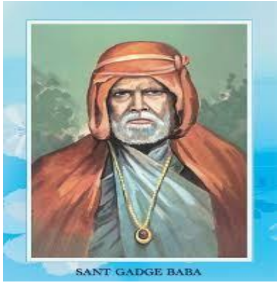
“EDUCATION THROUGH SELF-HELP IS OUR MOTTO”

RAYAT SHIKSHAN SANSTHA

SADGURU GADAGE MAHARAJ COLLEGE, KARAD

(An Autonomous College)



DEPARTMENT OF STATISTICS

Project report on,

“A Detection of Thyroid disease using machine learning techniques”

Submitted by,

Mr. Dhananjay Krushnat Gurav (M. Sc. II)

CERTIFICATE

This is to certify that the project Report entitled A study on " A Detection of thyroid disease by using machine learning techniques" being submitted by Mr. Dhananjay Gurav as partial fulfillment for M.Sc-II in Statistics of Sadguru Gadage Maharaj College, Karad is a record of bonafide work carried out by him under my supervision and guidance.

To the best of our knowledge the matter presented in this project report in original and has not been submitted elsewhere for any other purpose.

Place: Karad **Date:**

Teacher in-charge Examiner PG Co-ordinator Head

Department Statistics

ACKNOWLEDGEMENT

I have great pleasure in presenting this report of the successful completion of my project viz. “A Detection of thyroid disease by using machine learning techniques”.

I take this opportunity to express my great sense of gratitude to my Guide Dr. Mrs. Chavan R.V. For granting me permission to undertake this project for their constant encouragement, guidance and inspiration without which I could not have completed this task.

I would like to extend my sincere thanks to Mrs. Mahajan S.V. (Head Department of Statistics), Dr. Mrs. Patil S.P. and Miss. Patil R.D. for their guidance and kind operation in this project study.

Yours Sincerely,
Dhananjay Krushnat Gurav
M.Sc-II
Department of Statistics

INDEX

1)Introduction:

The Thyroid gland is a vital hormone gland that plays a major role in the growth and development of the human body. It helps to regulate many body functions by constantly releasing a steady amount of thyroid hormones into the bloodstream. The thyroid gland and its hormones affect almost every organ system of the human body hence proper working of our thyroid gland is necessary and important But because of today's Modern and busy lifestyle the ratio of thyroid disorders is increasing day by day thus Thyroid disease has become very common, there are several types of 'Thyroid disorders/disease ' based on different conditions. Different thyroid conditions have different symptoms including [10]:

- Slow or rapid heart rate.
- Unexplained weight loss or weight gain.
- Difficulty tolerating cold or heat.
- Depression or anxiety.
- In women, Irregular menstrual periods

these are the primary signs of having thyroid disorder. Sometimes detection of Thyroid disease can become a confusing task because many symptoms are easily confused with other illness conditions. Therefore doctor or Healthcare provider conducts a simple blood test because several hormones can be measured in the blood to determine how the thyroid gland is functioning [6]. These mainly include thyroid hormones like 'Thyroxine (T4)',' Thyroid-Stimulating hormone (TSH)', 'Triiodothyronine (T3)' and Thyroxine utilization rate(T4U)' whose reading value will help the healthcare provider to give more reliable results about the presence or absence of disease.

Considering the complexity of detecting thyroid disease based on various factors, sometimes it becomes very difficult to perform a diagnosis just by looking at symptoms only experienced doctors can examine the case properly. To overcome this problem we can take the help of different computational methods available today which can act as evidence for results, Machine learning plays an important role to provide evidence for such results. The use of machine learning in the healthcare industry has made revolutionary changes over a period, it has enhanced the reliability of diagnostic procedures to a greater extent. by using data of patients obtained from different clinical tests and readings and fitting appropriate machine learning techniques/algorithms to data we can predict the disease more accurately which gives support for doctors' argument or statement on disease detection. Examine the different machine learning techniques for Thyroid disease detection, we have used different types of feature values to train the model and fitted the various types of machine learning algorithms for detecting thyroid disease.

