CHAPTER 1

INTRODUCTION

The human body is a complex organism composed of various systems that work together to maintain homeostasis and ensure proper functioning. Glands play a crucial role in the body by producing and secreting hormones, chemical messengers that regulate various bodily processes. One important gland in the body is the thyroid gland.

1.1 The Thyroid Gland

The thyroid gland is a butterfly-shaped gland located in the neck, just below the Adam's apple. It is part of the endocrine system and plays a vital role in regulating metabolism, growth, and development. The thyroid gland produces and releases hormones, primarily triiodothyronine (T3) and thyroxine (T4), which are responsible for controlling the body's metabolic rate.

Thyroid disease refers to any condition that affects the function of the thyroid gland, a small butterfly-shaped gland located in the neck. The thyroid gland produces hormones that regulate various body functions, including metabolism, growth, and development.

Thyroid disease can cause a wide range of symptoms, depending on the specific condition and the severity of the disease. Common symptoms include fatigue, weight changes, hair loss, mood changes, and changes in heart rate.

Treatment for thyroid disease varies depending on the specific condition and its severity. Treatment may include medication, surgery, or radiation therapy. With proper treatment and management, most people with thyroid disease can lead normal, healthy lives.

1.2 Types of Thyroid Diseases

Hypothyroidism: a condition in which the thyroid gland does not produce enough thyroid hormones. This can lead to symptoms such as fatigue, weight gain, constipation, depression, and sensitivity to cold.

Hyperthyroidism: a condition in which the thyroid gland produces too much thyroid hormone. This can lead to symptoms such as weight loss, rapid heart rate, anxiety, insomnia, and sensitivity to heat.

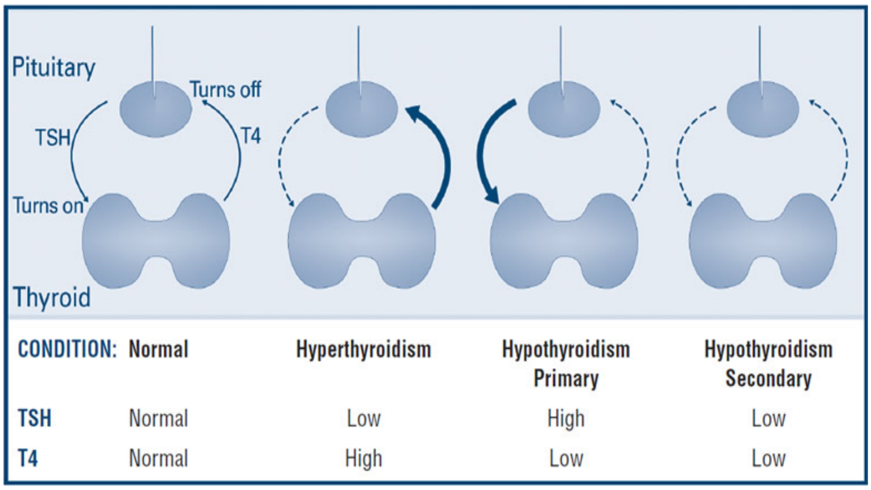
Thyroid nodules: growths on the thyroid gland that can be benign or cancerous. Thyroid nodules are relatively common, and most are benign. However, some nodules can be cancerous and require treatment.

Thyroiditis: inflammation of the thyroid gland, which can be caused by an infection or an autoimmune condition. There are several types of thyroiditis, including Hashimoto's thyroiditis (an autoimmune condition that leads to hypothyroidism), subacute thyroiditis (an inflammation of the thyroid gland that causes hyperthyroidism followed by hypothyroidism), and postpartum thyroiditis (an inflammation of the thyroid gland that occurs after giving birth).

Thyroid cancer: a type of cancer that affects the thyroid gland. Thyroid cancer is relatively rare but can be aggressive if not detected and treated early.

1.3 The Thyroid Gland Function

The major thyroid hormone secreted by the thyroid gland is thyroxine, also called T4 because it contains four iodine atoms. To exert its effects, T4 is converted to triiodothyronine (T3) by the removal of an iodine atom. This occurs mainly in the liver and in certain tissues where T3 acts, such as in the brain. The amount of T4 produced by the thyroid gland is controlled by another hormone, which is made in the pituitary gland located at the base of the brain, called Thyroid Stimulating Hormone (abbreviated TSH). The amount of TSH that the pituitary sends into the bloodstream depends on the amount of T4 that the pituitary sees. If the pituitary sees very little T4, then it produces more TSH to tell the thyroid gland to produce more T4. Once the T4 in the bloodstream goes above a certain level, the pituitary's production of TSH is shut off. In fact, the thyroid and pituitary act in many ways like a heater and a thermostat. When the heater is off and it becomes cold, the thermostat reads the temperature and turns on the heater. When the heat rises to an appropriate level, the thermostat senses this and turns off the heater. Thus, the thyroid and the pituitary, like a heater and thermostat, turn on and off. This is illustrated in the figure below.



CONDITION:	Normal	Hyperthyroidism	Hypothyroidism Primary	Hypothyroidism Secondary
TSH	Normal	Low	High	Low
T4	Normal	High	Low	Low

Figure 1.1: Classification of Thyroidism

1.4 Tests

Blood tests to measure these hormones are readily available and widely used, but not all are useful in all situations. Tests to evaluate thyroid function include the following:

1.4.1 TSH Tests

The best way to initially test thyroid function is to measure the TSH level in a blood sample. Changes in TSH can serve as an “early warning system” – often occurring before the actual level of thyroid hormones in the body becomes too high or too low. A high TSH level indicates that the thyroid gland is not making enough thyroid hormone (primary hypothyroid-ism). The opposite situation, in which the TSH level is low, usually indicates that the thyroid is producing too much thyroid hormone (hyperthyroidism). Occasionally, a low TSH may re-sult from an abnormality in the pituitary gland, which prevents it from making enough TSH to stimulate the thyroid (secondary hypothyroidism). In most healthy individuals, a normal TSH value means that the thyroid is functioning properly.

1.4.2 T4 Tests

T4 is the main form of thyroid hormone circulating in the blood. A Total T4 measures the bound and free hormone and can change when binding proteins differ (see above). A Free T4 measures what is not bound and able to enter and affect the body tissues. Tests measuring free T4 – either a free T4 (FT4) or free T4 index (FTI) – more accurately reflect how the thyroid gland is functioning when checked with a TSH.

The finding of an elevated TSH and low FT4 or FTI indicates primary hypothyroidism due to disease in the thyroid gland. A low TSH and low FT4 or FTI indicates hypothyroidism due to a problem involving the pituitary gland. A low TSH with an elevated FT4 or FTI is found in individuals who have hyperthyroidism.

1.4.3 T3 Tests

T3 tests are often useful to diagnosis hyperthyroidism or to determine the severity of the hyperthyroidism. Patients who are hyperthyroid will have an elevated T3 level. In some individuals with a low TSH, only the T3 is elevated and the FT4 or FTI is normal. T3 testing rarely is helpful in the hypothyroid patient, since it is the last test to become abnormal. Patients can be severely hypothyroid with a high TSH and low FT4 or FTI, but have a normal T3.

Free T3

Measurement of free T3 is possible, but is often not reliable and therefore not typically help-ful.

Reverse T3

Reverse T3 is a biologically inactive protein that is structurally very similar to T3, but the io-dine atoms are placed in different locations, which makes it inactive. Some reverse T3 is pro-duced normally in the body, but is then rapidly degraded. In healthy, nonhospitalized peo-ple, measurement of reverse T3 does not help determine whether hypothyroidism exists or not, and is not clinically useful.

1.5 Types of Thyroidism

1.5.1 Hyperthyroidism

The term hyperthyroidism refers to any condition in which there is too much thyroid hormone produced in the body. In other words, the thyroid gland is overactive. Another term that you might hear for this problem is thyrotoxicosis, which refers to high thyroid hormone levels in the blood stream, irrespective of their source.

Thyroid hormone plays a significant role in the pace of many processes in the body. These processes are called your metabolism. If there is too much thyroid hormone, every func-tion of the body tends to speed up. It is not surprising then that some of the symptoms of hy-perthyroidism are nervousness, irritability, increased sweating, heart racing, hand tremors, anxiety, difficulty sleeping, thinning of your skin, fine brittle hair and weakness in your mus-cles—especially in the upper arms and thighs. You may have more frequent bowel move-ments, but diarrhoea is uncommon. You may lose weight despite a good appetite and, for women, menstrual flow may lighten and menstrual periods may occur less often. Since hyper-thyroidism increases your metabolism, many individuals initially have a lot of energy. How-ever, as the hyperthyroidism continues, the body tends to break down, so being tired is very common.

Hyperthyroidism usually begins slowly but in some young patients these changes can be very abrupt. At first, the symptoms may be mistaken for simple nervousness due to stress. If you have been trying to lose weight by dieting, you may be pleased with your success until the hyperthyroidism, which has quickened the weight loss, causes other problems.

In *Graves' Disease* (also known as *Basedow's Disease*), which is the most common form of hyperthyroidism, the eyes may look enlarged because the upper lids are elevated.

Sometimes, one or both eyes may bulge. Some patients have swelling of the front of the neck from an enlarged thyroid gland (a goitre).

1.5.2 Hypothyroidism

Hypothyroidism is an underactive thyroid gland. Hypothyroidism means that the thyroid gland can't make enough thyroid hormone to keep the body running normally. People are hypothyroid if they have too little thyroid hormone in the blood. Common causes are auto-immune disease, such as Hashimoto's thyroiditis, surgical removal of the thyroid, and radiation treatment.

