

Classification of Differentiated Thyroid Cancer Recurrence Using Machine Learning Algorithms

submitted by

Manshi Shaw

Class Roll No: 002210503042

Exam Roll: MCA244041

Master of Computer Application

under the supervision of

Prof. (Dr.) Susmita Ghosh (De)

in the

Department of Computer Science and Engineering

Faculty of Engineering and Technology

Jadavpur University

2024

Faculty of Engineering and Technology
Jadavpur University

CERTIFICATE OF RECOMMENDATION

To Whom It May Concern

I hereby recommend that the project report entitled “Classification of Differentiated Thyroid Cancer Recurrence Using Supervised Machine Learning Algorithms” has been carried out by Manshi Shaw (Reg. No.: 163634 of 2022-23, Roll No.: 002210503042, Exam Roll: MCA244041), under my guidance and supervision and be accepted in partial fulfilment of the requirement for the degree of Master of COMPUTER APPLICATION in DEPARTMENT of COMPUTER SCIENCE and ENGINEERING, JADAVPUR UNIVERSITY during the academic year 2023-2024.

Prof. (Dr.) Susmita Ghosh (De)
Project Supervisor
Dept. of Computer Science and
Engineering
Jadavpur University

Prof. (Dr.) Nandini Mukhopadhyay
Head of the Department
Dept. of Computer Science & Engineering
Jadavpur University

Prof. Dipak Laha
Dean, Faculty Council
of Engineering & Technology
Jadavpur University, Kolkata-032

Faculty of Engineering and Technology
Jadavpur University

CERTIFICATE OF APPROVAL

This is to certify that the project report entitled “**Classification of Thyroid Cancer Recurrence Using Supervised Machine Learning Algorithms**” is a bonafide record of work carried out by MANSHI SHAW, in fulfillment of the requirements for the award of the degree of Master of Computer Application in the Department of Computer Science and Engineering, Jadavpur University. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the project report only for the purpose for which it has been submitted.

Signature of Examiner 1

Date :

Signature of Examiner 2

Date :

Faculty of Engineering and Technology
Jadavpur University

**DECLARATION OF COMPLIANCE WITH
ACADEMIC ETHICS**

I hereby declare that this project report entitled "Classification of Differentiated Thyroid Cancer Recurrence Using Supervised Machine Learning Algorithms" is an outcome of my genuine and independent work, as part of my Master of Computer Application program.

All information in this document has been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, all materials and results that are not original to this work have been fully cited and referenced.

Manshi Shaw

Registration No.: 163634 of 2022-23

Examination Roll No.: MCA244041

Thesis Title :- Classification of Differentiated Thyroid Cancer Recurrence Using Supervised Machine Learning Algorithms

Signature with Date

ACKNOWLEDGEMENT

With my most sincere and gratitude, I would like to thank Prof. (Dr.) Susmita Ghosh (De), Department of Computer Science & Engineering, my supervisor, for his overwhelming support throughout the duration of the project. His motivation always gave me the required inputs and momentum to continue with my work, without which the project work would not have taken its current shape. His valuable suggestions and numerous discussions have always inspired new ways of thinking. I feel deeply honored that I got this opportunity to work under him.

I would like to express my sincere thanks to all my teachers for providing sound knowledge base and cooperation.

I would like to thank all the faculty members of the Department of Computer Science & Engineering of Jadavpur University for their continuous support.

Last, but not the least, I would like to thank my batch mates for staying by my side when I needed them.

Manshi Shaw

Class Roll No.: 002210503042

Examination Roll No.: MCA244041

Department of Computer Science & Engineering
Jadavpur University

Contents

Chapters

Page No

1.Introduction.....	8
1.1 Scope of The Report.....	8
1.2 Organization of The Report.....	8-9
2.Literature Survey.....	10
3.Methodology.....	11
3.1 Introduction to learning.....	11
3.1.1 Supervised Learning.....	11
3.1.2 Unsupervised Learning.....	11
3.1.3 Semi-Supervised Learning.....	12
3.1.4 Reinforcement Learning.....	12
3.2 Classifier Used.....	13
3.3 Working Principle.....	14
3.3.1 K-Nearest Neighbors (KNN).....	14-16
3.3.2 Support Vector Machines (SVM).....	16-18
3.3.3 Decision Tree.....	19-21
4.Result.....	22
4.1 Introduction.....	22
4.2 Dataset Used.....	22-24
4.3 Parameter Used.....	25
4.4 Performance Metrics.....	25-26
4.5 Pre-processing.....	26
4.6 Results and Analysis.....	26-32
5. Conclusion and Future Scope.....	35
6.References.....	35

List of Tables

	Page No.
4.1 Results Obtained for 20 simulation using KNN classifier	27
4.2 Results Obtained for 20 simulation using SVM classifier	27
4.3 Results Obtained for 20 simulation using Decision Tree classifier	27
4.13 Classification report of KNN	33
4.14 Classification report of SVM	33
4.15 Classification report of Decision Tree	34

List of Figures

	Page No.
3.1 Flowchart of KNN classifier	15
3.2 Flowchart of SVM classifier	17
3.3 Flowchart of Decision Tree	20
4.4 KNN confusion matrix for 0.1 Test Size	28
4.5 KNN confusion matrix for 0.2 Test Size	28
4.6 KNN confusion matrix for 0.3 Test Size	29
4.7 SVM confusion matrix for 0.1 Test Size	29
4.8 SVM confusion matrix for 0.2 Test Size	30
4.9 SVM confusion matrix for 0.3 Test Size	30
4.10 Decision Tree confusion matrix for 0.1 Test Size	31
4.11 Decision Tree confusion matrix for 0.2 Test Size	31
4.12 Decision Tree confusion matrix for 0.3 Test Size	32

Chapter 1: Introduction

The primary objective of this project is to train machine learning models to predict the chances of recurrence in patients diagnosed with well-differentiated thyroid cancer. While thyroid cancer mortality remains low, the risk of recurrence is a significant concern [1]. Differentiated Thyroid Cancer is a kind of cancer that affects the endocrine system [2]. Thyroid cancer is characterized by clinical symptoms such as dysphagia, throat constriction, painful throat, hoarseness, coughing, weight loss, weakness, and excessive perspiration [3]. The number of thyroid cancer cases has steadily increased in recent years.

Therefore, creating a model that allows healthcare practitioners to accurately predict a patient's risk of recurrence is crucial. This capability enables personalized treatment plans and adjustments to follow-up schedules. Effective risk stratification is especially important for deciding when to start radioiodine therapy.

1.1 Scope of The Report:

This report deploys the application of various supervised machine learning models on the provided dataset to detect the recurrence of Differentiated Thyroid Cancer. The scope of this report encompasses data collection, pre- processing, model training and evaluation and the implementation of parameter tuning methods to find the best parameter which suits the model. The report also details the development and reckoning the effectiveness of different machine learning models applied on the dataset. Along with that it compares the performance of different models trained on the provided dataset.

1.2 Organization of The Report:

The thesis is organized as follows:

Chapter 1 contains Introduction to the report followed by scope of the report and the organization of the report.

Chapter 2 contains Literature Survey where a basic survey has been done on Differentiated Thyroid Cancer Recurrence. This literature survey highlights the significance of differentiated thyroid cancer (DTC) recurrence, emphasizing its impact on patient management and outcomes despite a generally favorable prognosis. Recent studies have focused on using predictive analytics, particularly supervised machine learning (ML), to identify risk factors such as tumor size, histopathological features, and genetic mutations associated with recurrence.

Supervised Machine Learning models, trained on large datasets of Differentiated Thyroid Cancer Recurrence patients, can uncover complex patterns and interactions among variables influencing recurrence risk. Decision trees, support vector machines, and K-nearest neighbor models have been employed to predict recurrence more accurately than traditional statistical methods.

Chapter 3 delves into the Methodology section, detailing the working principles, algorithms, and flowcharts for three classification algorithms: K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Decision Tree. This section provides a comprehensive understanding of how each algorithm operates and its specific steps in classifying data, aiding in the replication and implementation of the classification task.

While chapter 4 contains Dataset used, Analysis of results, Parameter considered for parameter tuning, different performance matrices, steps for data pre-processing, accuracy score, plotting of confusion matrix. Chapter 5 of the project encompasses the Conclusion and Future Scope, as well as a list of references and citations used.

Chapter 2: Literature Survey

The aim of this chapter is to provide a thorough review of current research and equip readers with the knowledge necessary to spot relevant methodologies, hypotheses, and gaps or limitations in existing research work in this domain.

Shiva Borzooei, Giovanni Brganti, Mitra Golparian, Jerome R. Lechein and Aidin Tarokhian [1] predicts the recurrence of Differentiated Thyroid Cancer using machine learning models Classic (K-nearest neighbors, support vector machines (SVM), tree-based models) and artificial neural networks (ANN) were trained on three distinct combinations of features. A data set with all features excluding American Thyroid Association (ATA) risk score (12 features), another with ATA risk alone, and a third with all features combined (13 features). 283 patients were allocated for the training process, and 100 patients were reserved for the validation of stage.

When using the full data set for training, the models showed the following sensitivity, specificity and AUC, respectively: SVM (99.33%, 97.14%, 99.71), K-nearest neighbors (83%, 97.14%, 98.44), Decision Tree (87%, 100%, 99.35), Random Forest (99.66%, 94.28%, 99.38), ANN (96.6%, 95.71%, 99.64). When the ATA risk data from the training set were removed, the models' ability to correctly identify negatives (specificity) improved, but their ability to correctly identify positives (sensitivity) worsened. On the other hand, training the models only with ATA risk data increased sensitivity but decreased specificity.