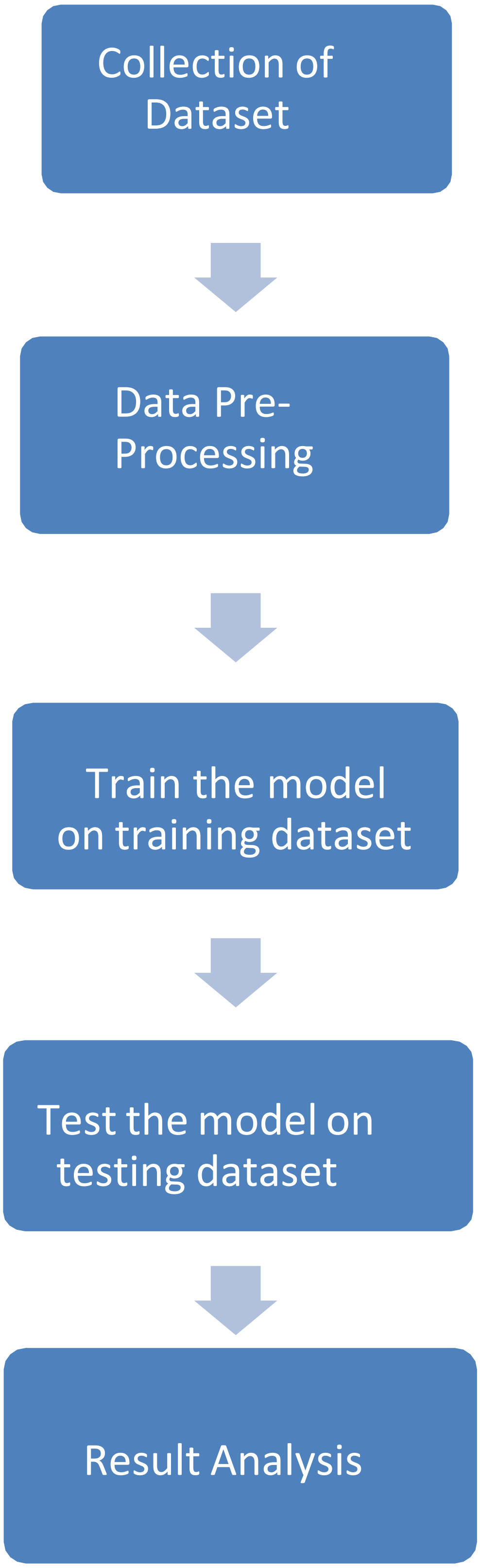
2)Objective:

- To Use the data collected on patients and construct a machine-learning model by using different algorithms which will predict,

  “Whether the patient has thyroid disease or not.”

  So that the healthcare providers will get some evidence of their diagnosis and can provide statements on disease detection more confidently.

3)Methodology:


Collection of Dataset
Data Pre-Processing
Train the model on training dataset
Test the model on testing dataset
Result Analysis

4) Tools and Techniques used:

- Tools: Jupyter Notebook

- Techniques:
  i)Logistic Regression
  ii)Decision Tree Classifier
  iii) Support Vector Classifier
  iv) K-Nearest Neighbour
  v) Random Forest Classifier

5) Data Source:

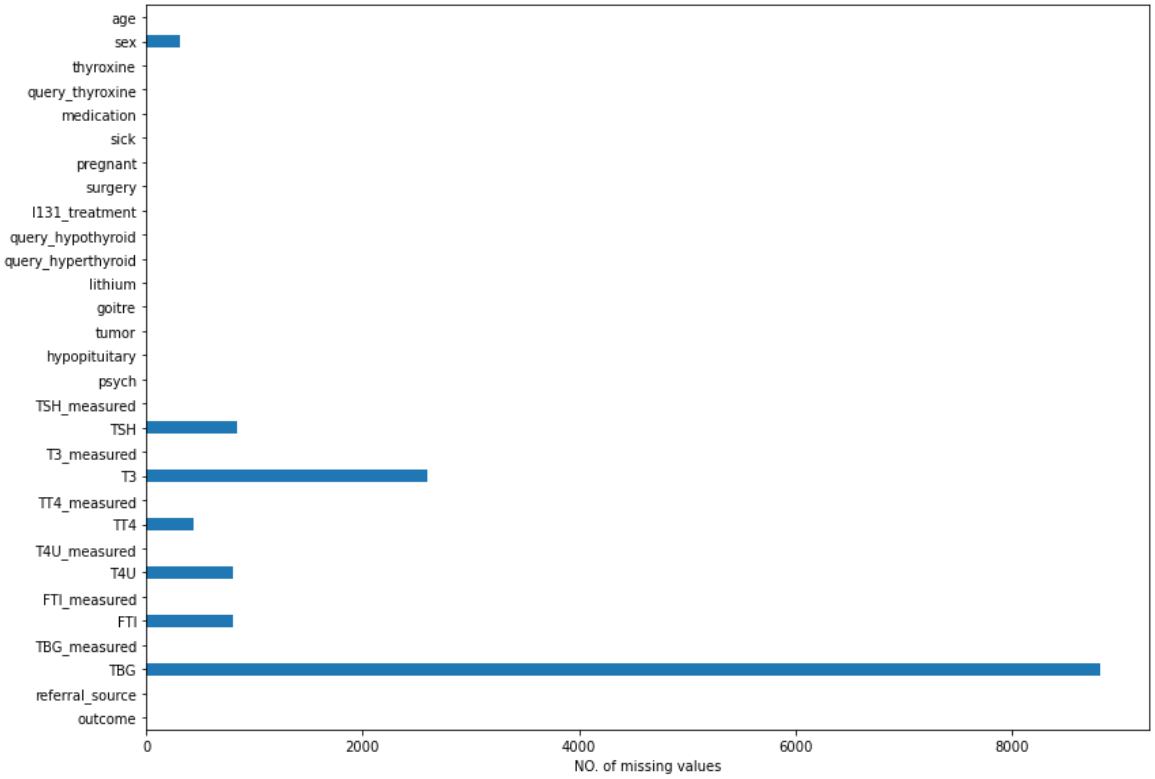
We have collected dataset for this project from machine learning UCI repository

