. Each leaf node is presented as an if/then rule, and the cases which get satisfy this condition statement are placed in the node. Data like continuous, ordinal, categorical, and binary can also be handled by this algorithm. It is also useful for detecting important interactions and identifying outliers. It also has the feature of handling missing data by identifying surrogate splits in the modeling process.

4) Random Forest :- This algorithm works with the decision tree algorithm , it works as a supervised machine learning algorithm . This model can be used in various industries such as e-commerce and banking to predict behaviour and outcomes.

This consist of many decision tree algorithms. The 'forest' which is generated by the Random Forest is trained through different algorithm that is bagging or bootstrap aggregating. The results produced through random forest algorithm is based on the prediction of the decision tree. it works by taking the average or mean of the output from various trees. Hence, we can say the increment in the number of trees will lead to increment in the precision of the outcome. It solves many issues like issue of overfitting in decision trees, even without hyper-parameter tuning it can produce a reasonable prediction. We can say that it is more accurate than the decision tree algorithm.

5) Bagging :- Bootstrap aggregating which is also known as bagging , is an ensemble learning method which helps in improving the accuracy of machine learning algorithms. And the performance . Bagging also helps in avoiding overfitting of the data. It is also used to deal with bias-variance trade-offs as it reduces the variance of prediction model. It can be used in both regression and classification model. This process starts with selecting a random sample from the training dataset with or without replacement. The next step takes place by creating a model using sample observations after a subset of m feature is chosen randomly. Bagging deals with higher dimensional data efficiently . This bootstrap Aggregation algorithm mainly creates multiple different model from single training dataset.

6) Gradient Boosting Algorithm: - it is one of the most powerful known algorithms; it is based on Machine Learning boosting Technique. Its working includes the building of the models continuously and combine the previous model with the next best precious model, it reduces the error in the prediction.

7) AdaBoost Algorithm: - AdaBoost means Adaptive Boosting basically used as an Ensemble method in Machine Learning, The weights in this technique are reassigned to each instance hence it is known as adaptive boosting, higher weights are assigned to incorrectly classified instance.

Architecture of the model: Gini impurity is the criterion used to measure the quality of a split. Best strategy is used to choose the split at each node. The maximum depth of the tree is none, therefore the nodes are expanded until all leaves are pure. The minimum number of samples required to split an internal node is 2. The minimum number of samples required to be at a leaf node is 1. The minimum weighted fraction of the sum total of weights (of all the input samples) required to be at a leaf node is 0.0, therefore the samples have equal weight. max_features which determines the number of features to consider when looking for the best split is set to "auto", which implies max_features=sqrt(n_features). random_state which controls the randomness of the estimator is set to 1 to keep the results consistent. Growing the tree with max_leaf_nodes=2 in best-first fashion.

D. Proposed Method

We are training various multi-class classification models - Linear Regression, Naive Bayes Classifier, Decision Tree Classifier, Random Forest Classifier, Gradient Boost Classifier, Adaboost Classifier, Bagging Classifier - and improving them using methods like hyperparameter tuning, stratified k-fold cross validation. Then we are evaluating the predictions made by these optimized models on testing data using multiple metrics - accuracy score, precision, recall and f1 score.