Karli Eka Setiawan [2] predicting the presence and absence of recurrence of DTC using 23 different machine learning models, such as SVM with various kernels (linear, polynomial, and radial basis functions), logistic regression, Naïve Bayes, Decision Tree, K-Nearest Neighbor, ensemble bagging using various base learners, random forest, ensemble stacking using various base learners, ensemble boosting using various base learners with AdaBoost, and Gradient Boosting Machine. Without using feature selection, the best machine learning model was Random Forest, achieving 94% precision, recall, and F1-score in predicting the presence of DTC recurrence, and 98% precision, recall, and F1-score in predicting the absence of DTC recurrence.

Cao et al. conducted a review on using machine learning to detect DTC. They used radiomics, a technique that extracts a lot of detailed information from one or more medical images, to predict the presence of DTC in those images [4].

Chapter 3: Methodology

3.1 Introduction to learning:

In the field of machine learning, "learning" refers to the process by which a computer system or model gains knowledge and makes predictions or decisions based on data. This involves the model identifying patterns and relationships within the data to make accurate predictions or decisions.

Machine learning (ML) is an application of artificial intelligence (AI) that allows computers to automatically learn and improve from past experiences without needing to be specifically programmed for those tasks. It helps them recognize patterns and make predictions with little human help.

3.1.1 Supervised Learning:

In supervised learning models are trained using labelled dataset, and on basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output. The objective is for the model to learn the mapping or relationship between the input and output. Once trained, the model can make predictions or classify new, unseen data accurately.

Classification: It predicts discrete labels (e.g., spam or not spam).

Regression: It Predicts continuous values (e.g., Prediction of values in Business etc.)

Examples: Support Vector Machine (SVM), K-Nearest Neighbourhood, Decision Tree, Random Forest , Linear Regression.

2.Application: Common applications includes sentiment analysis, image classification, spam filtering, Pattern detection etc.

3.1.2 Unsupervised Learning:

Unsupervised learning is a type of machine learning in which models are trained using unlabelled dataset and are allowed to act on that data without any supervision. The main aim of the unsupervised learning algorithm is to group or categories the dataset according to the similarities, patterns, and differences. Machines are instructed to find the hidden patterns from the input dataset.

1. The model finds the underlying structure of the dataset.
2. Application: common application includes recommendation system, network analysis, anomaly detection.

3.1.3 Semi-Supervised Learning:

Semi-Supervised learning is a type of Machine Learning algorithm that lies between Supervised and Unsupervised machine learning. It represents the intermediate ground between Supervised (With Labelled training data) and Unsupervised learning (with no labelled training data) algorithms and uses the combination of labelled and unlabelled datasets during the training period.

1. Although Semi-supervised learning is the middle ground between supervised and unsupervised learning and operates on the data that consists of a few labels, it mostly consists of unlabelled data.
2. To overcome the drawbacks of supervised learning and unsupervised learning algorithms, the concept of Semi-supervised learning is introduced.

3.1.4 Reinforcement Learning:

Reinforcement learning works on a feedback-based process, in which an AI agent (A software component) automatically explore its surrounding by hitting & trail, taking action, learning from experiences, and improving its performance. Agent gets rewarded for each good action and get punished for each bad action; hence the goal of reinforcement learning agent is to maximize the rewards.

1. In reinforcement learning, there is no labelled data like supervised learning and agents learn from their experiences only.
2. Reinforcement learning is often used in applications where decision-making and sequential actions are critical.

3.2 Classifier Used:

In the Classification of Differentiated Thyroid Cancer Recurrence using Supervised machine learning algorithms we have used these following machine learning classifiers known for their effectiveness in classification task:

1.K-Nearest Neighbors (KNN): The k-nearest neighbours (KNN) algorithm is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. KNN is a lazy learner, which means that it does not explicitly learn a model during the training phase. Instead, it memorizes the training dataset.

2.Support Vector Machines (SVM): Support Vector Machine (SVM) is a powerful machine learning algorithm used for linear or nonlinear classification, regression. SVM algorithm is very effective as we try to find the maximum separating hyperplane between the different classes available in the target feature. The main goal of SVM is to find the best boundary, or hyperplane, that separates different classes in the data. This boundary is chosen to maximize the distance between the classes, making it easier to classify new data points correctly.

SVM can be of two types:

Linear SVM: Linear SVM is used for linearly separable data, which means that a dataset can be classified into two classes by using a single straight line,

Non-linear SVM: Non-Linear SVM is used for non-linearly separated data, which means that a dataset cannot be classified into two classes by using a single straight line.

3.Decision Tree (DTs): Decision Trees are a non-parametric supervised learning method used for classification. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. It works by splitting data into branches based on feature values, leading to a tree-like structure. The tree starts with a root node, which represents the entire dataset.

A decision tree is a flowchart-like tree structure, where each non-leaf node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label. The topmost node in a tree is the root node [5].

3.3 Working principle:

3.3.1 K-Nearest Neighbors (KNN): KNN classifier is a machine learning algorithm mostly used for classification problems. It works by finding the K nearest points in the training dataset and uses their class to predict the class or value of a new data point. The corresponding algorithm is given below and the flowchart is shown in Fig. 3.1.

Algorithm:

Step-1: Select the number K of the neighbors

Step-2: Calculate the Euclidean distance of K number of neighbors.

Distance Calculation: When a new input needs to be classified or predicted, the algorithm calculates the distance between this test instance and all instances in the training dataset. Common distance metrics include:

- Euclidean distance: the distance between two points (x_1, y_1) and (x_2, y_2) is $d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$.
- Manhattan distance: Distance = $|x_1 - x_2| + |y_1 - y_2|$
- Minkowski distance: Generalized form of Euclidean and Manhattan distances

Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.

Step-4: Among these k neighbors, count the number of the data points in each category.

Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.

Flowchart of KNN:

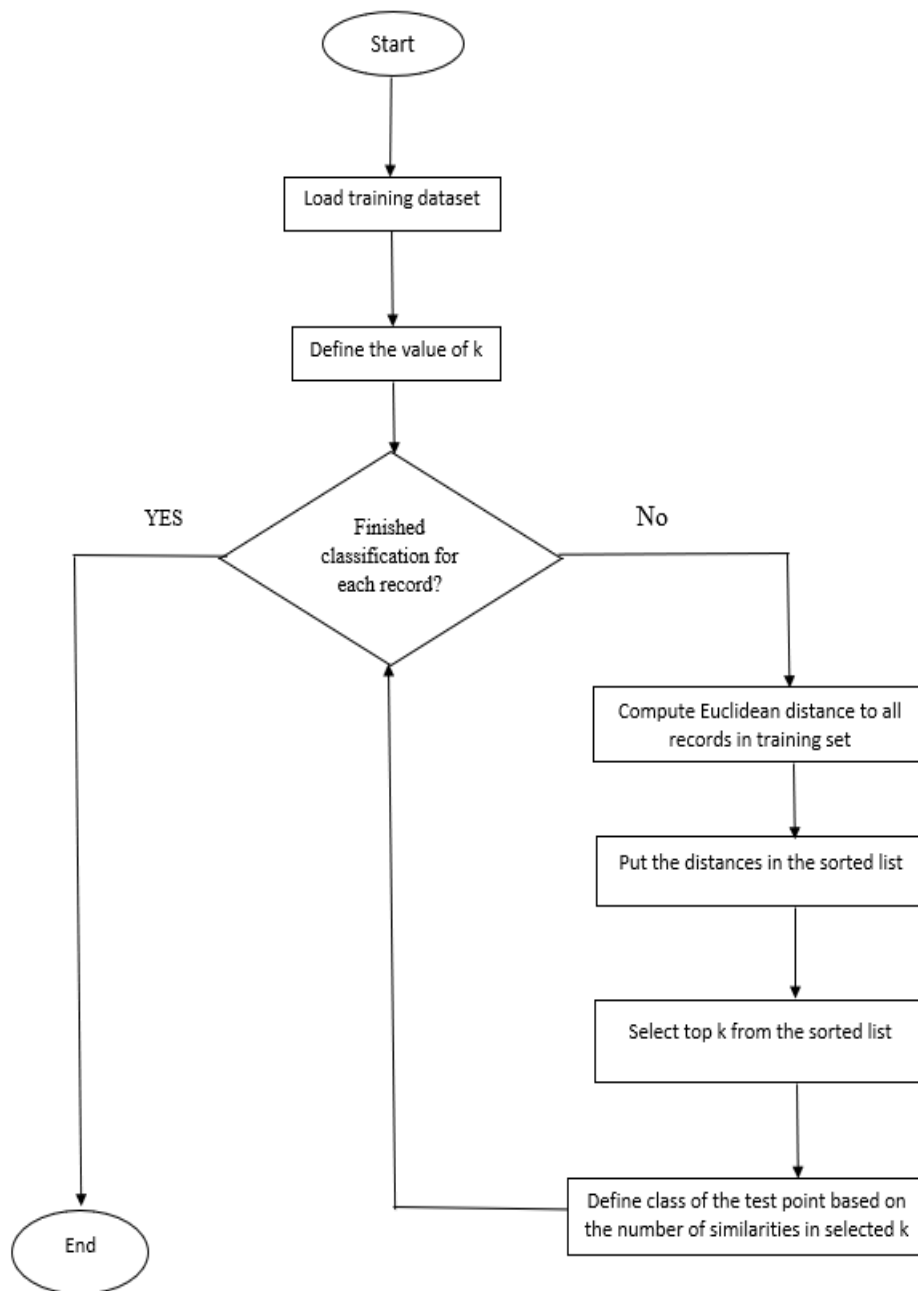


Fig. 3.1: Flowchart of KNN classifier

Advantages of KNN Algorithm:

1. **Simplicity:** Easy to understand and implement.
2. **No Training:** No explicit training phase, which can be advantageous for quickly updating the model with new data.
3. **Robust:** It is robust to the noisy training data. It can be more effective if the training data is large.

