Final Project Report on

Cross Validation of Different Classifiers

Submitted by

GROUP E

C201221	Fahmida Akter
C201224	Sabrina Mostary
C201242	Farida Nusrat
C201286	Sabrina Jahan

Course Code: CSE-4878

Course Title: Machine Learning and Data Mining Lab

Submitted to

Mrs. Subrina Akter

Assistant Professor

Dept. of CSE, IIUC

Date of Submission: 18/05/2024

Department of Computer Science and Engineering

International Islamic University Chittagong

TABLE OF CONTENTS

INTRODUCTION

In today's time, it has become important for a machine to give the fastest and reliable results to cope up with the speed of the everchanging world. So, machines are given a machine learning model to look upon and replicate the results. We can't just fit the model to our training data and expect it to perform correctly with real data it's never seen before. It needs some assurance that the model has correctly identified the majority of the patterns in the data and is not very sensitive to noise, or that it has low bias and variance.

Cross-validation is an approach for estimating machine learning model performance (or accuracy) through the expertise of statistics. It is a technique in which we train our model using a subset of the data set and then assess it using the other portion.

Our primary goal is for the model to perform well on real-world data yet, while the training dataset is also real-world data, it only represents a small portion of all the data points available. That is why we use unseen data or what is called test data to determine a model's true ability of prediction. In our project we applied a total of five classifiers (Decision Tree, Support Vector Machine, K-nearest neighbors, Artificial Neural Network) on our Thyroid Disease dataset and to determine the performance of each classifier we used Stratified k fold validation. Also, four common clustering algorithms (K-means, K-medoids, DBSCAN, Hierarchical) were used to convert the original dataset to a clustered supervised one and the same classifier and cross validation process was followed. In the end, we compared the performance of each classifier on different datasets.

OBJECTIVE

When we train a model on the training set, it tends to overfit most of the time, thus we utilize regularization approaches to avoid this. But if we only have a few training instances, we must be cautious while lowering the number of training samples and conserving them for testing. The easiest method to enhance the system's performance without sacrificing too much is to verify it using a tiny portion of the training data, since this will give us an indication of the model's capacity to predict unknown data. That is the main purpose of cross validation, enhancing performance of models without losing much data. Clustering is used to separate data with same feature or similar relation. It helps to divide data in different groups of clusters where each cluster represent a certain characteristic. That is why, we also used clustered data to determine what impact it has when applied different classification model. So that we could determine which clustering algorithm better suited our dataset.

DATA COLLECTION

The dataset used for the project is "**Thyroid Disease Analysis"** dataset which captures various attributes related to thyroid conditions for medical diagnosis. It includes demographic information such as age and sex. Medical history features encompass intake of thyroxine, antithyroid medications, and past surgeries. Patients' current health status regarding sickness, pregnancy, presence of goiter, tumor, or hypopituitary conditions is recorded. Additionally, it notes whether patients suspect hypothyroidism or hyperthyroidism. Laboratory results like TSH, T3, TT4, T4U, and FTI levels are included if measured. Binary classification denotes the presence or absence of hyperthyroidism. Referral sources are indicated as well. This dataset is collected from Kaggle.com, owner named Prakhar Kapoor. The source link of the dataset and overview is given below:

Source: https://www.kaggle.com/datasets/kapoorprakhar/thyroid-disease-patient-dataset

Overview:

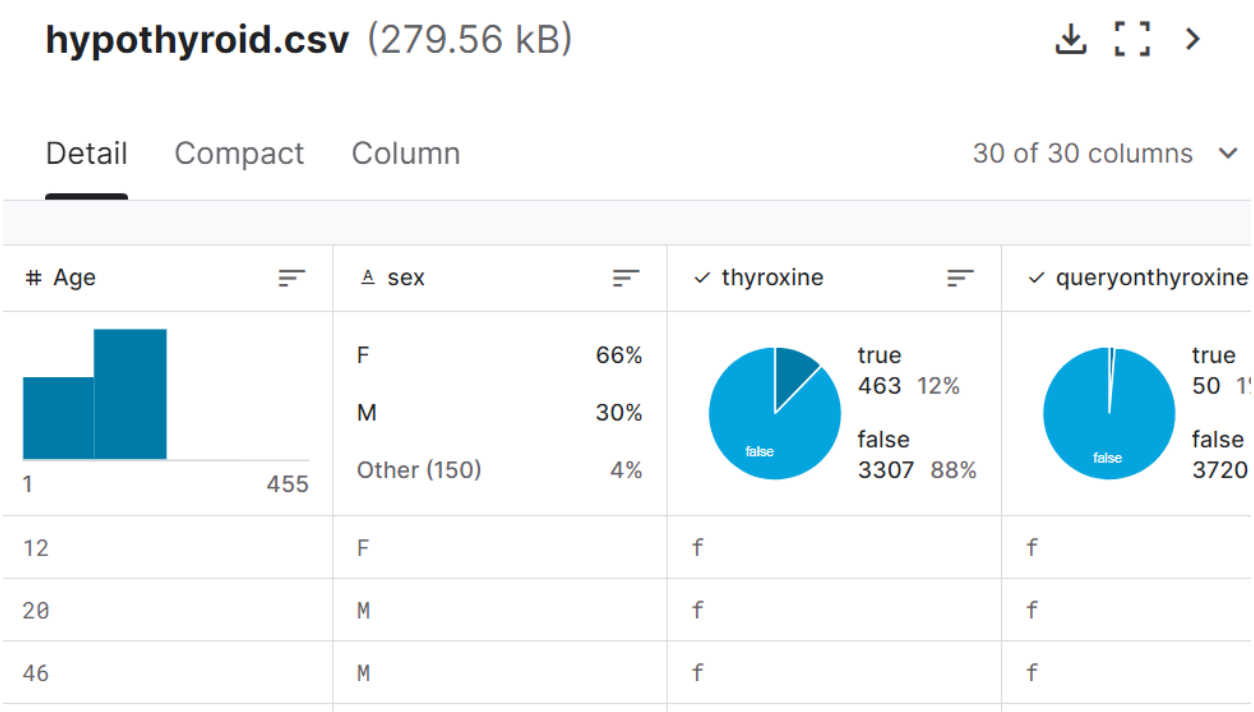


Figure 1

DATA ANALYSIS

The dataset has a total of 3770 observations and 30 attributes which means it describes the condition of 3770 patients giving 30 features along with the result which defines whether a patient has thyroid disease or not. To give more detailed information we used the tables below:

Feature Analysis Table

NO	Columns	Data type	Description	Unique Value
1	age	int64	age of the patient.	[12, 20, 46, 70, 18, 59, 80, 66..]
2	sex	object	sex patient identifies.	[F, M, nan]
3	on_thyroxine	object	whether patient is on thyroxine.	[f, t]
4	query_on_thyroxine	object	unique id of the patient.	[f, t]
5	on_antithyroid_meds	object	whether patient is on antithyroid meds.	[f, t]
6	sick	object	whether patient is sick .	[f, t]
7	pregnant	object	whether patient is pregnant.	[f, t]
8	thyroid_surgery	object	whether patient has undergone thyroid surgery .	[f, t]
9	I131_treatment	object	whether patient is undergoing I131 treatment .	[f, t]
10	query_hypothyroid	object	whether patient believes they have hypothyroid.	[f, t]
11	query_hyperthyroid	object	whether patient believes they have hyperthyroid.	[f, t]
12	lithium	object	whether patient has lithium.	[f, t]
13	goitre	object	whether patient has goitre.	[f, t]
14	tumor	object	whether patient has tumor.	[f, t]
15	hypopituitary	object	whether patient has hyperpituitary gland.	[f, t]
16	psych	object	whether patient has psych .	[f, t]
17	TSH_measured	object	whether TSH was measured in the blood .	[f, t]
18	TSH	float64	TSH level in blood from lab work .	[1.3, 4.1, 0.98, 0.16, 0.72, 0.03..]
19	T3_measured	object	whether T3 was measured in the blood.	[f, t]
20	T3	float64	T3 level in blood from lab work .	[2.5, 2.0, 0.0, 1.9, 1.2, 0.6…]
21	TT4_measured	object	whether TT4 was measured in the blood.	[f, t]
22	TT4	float64	TT4 level in blood from lab work .	[125.0, 102.0, 109.0…]
23	T4U_measured	object	whether T4U was measured in the blood .	[f, t]
24	T4U	float64	T4U level in blood from lab work .	[1.14, 0.0, 0.91, 0.87…]
25	FTI_measured	object	whether FTI was measured in the blood.	[f, t]
26	FTI	float64	FTI level in blood from lab work.	[109.0, 0.0, 120.0, 70.0, 141..]
27	TBG_measured	object	whether TBG was measured in the blood.	[f]
28	TBG	float64	TBG level in blood from lab work.	Nan
29	referral_source	object		[SVHC, other, STMW, SVHD]
30	Result	int64	hyperthyroidism medical diagnosis.	[P, N]