[8]. This directory consists of 9172 records, and each record has 29 attribute values (features) like patient's personal information like age, sex, and other information obtained from tests like 'TSH', 'T3', 'TT4', 'T4U' etc. out of 29 features 7 are of numerical type and 22 are categorical type. Outcome values of numerical features are present in discrete and continuous form Whereas categorical type variables possess the values in binary form e.g. 'Sex': M or F i.e. male or female, 'sick': t (True)or f (False).

6) EDA & Feature Engineering:

Before build the by using given dataset, we must convert it into suitable format, for this firstly we have to understand the data by exploring it. We have performed EDA & Feature engineering in following manner

- Missing values: If the dataset used for classification contains more missing values, then we may end up building a biased model hence it is important to handle them. Following Fig. 1 shows the proportion of missing values in dataset



(Fig. 1 Missing values)

- From the above figure we can see that there are missing values present in the data. In our study, we dealt with columns containing missing values in a different way depending on the condition such as,
  - In column ‘TBG’ 95% of observations were missing so we dropped this column from the dataset we have.
  - Columns ‘TSH’, ‘T3’, ‘TT4’, ‘T4U’, and ‘FTI’ were also missing some observations which are numerical types, but outliers were also present so we used the ‘Median’ value of the respective column to fill missing observations.

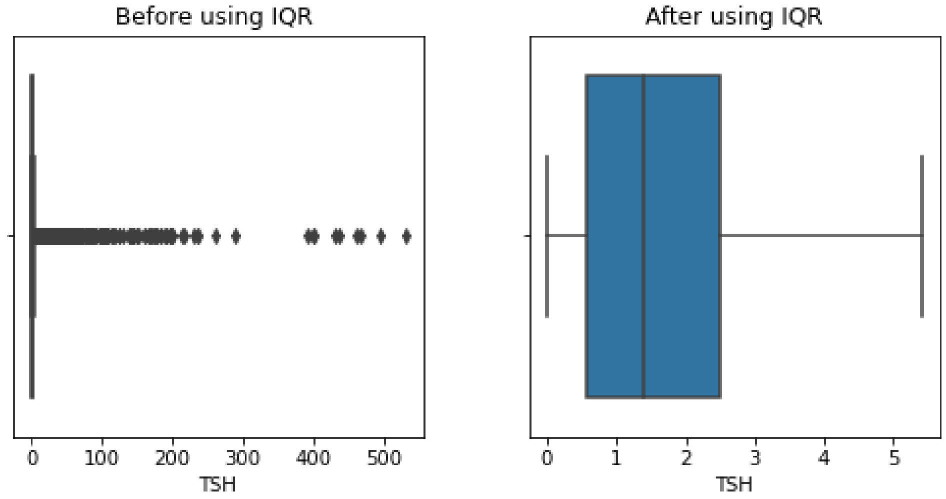
- The column named ‘sex’ is a categorical type column having missing values, since the column type is categorical so we used ‘Mode’ to fill them.

- B) Outlier Analysis:

The machine learning algorithms like ‘Logistic Regression’, ‘Support Vector Machine’, and ‘K-Nearest Neighbours’ are sensitive to outliers. Training the model with data containing outliers might provide false results by overfitting. In the dataset, we had the columns/features ‘TSH’, ‘T3’, ‘TT4’, ‘T4U’, and ‘FTI’ present with outliers. We dealt with them using different techniques available for handling outliers,

o ‘IQR’ Based Filtering:

