Thyroid Disease Treatment prediction with machine learning approaches

Aishwarya Sri Pati
Computer Science Department
University of Central Florida
Orlando, Florida
aishwaryasripati@knights.ucf.edu

***Abstract*—The initiation of thyroid illness is a challenging topic in medical study, yet it is a significant source of formation in medical diagnosis and prediction. One of the most crucial organs in our body is the thyroid gland. The thyroid's hormone releases are responsible for regulating metabolism. One of the two thyroid conditions that often affect people, hyperthyroidism and hypothyroidism produce thyroid hormones that control the body's metabolism. In order to make the data simple enough to do analytics to highlight the risk of patients developing thyroid, data purification methods were used. This study deals with the analysis and classification models that are employed in the thyroid disease based on the data acquired from the dataset taken from the UCI machine learning repository. Machine learning plays a significant part in the process of illness prediction. Making ensuring there is a solid knowledge foundation that can be ingrained and utilized as a hybrid model to solve complicated learning problems, like diagnosing and predicting illnesses, is crucial. We also suggested many machine learning methods for thyroid preventive diagnoses in this research. The estimated probability of a patient developing thyroid illness was predicted using machine learning algorithms, support vector machines (SVM), K-NN, and logistic regression.**

***Index Terms*—Thyroid,Thyroid hormones, Machine Learning, Classification,Logistic Regression, SVM, K Nearest Neighbours, Confusion Matrix**

I. Introduction

It's vital to note that illnesses of this kind mostly afflict women globally, particularly as they become older. Women are four to seven times as likely than males to have an eating problem, according to data from THYROID AWARE. than men's thyroid The International Thyroid Federation states in its publication [6] that symptoms of this disorder in people can include extreme fatigue, depression, anxiety, irritability, difficulty concentrating, headaches and migraines, lethargy, intolerance to cold or hot environments, alteration of menstruation, and even alteration in the autoimmune system. "Everyone has a thyroid gland, which is an organ that regulates metabolism and ensures that almost all of the body's organs operate properly. It is located in the lower portion of the anterior area of the neck. In response to the activity of thyroid-stimulating hormone (TSH), which is released by the pituitary gland located in the brain, this gland generates the thyroid hormones thyroxine (T4) and triiodothyronine (T3)."

Due to the above-mentioned conditions, it is imperative to suggest mechanisms that can help combat the issue, either to recommend treatments, discover the pathology in a person who is unaware of their health situation (it is estimated that only 50 percent of people who suffer from thyroid disorders are aware of their situation) [8], or in the best cases, predict it. Because of the aforementioned, doing this study requires incentive beyond the academic. It is suggested to employ technology techniques to explore the thyroid, as many individuals have done, in the instance It is specifically meant to employ machine learning in its supervised learning paradigm to be able to identify a person's potential thyroid condition assertively and swiftly based on information acquired from them.

A data collection that contains a history of the condition is required in order to implement the plan and build the project. In this instance, the data set that will be utilized is kept in a collection by the University of California, Irvine (UCI), which is a university in the United States. This collection is accessible in the university archive at the following address: https://archive.ics.uci.edu/ml/datasets/Thyroid+Disease, where you may discover several sets of thyroid-related data under the heading "Thyroid Disease Data Set".This dataset is supplied by the Garavan Institute and J. Ross Quinlan, New South Wales Institute, Syndney, Australia,1987.

In the project to be created, use will be made of the dataset that obtains the designation "allhypo," which has been accessible since 1987.

II. Motivation and Overview

Globally, thyroid disorders are on the rise, more so for women than for males. Humans' way of living is the primary source of this issue. The thyroid gland secretes thyroid hormone, which has a variety of uses including protein synthesis, calorie burning, and rate regulation of metabolism. Additionally, it controls the overproduction of hormones by other hormonal glands. Thyroid disorders come in two different forms. The first one, hyperthyroidism, is brought on by an abundance of thyroid hormone in the blood, while the second, hypothyroidism, is brought on by a deficiency of thyroid hormone in the blood. Both hyperthyroidism and hypothyroidism have adverse side effects on how the human body normally functions. For the majority of healthcare professionals, the most difficult duty in the area of medical science is making an early diagnosis of a health issue. It is necessary to

have someone with expertise and understanding, particularly in cases with thyroid disorders. Additionally, it could have symptoms with other medical conditions. Thyroid disorders may cause additional health issues and sometimes even death if they are not addressed at an early stage. Data mining methods play a significant role in the study of medical data for various disorders.