Disadvantages of KNN Algorithm:

1. **Computationally Intensive:** It needs a lot of computing power because it has to calculate the distance between the new data point and all the stored data points every time it makes a prediction.
2. **Memory Intensive:** It requires a lot of data storage since it has to keep all the training data.
3. **Curse of Dimensionality:** This implies the algorithm faces a hard time classifying the data points properly when the dimensionality is too high.

3.3.2. Support Vector Machines (SVM): SVM works by mapping data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. A separator between the categories is found, then the data are transformed in such a way that the separator could be drawn as a hyperplane. The main objective of SVM is to find the best boundary, or hyperplane, that separates the data into different classes. The corresponding algorithm is given below and the flowchart is provided in Fig. 3.2.

Algorithm:

1. **Input:** Training dataset with labelled data points.
2. **Initialization:** Choose a kernel function and regularization parameter.
 - **Linear Kernel :** For linearly separable data.
 - **Polynomial Kernel:** For data where the relationship between features is polynomial.
 - **Radial Basis Function (RBF) Kernel/Gaussian Kernel:** For data where the decision boundary is non-linear.
 - **Sigmoid Kernel:** Similar to neural networks with a certain type of activation function.

3. Training:

- Construct a hyperplane that maximizes the margin between classes.
- Identify support vectors (data points closest to the hyperplane).
- Optimize the margin and minimize classification errors.

4. Classification: For new data points:

- Determine which side of the hyperplane they fall on.
- Assign them to the corresponding class based on the side.

5. Output: Trained SVM model for classification tasks.

Flowchart of SVM:

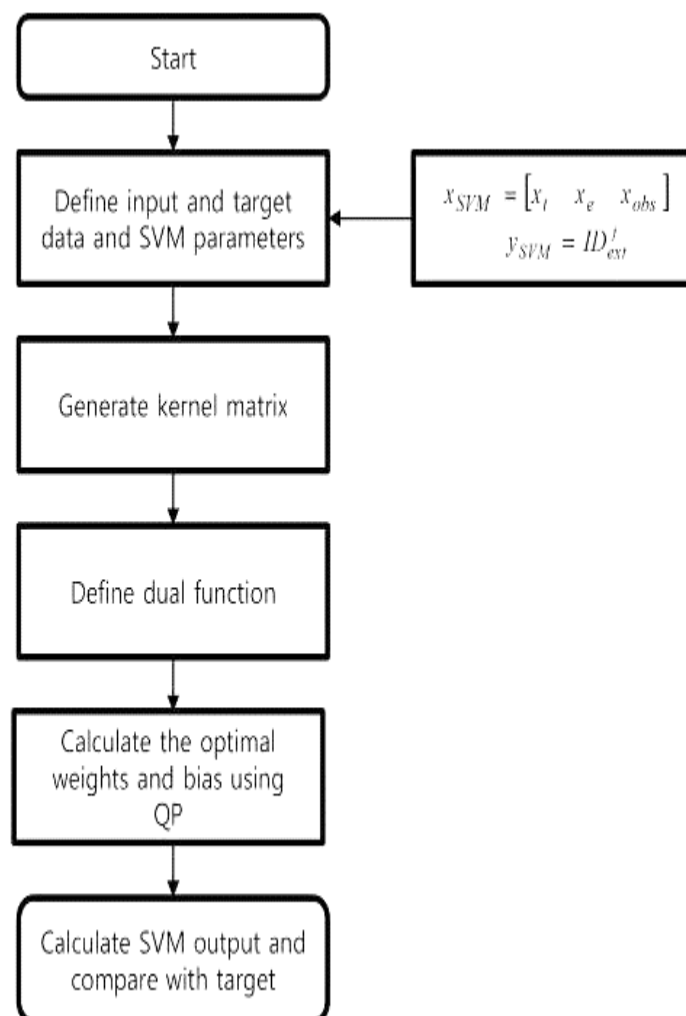


Fig. 3.2: Flowchart of SVM classifier

Advantages of SVM Algorithm:

1. Effective in High-Dimensional Spaces:

SVM works well in cases where the number of dimensions is greater than the number of samples.

2. Memory Efficient:

SVM uses a subset of training points (support vectors) in the decision function, making it memory efficient.

3. Versatile with Different Kernel Functions:

The ability to use custom kernel functions allows SVM to adapt to different types of data distributions (linear, non-linear, polynomial, radial basis function, etc.).

4. Clear Margin of Separation:

SVM provides a clear margin of separation between classes, which can improve the generalization of the classifier.

5. Robust to Overfitting:

Proper regularization parameters (like C) help prevent overfitting, especially in high-dimensional spaces.

Disadvantages of SVM:

1. Computationally expensive: SVMs can be computationally expensive for large datasets, as the algorithm requires solving a quadratic optimization problem.

2. Choice of kernel: The choice of kernel can greatly affect the performance of an SVM, and it can be difficult to determine the best kernel for a given dataset.

3. Sensitivity to the choice of parameters: SVMs can be sensitive to the choice of parameters, such as the regularization parameter, and it can be difficult to determine the optimal parameter values for a given dataset.

4. Not suitable for large datasets with many features: SVMs can be very slow and can consume a lot of memory when the dataset has many features.

3.3.3. Decision Tree: Decision tree learning employs a divide and conquer strategy by conducting a greedy search to identify the optimal split points within a tree. This process of splitting is then repeated in a top-down, recursive manner until all, or the majority of records have been classified under specific class labels. The attribute that maximizes information gain is chosen as the splitting criterion for building the decision tree. Information gain is used in both classification and regression decision trees. In classification, entropy is used as a measure of impurity, while in regression, variance is used as a measure of impurity. The corresponding algorithm is given below and the flowchart is visualized in Fig. 3.3.

Algorithm:

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) to split the dataset. Compute the ASM for each attribute. Choose the attribute with the highest ASM value as the best attribute for splitting.

Common ASMs include:

Information Gain (based on entropy)

Gini Index

Step-3: Divide the S into subsets that contains possible values for the best attributes identified in Step 2.

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.

Step-6: End.

Flowchart of Decision Tree:

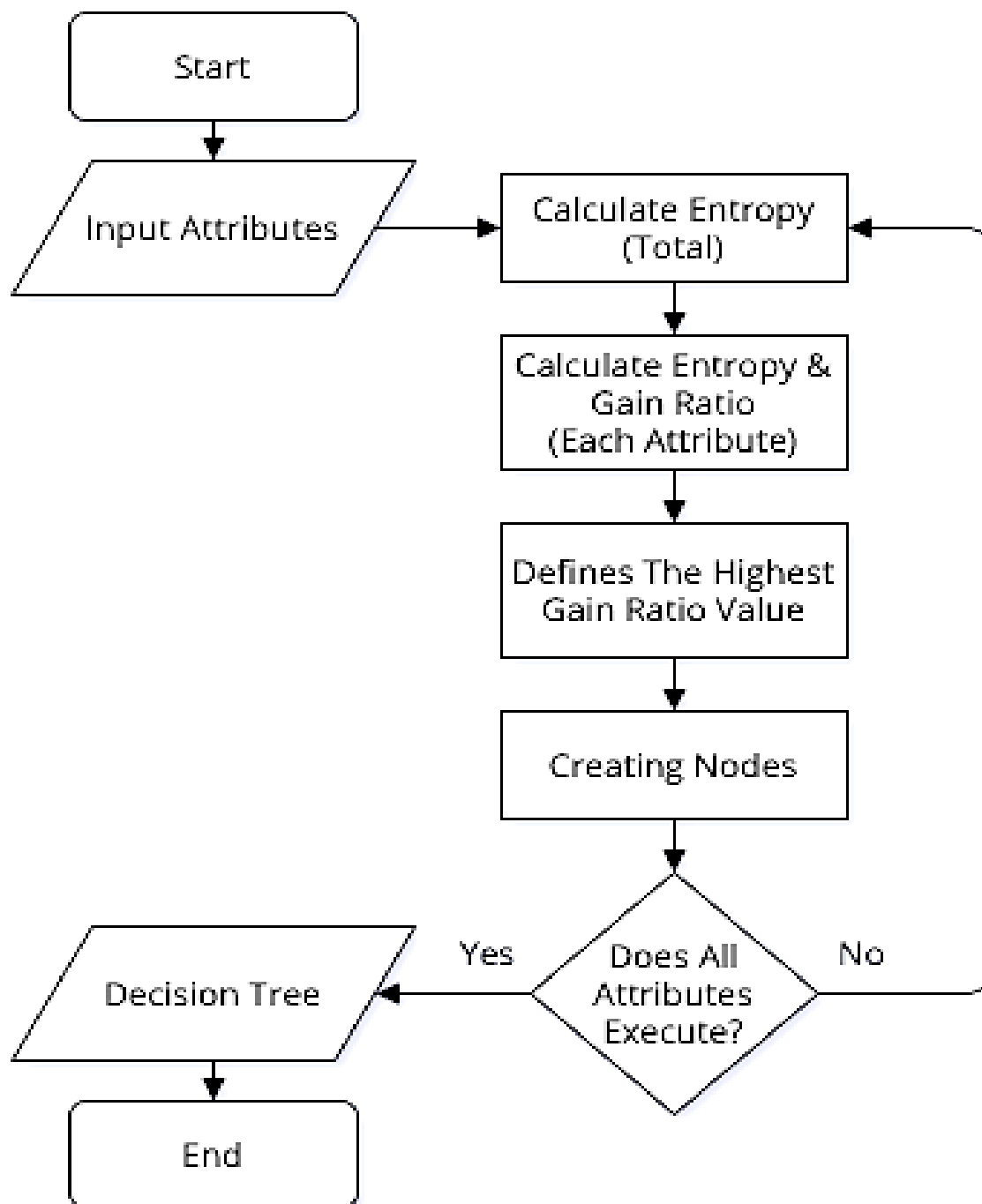


Fig. 3.3: Flowchart of Decision Tree classifier

Advantages of Decision Tree:

1.Easy to understand:

It is simple to understand as it follows the same process which a human follow while making any decision in real-life.