1. **Symptoms:** Hypothyroidism doesn't have any characteristic symptoms. There are no symptoms that people with hypothyroidism always have and many symptoms of hy-pothyroidism can occur in people with other diseases. One way to help figure out whether your symptoms are due to hypothyroidism is to think about whether you've always had the symptom (hypothyroidism is less likely) or whether the symptom is a change from the way you used to feel (hypothyroidism is more likely).

2. **Medical and family history:** You should tell your doctor:

   - about changes in your health that suggest that your body is slowing down;
   - if you've ever had thyroid surgery;
   - if you've ever had radiation to your neck to treat cancer;
   - if you're taking any of the medicines that can cause hypothyroidism — amiodarone, lithium, interferon alpha, interleukin-2, and maybe thalidomide;
   - whether any of your family members have thyroid disease.

3. **Physical exam:** The doctor will check your thyroid gland and look for changes such as dry skin, swelling, slower reflexes, and a slower heart rate.

4. **Blood tests:** There are two blood tests that are used in the diagnosis of hypothyroidism.

5. **TSH (thyroid-stimulating hormone) test:** This is the most important and sensitive test for hypothyroidism. It measures how much of the thyroid hor-mone thyroxine (T4) the thyroid gland is being asked to make. An abnormally high TSH means hypothyroidism: the thyroid gland is being asked to make more T4 be-cause there isn't enough T4 in the blood.

6. **T4 tests:** Most of the T4 in the blood is attached to a protein called thyroxine-binding globulin. The "bound" T4 can't get into body cells. Only about 1%–2% of T4 in the blood is unattached ("free") and can get into cells. The free T4 and the free T4 index are both simple blood tests that measure how much unat-tached T4 is in the blood and available to get into cells.

1.5.3 Goitre

The term "goitre" simply refers to the abnormal enlargement of the thyroid gland. It is important to know that the presence of a goitre does not necessarily mean that the thyroid gland is malfunctioning. A goitre can occur in a gland that is producing too much hormone (hyperthyroidism), too little hormone (hypothyroidism), or the correct amount of hormone (euthyroidism). A goitre indicates there is a condition present which is causing the thyroid to grow abnormally.

1.6 Effects of Thyroid Dysfunction on The Body

Thyroid dysfunction can have widespread effects on various body systems. Some of the notable impacts include:

Metabolism: Thyroid hormones regulate the body's metabolic rate. In hypothyroidism, the slowed-down metabolism can lead to weight gain, fatigue, and sluggishness. Conversely, hyperthyroidism accelerates the metabolic rate, causing weight loss and increased energy levels.

Cardiovascular System: Thyroid hormones affect heart rate, blood pressure, and cholesterol levels. Hypothyroidism can lead to bradycardia (slow heart rate), increased blood pressure, and elevated cholesterol levels. Hyperthyroidism, on the other hand, can cause tachycardia (rapid heart rate), increased blood pressure, and abnormal heart rhythms.

Central Nervous System: Thyroid hormones are essential for brain development and function. Thyroid dysfunction, particularly in infants and children, can impair cognitive development and lead to learning difficulties. In adults, both hypo- and hyperthyroidism can affect mood, causing depression or anxiety.

Reproductive System: Thyroid hormones play a role in regulating menstrual cycles and fertility. Thyroid disorders can disrupt these processes, leading to irregular periods, fertility problems, or complications during pregnancy.

Skeletal System: Thyroid hormones are necessary for maintaining bone health. Untreated thyroid disorders can result in decreased bone density (osteoporosis) and an increased risk of fractures.

1.7 Objective of The Study

This study attempts to classify / predict the presence of the thyroid disease according to the several predictors such as T4, TSH etc. In this study, the feature importance of the various predictors is determined as well.

1.8 Organization of The Report

The project is organized with four chapters and Bibliography section. Chapter 1 consists of the Introduction about the thyroid disease, classification, treatments, symptoms, diagnosis of the disease. Chapter 2 describes the data and the elucidated methodology used for the data analysis; appropriate machine learning techniques which are used to carry out the analysis are briefly explained in this chapter as well. In chapter 3, the data has been classified using the machine Learning techniques such as Logistic Regression, Random Forest, Decision Tree, K- nearest neighbor and comparison of the algorithms were made. In Chapter 4 contains the summary and concluding remarks which are made from the statistical data analysis. Bibliography lists the books, materials referred to carry out this work.

CHAPTER 2

MATERIALS AND METHODS

Machine learning is a field of artificial intelligence (AI) that focuses on developing algorithms and models that enable computers to learn and make predictions or decisions with-out being explicitly programmed. It involves creating mathematical models and algorithms that learn from and make predictions or decisions based on data.

In traditional programming, humans write explicit instructions for the computer to fol-low. However, in machine learning, instead of providing specific instructions, we feed the computer with a large amount of data and let it discover patterns, relationships, and insights on its own.

There are various types of machine learning algorithms, including supervised learning, unsupervised learning, and reinforcement learning. Supervised learning involves training a model on labelled examples, where the desired output is known. Unsupervised learning in-volves finding patterns and structures in unlabelled data. Reinforcement learning focuses on training models to make sequential decisions based on rewards and punishments.

Machine learning has applications in numerous fields, such as image and speech recognition, natural language processing, recommendation systems, fraud detection, medical diagnosis, and autonomous vehicles, among others. It continues to advance rapidly, enabling computers to learn and adapt from data to solve increasingly complex problems.

Classification is a fundamental task in machine learning that involves categorizing or assigning input data into predefined classes or categories based on their features. It is a supervised learning technique where the algorithm learns from labelled examples to predict the class or category of unseen data.

2.1 Logistic Regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical de-pendent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.

In Logistic regression, instead of fitting a regression line, we fit an ”S” shaped logistic function, which predicts two maximum values (0 or 1). The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc. Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using con-tinuous and discrete datasets. Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

2.1.1 Assumptions

- The dependent variable must be categorical in nature.
- The independent variable should not have multi-collinearity.

2.1.2 Mathematical Model

One of the generalized linear models (GLM) is the logistic regression (LR) model. It is a model for binary variables in which the response indicates whether an event was successful or unsuccessful.

$$p_i = f(y|x) = \frac{1}{1 + exp\{-(\beta_0 + \beta_1 x_i)\}} \tag{2.1}$$

Where p_i is the probability of success and, $P(y_i = 1) = p_i$ and $P(y_i = 0) = q_i = 1 - p_i$, $0 \leq p_i \leq 1$.

Additionally, the model's parameters β_0 and β_1 are s_i, an independent variable, and exp, a mathematical constant known as Euler's number that is approximately equivalent to 2.78.

Multiple independent variables, which can be continuous or categorical, can be included in a LR. In that case, the multiple logistic regression model would be:

$$p_i = \frac{1}{1 + exp\{-z_i\}} = \frac{exp\{z_i\}}{1 + exp\{z_i\}} \tag{2.2}$$

Where, $z_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_n x_n$ and $\frac{p_i}{1-p_i} = \frac{1+exp\{z_i\}}{1+exp\{z_i\}=exp\{z_i\}}$. Where, $\frac{p_i}{1-p_i}$ is the odd ratio is derived by dividing the probability of an event happening by the probability that it won't. The odd ratio is a solution for the probability's upper and lower bounds where $0 <$ odd ratio $< \infty$.

$$p_i = \frac{odds}{1 + odds} = \frac{exp\{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_n x_n\}}{1 + exp\{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_n x_n\}} \tag{2.3}$$

$$\text{log}it = In(odds) = In\left(\frac{p_i}{1 - p_i}\right) = z_i \tag{2.4}$$

$$\text{log}it = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_n x_n, \ -\infty < \text{log}it < \infty \tag{2.5}$$

Therefore, logit is a linear function in independent variables $x_j, 1 \leq j \leq n$

2.2 Decision Tree Algorithm

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 represent the features of a dataset, branch-es 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 a graph-ical representation for getting all the possible solutions to a problem/decision based on given conditions.

2.2.1 Terminologies used in the Decision Tree algorithm

Root Node: Root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.

Leaf Node: Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node.

Splitting: Splitting is the process of dividing the decision node/root node into sub nodes according to the given conditions.

Branch/Sub Tree: A tree formed by splitting the tree.

Pruning: Pruning is the process of removing the unwanted branches from the tree.

Parent/Child node: The root node of the tree is called the parent node, and other nodes are called the child nodes.

2.2.2 Work Flow of Decision Tree

Step 1: Begin the tree with the root node, says S, which contains the complete dataset.

Step 2: Find the best attribute in the dataset using Attribute Selection Measure (ASM).

Step 3: Divide the S into subsets that contains possible values for the best attributes.

Step 4: Generate the decision tree node, which contains the best attribute.