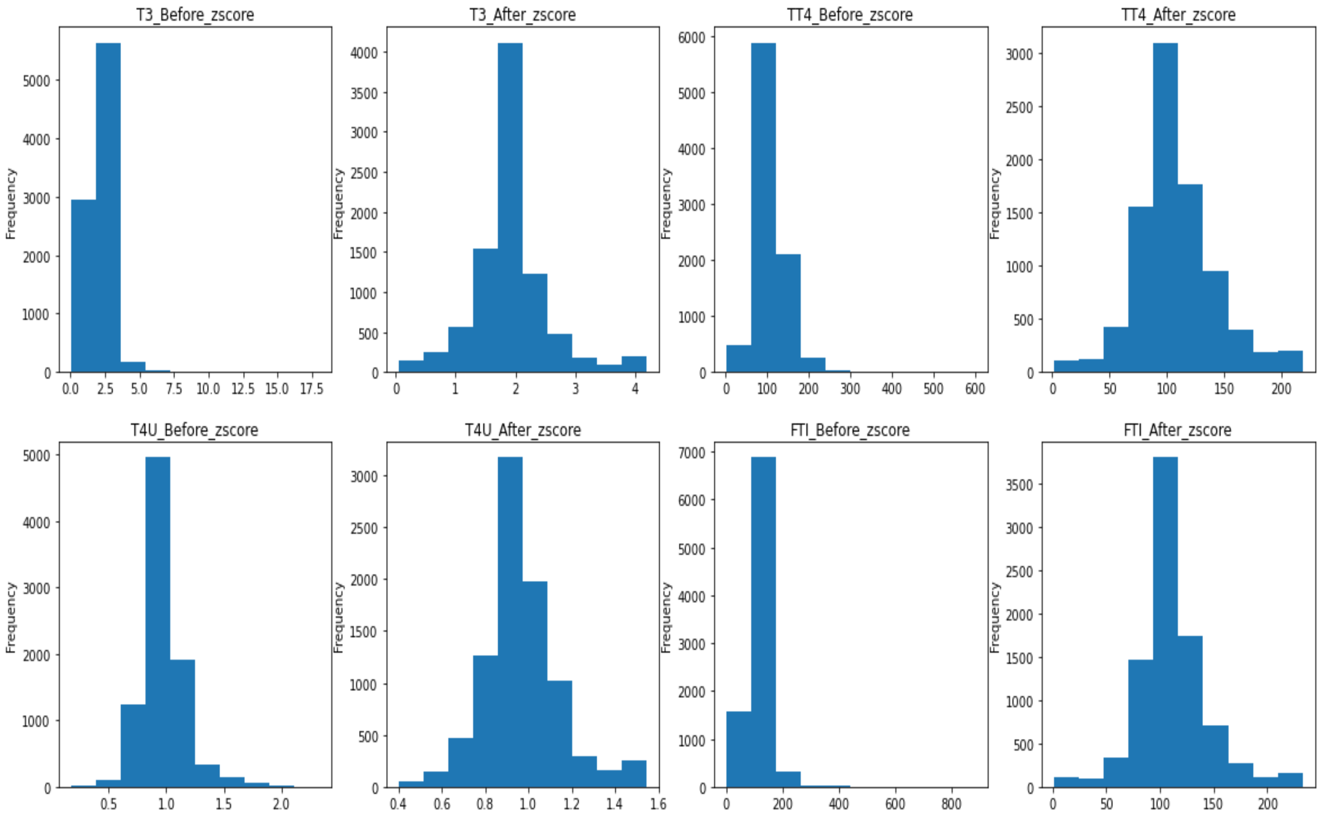
This method is used when the data distribution is skewed. In our dataset, the feature ‘TSH’ is right skewed so we used IQR Based Method for detecting and capping the outliers.Fig.2 shows boxplot for TSH before and after using IQR technique.



(Fig.2 Outlier handling using IQR method)

o Z-Score Treatment :

This method is used for normally distributed features, here in our case the features ‘T3’, ‘TT4’, ‘T4U’, and ‘FTI’ are normally distributed therefore we detect and capped outliers by using ‘Empirical relations.Fig.3 shows Z-Score treatment for features.