https://github.com/DATA_MINING_FINAL_PROJECT

III. Related Work

It was revealed that many individuals have worked on the same topic using various machine learning architectures throughout the process of collecting academic material to assess the status of the work. The following is a list of the five most representative scholarly papers that have been released. They are

A. Predicting Thyroid Disease Using Linear Discriminant Analysis (LDA) Data Mining Technique[1]

G. Rasitha Banu used the Linear Discriminant Analysis (LDA) function in the Weka tool to complete her work. Using cross validation and k=6, her prior method had a 99.62 percent accuracy rate for the classification challenge.

B. Predictive Data Mining for Diagnosis of Thyroid Disease using Neural Network[3]

Prerana, Parveen Sehgal, Khushboo Taneja, developed using the back propagation algorithm an accurate method for thyroid detection. Utilizing back propagation of error, an artificial neural network is created to determine the first thyroid prediction. Following training, ANN is tested using experimental data, but not the same training sets. Unsupervised learning and supervised learning are two methods for completing the training. The MATLAB Neural Network Toolbox Software is used to conduct the experiment and provide the results. This performs better than the straightforward gradient descent approach.

C. Efficient Thyroid Disease Classification Using Differential Evolution Systems[4]

The research conducted by K. Geetha and S. Santosh focuses on the use of a hybrid algorithm known as Differential Evolution, a technique from the family of evolutionary algorithms that is used to produce subsets of the parent records' children records. The T test, which calculates the likelihood of misclassification, was used to assess the team's classification accuracy, which was 99.89

D. Thyroid Data Prediction using Data Classification Algorithm[5]

Ammulu and venugopal has done research on hypothyroid dataset. The hypothyroid dataset is subjected to the data mining approach in order to identify the positive and negative instances across the board. To improve therapy, decision-making, and illness diagnosis, datasets are categorized. In their study, the random forest method from data mining is used to predict hypothyroid disease. Comparing the random forest with LDA method leads to better accuracy, precision, recall, and F-measure. Future research focuses on verifying several illness datasets concurrently, including those for heart disease, diabetes, and other conditions.

IV. Dataset Description

The webpage (http://repository.seasr.org/Datasets/UCI/) is where the data set utilized for experimental purposes may be downloaded. 3772 occurrences make up the data set, of which 3481 fall into the negative group, 194 into replacement therapy,95 into the underreplacment, and 2 into the overreplacement. There are 29 characteristics total, including the class, the final property, which will be utilized to categorize the data. Below is a display of the data set's details.

Attribute Name	***Value Type***
age	continuous
sex	M,F
on thyroxine	t,f
query on thyroxine	t,f
on antithyroid medication	t,f
sick	t,f
pregnant	t,f
thyroid surgery	t,f
I131 treatment	t,f
query hypothyroid	t,f
query hyperthyroid	t,f
lithium	t,f
goitre	t,f
tumor	t,f
hypopituitary	t,f
psych	t,f
TSH measured	t,f
TSH	continuous
T3 measured	t,f
T3	continuous
TT4 measured	t,f
TT4	continuous
T4U measured	t,f
T4U	continuous
FTI measured	t,f
FTI	continuous
TBG measured	t,f
TBG	continuous
referral source	SVHC,other,SVI,STMW,SVHD

V. Exploratory Analysis

A. Preprocessing

In the first step Since the target value column contains both categorical and numerical data, we divided the column and removed the unnecessary columns.

The data description shows clearly that there are no missing values. However, if incorrect entries like "?" are used to fill in the missing values in the collection. I then substituted these values with "nan" and once again checked for missing values.

We found that many columns, including age, sex, TSH, T3, TT4, T4U, FTI, and TBG, had several missing data. About 95 percent of the data in TBG are missing. Therefore, there is no need in using the columns, therefore we have removed that column.

Additionally, while viewing the dataset, it is possible to notice that certain columns include true and false values that simply indicate if the subsequent column has values or not.

All of these values have been changed to ”nan.” Before applying any imputation methods, convert the categorical values to numerical ones.