Step 5: Recursively make new decision trees using the subsets of the dataset created in Step - 3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

2.2.3 Attribute Selection Measures

While implementing a Decision tree, the main issue arises that how to select the best attribute for the root node and for sub nodes. So, to solve such problems there is a technique which is called as Attribute selection measure or ASM. By this measurement, we can easily select the best attribute for the nodes of the tree. There are two popular techniques for ASM, which are:

- Information Gain
- Gini Index

2.2.3.1 Information Gain

Information gain is the measurement of changes in entropy after the segmentation of a dataset based on an attribute. It calculates how much information a feature provides us about a class. According to the value of information gain, we split the node and build the decision tree. A decision tree algorithm always tries to maximize the value of information gain, and a node/attribute having the highest information gain is split first. It can be calculated using the below formula:

$$\text{Information Gain} = \text{Entropy(S)} - [(\text{Weighted Avg}) * \text{Entropy(each feature)}] \tag{2.6}$$

Entropy: Entropy is a metric to measure the impurity in a given attribute. It specifies randomness in data. Entropy can be calculated as:

$$\text{Entropy(s)} = -P(\text{yes})\ log\ 2\ P(\text{yes}) - P(\text{no})\ log\ 2\ P(\text{no}) \tag{2.7}$$

Where,

S= Total number of samples

P (yes) = probability of yes

P (no) = probability of no

2.2.3.2 Gini Index

Gini index is a measure of impurity or purity used while creating a decision tree in the CART(Classification and Regression Tree) algorithm. An attribute with the low Gini index should be preferred as compared to the high Gini index. It only creates binary splits, and the CART algorithm uses the Gini index to create bina-ry splits. Gini index can be calculated using the below formula:

$$\text{Gini Index} = 1 - \sum_j P_j^2 \tag{2.8}$$

2.2.4 Pruning: Getting an Optimal Decision tree

Pruning is a process of deleting the unnecessary nodes from a tree in order to get the optimal decision tree.

A too large tree increases the risk of overfitting, and a small tree may not capture all the im-portant features of the dataset. Therefore, a technique that decreases the size of the learning tree without reducing accuracy is known as Pruning. There are mainly two types of trees pruning technology used:

- Pre-Pruning
- Post Pruning.

2.3 Random Forest algorithm

A random forest (RF) model introduced by Bierman is a collection of tree predictors. Each tree is grown according to the following procedure:

The bootstrap phase: select randomly a subset of the training dataset a local training set for growing the tree. The remaining samples in the training dataset form a so called out-of-bag (OOB) set and are used to estimate the RF's goodness-of-fit.

The growing phase: grow the tree by splitting the local training set at each node according to the value of one variable from a randomly selected subset of variables (the best split) using classification and regression tree (CART) method. Each tree is grown to the largest extent possible. There is no pruning.

The bootstrap and growing phases require an input of random quantities. It is assumed that these quantities are independent between trees and identically distributed. Consequently, each tree can be viewed as sampled independently from the ensemble of all tree predictors for a given training dataset.

For prediction, an instance is run through each tree in a forest down to a terminal node which assigns it a class. Predictions supplied by the trees undergo a voting process: the forest returns ca class with the maximum number of votes. Draws are resolved through a random selection.

To present our feature contribution procedure in the following section, we have to develop a probabilistic interpretation of the forest prediction process. Denote by $C = \{C_1, C_2, ..., C_K\}$ the set of classes and by $\triangle_K$ the set

$$\triangle_K = (p_1, ..., p_K) : \sum_{k=1}^{K} p_k = 1 \text{ and } p_k \geq 0.$$

An element of $\triangle_K$ can be interpreted as a probability distribution over C. Let e_k be an element of $\triangle_K$ with 1 at position $k–a$ probability distribution concentrated at class C_k. If a tree t predicts that an instance i belongs to a class C_k then we write $\widehat{Y}_{i,t} = e_k$. This provides a mapping from predictions of a tree to the set $\triangle_K$ of probability measures on C. Let

$$\widehat{Y}_i = \frac{1}{T}\sum_{t=1}^{T}\widehat{Y}_{i,t}$$

where T is the overall number of trees in the forest. Then $\widehat{Y}_i \in \triangle_K$ and the prediction of the random forest for the instance i coincides with a class C_k for which the k-th coordinate of $\widehat{Y}_i$ is maximal.

2.4 Support Vector Machines

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in 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:

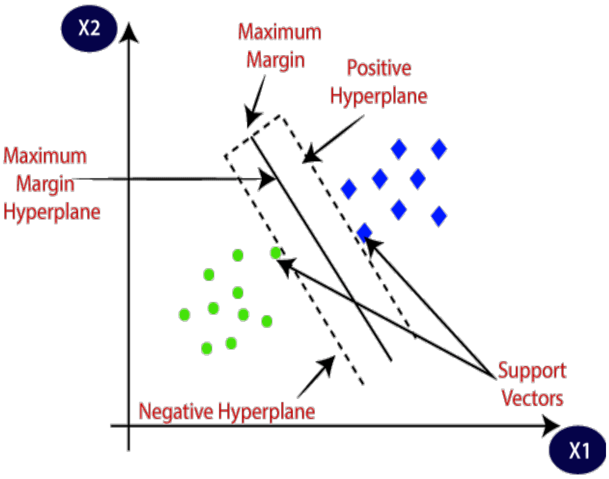


Figure 2.1: Classification diagram using Support Vector Machine

2.4.1 Hyperplane

There can be multiple lines/decision boundaries to segregate the classes in n-dimensional space, but we need to find out the best decision boundary that helps to classify the data points. This best boundary is known as the hyperplane of SVM.

The dimensions of the hyperplane depend on the features present in the dataset, which means if there are 2 features (as shown in image), then hyperplane will be a straight line. And if there are 3 features, then hyperplane will be a 2-dimension plane.

We always create a hyperplane that has a maximum margin, which means the maximum distance between the data points.

2.4.2 Support Vectors

The data points or vectors that are the closest to the hyperplane and which affect the position of the hyperplane are termed as Support Vector. Since these vectors support the hyperplane, hence called a Support vector.

2.4.3 Types of SVM

SVM can be of two types:

- Linear SVM
- Non-linear SVM

2.4.3.1 Linear SVM

Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, and classifier is used called as Linear SVM classifier.

2.4.3.2 Non-linear SVM

Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data and classifier used is called as Non-linear SVM classifier.

2.4.4 Mathematical Form of SVM

In the simplest pattern recognition tasks, support vector machines use a linear separating hyperplane to create a classifier with a maximal margin. Let us consider a binary classification task with the training data set $(x_i, y_i)_{i=1..N}, x_i \in R^m, y_i = \{-1, +1\}$, and let the decision function be $f(x) = sign(w.x + b)$. This optimal hyperplane can be determined as follow:

minimize $\langle w.w \rangle$

subject to $y_i \langle w.x_i + b \rangle \geq +l \forall_i$

Introducing Lagrange multipliers to solve this problem of convex optimization and making some substitutions, we arrive at the Wolfe dual of the optimization problem:

maximize $W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,k=1}^{m} \alpha_i y_i \alpha_k y_k \langle x_i.x_k \rangle$

subject to $\alpha_i \geq 0 \ \forall i, \sum_{i=1}^{m} \alpha_i y_i = 0$ The hyperplane decision function can thus be written as

$$f(x) = sign\left(\sum_{i=1}^{sv} \alpha_i y_i \langle x_i.x_k \rangle + b\right) \tag{2.9}$$

In cases when given classes cannot be linearly separated in the original input space, the SVM first projects the input space to another higher-dimensional dot product space F, called feature space, via a nonlinear map

$$\phi : R^m \to F^d (d >> m)$$

In this new space the optimal hyperplane is derived. Nevertheless, by using kernel functions which satisfy the Mercer' theorem, it is possible to carry out all the necessary operations in the input space by using

$$\phi(x_i).\phi(x_j) = K(x_i.x_j) \tag{2.10}$$

The decision function is formulated in terms of these kernels:

$$f(x) = sign\left(\sum_{i=1}^{sv} \alpha_i y_i K(x_i, x_j) + b\right) \tag{2.11}$$

In the equations (2.6) and (2.8) just a few of the training patterns have a non-zero weight α_i. These elements lie on the margin and they are known as support vectors (SV). They are the training samples that define the optimal separating hyperplane and are the most difficult patterns to classify. Informally speaking, they are the most informative samples for the classification task

2.5 Evaluation Measures of the Performance of the Classifiers

2.5.1 Accuracy

It is the ratio of number of correct predictions to the total number of input samples.

$$Accuracy = \frac{Number of correct predicitons}{Total number of predicitons made}$$

It works well only if there are equal number of samples belonging to each class.

2.5.2 Sensitivity

Sensitivity corresponds to the proportion of positive data points that are correctly considered as positive, with respect to all positive data points.

$$sensitivity = \frac{TruePositive}{FalseNegative + TruePositive}$$

2.5.3 Specificity

Specificity corresponds to the proportion of negative data points that are correctly considered as negative, with respect to all negative data points.

$$specificity = \frac{TrueNegative}{TrueNegative + FalsePositive}$$

2.5.4 Precison

Precision is the ratio of true positive predictions to the total number of positive predictions made by the model. It measures how accurate the positve predictions are.

$$Precision = \frac{TruePositive}{(TruePositive + FalsePositive)}$$

2.5.5 Recall

Recall is the ratio of true positive predictions to the total number of actual positive in the dataset. It measures how well the models is able to identify all postive instances.

$$Recall = \frac{TruePositive}{(TruePositive + FalseNegative)}$$

2.5.6 F1_Score

F1_score is the harmonic mean of precision and recall. It balances both precision and recall, and provides a single metric that summarizes the overall performance of the model.

$$F1\_Score = \frac{2 \times (Precision \times Recall)}{(Precision + Recall)}$$

2.6 Receiver Operating Characteristic Curve

The Receiver Operating Characteristic (ROC) curve is a graphical representation of the performance of a binary classification model. It illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate (1 - specificity) at different classification thresholds.

To construct an ROC curve, the following methodology will be followed:

1. Develop and train a binary classification model using a suitable algorithm (e.g., logistic regression, support vector machines, random forests, etc.) on the training data. Ensure that data is labeled with known ground truth for both positive and negative instances.