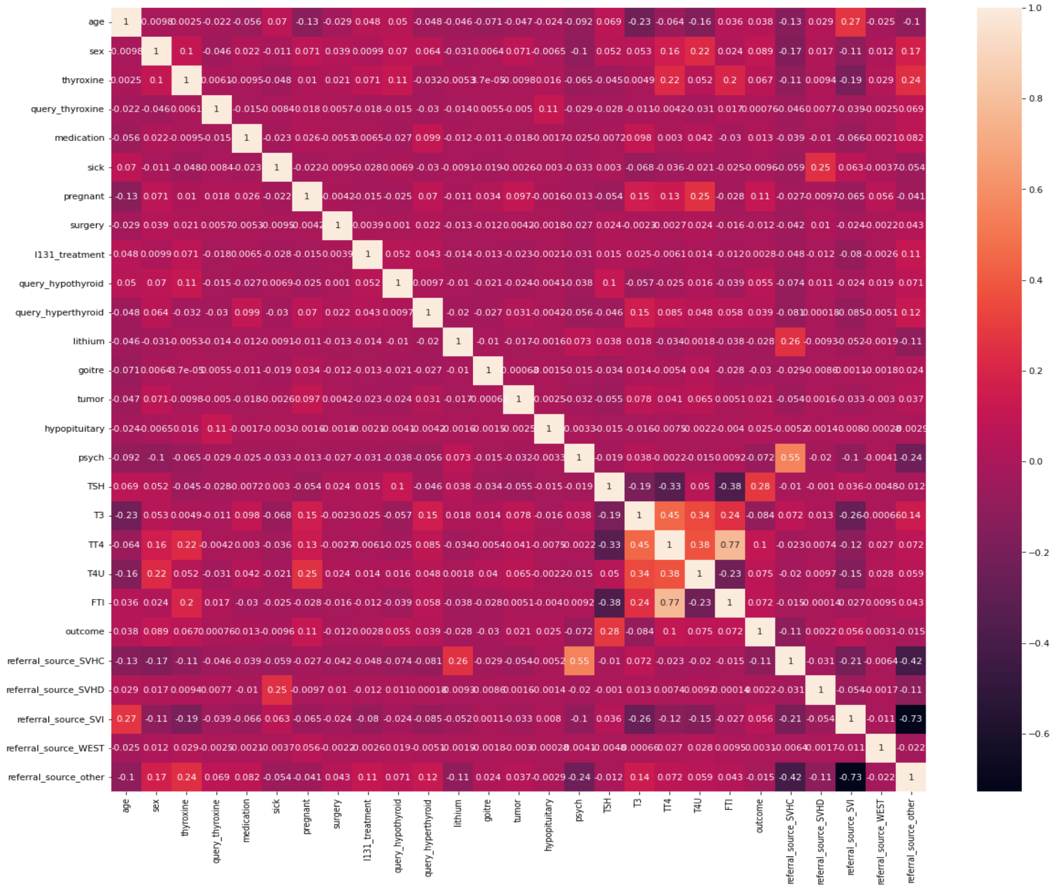


(Fig.3 Outliers handling using Z-Score treatment)

- **Feature Selection:**

Building a machine learning model with all features may lead to overfitting and getting low accuracy on test data hence we must select only those features that are contributing more to the dependent variable important for the prediction. In this project, we checked with the following methods to select independent features.

i) Feature Selection using correlation method Fig.4 shows the correlation of each independent variable with the dependent variable.



(Fig.4 Heatmap to check Correlation among features)

From the above figure, we can see that None of the independent variables is highly correlated with the dependent variable(outcome). There might be a spurious correlation among these so we will not use this method for selecting features

ii)SelectKBest**(**chi2**):** ‘SelectKBest(chi2)’ provides F-Score and P-value for each feature. The feature with lowest p-value is more important for the dependent variable. based on which the top 4 most important features are ‘TSH’, ‘TT4’, ‘FTI’, and ‘age’.

Feature	p-value
TSH	1.319414e-182
TT4	9.891485e-172
FTI	1.043757e-74
AGE	6.145162e-23

we will use these 4 features to predict our outcome variable. After analysing the previously done research work, we found that features ‘T3’ and ‘T4U’ are also used frequently to detect thyroid disease, therefore, we will use total 6 features to make predictions. The final set of features selected in our study is ‘TSH’, ‘TT4’, ‘FTI’, ‘age’, ‘T3’, and ‘T4U’.

7) Model Building: Once we are done with EDA&Feature Engineering next step is to fit the data with selected features to different machine learning algorithms and find the best model to perform a task. In this project, our objective is to detect “whether the patient has thyroid disease or not”. To achieve this we will use following classification techniques.

i) Logistic Regression

ii) Decision Tree

iii) Support Vector Machine

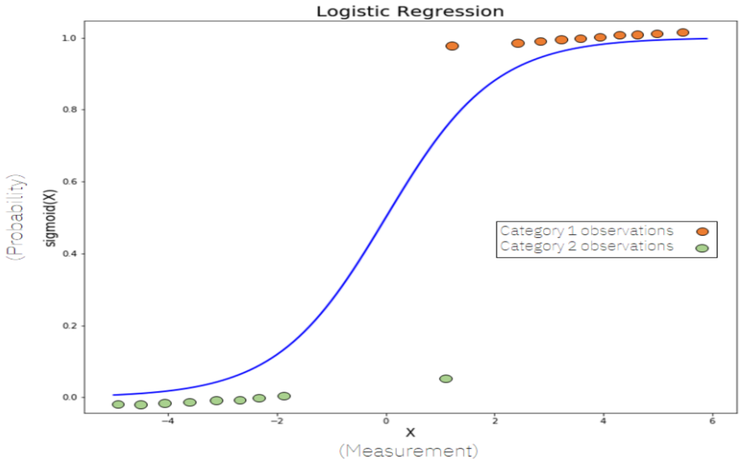
iv) KNN

v) Random Forest

i) Logistic regression: Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable isdichotomous, which means there would be only two possible classes.