Using mapping for columns with two unique values and obtain dummies when there are more than two values since there are only two categories will result in two columns that have extremely high correlation because they both describe the same thing. Therefore, we had to remove one of the columns. For such columns, mapping is used.

Now there are 4 unique categories in our ouptut class as well. With our Output class, utilizing get dummies makes no sense; instead, simply map them using the LabelEncoder method.After observing the result, we understand there is no more encoding needed for categorical values. We used KNN Imputer to impute the missing values.After doing that,the dataset has no missing values. We’ll now analyze the distribution of the continuous dataset.It shows in fig1.

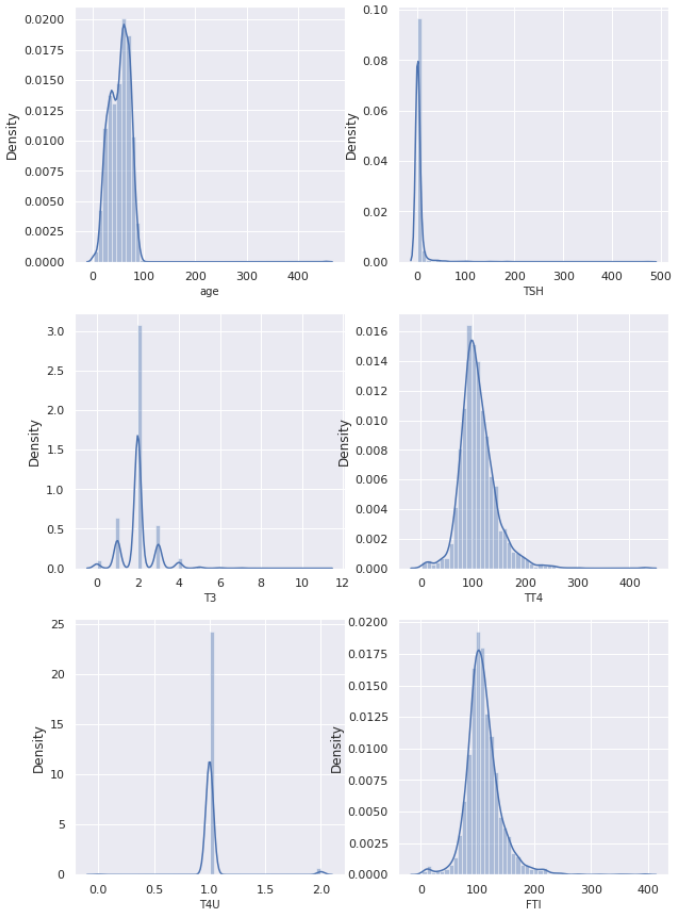

Fig. 1. Distribution of Continuous Dataset

Age, TSH, and T3 graphs seem to be heavily skewed to the left. Let’s change the data a little bit and see if the graphic becomes better.To handle an exception while trying to get the log of ”0,” let’s add 1 to each value in the column before doing the log transformation.The other columns seem OK after log transformation, but ”TSH” exhibits an odd trend.This column is being dropped since it won’t provide much information.It shows in fig2.