2. Use the trained model to generate predictions on a separate validation dataset or through cross-validation on the training data. Obtain the predicted probabilities or scores for each instance in the dataset.

3. Vary the classification threshold to convert the predicted probabilities into binary class labels (positive or negative). The balance between the true positive rate and the false positive rate is controlled by adjusting the threshold.

4. For each threshold value, calculate the true positive rate (TPR) and the false positive rate (FPR). TPR is the proportion of true positives correctly classified as positive, while FPR is the proportion of true negatives incorrectly classified as positive.

   TPR = True Positives / (True Positives + False Negatives)

   FPR = False Positives /(False Positives + True Negatives)

5. Plot the obtained TPR values on the y-axis against the corresponding FPR values on the x-axis. Each point on the curve represents a different classification threshold. Connect the points to visualize the ROC curve.

6. The AUC (Area Under the Curve) represents the overall performance of the classification model. It measures the probability that a randomly chosen positive instance is ranked higher than a randomly chosen negative instance. A higher AUC indicates better model performance.

CHAPTER 3

RESULTS AND DISCUSSION

This study examines the factors that influence a presence of Thyroid. The data used in the study was obtained from the UCI website `https://archive.ics.uci.edu/ml/datasets.php`. The dataset consists of 27 variables, including age, sex, on thyroxine, query on thyroxine, on antithyroid medication, sick, pregnant, thyroid surgery, I131 treatment, query hypothyroid, query hyperthyroid, lithium, goiter, tumor, hypopituitary, psych, TSH measured, TSH, T3 measured, T3, TT4 measured, TT4, T4U measured, T4U, FTI measured, FTI. Of these variables, 26 are independent and one is dependent (Binary Class). The study aims to identify which independent variables have an impact on the dependent variable.

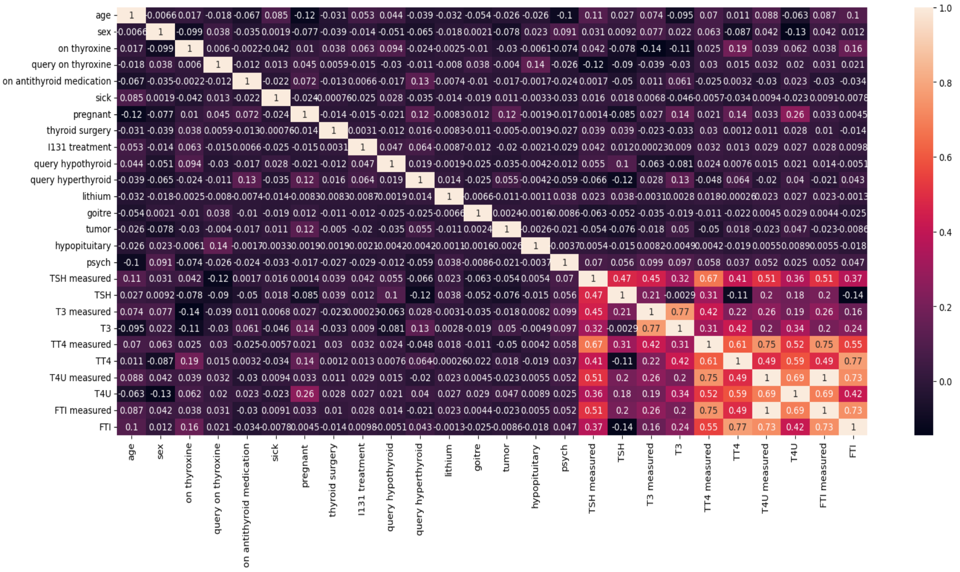


Figure 3.1: Heat Map representing the correlation between the variables

The heatmap (Figure 3.1) represents the most effective attributes of the classification report with various primary colors. A dark color is assigned the lowest number, while the brightest color is assigned for highest number which is considered for the most effective correlation between other variables.

The value -0.0066 indicates that there is a weak negative correlation between age and sex. 0.017 indicates that the there is a week positive correlation between age and on thyroxine value. The above Figure 3.1 is interpreted in this manner.

The presence of a target variable indicates that a supervised learning algorithm is appropriate. Since the dependent variable is categorical, binary supervised classification algorithms are necessary. The following algorithms were applied: Logistic Regression, Random Forest, Decision tree, and Support Vector Machine. The goal of utilizing these algorithms is to determine which one produces the most accurate predictions for the dependent variable.

3.1 Logistic Regression

As the target variable in this study is binary, logistic regression is a suitable algorithm to use. Additionally, the data satisfies the binary assumption. The next step is to check the existence of multicollinearity, which can be done by examining the variance inflation factor (VIF) of the variables. VIF values for the dependent variables are shown below in Table 3.1.

Table 3.1: VIF for the independent variables

	Feature	**VIF**
0	Age	8.184268
1	Sex	1.647187
2	on thyroxine	1.279776
3	query on thyroxine	1.080715
4	on antithyroid medication	1.043379
5	Sick	1.059354
6	Pregnant	1.186259
7	thyroid surgery	1.028858

Continued on next page

Table 3.1: VIF for the independent variables (Continued)

8	I131 treatment	1.035491
9	query hypothyroid	1.105896
10	query hyperthyroid	1.144024
11	Lithium	1.010510
12	Goitre	1.023528
13	Tumour	1.055247
14	Hypopituitary	1.025723
15	Psych	1.101228
16	TSH measured	25.181180
17	TSH	5.584785
18	T3 measured	41.830259
19	TT4 measured	9.698992
20	T3	182.747050
21	TT4	173.418236
22	T4U measured	1736.135256
23	T4U	69.868389
24	FTI measured	1903.733944
25	FTI	165.947058

From the above table, Age has the value of VIF = 8.184268. The Age variable has a moderate VIF, indicating some correlation with other variables in the model but not excessively so. The variable sex has the VIF value = 1.647187. The Sex variable has a relatively low VIF, indicating a low correlation with other variables in the model. Most of the other variables have VIF values close to 1, indicating no significant multicollinearity issues.

T3, TT4, T4U, and FTI: These variables have very high VIF values (T3: 182.747050, TT4: 173.418236, T4U: 69.868389, FTI: 165.947058). These high VIF values suggest strong multicollinearity among these variables, indicating they are highly correlated with each other.

After checking for multicollinearity using the VIF, it was determined that seven variables exhibited severe multicollinearity. These variables (TSH measured, T3 measured, T3, TT4, T4U measured, T4U, FTI measured, FTI) were removed from the dataset as they were not found to have a significant impact on the logistic regression model.

The variables having low VIF values are effective for the logistic Model.

Table 3.2: Results of Logistic Regression

	Coefficient	**S.E.**	**Z**	**P>|Z|**	**[0.025**	**0.975]**
Constant	19.3418	1.296	14.925	0.000	16.802	21.882
Age	0.2250	0.710	0.317	0.752	-1.167	1.617
Sex	0.3376	0.351	0.963	0.335	-0.349	1.025
On thyroxine	9.1604	0.970	9.448	0.000	7.260	11.061
Query on thyroxine	0.5367	1.677	0.320	0.749	-2.750	3.823
On antithyroid medication	4.3093	2.591	1.663	0.096	-0.769	9.388
Sick	-0.0617	0.696	-0.089	0.929	-1.426	1.302
Pregnant	14.7451	7407.868	0.002	0.998	-1.45e+04	1.45e+04
Thyroid surgery	10.9531	1.508	7.265	0.000	7.998	13.908
I131 treatment	0.2361	0.927	0.255	0.799	-1.581	2.054
Query hypothyroid	-0.1697	0.536	-0.316	0.752	-1.221	0.882
Query hyperthyroid	0.1721	0.781	0.220	0.826	-1.358	1.702
Lithium	0.8124	1.480	0.549	0.583	-2.089	3.714
Goitre	7.5313	192.089	0.039	0.969	-368.955	384.018
Tumor	-0.4379	1.049	-0.417	0.676	-2.494	1.619
Hypopituitary	4.8720	4.61e+04	0.000	1.000	-9.04e+04	9.04e+04
Psych	-0.1217	0.753	-0.162	0.872	-1.597	1.353
TSH	-0.1179	0.008	-14.968	0.000	-0.133	-0.102

The logistic regression table displays the estimated coefficients for each independent variable in a logistic regression model, along with their standard errors, z-scores, p-values, and confidence intervals. The logistic regression model is used to predict the probability of a binary outcome based on one or more predictor variables.

In this table, the constant term represents the intercept of the logistic regression model, and its coefficient (19.3418) indicates that when all other predictors are set to zero, the log-odds of the outcome is expected to be 19.3418. The coefficients of the remaining variables represent the estimated effect of each predictor on the log-odds of the outcome, while holding all other predictors constant.

Based on the p-values, the variables ”on thyroxine”, ”thyroid surgery”, and ”TSH” appear to be significant predictors of the outcome, since their p-values are less than 0.05. This means that there is strong evidence to suggest that these variables are associated with the outcome, after accounting for the effects of other predictors.

On the other hand, the variables ”age”, ”sex”, ”query on thyroxine”, ”on antithyroid medication”, ”sick”, ”pregnant”, ”I131 treatment”, ”query hypothyroid”, ”query hyperthyroid”, ”lithium”, ”goitre”, ”tumour”, ”hypopituitary”, and ”psych” do not appear to be significant predictors of the outcome, since their p-values are greater than 0.05. However, it’s important to note that this interpretation assumes that the logistic regression model is an appropriate model for the data, and that the assumptions of the model have been met.