2.Effective to decision -related problem:

It can be very useful for solving decision-related problems. It helps to think about all the possible outcomes for a problem.

3.Requires Little Data Pre-processing:

Unlike some other algorithms, decision trees do not require normalization or scaling of data.

4.Handles Both Numerical and Categorical Data:

Decision trees can be used for both regression and classification tasks, accommodating a variety of data types.

5.Feature Selection:

They perform implicit feature selection, as the most important features are used at the top of the tree.

Disadvantages of Decision Tree:

1.Overfitting:

Decision trees can easily become too complex and overfit the training data, capturing noise rather than the underlying pattern. Pruning methods or setting constraints like maximum depth can help mitigate this.

2.Instability:

Small changes in the data can result in a completely different tree structure, making them unstable. This can be addressed by using ensemble methods like Random Forests.

3.Biased with Imbalanced Data:

If some classes are more frequent than others, decision trees can become biased towards the more frequent classes. Techniques like balanced class weights or sampling methods can help.

Chapter 4: Results

4.1 Introduction:

In this section, we review the results of our classification model evaluations and provide visualizations to better understand their performance. We have analyzed three different machine learning classifiers: K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Decision Tree (DT). We have run 20 simulation for each classifier to make sure our evaluation is thorough.

For each classifier, we closely look at their accuracy to assess their predictive abilities. Additionally, we explore their performance in detail using confusion matrices and classification reports. These metrics help us understand how well the models can correctly classify and distinguish between different classes.

To make this analysis more intuitive, we use heat maps to visualize the confusion matrices. These visual tools give a clearer picture of the models' predictive accuracy and highlight any tendencies to misclassify certain classes. By combining these quantitative metrics with visual aids, we can make more informed decisions about which models are best suited for our dataset and problem.

4.2 Dataset Used:

Differentiated Thyroid Cancer Recurrence dataset which we used is collected from [UCI Machine Learning Repository](#). The dataset has 383 instances and 16 features. It has 2 class labels yes and no. In this dataset each instance represents individual patients[6].

The dataset suffers from class imbalance where the number of recurred cases is less than the number of non-recurred cases. The majority of patients have not experienced a recurrence of thyroid cancer.

The collected patient data included the variables age at diagnosis in years, gender (Female and Male), current smoking status, past smoking history, history of radiation therapy to the head and neck region, thyroid function (classified as euthyroid, clinical or subclinical hypo/hyperthyroidism), presence of goiter (diffuse, single nodular goiter on the left or right lobe, multinodular, or normal), presence of adenopathy on physical examination (no adenopathy, anterior right, anterior left, bilateral, posterior, or extensive locations, pathological subtype of cancer (papillary, micropapillary, follicular, Hurthel cell), focality (unifocal, multifocal), risk assessment according to ATA guidelines (low, intermediate, high), TNM staging (individual T, N, and M scores, and final stage), initial treatment response (excellent, biochemical incomplete, structurally incomplete, indeterminate), and recurrence status that is Whether the cancer has come back, either locally or in distant sites [1].

Dataset:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Age	Gender	Smoking	Hx Smokin	Hx Radioti	Thyroid Function	Physical Examination	Adenopatt	Pathology	Focality	Risk	T	N	M	Stage	Response	Recurred
2	27	F	No	No	No	Euthyroid	Single nodular goiter-left	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Indeterminate	No
3	34	F	No	Yes	No	Euthyroid	Multinodular goiter	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
4	30	F	No	No	No	Euthyroid	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
5	62	F	No	No	No	Euthyroid	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
6	62	F	No	No	No	Euthyroid	Multinodular goiter	No	Micropapillary	Multi-Focal	Low	T1a	N0	M0	I	Excellent	No
7	52	M	Yes	No	No	Euthyroid	Multinodular goiter	No	Micropapillary	Multi-Focal	Low	T1a	N0	M0	I	Indeterminate	No
8	41	F	No	Yes	No	Clinical Hyperthyroidism	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
9	46	F	No	No	No	Euthyroid	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
10	51	F	No	No	No	Euthyroid	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
11	40	F	No	No	No	Euthyroid	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
12	75	F	No	No	No	Euthyroid	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
13	59	F	No	No	No	Euthyroid	Single nodular goiter-left	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
14	49	F	No	No	No	Euthyroid	Multinodular goiter	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
15	50	F	No	No	No	Clinical Hyperthyroidism	Multinodular goiter	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
16	76	F	No	No	No	Clinical Hypothyroidism	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
17	42	F	No	No	No	Euthyroid	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Indeterminate	No
18	40	F	No	Yes	No	Euthyroid	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
19	44	F	No	No	No	Euthyroid	Multinodular goiter	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Indeterminate	No
20	43	F	No	No	No	Euthyroid	Multinodular goiter	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
21	52	F	No	No	No	Euthyroid	Multinodular goiter	No	Micropapillary	Multi-Focal	Low	T1a	N0	M0	I	Indeterminate	No
22	41	F	No	No	No	Euthyroid	Single nodular goiter-left	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No

23	44	F	No	No	No	Euthyroid	Single nodular goiter-left	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
24	36	F	No	No	No	Euthyroid	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
25	70	F	No	No	No	Euthyroid	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
26	60	F	No	No	No	Euthyroid	Multinodular goiter	No	Micropapillary	Multi-Focal	Low	T1a	N0	M0	I	Excellent	No
27	33	F	No	No	No	Euthyroid	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
28	43	M	No	No	No	Subclinical Hyperthyroid	Single nodular goiter-left	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
29	26	M	No	No	No	Euthyroid	Single nodular goiter-left	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
30	41	F	No	No	No	Euthyroid	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
31	37	F	No	Yes	No	Subclinical Hypothyroid	Normal	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
32	37	F	No	No	No	Euthyroid	Single nodular goiter-left	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
33	30	F	No	No	No	Clinical Hyperthyroidism	Normal	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
34	36	F	No	No	No	Euthyroid	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
35	55	F	No	No	No	Euthyroid	Multinodular goiter	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
36	52	M	Yes	No	No	Euthyroid	Single nodular goiter-left	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
37	37	F	No	No	No	Subclinical Hypothyroid	Single nodular goiter-left	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
38	31	F	No	No	No	Clinical Hyperthyroidism	Diffuse goiter	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
39	43	M	No	No	No	Euthyroid	Multinodular goiter	No	Micropapillary	Multi-Focal	Low	T1a	N0	M0	I	Excellent	No
40	40	F	No	No	No	Euthyroid	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
41	34	F	No	No	No	Euthyroid	Diffuse goiter	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
42	51	F	No	No	No	Euthyroid	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
43	45	F	No	No	No	Euthyroid	Single nodular goiter-left	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
44	20	F	No	No	No	Euthyroid	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
45	38	F	No	No	No	Euthyroid	Multinodular goiter	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
46	38	F	No	No	No	Clinical Hyperthyroidism	Diffuse goiter	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
47	33	F	No	No	No	Euthyroid	Multinodular goiter	No	Micropapillary	Uni-Focal	Low	T1a	N0	M0	I	Excellent	No
48	31	F	No	No	No	Clinical Hypothyroidism	Multinodular goiter	No	Micropapillary	Uni-Focal	Intermediate	T1a	N0	M0	I	Excellent	No
49	21	F	No	No	No	Euthyroid	Multinodular goiter	No	Micropapillary	Multi-Focal	Intermediate	T1a	N0	M0	I	Excellent	No

48	31 F	No	No	No	Clinical Hypothyroidism	Multinodular goiter	No	Micropapillary	Uni-Focal	Intermediate	T1a	N0	M0	I	Excellent	No
49	31 F	No	No	No	Euthyroid	Multinodular goiter	Right	Papillary	Multi-Focal	Intermediate	T1a	N1b	M0	I	Excellent	No
50	26 F	No	No	No	Euthyroid	Normal	Extensive	Papillary	Uni-Focal	Intermediate	T1a	N1b	M0	I	Structural Incon	Yes
51	29 F	No	No	No	Euthyroid	Multinodular goiter	No	Follicular	Uni-Focal	Low	T1b	N0	M0	I	Excellent	No
52	43 F	No	No	No	Euthyroid	Single nodular goiter-right	No	Micropapillary	Uni-Focal	Low	T1b	N0	M0	I	Excellent	No
53	30 F	No	No	No	Subclinical Hyperthyroidism	Multinodular goiter	No	Papillary	Multi-Focal	Low	T1b	N0	M0	I	Excellent	No
54	25 F	No	No	No	Euthyroid	Multinodular goiter	No	Papillary	Multi-Focal	Low	T1b	N0	M0	I	Excellent	No
55	27 F	No	No	No	Euthyroid	Multinodular goiter	No	Papillary	Uni-Focal	Low	T1b	N0	M0	I	Excellent	No
56	25 F	No	No	No	Euthyroid	Single nodular goiter-left	No	Papillary	Uni-Focal	Low	T1b	N0	M0	I	Excellent	No
57	21 F	No	No	No	Euthyroid	Single nodular goiter-right	No	Papillary	Uni-Focal	Low	T1b	N0	M0	I	Indeterminate	No
58	43 F	No	Yes	No	Euthyroid	Single nodular goiter-right	No	Papillary	Uni-Focal	Low	T1b	N0	M0	I	Excellent	No
59	23 F	No	No	No	Euthyroid	Single nodular goiter-left	No	Papillary	Uni-Focal	Low	T1b	N0	M0	I	Excellent	No
60	23 F	No	No	No	Clinical Hyperthyroidism	Single nodular goiter-left	No	Papillary	Uni-Focal	Low	T1b	N0	M0	I	Excellent	No
61	43 F	No	No	No	Euthyroid	Single nodular goiter-left	No	Papillary	Uni-Focal	Low	T1b	N0	M0	I	Indeterminate	No
62	24 M	No	No	No	Euthyroid	Single nodular goiter-left	No	Papillary	Uni-Focal	Low	T1b	N0	M0	I	Excellent	No
63	35 F	No	No	No	Euthyroid	Single nodular goiter-right	No	Papillary	Uni-Focal	Low	T1b	N0	M0	I	Excellent	No
64	54 F	No	No	No	Euthyroid	Single nodular goiter-left	No	Papillary	Uni-Focal	Low	T1b	N0	M0	I	Excellent	No
65	54 F	No	No	No	Euthyroid	Multinodular goiter	No	Papillary	Uni-Focal	Low	T1b	N0	M0	I	Indeterminate	No
66	22 F	No	No	No	Subclinical Hyperthyroidism	Single nodular goiter-right	No	Papillary	Uni-Focal	Low	T1b	N0	M0	I	Excellent	No
67	38 F	No	No	No	Euthyroid	Multinodular goiter	No	Papillary	Multi-Focal	Low	T1b	N0	M0	I	Excellent	No
68	35 F	No	No	No	Euthyroid	Single nodular goiter-right	No	Papillary	Uni-Focal	Low	T1b	N0	M0	I	Excellent	No
69	51 F	No	No	No	Euthyroid	Single nodular goiter-left	No	Papillary	Uni-Focal	Low	T1b	N0	M0	I	Excellent	No
70	22 M	No	No	No	Euthyroid	Multinodular goiter	No	Papillary	Multi-Focal	Low	T1b	N0	M0	I	Excellent	No
71	51 F	No	No	No	Euthyroid	Single nodular goiter-left	No	Papillary	Uni-Focal	Low	T1b	N0	M0	I	Excellent	No
72	40 F	No	No	No	Euthyroid	Single nodular goiter-right	No	Papillary	Uni-Focal	Low	T1b	N0	M0	I	Excellent	No
73	69 F	No	No	No	Euthyroid	Single nodular goiter-left	No	Papillary	Uni-Focal	Low	T1b	N0	M0	I	Excellent	No
74	21 F	No	No	No	Euthyroid	Single nodular goiter-right	No	Papillary	Uni-Focal	Low	T1b	N0	M0	I	Excellent	No