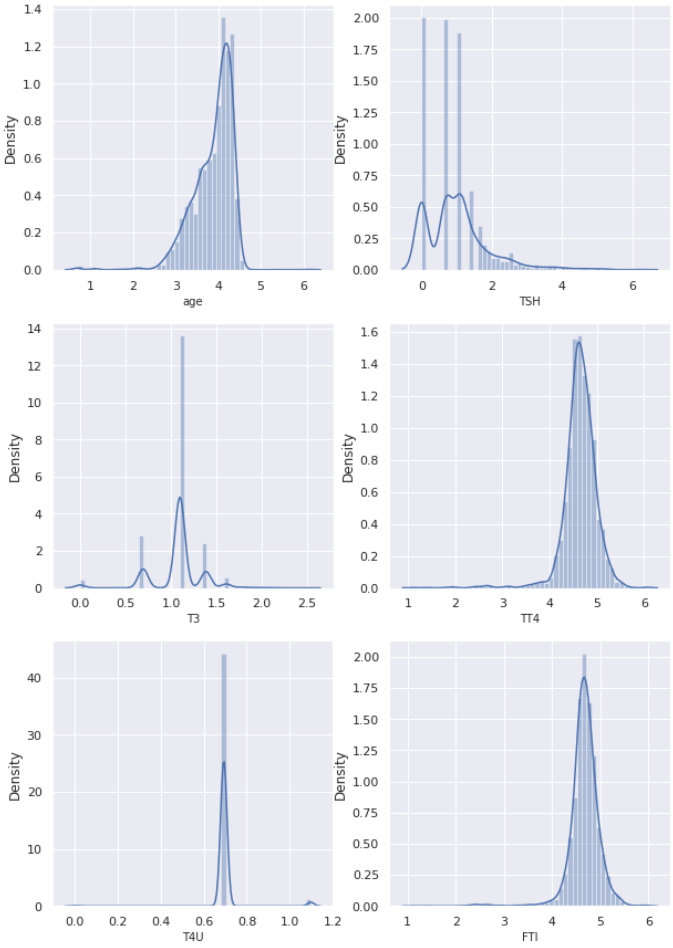

Fig. 2. Distribution of Continuous Dataset by applying log function

B. Handling imbalanced dataset

We have examined the dataset’s balance in relation to the specified target classes. We discovered a very unbalanced dataset which shows in fig 3. We thus use imbalanced-learn to deal with unbalanced data. An algorithm for unbalanced learning is RandomOverSampler.

Imbalanced data are those datasets that have an unequal distribution of observations across the target class, i.e., one class label has an extremely high number of data points whereas the other has an extremely low number. Another example of unbalanced data is the diagnosis of diseases.

1) Method to address the imbalanced dataset issue.:

- Select the Correct Evaluation Metric. A better statistic would be the F1 score for a class dataset that is unbalanced.
- The minority or majority class is up- or down- sampled using this method. When working with an unbalanced

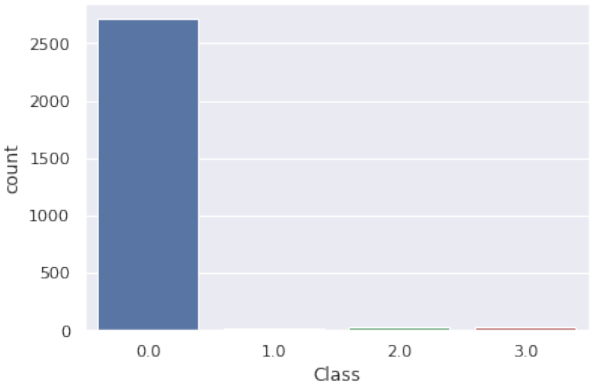

Fig. 3. Unbalanced Data

dataset, we may use replacement to oversample the minority class. Oversampling is the term used for this method. Undersampling is the process of randomly deleting rows from the majority class in order to match them with the minority class.

- Another method for oversampling the minority class is called Synthetic Minority Oversampling Technique, or SMOTE. Many times, just adding duplicate minority class entries to the model doesn't provide any new data. SMOTE creates new instances by synthesizing the data already available.
- The only difference between a BalancedBaggingClassifier and a Sklearn classifier is the added balancing. The training set is balanced at the moment of fit for a specific sampler as an extra step. The "sampling strategy" and "replacement" special parameters are required by this classifier.
- This default threshold may not function well for situations with unbalanced classes. In order for the threshold to effectively divide two groups, we must adjust it to the ideal value.

After applying the randomoversample algorithm the data has become balanced.

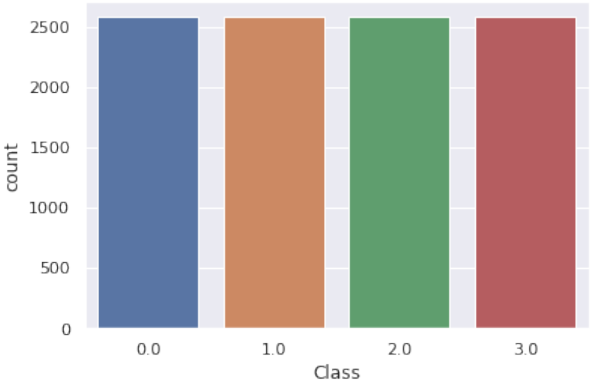

Fig. 4. Balanced Data

C. Feature Selection

When we get a dataset, we often discover a wealth of characteristics. It's possible that not all of the dataset's attributes will be helpful in creating a machine learning model that can make the required prediction. Utilizing some of the characteristics can possibly result in inaccurate forecasts. Therefore, choosing the right features is crucial when creating a machine learning model.