The logistic Regression equation based on our dataset is,

logit(p) = 19.3418 + 0.2250 * Age + 0.3376 * Sex + 9.1604 * On thyroxine + 0.5367 * Query on thyroxine + 4.3093 * On antithyroid medication + (-0.0617) * Sick + 14.7451 * Pregnant + 10.9531 * Thyroid surgery + 0.2361 * I131 treatment + (-0.1697) * Query hypothyroid + 0.1721 * Query hyperthyroid + 0.8124 * Lithium + 7.5313 * Goitre + (-0.4379) * Tumor + 4.8720 * Hypopituitary + (-0.1217) * Psych + (-0.1179) * TSH

The odds ratios are obtained by exponentiating the coefficients. The odds ratio is a measure of the effect of a one-unit increase in the predictor variable on the odds of the response variable being 1 (compared to being 0).

Interpreting the coefficients and odds ratios can provide insights into the relationship between each feature and the target variable. In this case, features with higher absolute coefficients or odds ratios are more influential in predicting the target variable.

Table 3.3: Odds ratio for the selected Features

Features Name	**Odd ratio**
Age	1.00
Sex	1.07
on thyroxine	19.52
query on thyroxine	0.91
on antithyroid medication	1.56
Sick	0.96
Pregnant	1.11
thyroid surgery	12.34
I131 treatment	1.17
query hypothyroid	0.94
query hyperthyroid	1.38
Lithium	1.26
Goitre	1.11
Tumor	0.71
Hypopituitary	1.00
Psych	0.91
TSH	0.94
T3	1.02

From the above table, the following results can be obtained by each row. The odds ratio of 1.00 suggests that age has no effect on the odds of the outcome variable, as the odds remain unchanged for each unit increase in age. The odds ratio of 1.07 indicates that being female (sex = 1) is associated with a 7% increase in the odds of the outcome variable compared to being male (sex = 0). The odds ratio of 19.52 suggests that individuals on thyroxine have significantly higher odds of the outcome variable compared to those not on thyroxine. The odds ratio of 0.91 indicates that individuals with queries about thyroxine have slightly lower odds of the outcome variable compared to those without such queries, although the effect is not statistically significant. The odds ratio of 1.56 suggests that individuals on antithyroid medication have 56% higher odds of the outcome variable compared to those not on this medication. The odds ratio of 0.96 suggests that being sick has a negligible effect on the odds of the outcome variable. The odds ratio of 1.11 suggests that pregnant individuals have slightly higher odds of the outcome variable compared to non-pregnant individuals, although the effect is not statistically significant. The odds ratio of 12.34 indicates that individuals who have undergone thyroid surgery have significantly higher odds of the outcome variable compared to those without the surgery. The odds ratio of 1.17 suggests that individuals undergoing I131 treatment have slightly higher odds of the outcome variable compared to those without this treatment, although the effect is not statistically significant. The odds ratio of 0.94 suggests that individuals with queries about hypothyroidism have slightly lower odds of the outcome variable compared to those without such queries, although the effect is not statistically significant. The odds ratio of 1.38 indicates that individuals with queries about hyperthyroidism have slightly higher odds of the outcome variable compared to those without such queries, although the effect is not statistically significant. The odds ratio of 1.26 suggests that individuals taking lithium have slightly higher odds of the outcome variable compared to those not taking lithium, although the effect is not statistically significant. The odds ratio of 1.11 suggests that individuals with goitre have slightly higher odds of the outcome variable compared to those without goitre, although the effect is not statistically significant. The odds ratio of 0.71 indicates that individuals with a tumour have significantly lower odds of the outcome variable compared to those without a tumour. The odds ratio of 1.00 suggests that hypopituitary status does not have a significant effect on the odds of the outcome variable. The odds ratio of 0.91 suggests that individuals with a psychiatric condition have slightly lower odds of the outcome variable compared to those without such a condition, although the effect is not statistically significant. The odds ratio of 0.94 suggests that for each unit increase in TSH (thyroid-stimulating hormone), the odds of the outcome variable decrease by approximately 6%. The odds ratio of 1.02 suggests that for each unit increase in T3, the odds of the outcome variable increase by approximately 2%.

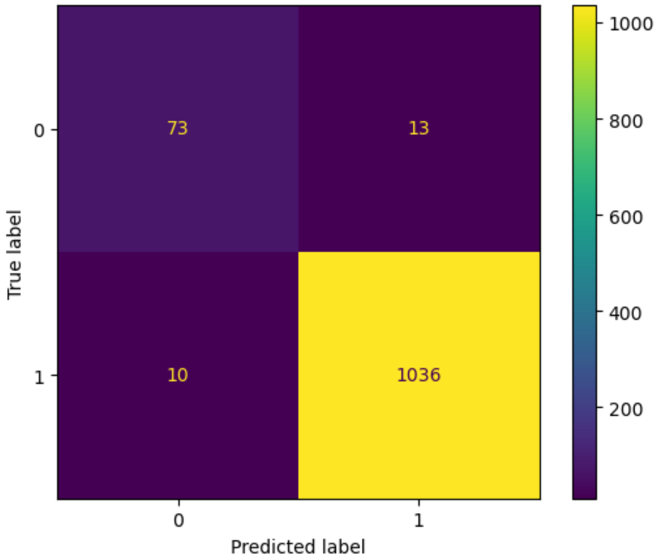


Figure 3.2: Confusion matrix of Logistic Regression.

The above Confusion matrix (Figure 3.2) illustrates the predicted results, as well as the model performance evaluation. According to the Logistic Regression Classifier, the total number of correct predictions is 1109 while the total number of incorrect predictions is 23. It has a 97% accuracy and test F1 score 92%.

3.2 Random Forest

The obtained random forest is shown below.

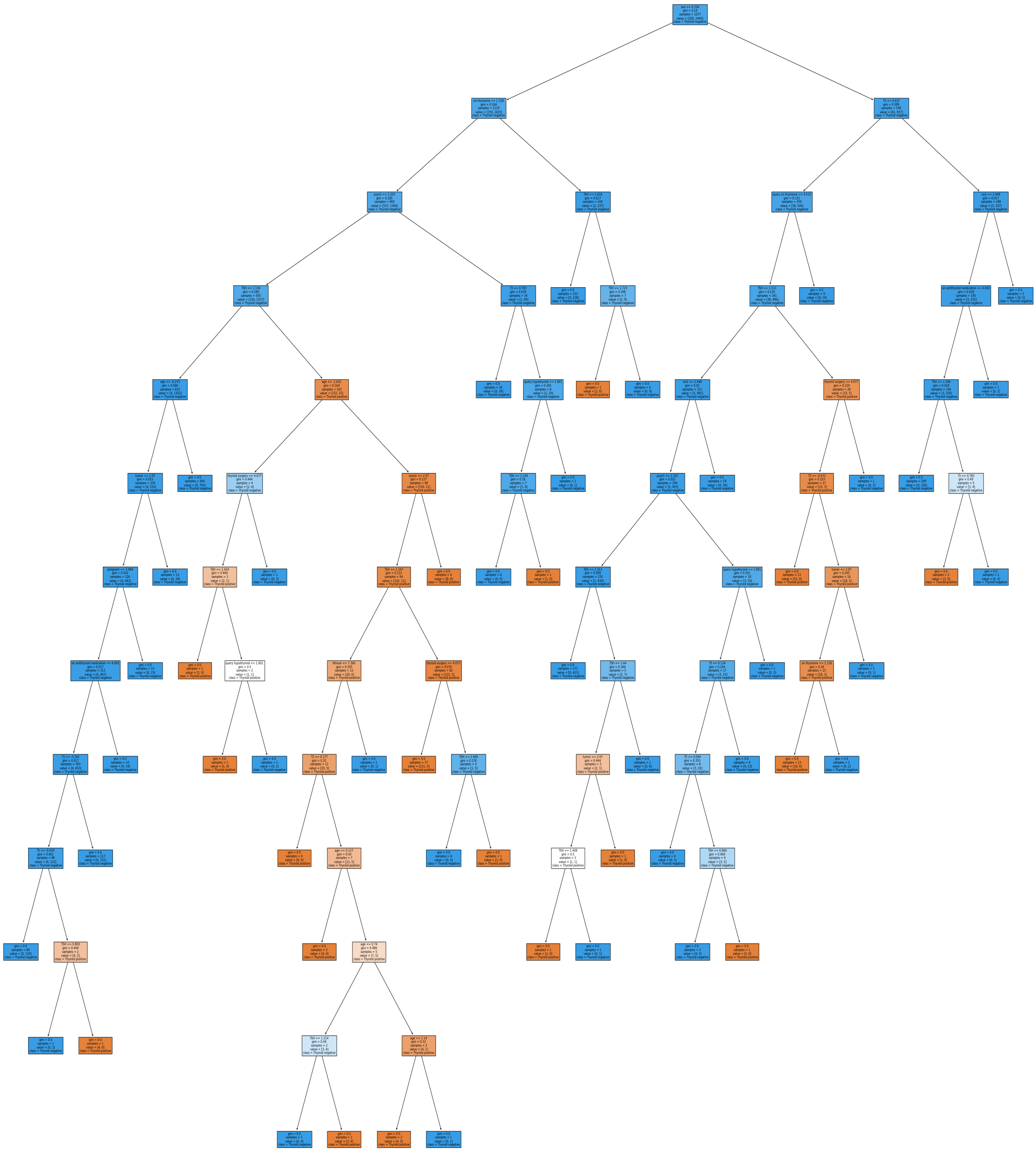


Figure 3.3: Random Forest Tree

The above mentioned Random Forest Trees is one among the 10 random forest trees. The random forest tree is constructed using the bagging method. Bagging is the combination of bootstrapping and aggregation. The trained tree will be used to predict the

Thyroid disease. The new datapoint will be considered and compared with the trained model.