Table 1

Characteristic Table

Dataset Characteristics:	Attribute Characteristics:	Number of Instances:	Number of Attributes:
Multivariate, Imbalanced	Numerical & Categorical	3770	30

Table 2

Distribution of Result Class:

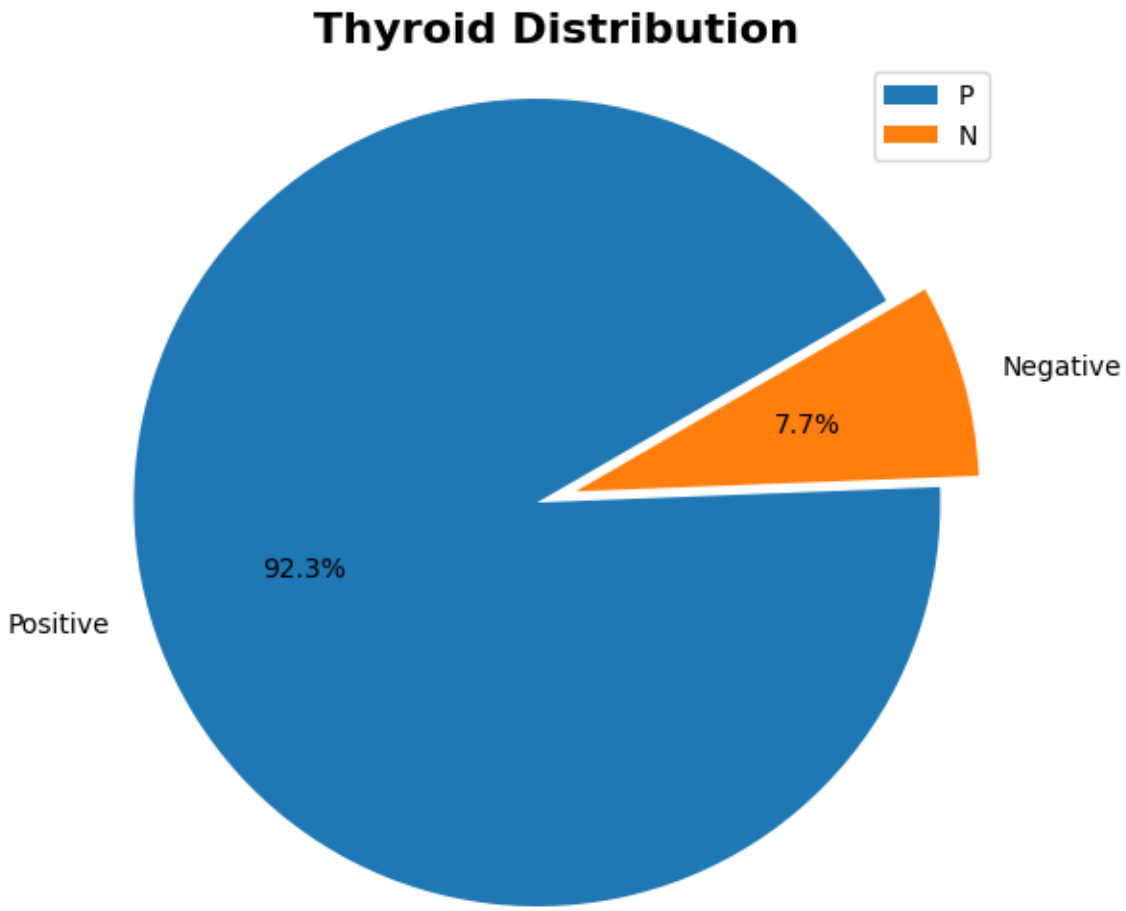


Figure 2

From the above pie chart, it seen that the dataset has two class: 'Positive' and 'Negative' with 92.3% and 7.7% ratio respectively. That means our dataset is highly imbalanced where Positive class is dominating.

DATA PREPROCESSING

Missing Value Handling: Before we start applying any model, first we must check and handle missing values, duplicate values and drop irrelevant data.