We used correlation to carry out feature selection for our project. High correlation features are more linearly dependant and hence virtually equally affect the dependent variable. We may thus exclude one of the two characteristics when there is a substantial correlation between two features. We can see target class is negatively corelated with TT4 and FTI.

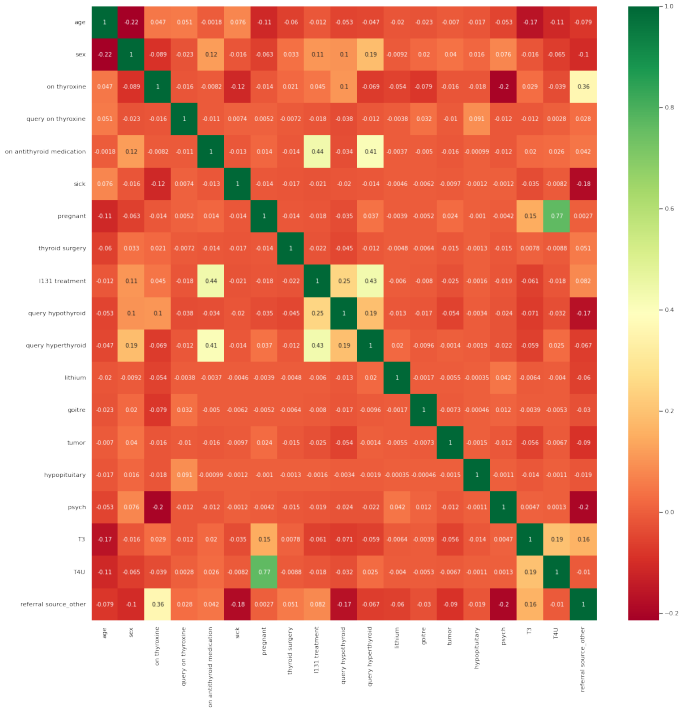

Fig. 5. Heat Map of the Dataset

D. Checking the VIF score for multicollinearity

The features that show the data's multicollinearity are referral source STMW, referral source SVHC, referral source SVHD referral source SVI, referral source other, TT4, and FTI. By removing the first four features—referral source STMW, referral source SVHC, referral source SVHD referral source SVI, referral source other, TT4 and FTI—multicollinearity is prevented. NOTE: In this case, we performed clustering on the features FTI and TT4, but since the dataset is small and the VIF value is high, I am dropping those. Additionally, we can see from the corelation matrix that no other features have positive corelation, so we can either consult with doctors to find out which prominent features they would recommend we perform clustering on in order to improve the model, or we can keep those two features in place without dropping the

rest.In order to avoid multicollinearity, I have at last decided to eliminate the columns.

Now we can see that multicollinearity is not present in the data proceding ahead with training the model using different algorithms

VI. Model Analysis

A. Logistic Regression Model

Despite its name, logistic regression is more of a classification model than a regression model. For situations involving binary and linear classification, logistic regression is a straightforward and more effective approach. It's a classification model that's incredibly simple to implement and performs well with linearly separable classes. It is a widely used categorization method in business. Similar to the Adaline and perceptron, the logistic regression model is a statistical technique for binary classification that may also be used to multiclass classification. A highly efficient logistic regression implementation that handles multiclass classification tasks is available in Scikit-learn.

We got accuracy accuracy and precision and recall which is about 73 percent

	precision	recall	f1-score	support
0.0	0.48	0.37	0.42	498
1.0	0.58	0.53	0.55	522
2.0	0.60	0.49	0.54	494
3.0	0.68	1.00	0.81	550
accuracy			0.61	2064
macro avg	0.59	0.60	0.58	2064
weighted avg	0.59	0.61	0.59	2064

0.7324437809163472

Fig. 6. Classification Matrix of Logistic Regression

Confusion Matrix is, in fact, a performance indicator for a classification issue using machine learning, the output of which might be two or more classes. There are four possible anticipated and actual value combinations in the table. Recall, precision, specificity, accuracy, and—most importantly—AUC-ROC curves may all be measured with great success with this tool.It Shows in fig 7.