Random forests would provide a measure of feature importance, which indicates how much each feature contributes to the classification accuracy. This information can be useful for understanding the underlying patterns in the data and identifying which features are most relevant for the classification task.

From the below table, we can determine the importance of each feature on the dependent variable.

Table 3.4: Features importance contributes table

Name of the Feature	**Percentage of Importance**
Age	4.12%
Sex	0.46%
On thyroxine	5.64%
Query on Thyroxine	0.08%
On antithyroid medication	0.12%
Sick	0.15%
Pregnant	0.11%
Thyroid surgery	1.93%
I131 Treatment	0.04%
Query Hypothyroid	0.72%
Query hyperthyroid	0.16%
Lithium	0.13%
Goiter	0.01%
Tumor	0.19%
Hypopituitary	0.00%
Psych	0.10%
TSH	81.45%
T3	4.56%

The feature importance scores would provide insights into the most important features for the classification task. In this case, features with higher importance scores are more influential in predicting the target variable. By examining these features, we could gain insights into the underlying patterns in the data and identify potential factors that contribute to the classification task.

Confusion matrix (Figure 3.4) illustrates the predicted results, as well as the model performance evaluation. According to the Random Forest Classifier, the total number of correct prediction is 1120 while the total number of incorrect predictions is 12. It has 98% accuracy and test F1 score 96%.

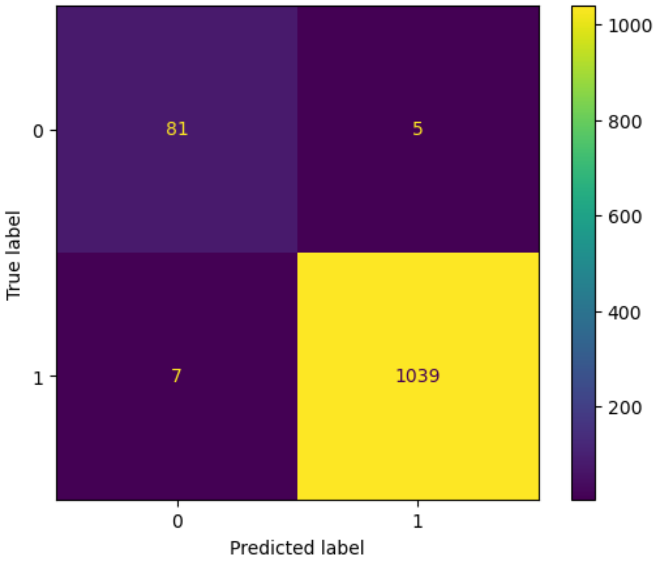
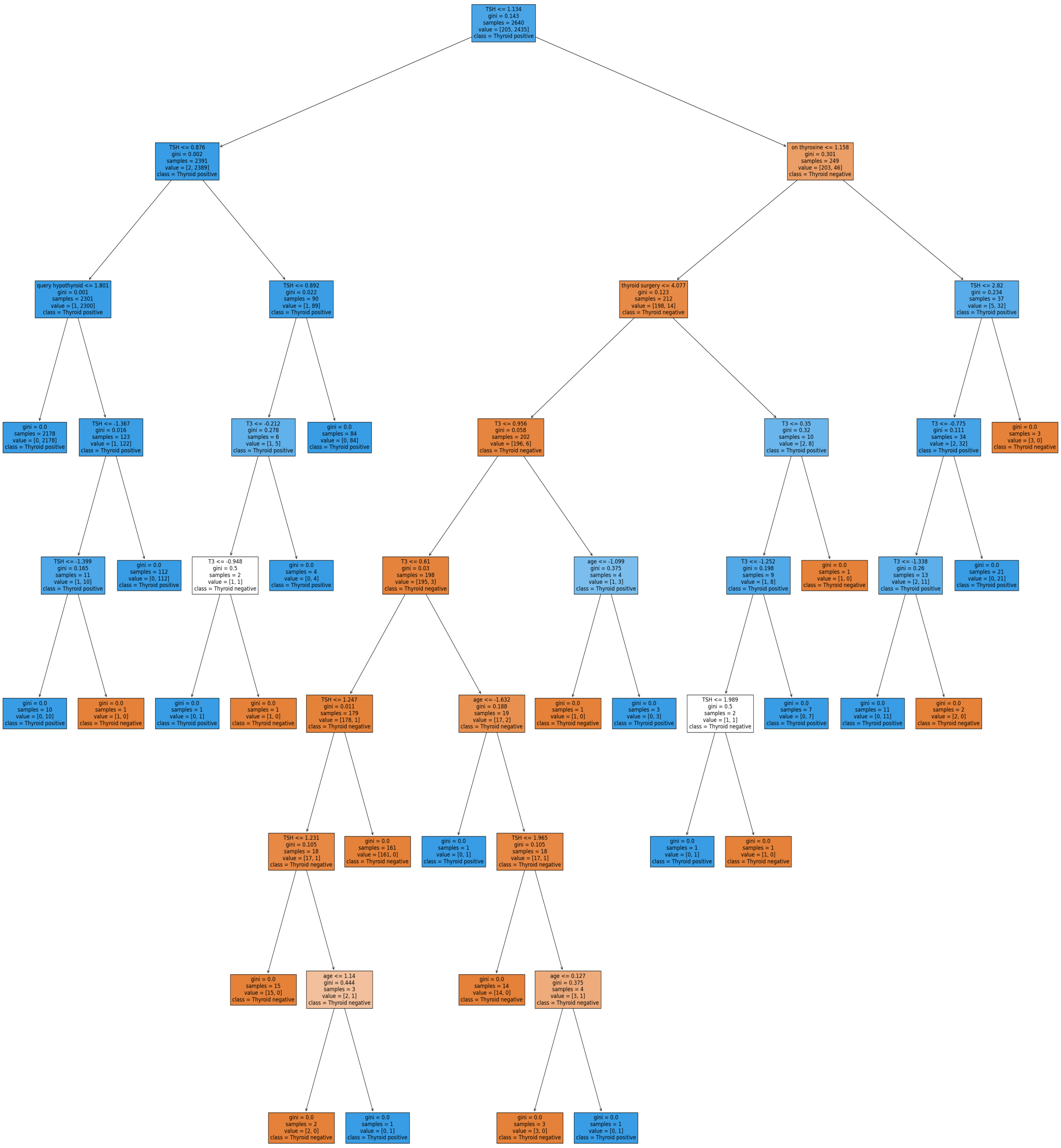


Figure 3.4: Confusion matrix of Random Forest.

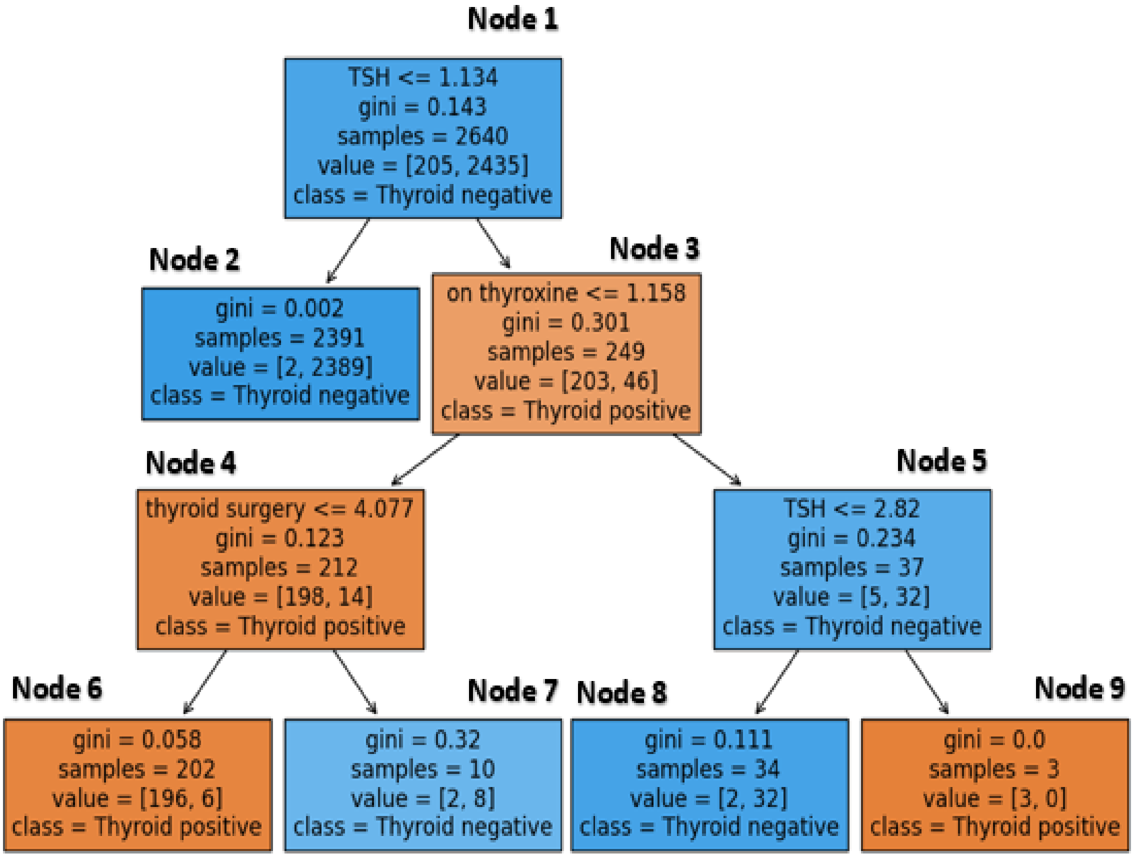
3.3 Decision Tree Classifier

The below decision tree depicts the classification of the data. Each node in the decision tree represents a split on a particular feature, and the edges represent the possible outcomes of the split. The leaves represent the final predictions such as the occurrence of the thyroid is positive or negative. The Gini index is a metric used to measure the impurity of a node in a decision tree. The impurity of a node is a measure of how mixed the labels are in that node. A node with low impurity has mostly labels of the same class, while a node with high impurity has roughly equal numbers of labels of different classes. The Gini Index within the nodes indicates the level of pureness of each node. Features with higher Gini importance scores are more important in predicting the existence of thyroid.