NO	Attributes	Missing Value Count	Missing value after Handling
1.	Age	0	0
2.	sex	150	0
3.	sick	0	0
4.	pregnant	0	0
5.	thyroid_surgery	0	0
6.	goitre	0	0
7.	tumor	0	0
8.	T3	0	0
9.	TT4	0	0
10.	T4U	0	0
11.	FTI	0	0
12.	TBG	3770	0
13.	Result	0	0

Table 3

Normalization: After dropping irrelevant features and label encoding necessary categorical columns we get:

Age	sex	sick	pregnant	thyroidsurgery	goitre	tumor	T3	T4	T4U	FTI	Result
12	0	0	0	0	0	0	2.5	125.0	1.14	109.0	1
20	1	0	0	0	0	0	2.0	102.0	0.00	0.0	1
46	1	0	0	0	0	0	0.0	109.0	0.91	120.0	1
70	0	0	0	0	0	0	1.9	175.0	0.00	0.0	1
70	0	0	0	0	0	0	1.2	61.0	0.87	70.0	1

Table 4

CROSS VALIDATION

Stratified K Fold Validation: Stratified sampling is a sampling technique where the samples are selected in the same proportion as they appear in the dataset. Implementing the concept of stratified sampling in cross-validation ensures the training and test sets have the same proportion of the feature of interest as in the original dataset. Doing this with the target variable ensures that the cross-validation result is a close approximation of generalization error. In our cross validation process we used 5 folds to validate the dataset. This approach is suitable for our dataset since it is an imbalanced dataset so there is a chance of having no minority class in some folds. The figure below represents our idea of using this approach as each folds contains same ratio of both classes:

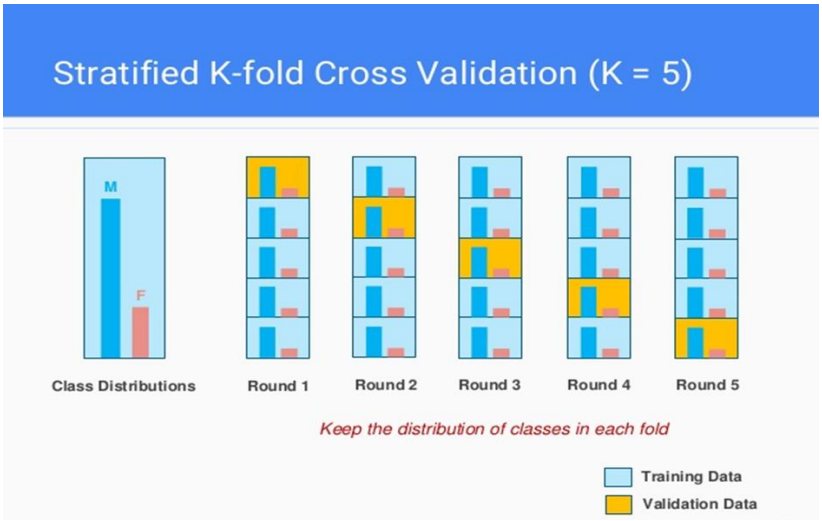


Figure 3

Classifiers: We applied 5 classifiers to perform cross validation:

Decision Tree: A decision tree is a supervised machine learning algorithm used for both classification and regression tasks. As the tree grows, it creates a hierarchy of decision rules that partition the feature space into regions corresponding to each class, allowing it to make accurate predictions for new data.

SVM: A support vector machine (SVM) is a supervised machine learning algorithm that classifies data by finding an optimal line. It supports binary classification and separates data points into two classes.

KNN: The KNN classifier in Python is a machine learning model that assigns a class label to a data point based on the majority class of its k nearest neighbors. The main advantage of KNN over other algorithms is that KNN can be used for multiclass classification.

Naïve Bayes: Naive Bayes is a statistical classification technique based on Bayes Theorem. It is one of the simplest supervised learning algorithms. It is a classification algorithm for binary (two-class) and multiclass classification problems.

ANN: An Artificial Neural Network (ANN) is a machine learning model inspired by the human brain's neural structure. It comprises interconnected nodes (neurons) organized into layers.

Original Dataset

Dataset Splitting: Using 5 folds splits our dataset where train set contains 2911 data and test set contains 728 data but last fold contains 727 data. Each fold maintains the class ratio which is 92.3% for class "1" and 7.7% for class "0".