True positive rate (TPR) and false positive rate (FPR) are commonly shown on the Y and X axes, respectively, of ROC curves. As a result, the top left corner of the plot, which has an FPR of 0 and a TPR of 1, is the "ideal" location. Despite the fact that this is not particularly practical, it does imply that a greater area under the curve (AUC) is often preferable. Due to the fact that it is optimal to maximize the TPR while reducing the FPR, the "steepness" of ROC curves is also crucial.It shows in Fig 8.

B. Support Vector Machine

A supervised machine learning approach called "Support Vector Machine" (SVM) may be used to classification or regression problems. However, categorization issues are where

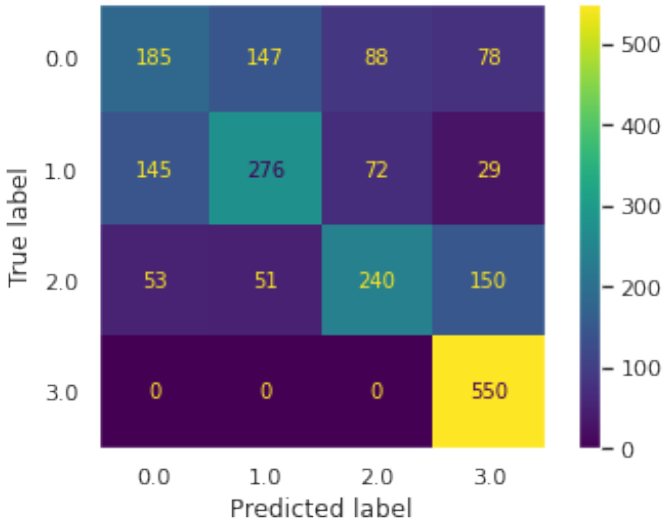

Fig. 7. Confusion Matrix of Logistic Regression

Fig. 8. ROC of Logistic Regression

it is most often utilized. When using the SVM method, each data point is represented as a point in n-dimensional space (where n is the number of features you have), with each feature's value being the value of a certain coordinate.We got accuracy accuracy and precision and recall which is about 76 percent

K Nearest Neighbours One of the simplest machine learning algorithms, based on the supervised learning method, is K-Nearest Neighbour. The K-NN method makes the assumption that the new case and the existing cases are comparable, and it places the new instance in the category that is most like the existing categories. A new data point is classified using the K-NN algorithm based on similarity after all the existing data has been stored. This implies that utilizing the K-NN method, fresh data may be quickly and accurately sorted into a suitable category. We got accuracy accuracy and precision and recall which is about 93 percent

	precision	recall	f1-score	support
0.0	0.50	0.54	0.52	498
1.0	0.62	0.42	0.50	522
2.0	0.59	0.62	0.61	494
3.0	0.85	1.00	0.92	550
accuracy			0.65	2064
macro avg	0.64	0.65	0.64	2064
weighted avg	0.64	0.65	0.64	2064