73	69 F	No	No	No	Euthyroid	Single nodular goiter-left	No	Papillary	Uni-Focal	Low	T1b	N0	M0	I	Excellent	No
74	31 F	No	No	No	Euthyroid	Single nodular goiter-right	No	Papillary	Uni-Focal	Low	T1b	N0	M0	I	Excellent	No
75	29 F	No	No	No	Euthyroid	Single nodular goiter-right	No	Papillary	Uni-Focal	Low	T1b	N0	M0	I	Excellent	No
76	30 F	No	No	No	Euthyroid	Single nodular goiter-right	No	Papillary	Uni-Focal	Low	T1b	N0	M0	I	Excellent	No
77	28 F	No	No	No	Euthyroid	Multinodular goiter	No	Papillary	Uni-Focal	Low	T1b	N0	M0	I	Excellent	No
78	22 F	No	No	No	Clinical Hypothyroidism	Single nodular goiter-right	No	Papillary	Uni-Focal	Low	T1b	N0	M0	I	Excellent	No
79	29 F	No	No	No	Euthyroid	Single nodular goiter-right	No	Papillary	Uni-Focal	Low	T1b	N0	M0	I	Excellent	No
80	35 F	No	No	No	Euthyroid	Single nodular goiter-left	No	Papillary	Uni-Focal	Low	T1b	N0	M0	I	Excellent	No
81	50 F	No	No	No	Euthyroid	Multinodular goiter	No	Papillary	Multi-Focal	Low	T1b	N0	M0	I	Excellent	No
82	27 F	No	No	No	Clinical Hyperthyroidism	Diffuse goiter	No	Papillary	Uni-Focal	Low	T1b	N0	M0	I	Excellent	No
83	17 F	No	Yes	No	Euthyroid	Single nodular goiter-right	No	Papillary	Uni-Focal	Low	T1b	N0	M0	I	Excellent	No
84	27 F	No	No	No	Subclinical Hyperthyroidism	Single nodular goiter-right	No	Papillary	Uni-Focal	Low	T1b	N0	M0	I	Excellent	No
85	40 F	No	No	No	Euthyroid	Single nodular goiter-right	No	Papillary	Uni-Focal	Intermediate	T1b	N0	M0	I	Excellent	No
86	33 F	No	No	No	Euthyroid	Single nodular goiter-left	No	Papillary	Uni-Focal	Intermediate	T1b	N1b	M0	I	Indeterminate	No
87	25 F	No	Yes	Yes	Euthyroid	Single nodular goiter-right	No	Papillary	Uni-Focal	Low	T1b	N0	M0	I	Excellent	No
88	73 F	No	No	No	Euthyroid	Single nodular goiter-right	No	Papillary	Uni-Focal	Low	T1b	N0	M0	I	Excellent	No
89	36 F	No	No	No	Euthyroid	Multinodular goiter	No	Papillary	Multi-Focal	Low	T1b	N0	M0	I	Indeterminate	Yes
90	35 F	No	No	No	Euthyroid	Multinodular goiter	Right	Papillary	Multi-Focal	Intermediate	T1b	N1b	M0	I	Structural Incon	Yes
91	31 M	Yes	No	No	Euthyroid	Multinodular goiter	Right	Papillary	Multi-Focal	Intermediate	T1b	N1b	M0	I	Structural Incon	Yes
92	18 F	No	No	No	Euthyroid	Single nodular goiter-right	Right	Papillary	Uni-Focal	Intermediate	T1b	N1b	M0	I	Structural Incon	Yes
93	62 F	No	No	No	Euthyroid	Single nodular goiter-right	No	Follicular	Multi-Focal	Intermediate	T1b	N0	M0	I	Structural Incon	Yes
94	62 F	No	No	No	Euthyroid	Single nodular goiter-right	No	Hurthel cell	Uni-Focal	Low	T2	N0	M0	I	Excellent	No
95	39 F	No	No	No	Euthyroid	Diffuse goiter	No	Hurthel cell	Uni-Focal	Low	T2	N0	M0	I	Indeterminate	No
96	37 F	No	No	No	Euthyroid	Single nodular goiter-left	No	Hurthel cell	Multi-Focal	Low	T2	N0	M0	I	Excellent	No
97	26 F	No	Yes	No	Subclinical Hyperthyroidism	Single nodular goiter-right	No	Hurthel cell	Uni-Focal	Low	T2	N0	M0	I	Excellent	No
98	31 F	No	No	No	Euthyroid	Single nodular goiter-left	No	Hurthel cell	Uni-Focal	Low	T2	N0	M0	I	Indeterminate	No
99	24 F	No	No	No	Euthyroid	Single nodular goiter-right	No	Follicular	Uni-Focal	Low	T2	N0	M0	I	Excellent	No

4.3 Parameter Used:

1.K-Nearest Neighbours (KNN): Euclidean distance was chosen as the distance metric for classification. The hyper parameter 'n_neighbors' was tuned to find the optimal value. The value n_neighbors for which we get best accuracy is 7.

2.Support Vector Machines (SVM): SVM hyper-parameter tuning was performed with different kernel functions (e.g. linear, radial basis function, polynomial, sigmoid) using grid search cv. Regularization parameter 'C' was tuned for optimal performance. The best result is obtained for linear kernel and C value 1.

3.Decision Tree: For parameter tuning we have considered max_depth values as. The parameter was optimized through grid search and cross-validation and we obtained best results for max_depth value 3.

4.4 Performance Metrics:

Accuracy: Accuracy is used to measure the performance of the model. It is the ratio of Total correct instances to the total instances.

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

Where,

True Positive(TP): The prediction outcome is true, and it is true in reality, also.

True Negative(TN): In this case, the prediction outcome is false, and it is false in reality, also.

False Positive(FP): The prediction outcomes are true, but they are false in actuality.

False Negative(FN): In this case, predictions are false, and they are true in actuality.

Precision: It is a measure of how accurate a model's positive predictions are. It is defined as the ratio of true positive predictions to the total number of positive predictions made by the model.

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall: It is the ratio of the number of true positive (TP) instances to the sum of true positive and false negative (FN).

$$\text{Recall} = \frac{TP}{TP+FN}$$

F1-Score: This is used to evaluate the overall performance of a classification model. It is the harmonic mean of precision and recall.

$$\text{F1-Score} = 2 * (\text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall}))$$

4.5 Pre-Processing:

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. This refers to the transformations applied to our data before feeding it to the algorithm.

The dataset which we have used has no missing values. Most of the features of the dataset are categorical hence the dataset needs to be pre-processed for working. We have used label encoding technique for transforming data into numeric forms and for scaling the data we used MinMaxScaler.

4.6 Results and Analysis:

Standard deviation: Standard deviation is a way to measure how spread out the values in a set of data are. It shows how much the numbers in the data set differ from the average (mean) value.

The Dataset is split into 90:10, 80:20 and 70:30 where 90%, 80%, and 70% of data were used for training purposes while the remaining 10%, 20%, and 30% were used for testing purposes. Each of the classifier runs 20 times and the results obtained including maximum accuracy, average accuracy and standard deviation are shown for all three classifier (KNN, SVM and Decision Tree) are shown in the Table 4.1- 4.3.

Confusion Matrix: The confusion matrix helps assess classification model performance in machine learning by comparing predicted values against actual values for a dataset. This is a very popular measure used while solving classification problems. Confusion matrix can be applied to binary classification as well as for multiclass classification problems.

The output “TN” stands for True Negative which shows the number of negative examples classified accurately. Similarly, “TP” stands for True Positive which indicates the number of positive examples classified accurately. The term “FP” shows False Positive value, i.e., the number of actual negative examples classified as positive; and “FN” means a False Negative value which is the number of actual positive examples classified as negative.

Confusion matrix obtained for highest accuracy after 20 simulation for each of the three different classifiers KNN, SVM and Decision Tree have been put in fig 4.4- 4.12.