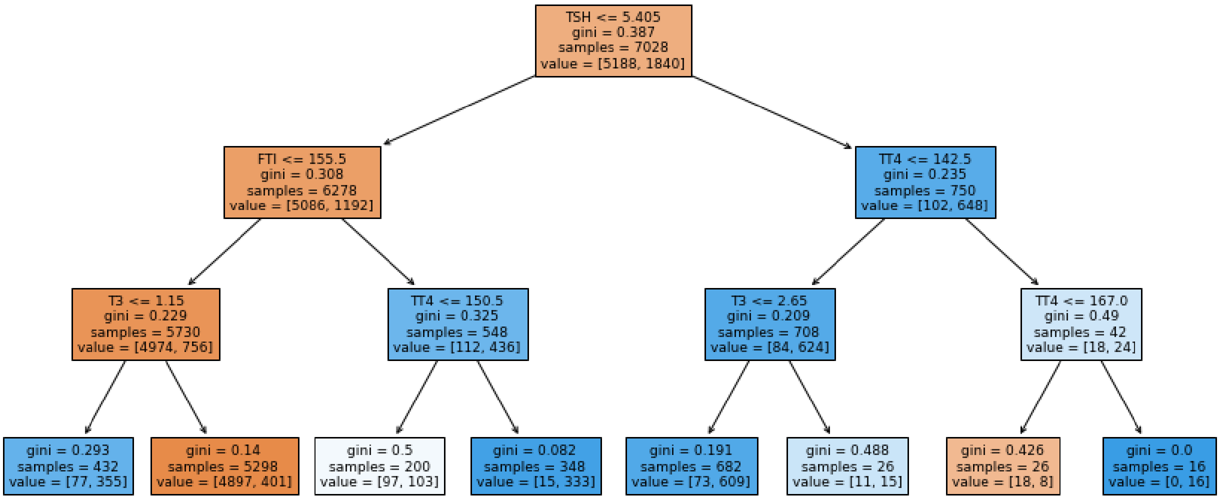
In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).

Mathematically, a logistic regression model predicts P(Y=1) as a function of X.It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection, heart failure prediction etc.



ii) Decision Tree: Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes representthe features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset.

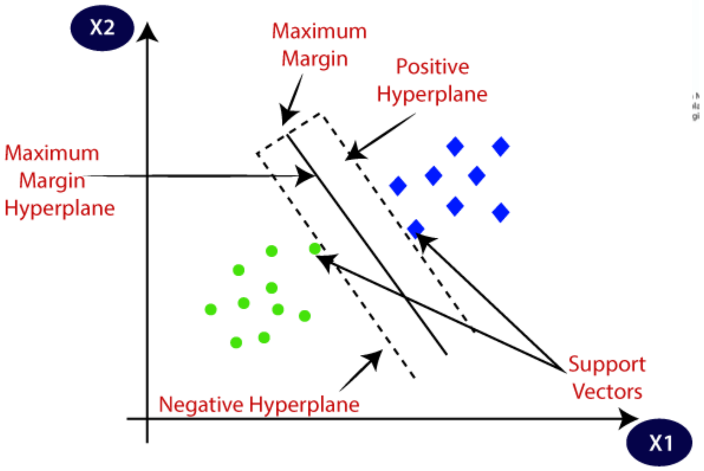
It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. Below diagram explains the structure of a decision tree obtained from training data :