0.7648566792213148

Fig. 9. Classification Matrix of SVM

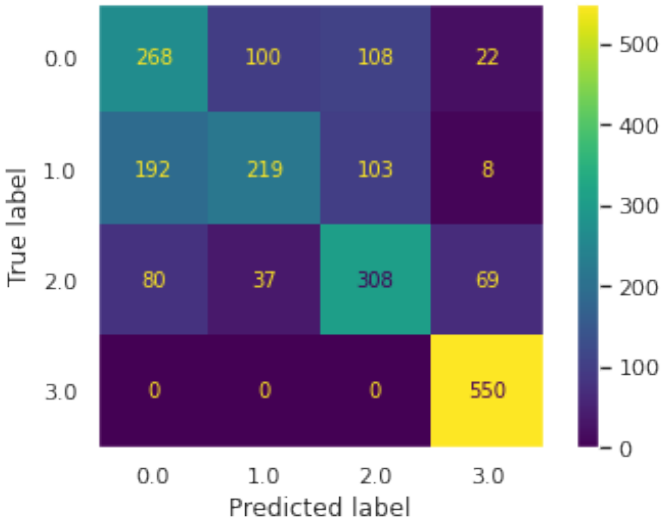

Fig. 10. Confusion Matrix of SVM

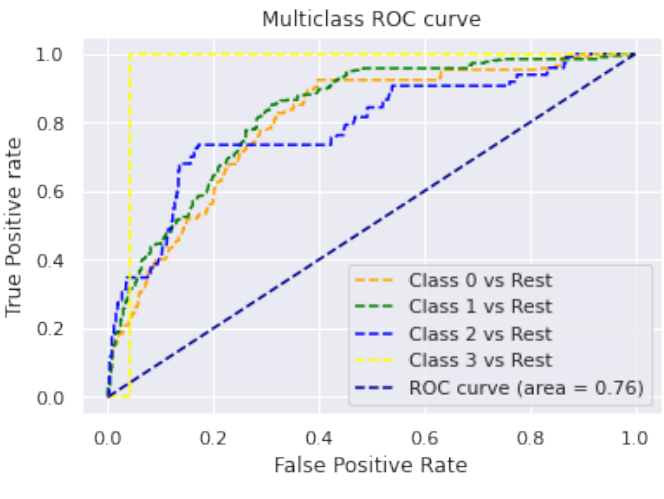

Fig. 11. ROC of SVM

	precision	recall	f1-score	support
0.0	0.78	0.92	0.84	498
1.0	0.95	0.72	0.82	522
2.0	0.91	0.97	0.94	494
3.0	1.00	1.00	1.00	550
accuracy			0.90	2064
macro avg	0.91	0.90	0.90	2064
weighted avg	0.91	0.90	0.90	2064

0.9354741629718791

Fig. 12. Classification Matrix of KNN

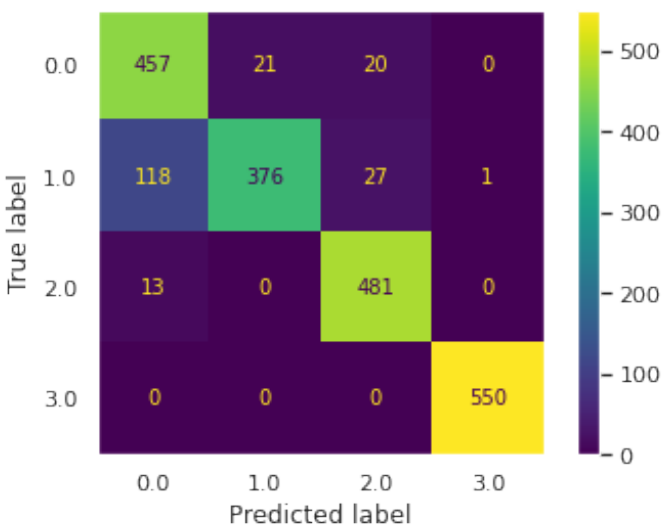

Fig. 13. Confusion Matrix of kNN

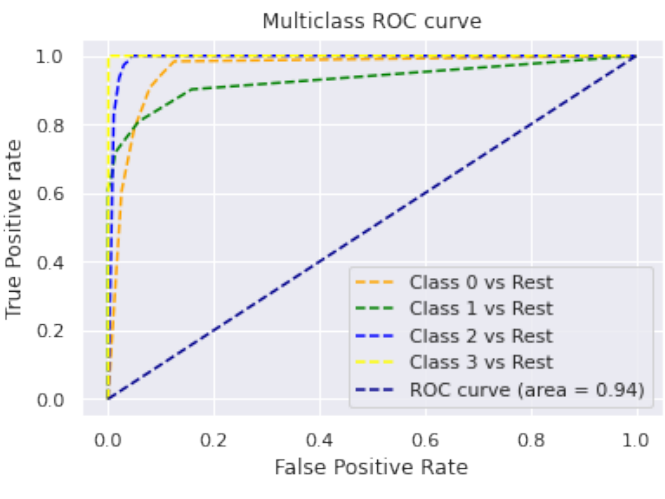

Fig. 14. ROC of KNN

C. Hyperparameter Tuning

A mathematical model containing a number of parameters that must be learnt from the data is referred to as a machine learning model. We may fit the model parameters by using existing data to train a model. Hyperparameters, on the other hand, are a different kind of parameter that cannot be directly learnt by routine training. Usually, they are established before to the start of the program itself. These parameters describe crucial model characteristics including complexity and learning rate. Models may have a large number of hyperparameters, and determining the ideal set of parameters may be approached as a search issue. The following are the top two methods for hyperparameter tuning:GridSearchCV,RandomizedSearchCV. After comparing Logistic regressor , SCV and KNN Classfier I have got the best results in case of KNN Classifier Now performing Hyperparamter tuning on it using GridSerchCV and RandomisedSearchCV.