Test Size	Maximum accuracy	Mean accuracy	Standard deviation
0.1	100	91.94	4.13
0.2	96.10	91.77	2.42
0.3	95.65	91.47	2.27

Table 4.1: Results obtained for 20 simulation using KNN classifier

Test Size	Maximum accuracy	Mean accuracy	Standard deviation
0.1	100	95.38	3.85
0.2	97.40	95.06	1.40
0.3	98.26	94.65	1.66

Table 4.2: Results obtained for 20 simulation using SVM classifier

Test Size	Maximum accuracy	Mean accuracy	Standard deviation
0.1	100	95.26	3.17
0.2	98.70	95.26	1.85
0.3	96.52	94.35	1.71

Table 4.3: Results obtained for 20 simulation using Decision Tree Classifier

Confusion Matrix:

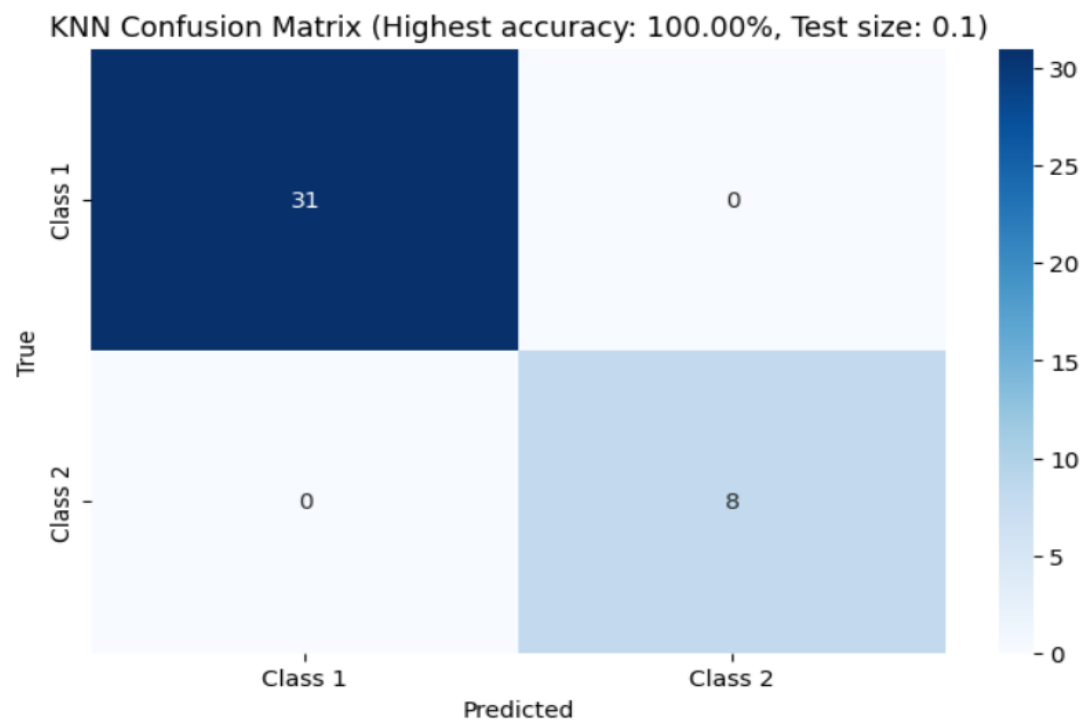


Fig.4.4: KNN confusion matrix for 0.1 Test size

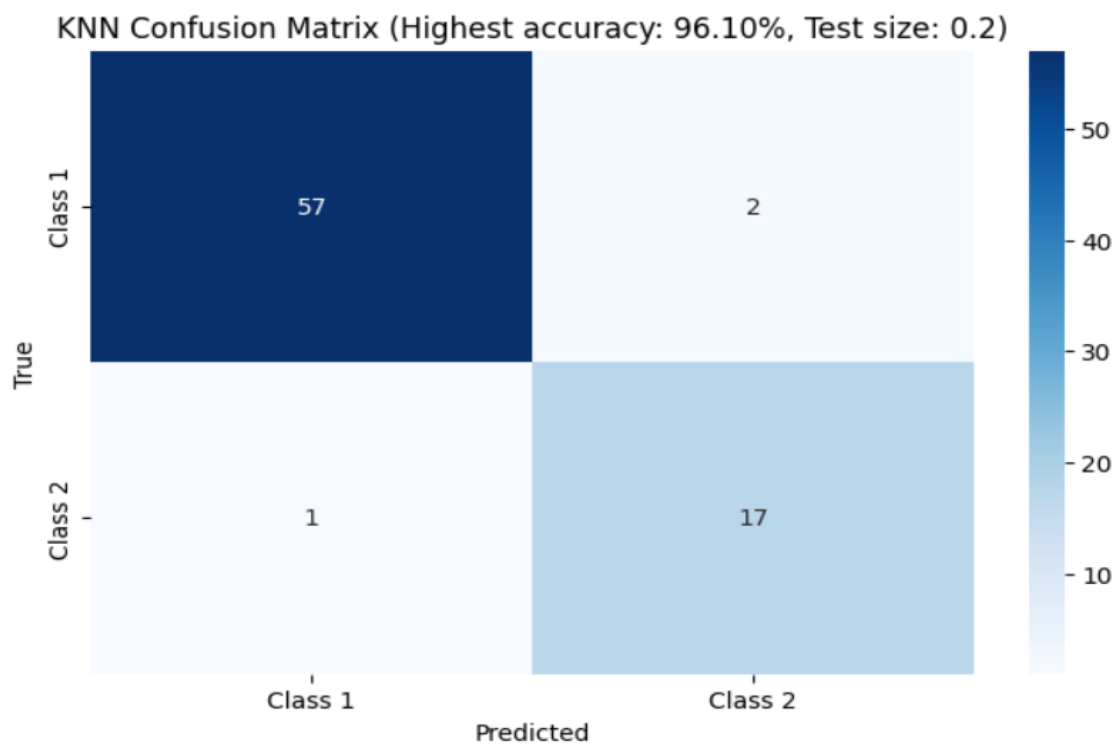


Fig.4.5: KNN confusion matrix for 0.2 Test size

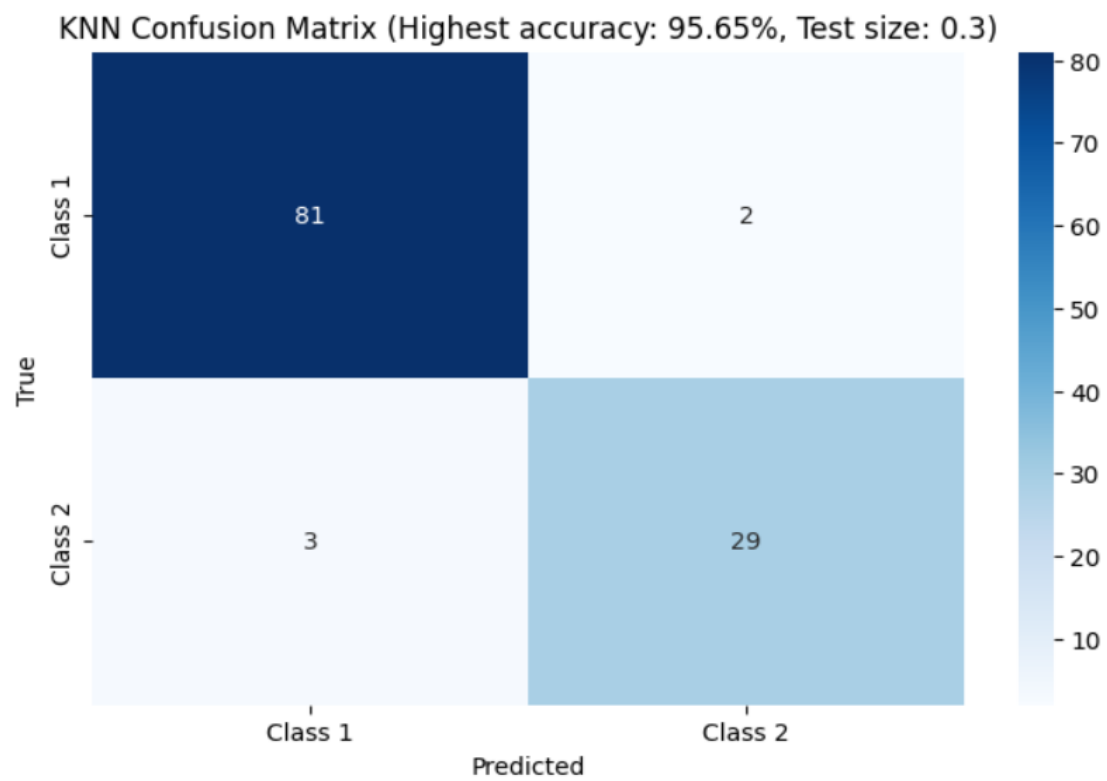


Fig.4.6: KNN confusion matrix for 0.3 Test size

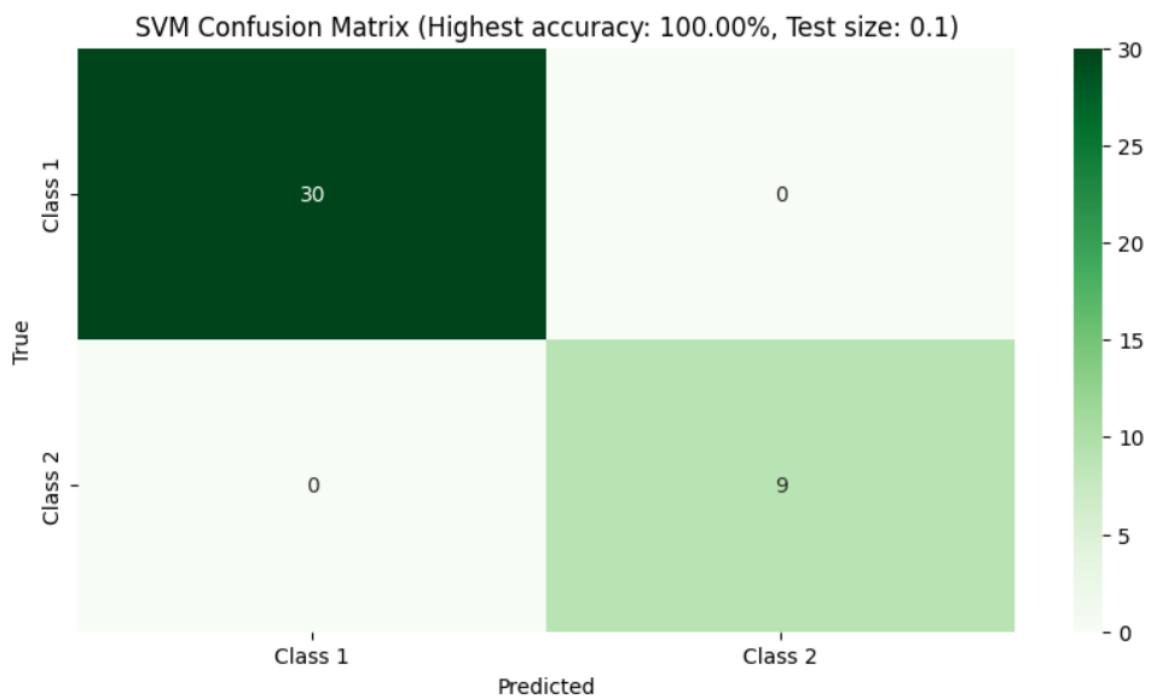


Fig.4.7: SVM confusion matrix for 0.1 Test size

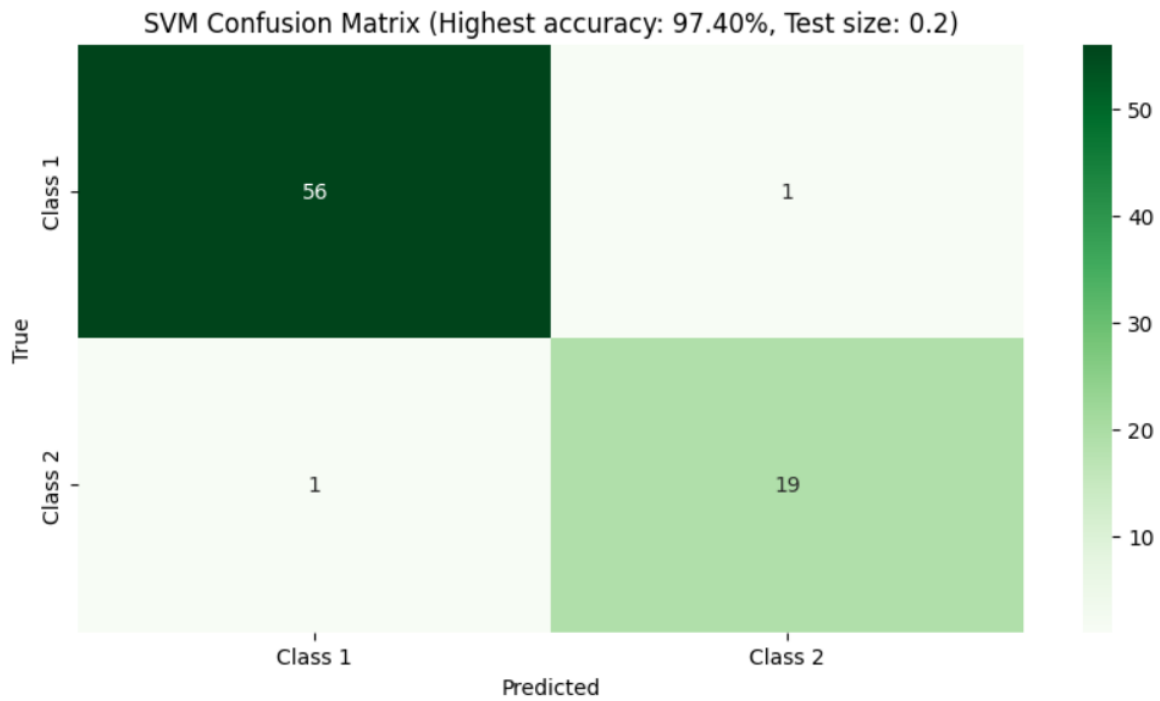


Fig.4.8: SVM confusion matrix for 0.2 Test size

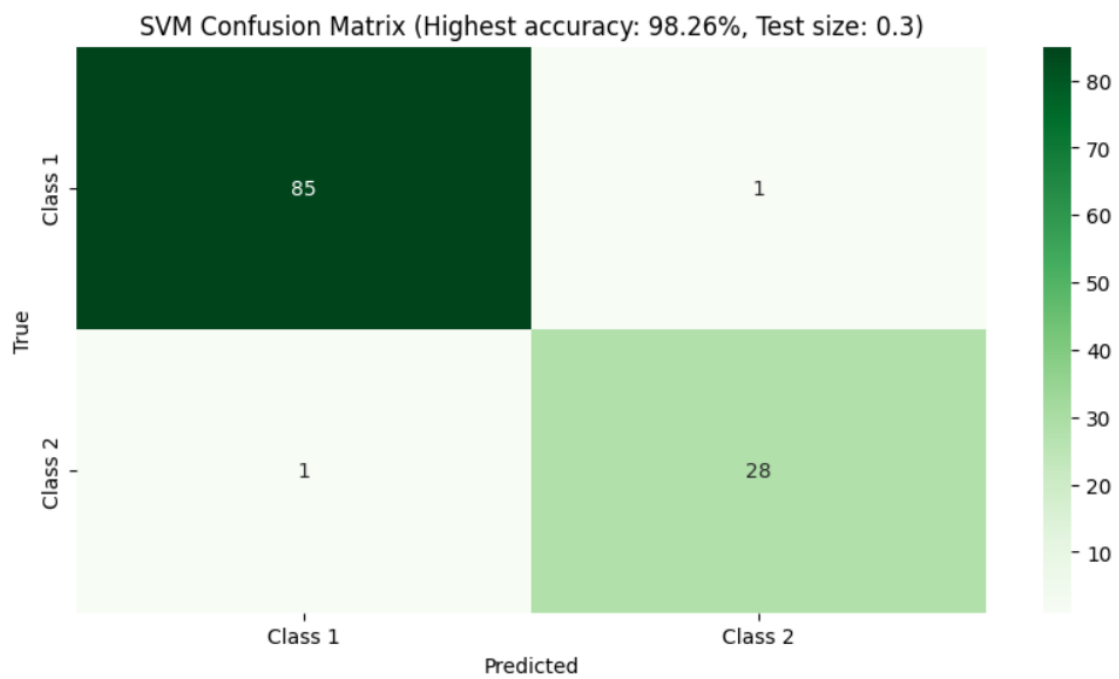


Fig.4.9: SVM confusion matrix for 0.3 Test size

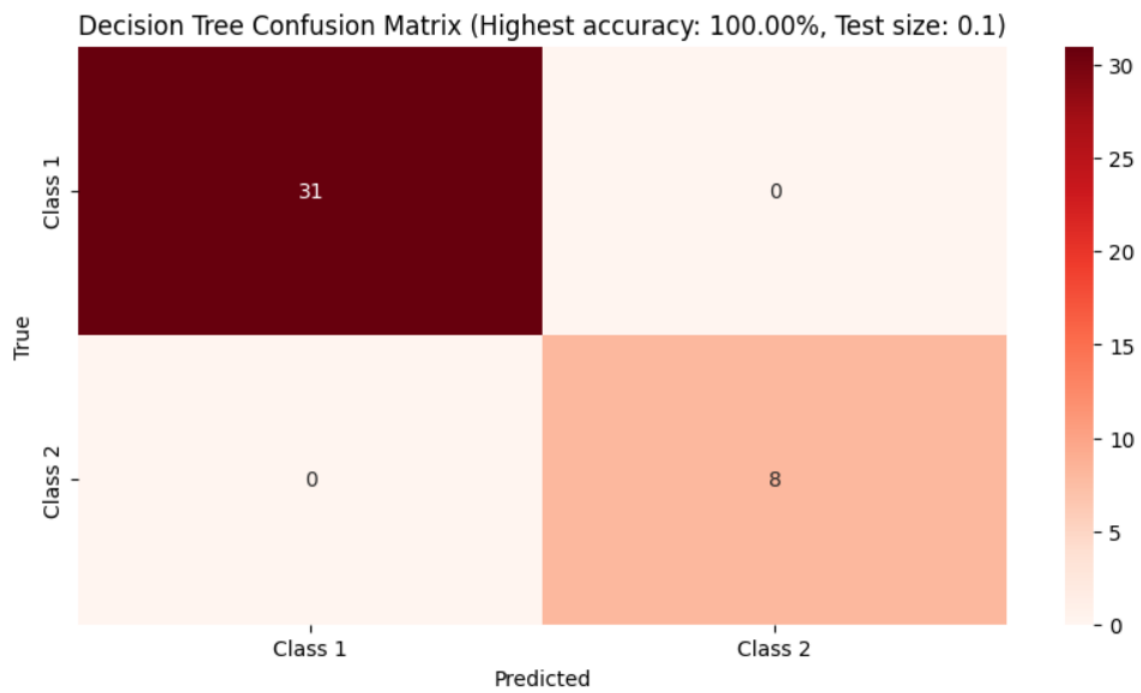


Fig.4.10: Decision Tree confusion matrix for 0.1 Test size

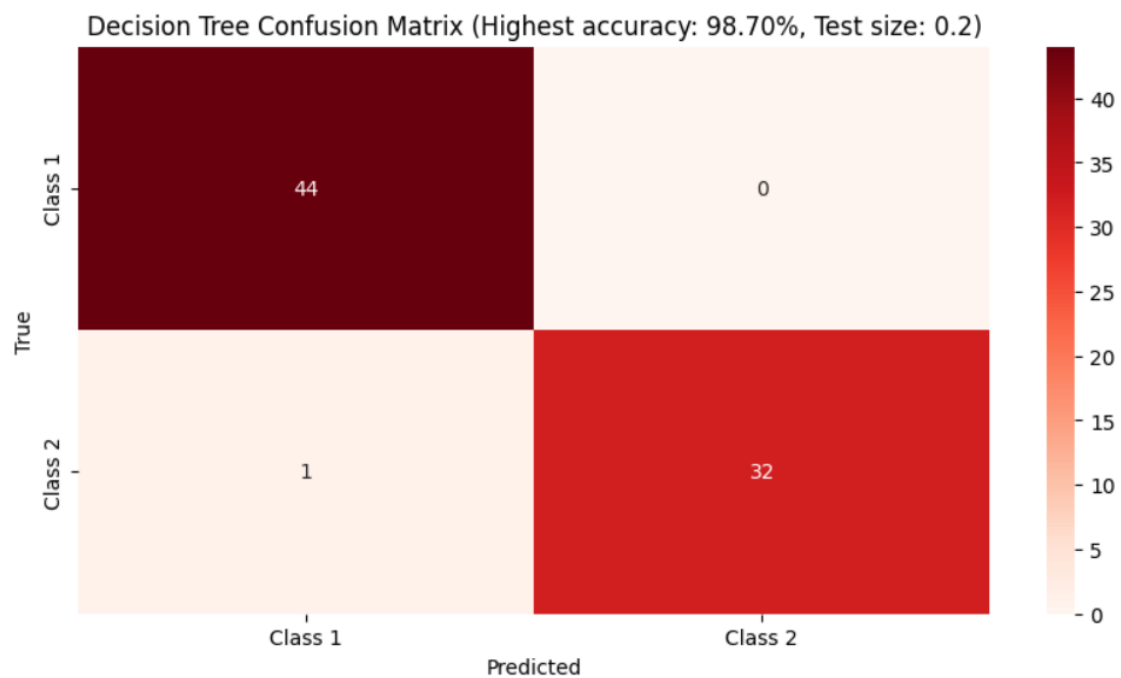


Fig.4.11: Decision Tree confusion matrix for 0.2 Test size

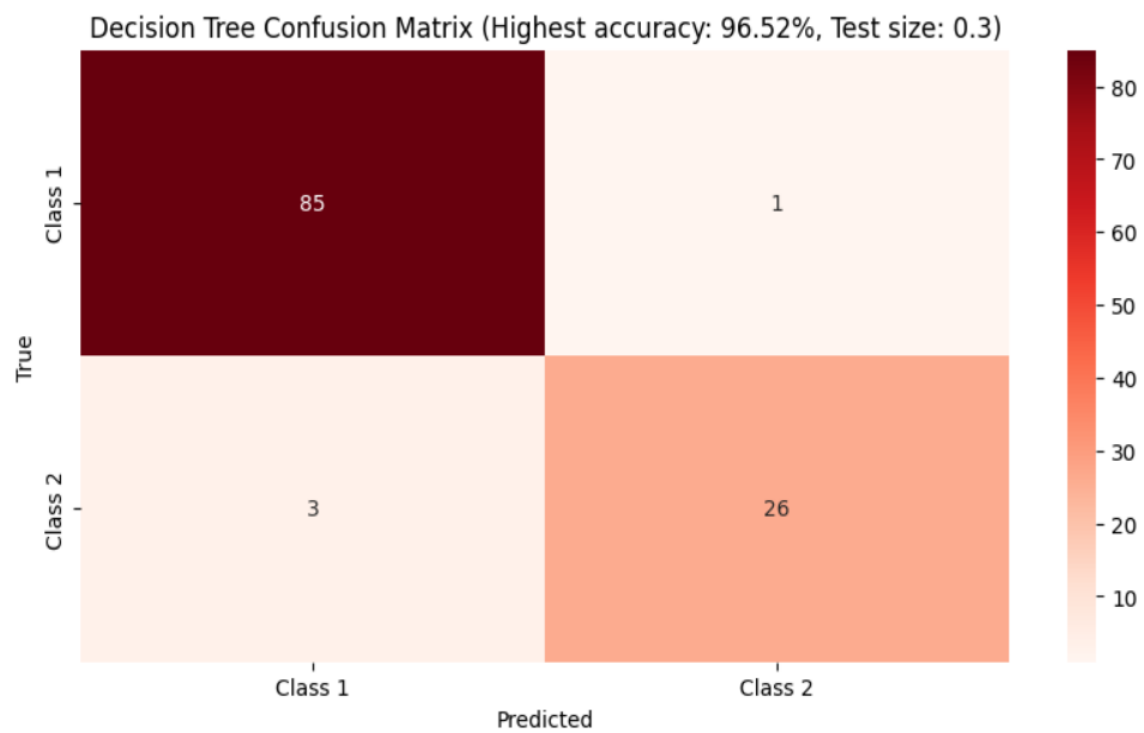


Fig.4.12: Decision Tree confusion matrix for 0.3 Test size

Classification Report:

Test size	Class Labels	precision	recall	f1-score	support
0.1	0	0.97	1.00	0.98	28
	1	1.00	0.91	0.95	11
	accuracy			0.97	39
	macro avg	0.98	0.95	0.97	39
	weighted avg	0.98	0.97	0.97	39
0.2	0	0.93	0.98	0.96	56
	1	0.94	0.81	0.87	21
	accuracy			0.94	77
	macro avg	0.94	0.90	0.91	77