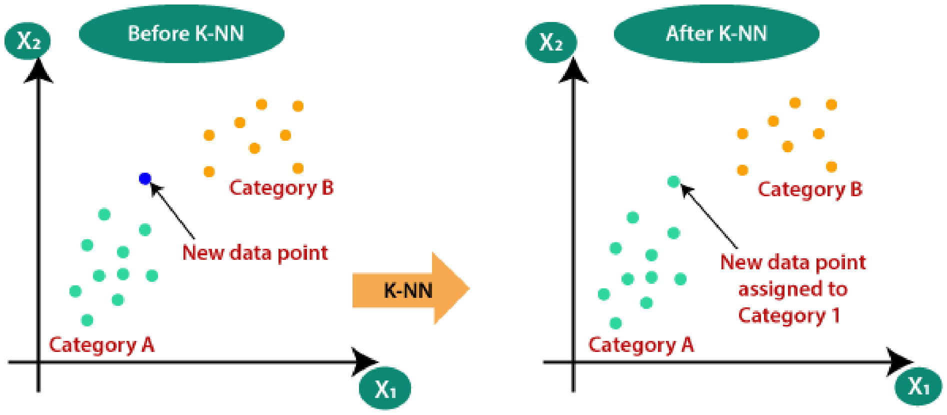
(Decision Tree Structure obtained from training the model)

Interpretation: From the above decision tree we observed that the attribute named as ‘TSH’ is selected as a Root Node, the criterion used here is ‘gini’.

iii)Support Vector Machine : SVM is one of the most popular SupervisedLearning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.
The goal of the SVM algorithm is to create the best line or decision boundary that cansegregate n-dimensional space into classes so that we can easily put the new data pointin the correct category in the future. This best decision boundary is called a hyperplane.SVM chooses the extreme points/vectors that help in creating the hyperplane.These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:



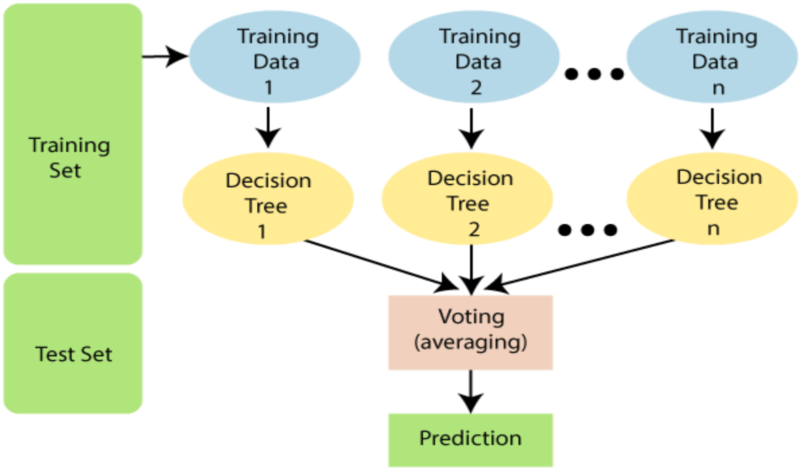
iv) K-Nearest Neighbour: KNN is one of the simplest Machine Learning algorithms based on Supervised Learning technique. It assumes the similarity betweenthe new case/data and available cases and put the new case into the category that is most similar to the available categories. It stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. It can be usedfor Regression as well as for Classification but mostly it is used for the Classification problems. It is a non-parametric algorithm, which means it does not make any assumption on underlying data.
KNN algorithm at the training phase just stores the dataset and when it gets newdata, then it classifies that data into a category that is much similar to the new data.



v)Random Forest: Random Forest Classifier is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a processof combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve thepredictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.



7) Results:

ACCURACY: Accuracy is used to find the correct values; it is the sum of all true values divided by total values

$$\text{Accuracy}=\frac{TP+TN}{TP+FP+FN+TN}$$

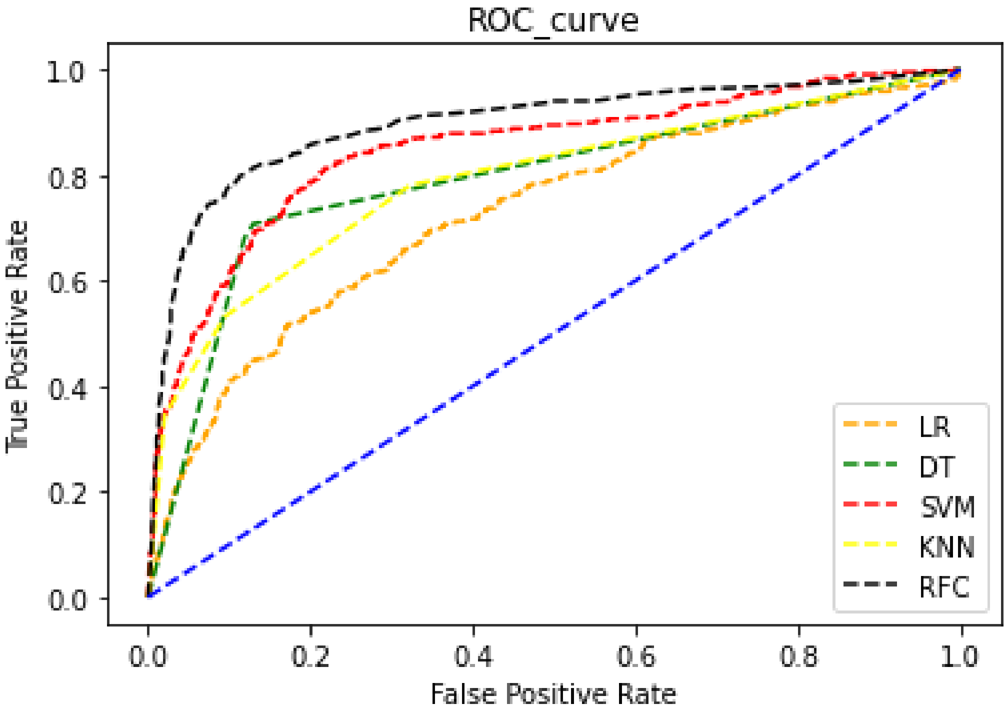
We have trained machine learning models mentioned above on the dataset with selected features ‘TSH’, ‘TT4’, ‘FTI’, ‘age’, ‘T3’ and ‘T4U’. We used 80% data for training the models and 20% for testing purposes. Accuracy has been calculated for each model by using the formula,