TSH <= 1.134 gini = 0.143 samples = 2640 value = [205, 2435] class = Thyroid positive
TSH <= 0.876 gini = 0.002 samples = 2391 value = [2, 2389] class = Thyroid positive
on thyroxine <= 1.158 gini = 0.301 samples = 249 value = [203, 46] class = Thyroid negative
query hypothyroid <= 1.801 gini = 0.001 samples = 2301 value = [1, 2300] class = Thyroid positive
TSH <= 0.892 gini = 0.022 samples = 90 value = [1, 89] class = Thyroid positive
thyroid surgery <= 4.077 gini = 0.123 samples = 212 value = [198, 14] class = Thyroid negative
TSH <= 2.82 gini = 0.234 samples = 37 value = [5, 32] class = Thyroid positive
gini = 0.0 samples = 2178 value = [0, 2178] class = Thyroid positive
TSH <= -1.367 gini = 0.016 samples = 123 value = [1, 122] class = Thyroid positive
T3 <= -0.212 gini = 0.278 samples = 6 value = [1, 5] class = Thyroid positive
gini = 0.0 samples = 84 value = [0, 84] class = Thyroid positive
T3 <= 0.956 gini = 0.058 samples = 202 value = [196, 6] class = Thyroid negative
T3 <= 0.35 gini = 0.32 samples = 10 value = [2, 8] class = Thyroid positive
T3 <= -0.775 gini = 0.111 samples = 34 value = [2, 32] class = Thyroid positive
gini = 0.0 samples = 3 value = [3, 0] class = Thyroid negative
TSH <= -1.399 gini = 0.165 samples = 11 value = [1, 10] class = Thyroid positive
gini = 0.0 samples = 112 value = [0, 112] class = Thyroid positive
T3 <= -0.948 gini = 0.5 samples = 2 value = [1, 1] class = Thyroid negative
gini = 0.0 samples = 4 value = [0, 4] class = Thyroid positive
T3 <= 0.61 gini = 0.03 samples = 198 value = [195, 3] class = Thyroid negative
age <= -1.099 gini = 0.375 samples = 4 value = [1, 3] class = Thyroid positive
T3 <= -1.252 gini = 0.198 samples = 9 value = [1, 8] class = Thyroid positive
gini = 0.0 samples = 1 value = [1, 0] class = Thyroid negative
T3 <= -1.338 gini = 0.26 samples = 13 value = [2, 11] class = Thyroid positive
gini = 0.0 samples = 21 value = [0, 21] class = Thyroid positive
gini = 0.0 samples = 10 value = [0, 10] class = Thyroid positive
gini = 0.0 samples = 1 value = [1, 0] class = Thyroid negative
gini = 0.0 samples = 1 value = [0, 1] class = Thyroid positive
gini = 0.0 samples = 1 value = [1, 0] class = Thyroid negative
TSH <= 1.247 gini = 0.011 samples = 179 value = [178, 1] class = Thyroid negative
age <= -1.632 gini = 0.188 samples = 19 value = [17, 2] class = Thyroid negative
gini = 0.0 samples = 1 value = [1, 0] class = Thyroid negative
gini = 0.0 samples = 3 value = [0, 3] class = Thyroid positive
TSH <= 1.989 gini = 0.5 samples = 2 value = [1, 1] class = Thyroid negative
gini = 0.0 samples = 7 value = [0, 7] class = Thyroid positive
gini = 0.0 samples = 11 value = [0, 11] class = Thyroid positive
gini = 0.0 samples = 2 value = [2, 0] class = Thyroid negative
TSH <= 1.231 gini = 0.105 samples = 18 value = [17, 1] class = Thyroid negative
gini = 0.0 samples = 161 value = [161, 0] class = Thyroid negative
gini = 0.0 samples = 1 value = [0, 1] class = Thyroid positive
TSH <= 1.965 gini = 0.105 samples = 18 value = [17, 1] class = Thyroid negative
gini = 0.0 samples = 1 value = [0, 1] class = Thyroid positive
gini = 0.0 samples = 1 value = [1, 0] class = Thyroid negative
gini = 0.0 samples = 15 value = [15, 0] class = Thyroid negative
age <= 1.14 gini = 0.444 samples = 3 value = [2, 1] class = Thyroid negative
gini = 0.0 samples = 14 value = [14, 0] class = Thyroid negative
age <= 0.127 gini = 0.375 samples = 4 value = [3, 1] class = Thyroid negative
gini = 0.0 samples = 2 value = [2, 0] class = Thyroid negative
gini = 0.0 samples = 1 value = [0, 1] class = Thyroid positive
gini = 0.0 samples = 3 value = [3, 0] class = Thyroid negative
gini = 0.0 samples = 1 value = [0, 1] class = Thyroid positive

Since the data overfitted in the Decision Tree Algorithm, we have to proceed with the Pruned Decision Tree. Overfitting tends to occur when the model overly complex. For instance, in decision trees, an excessively deep tree with many splits and leaf nodes can be a sign of overfitting.

3.3.1 Pruned Decision Tree



In the Primary split, TSH predictor is considered as the root node. A total of 2640 samples were considered. and 2435 of them are predicted as positive and the split has a Gini index of 0.143.

The secondary split consists of 2 nodes, one node is the terminal node another one is decision node. Node 2 arises as terminal nodes of Root node with a gini index value 0.002. Next node is a decision node Leading other variable Node 3 has On Thyroxine as the variable with gini index value of around 1.158. Total of 249 samples where considered; where 203 are classified as positive and 46 are classified as negative.

The third split gives rise to two Decision node. Node 4 and Node 5 arises as result of splitting node number 3. Node 4 has Thyroid surgery as the predictor variable with a gini index value of 0.123 and total of 212 sample were considered; where 198 were classified as positive and 14 were classified as negative. Node 5 has again TSH as the predictor variable with a gini index value of 2.82 and total of 37 sample where considered of which 5 were classified as negative and 32 were classified as positive.

The fourth split has 4 nodes and all the 4 nodes are terminal nodes. Nodes 6 and node 7 are terminal nodes of node 4 with a gini index of 0.058 and 0.032 respectively. Nodes 8 and 9 are terminal nodes of node 5 with gini index values of 0.111 and 0.0 respectively. The model also only took only 3 predictor variable (TSH, on thyroxine, thyroid surgery) in constructing the tree.

There are a total of 3 Decision nodes and 5 terminal nodes, blue coloured nodes indicate that they are thyroid positive and the orange shades indicate that they are thyroid negative.

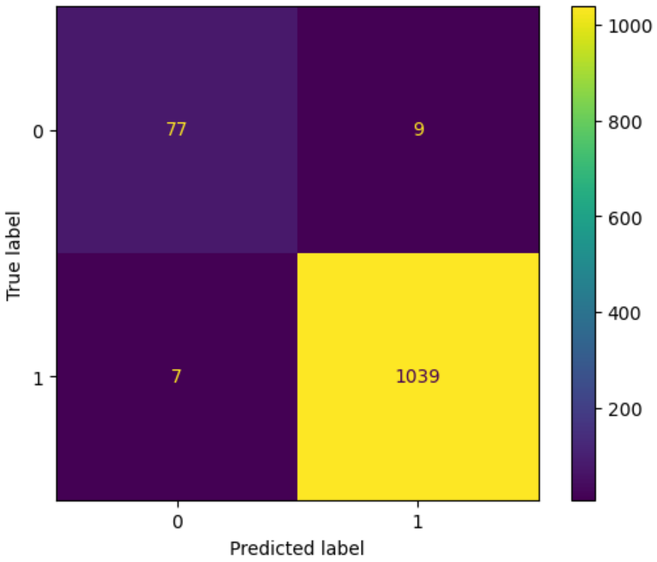


Figure 3.5: Confusion matrix of Decision tree.

Confusion matrix (Figure 3.5) illustrates the predicted results, as well as the model performance evaluation. According to the Decision Tree Classifier, the total number of correct predictions is 1116 while the total number of incorrect predictions is 16. It has a 98% accuracy and test F1 score 94%.

3.4 Support Vector Machine

As by the definition, SVMs aim to find a hyperplane that maximizes the margin between different classes. The margin is defined as the distance between the hyperplane and the nearest data points from each class. Here the data is classified and the margin between the two classes are maximized. The data points belong to the target variable such as the presence of Thyroid is classified according to the class.