	weighted avg	0.94	0.94	0.93	77
0.3	0	0.91	0.97	0.94	87
	1	0.87	0.71	0.78	28
	accuracy			0.90	115
	macro avg	0.89	0.84	0.86	115
	weighted avg	0.90	0.90	0.90	115

Table 4.13: Classification report of KNN

Test size	Class Labels	precision	recall	f1-score	support
0.1	0	0.83	0.95	0.89	21
	1	0.93	0.78	0.85	18
	accuracy			0.87	39
	macro avg	0.88	0.87	0.87	39
	weighted avg	0.88	0.87	0.87	39
0.2	0	0.95	0.98	0.97	58
	1	0.94	0.84	0.89	19
	accuracy			0.95	77
	macro avg	0.95	0.91	0.93	77
	weighted avg	0.95	0.95	0.95	77
0.3	0	0.99	0.99	0.99	86
	1	0.97	0.97	0.97	29
	accuracy			0.98	115
	macro avg	0.98	0.98	0.98	115
	weighted avg	0.98	0.98	0.98	115

Table 4.14: Classification report of SVM

Test size	Class label	precision	recall	f1-score	support
0.1	0	1.00	1.00	1.00	28
	1	1.00	1.00	1.00	11
accuracy				1.00	39
macro avg		1.00	1.00	1.00	39
weighted avg		1.00	1.00	1.00	39
0.2	0	0.97	1.00	0.98	60
	1	1.00	0.88	0.94	17
accuracy				0.97	77
macro avg		0.98	0.94	0.96	77
weighted avg		0.97	0.97	0.97	77
0.3	0	0.98	0.98	0.98	84
	1	0.94	0.94	0.94	31
accuracy				0.97	115
macro avg		0.96	0.96	0.96	115
weighted avg		0.97	0.97	0.97	115

Table 4.15: Classification report of Decision Tree

Chapter 5: Conclusion and Future Scope

In conclusion, we have shown accuracy, mean accuracy, standard deviation for all the classifier KNN, SVM and Decision Tree. The choice of the test size impacts model performance, balancing accuracy and robustness is crucial.

The projects shows that traditional machine learning models are effective in predicting the risk of recurrence and categorizing patients with differentiated thyroid cancers. Furthermore, it emphasizes that machine learning techniques can improve current rule-based clinical tools, like the ATA risk prediction model, by making them more precise and easier to understand.