The accuracy obtained for each model is given below,

Sr. No.	Classifier	Accuracy (%)
1	Logistic Regression	77.70
2	KNN	81.28
3	SVC	82.19
4	Decision Tree	91.03
5	Random Forest	93.80

From the above table, we observed that the Decision Tree and Random Forest classifier are the models with maximum accuracy of 91.03 %, and 93.80 % respectively. To choose the best model we will take the help of AUC and ROC curves.Below tables shows the AUC score obtained for each model.

Sr. No.	Classifier	AUC
1	Logistic Regression	73.54
2	KNN	79.27
3	SVC	84.90
4	Decision Tree	88.80
5	Random Forest	97.20



By looking at the accuracy score, AUC and Roc curve, now we can clearly say that the Random Forest classifier is the best model we can use for Thyroid Detection.

8) Conclusion and Future Scope:

The ‘Thyroid’ gland is important for the overall efficient functioning of the human body. Hence, if a patient has a thyroid disorder, he should get treatment at an early stage of the disease so that it will be less difficult to recover from it. But as we saw above detection of Thyroid disease can be tricky because symptoms are confusing with other diseases.

We build a machine learning model which predicts ‘Whether the patient has a Thyroid disorder or not?’ So that healthcare providers will get some support for their diagnosis. We found that ‘Random Forest’ is providing higher accuracy 93.8% for the detection of Thyroid disease.

The future scope for this study is that at the current stage, the model is just able to predict ‘Whether the patient has thyroid disease or not?’, If the patient has a disease, then the model does not give an idea about a particular type of disorder the patient is likely to have So if we further develop our model in future then we will be able to predict the exact type of thyroid disorder a patient is diagnosed with.

9) References:

[1] Tyagi, A., Mehra, R., & Saxena, A. (2018, December). Interactive thyroid disease prediction system using machine learning technique. In *2018 Fifth international conference on parallel, distributed and grid computing (PDGC)* (pp. 689-693). IEEE.

[2] Al-muwaffaq, I., & Bozkus, Z. (2016). MLTDD: use of machine learning techniques for diagnosis of thyroid gland disorder. *Comput Sci Inf Technol*, 67-3.

[3] Aversano, L., Bernardi, M. L., Cimitile, M., Iammarino, M., Macchia, P. E., Nettore, I. C., & Verdone, C. (2021). Thyroid disease treatment prediction with machine learning approaches. *Procedia Computer Science*, *192*, 1031-1040.

[4] Vasan, C. R. C., DSU, B., MS, C., & Devikarani, H. S. (2018). Thyroid detection using machine learning. *Computing (PDGC-2018)*, *20*, 22.

[5] Ladenson, P. W., Singer, P. A., Ain, K. B., Bagchi, N., Bigos, S. T., Levy, E. G., ... & Daniels, G. H. (2000). American Thyroid Association guidelines for detection of thyroid dysfunction. *Archives of internal medicine*, *160*(11), 1573-1575.

[6] Thyroid - Wikipedia

[7] https://www.hopkinsmedicine.org/health/conditions-and-diseases/disorders-of-the-thyroid

[8] https://archive.ics.uci.edu/ml/datasets/thyroid

[9] https://labpedia.net/thyroid-part-1-thyroid-function-test-thyroid-hormones-t4-t3-tsh-metabolism-and-role-of-tsh/

[10] https://my.clevelandclinic.org/health/diseases/8541-thyroid-disease