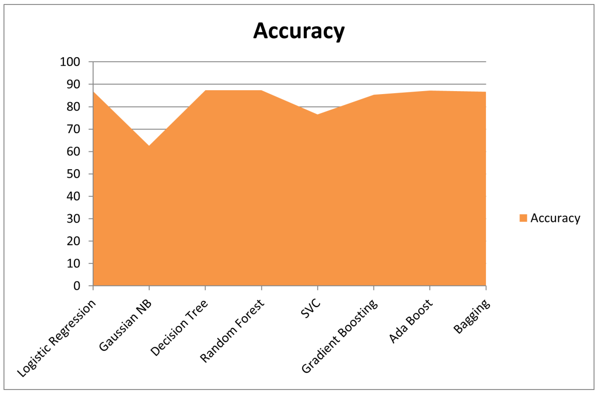


Fig. 3. Accuracy obtained after training

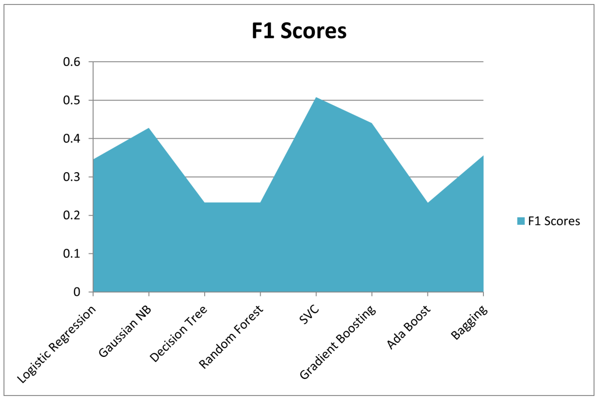


Fig. 4. F1 Scores obtained

IV. Result and Analysis

A. Hyperparameter Analysis

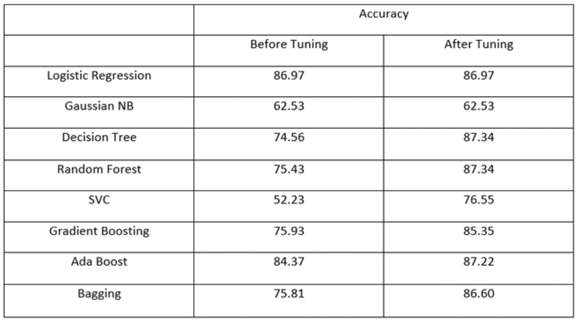
On training the data across several models such as Logistic Regression,Naive Bayes,Decision Tree,Random Forest,SVM (Support Vector Machine),Gradient Boost,Adaboost and Bagging, the first time around the results that were obtained were not very good.On using the DecisionTreeClassifier(), the accuracy of the default Decision Tree Classifier model came out to be 74.56%. This was not an acceptable result that we were looking for hence, we changed the parameters further till we improved the accuracy of the model.

We modified the algorithm again and this time we looked upon a better result: DecisionTreeClassifier(max_features='auto', max_leaf_nodes=2, random_state=1), the accuracy of the Decision Tree Classifier after hyperparameter tuning came out to be 87.34%.

B. Comparison

V. Conclusion

As we all know of the fatality of the thyroid disease, it is very important to not sideline or overlook this problem as the majority suffers from it. We, the students of Bennett



	Accuracy	
	Before Tuning	After Tuning
Logistic Regression	86.97	86.97
Gaussian NB	62.53	62.53
Decision Tree	74.56	87.34
Random Forest	75.43	87.34
SVC	52.23	76.55
Gradient Boosting	75.93	85.35
Ada Boost	84.37	87.22
Bagging	75.81	86.60

Fig. 5. Table 1. Comparison Table with respect to previous algorithm

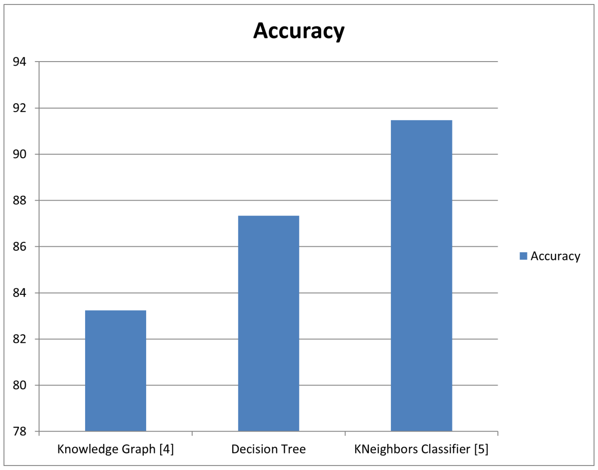


Fig. 6. Comparision of Prediction Accuracy with other related works

University, took the initiative to find a classification technique for this problem using machine learning. In our journey through the project, we collected and worked on the data and then we used normalization followed by data splitting. Then, we trained several classification models: Logistic Regression,Naive Bayes,Decision Tree,Random Forest,SVM (Support Vector Machine),Gradient Boost,Adaboost and Bagging. Later we did hyperparameter tuning to evaluate models. We then ended the project with visual metrics. The first time we got the accuracy of 74.54% and later on after modifying the algorithm, we got an improved accuracy of 87.34%. This final result was obtained using the Decision Tree Classifier after Hyperparameter tuning.

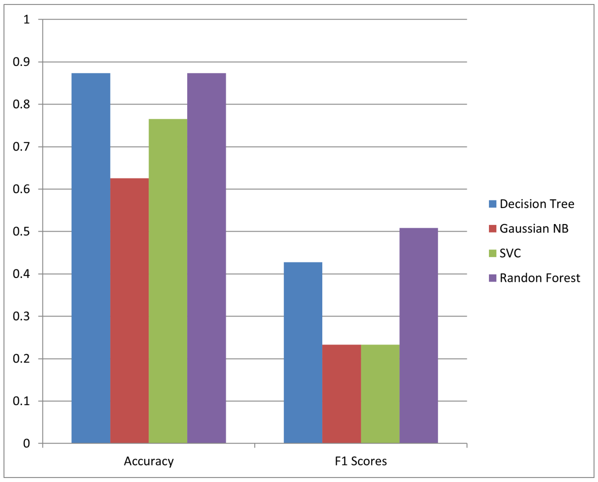


Fig. 7. Comparative Analysis

References

[1] Jerome M. Hershman, Euthyroid Sick Syndrome, Sept 2020.

[2] V. Vapnik, Estimation of Dependences Based on Empirical Data, Springer, New York, 2012.

[3] https://www.researchgate.net/publication/ 223190310_A_novel _hybrid_method_bas ed_on _artificial _immune _recognition_sys tem_AIRS _with_fuzzy_weighted_pre- processing _for_thyroid_disease _diagnosis

[4] https://ieeexplore.ieee.org/document/9167227/metrics#metrics

[5] https://iopscience.iop.org/article/10.1088/1742-6596/1963/1/012140/pdf