1) GridSearchCV: The machine learning model is assessed for a variety of hyperparameter values in the GridSearchCV technique. This method is known as GridSearchCV because it analyzes a grid of hyperparameter values to find the optimum set of hyperparameters.

	precision	recall	f1-score	support
0.0	0.82	0.90	0.86	498
1.0	0.94	0.81	0.87	522
2.0	0.92	0.96	0.94	494
3.0	1.00	1.00	1.00	550
accuracy			0.92	2064
macro avg	0.92	0.92	0.92	2064
weighted avg	0.92	0.92	0.92	2064

0.9455137658501863

Fig. 15. Classification Matrix of GridSerach CV

2) RandomizedSearchCV: The machine learning model is assessed for a variety of hyperparameter values in the GridSearchCV technique. This method is known as GridSearchCV because it analyzes a grid of hyperparameter values to find the optimum set of hyperparameters.

	precision	recall	f1-score	support
0.0	0.74	0.97	0.84	498
1.0	0.97	0.75	0.85	522
2.0	0.95	0.87	0.91	494
3.0	1.00	1.00	1.00	550
accuracy			0.90	2064
macro avg	0.91	0.90	0.90	2064
weighted avg	0.92	0.90	0.90	2064

0.9331810597152131

Fig. 16. Classification Matrix of RandomizedeSearch CV

After comparing Logistic regressor , SCV and KNN Classfier I have got the best results in case of KNN Classifier Now performing Hyperparamter tuning on it using GridSerchCV and RandomisedSearchCV

The best result was found in case of KNN classfier and it was furter hypertuned using the hyperparmeters and gridserchCV adn the AUC ROC sore came out to be 0.94 in that case

VII. Conclusion

The thyroid gland is the largest and most important endocrine gland. The hypothyroid dataset is subjected to the data mining approach in order to identify the positive and negative instances across the board. To improve therapy, decision-making, and illness diagnosis, datasets are categorized. In this study, the SVM, Knn, and Logistic Regression are used to predict hypothyroid disease. After applying hyperparameters to the KNN Classier and employing GridSearchCV and RandomisedSearchCV a, we obtained accuracy of roughly 95

Acknowledgment

I would want to thank my professor, Rui Xie, for giving us this project and for challenging us to become knowledgeable in data mining.

References

[1] Banu, Gulmohamed. (2016). Predicting Thyroid Disease using Linear Discriminant Analysis (LDA) Data Mining Technique. Communications on Applied Electronics. 4. 4-6. 10.5120/ cae2016651990.

[2] https://www.analyticsvidhya.com/blog/2021/06/5-techniques-to-handle-imbalanced-data-for-a-classification-problem.

[3] Prerana, Parveen Sehgal, Khushboo Taneja, "Predictive Data Mining for Diagnosis of Thyroid Disease using Neural Network", published in International Journal of Research in Management, Science Technology, Vol. 3, No. 2, April 2015.

[4] Geetha K., Santosh S. Eficient Thyroid Disease Classification Using Differential Evolution with SVM. Journal of Theoretical and Applied Information Technology. Vol.88. No.3. E-ISSN: 1817-3195

[5] mmulu K., Venugopal. (2017). Thyroid Data Prediction using Data Classification Algorithm. IJIRST –International Journal for Innovative Research in Science Technology. Vol.4. Issue 2. July 2017. ISSN (online): 2349-6010

[6] https://www.analyticsvidhya.com/blog/2021/06/5-techniques-to-handle-imbalanced-data-for-a-classification-problem/

[7] https://pypi.org/project/imbalanced-learn/

[8] https://archive.ics.uci.edu/ml/machine-learning-databases/thyroid-disease/

[9] https://towardsdatascience.com/feature-selection-correlation-and-p-value-da8921bfb3cf

[10] https://scikitlearn.org/stable/auto$_e xamples/model_s election/plot_r oc.html$