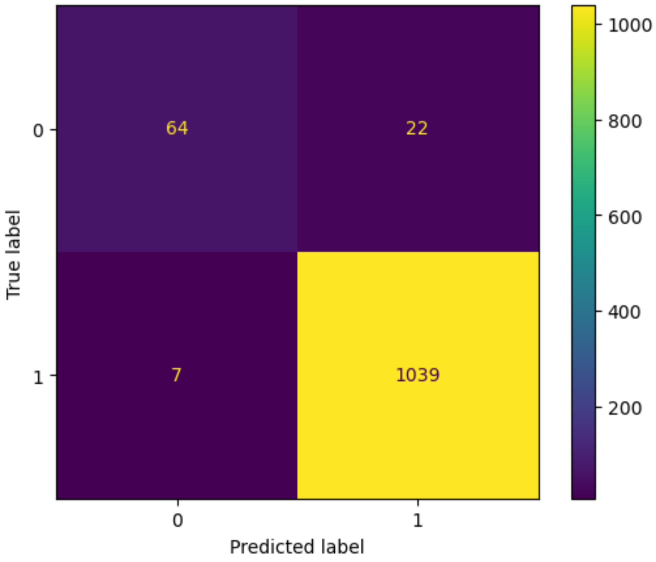


Figure 3.6: Confusion matrix of Support Vector Machine.

Confusion matrix (Figure 3.6) illustrates the predicted results, as well as the model performance evaluation. According to the Support vector machine, the total number of correct predictions is 1103 while the total number of incorrect predictions is 29. It has a 97% accuracy and test F1 score 90%.

3.5 Comparison of the performance measures of the Classifiers

Table 3.5: Evaluvation metrics of the classifier

Model	Accuracy	F1_score	Precision	Recall	Sensitivity	Specificity
Logistic Regression	0.979682	0.926463	0.933563	0.919638	0.990440	0.848837
Random Forest	0.989399	0.962646	0.957833	0.967584	0.993308	0.941860
Decision Tree	0.985866	0.949121	0.954039	0.944328	0.993308	0.895349
SVM	0.974382	0.900761	0.940337	0.868747	0.993308	0.744186

In the study, the four prediction algorithms are used i.e., Logistic regression, Random Forest, Decision tree and Support Vector machine. The performance between these algorithms is compared using the classification accuracy, F1-score, Precision, Recall, sensitivity and specificity. As compared to other models, Random Forest has achieved best accuracy results. This suggests that the Random Forest model is the most accurate and reliable model for the given classification task.

The Random Forest and Decision Tree models also performed well, with high scores for most of the metrics.

Accuracy is an important metric that indicates the proportion of correct predictions made by the model. The Random Forest model achieved the highest accuracy of 0.989399, meaning that it correctly classified almost all instances in the test dataset. The F1 score is a weighted average of precision and recall, and is a good measure of the balance between the two. The Random Forest model achieved the highest F1 score of 0.962646, indicating a good balance between precision and recall.

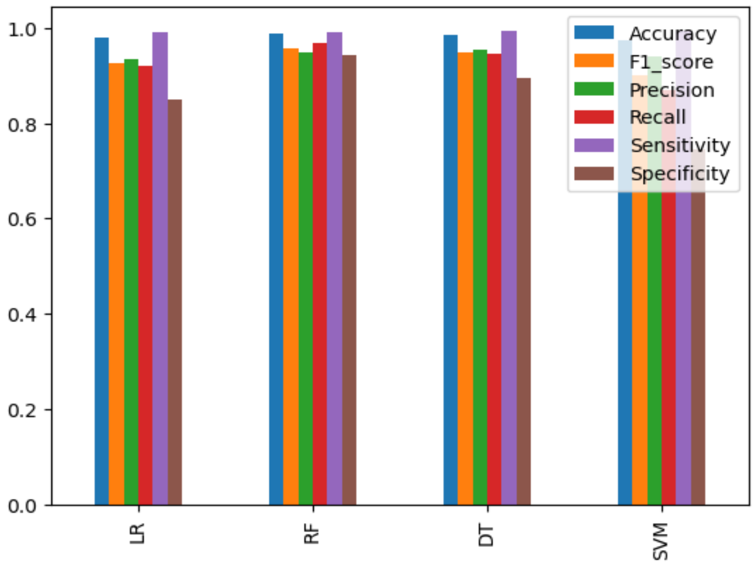


Figure 3.7: Bar chat diagram for 4 algorithms with six-evaluation metrics

The Random Forest model have high precision (0.957833) and recall (0.967584) scores, indicating its effectiveness in accurately classifying both positive and negative instances. Moreover, the model achieved the highest scores for both metrics, with a sensitivity of 0.993308 and a specificity of 0.941860. This suggests that the model can correctly identify positive instances while avoiding false positives. Based on its high scores for all six-evaluation metrics, the Random Forest model appears to be the most effective approach for the given classification task.

3.5.1 ROC Curve

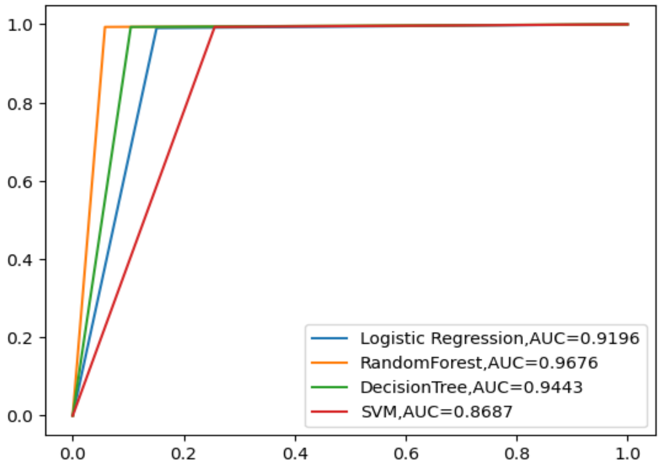


Figure 3.8: Receiver operating characteristic curve of classifier

When comparing the Area Under the Curve (AUC) values for different classifiers like logistic regression, random forest, decision tree, and support vector machine (SVM), a higher AUC indicates better performance in distinguishing between positive and negative instances.

With an AUC of 0.9676, the random forest classifier demonstrates excellent When comparing the Area Under the Curve (AUC) values for different classifiers like logistic regression, random forest, decision tree, and support vector machine (SVM), a higher AUC indicates better performance in distinguishing between positive and negative instances. performance in distinguishing between positive and negative instances. It provides the highest discriminatory power among the four classifiers you mentioned. The decision tree performs well with an AUC of 0.9443. Although slightly lower than the random forest, it still shows strong discriminatory ability. With an AUC of 0.9196, the logistic regression model performs reasonably well but falls slightly behind the decision tree. The SVM model has the lowest AUC of 0.8687, indicating comparatively weaker performance in distinguishing between positive and negative instances.

In summary, based on the AUC values, the random forest classifier is the most effective among the four models, followed by the decision tree, logistic regression, and SVM.

CHAPTER 4

SUMMARY AND CONCLUSION

This study examines the factors that influence the presence of Thyroid using a dataset obtained from the UCI website. The dataset consists of 27 variables, including age, sex, medication history, and various measurements related to thyroid function. The study aims to identify which independent variables have a significant impact on the dependent variable.

Initially, a heatmap is used to visualize the correlation between variables. After checking for multicollinearity using the VIF (Variance Inflation Factor), seven variables are removed from the dataset due to severe multicollinearity. These variables were found to have no significant impact on the logistic regression model.

A logistic regression model is then employed to predict the probability of a binary outcome based on the remaining independent variables. The logistic regression table provides estimated coefficients, standard errors, z-scores, p-values, and confidence intervals for each independent variable. The odds ratios, obtained by exponentiating the coefficients, measure the effect of a one-unit increase in a predictor variable on the odds of the response variable being positive.

A confusion matrix is presented to evaluate the performance of the logistic regression model, showing a total of 1109 correct predictions and 23 incorrect predictions. The model achieves 97% accuracy and a test F1 score of 92%.

To further explore feature importance, a Random Forest Classifier is employed. This algorithm provides insights into the most important features for the classification task. The Random Forest Classifier achieves 98% accuracy and a test F1 score of 96%, with a total of 1120 correct predictions and 12 incorrect predictions.

A decision tree is also constructed to classify the data. The Gini index, a measure of impurity, is used to evaluate the nodes' purity and splitting decisions. The decision tree structure is visualized, with nodes representing splits on specific features and leaves representing final predictions. The decision tree achieves 98% accuracy and a test F1 score of 94%, with 1116 correct predictions and 16 incorrect predictions. The support vector machine achieves 97% accuracy and test F1 score 90% with 1103 correct predictions and 29 incorrect predictions.

Comparing the performance of the four prediction algorithms (Logistic Regression, Random Forest, Decision Tree, Support Vector Machine), Random Forest demonstrates the highest accuracy. The Random Forest model achieves an accuracy of 98.94%, an F1 score of 96.26%, precision of 95.78%, recall of 96.76%, sensitivity of 99.33%, and specificity of 94.19%.

Based on the evaluation of the classification algorithms, the Random Forest model emerges as the most accurate and reliable approach for predicting the presence of Thyroid. It achieves high scores for accuracy, F1 score, precision, recall, sensitivity, and specificity. The Random Forest model effectively classifies both positive and negative instances, demonstrating its ability to correctly identify positive cases while minimizing false positives. Therefore, the Random Forest algorithm is recommended for further use in predicting thyroid presence based on the provided dataset.

REFERENCES

1. Géron, A. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, 2019.

2. Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

3. Kowalczyk, A. *Support Vector Machines Succinctly*. Syncfusion Inc., 2018.

4. Kuhn, M. and Johnson, K. *Applied Predictive Modeling*. Springer, 2013.