Future directions: While our project achieved promising results, there is room for improvement. Future research avenues include exploring data balancing methods and delving into deep learning techniques.

In data science, various techniques can enhance the performance of machine learning models. Two such techniques are feature importance and dimensionality reduction. These methods help improve model accuracy and efficiency. Additionally, in machine learning and deep learning research, having access to large datasets can significantly boost the predictive quality of models.

References:

[1] S. Borzooei, G. Brganti, M. Golparian, J. R. Lechein and A. Tarokhian, “Machine learning for risk stratification of thyroid cancer patients: a 15-year cohort study”, *European Archives of Oto-Rhino-Laryngology*, vol.281, pp. 2095–2104, 2024.

<https://doi.org/10.1007/s00405-023-08299-w>

[2] K. E. Setiawan, “Predicting recurrence in differentiated thyroid cancer: a comparative analysis of various machine learning models including ensemble methods with chi-squared feature selection”, *Communications in Mathematical Biology and Neuroscience*, vol. 2024, 2024.

[3] Y.H. Sharifovna, “Thyroid cancer diagnostics, classification, staging”, *Ijtimoiy Fanlarda Innovasiya Onlayn Ilmiy Jurnal*, vol. 1, pp. 63-69, 2024.

[4] Y. Cao, X. Zhong, W. Diao, et al., “Radiomics in differentiated thyroid cancer and nodules: explorations, application, and limitations”, *Cancers* 13, vol. 2436, 2021.

<https://doi.org/10.3390/cancers13102436>.

[5] J. Han and M. Kamber, “Data Mining Concepts and Techniques ”, Morgan Kaufmann Publishers,2006.

[6] S. Borzooei and A. Tarokhian,”Differentiated Thyroid Cancer Recurrence”, 2023.

<https://doi.org/10.24432/C5632J>