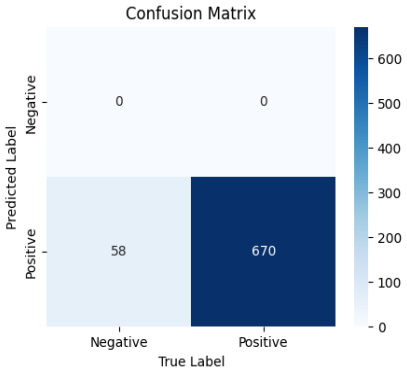
Decision Tree

Accuracy Table:

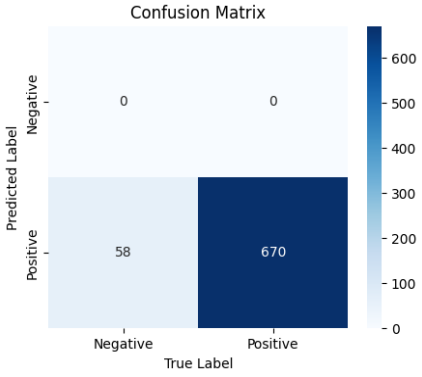
Fold	Accuracy
CV1	92.03%
CV2	92.03%
CV3	92.03%
CV4	91.90%
CV5	92.02%
Average	**92.00%**

Table 5

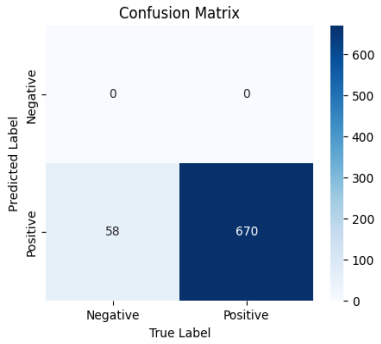
Fold 1



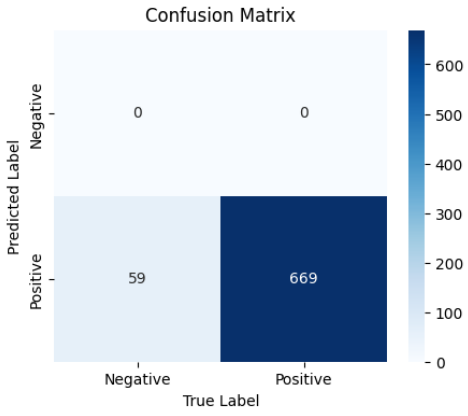
Fold 2



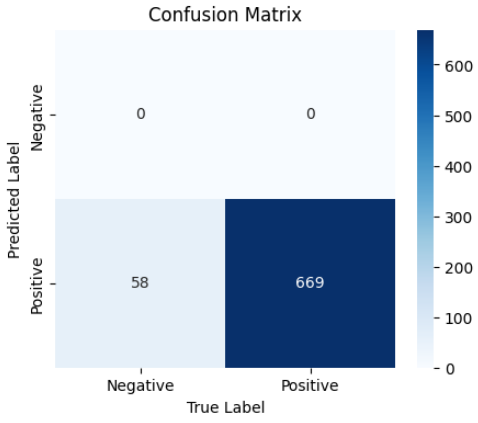
Fold 3



Fold 4



Fold 5



Classification Report for Class "1":

Fold	Precision	Recall	F1 score	Support
CV1	0.92	1.00	0.96	670
CV2	0.92	1.00	0.96	670
CV3	0.92	1.00	0.96	670
CV4	0.92	1.00	0.96	669
CV5	0.92	1.00	0.96	669

Table 6

Classification Report for Class "0":

Fold	Precision	Recall	F1 score	Support
CV1	0.00	0.00	0.00	58
CV2	0.00	0.00	0.00	58
CV3	0.00	0.00	0.00	58
CV4	0.00	0.00	0.00	59
CV5	0.00	0.00	0.00	58

Table 7

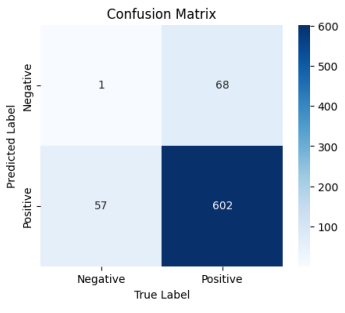
Support Vector Machine

Accuracy Table:

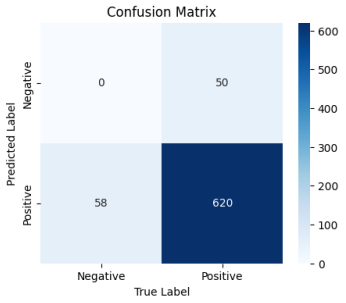
Fold	Accuracy
CV1	82.83%
CV2	85.16%
CV3	83.79%
CV4	83.79%
CV5	86.11%
Average	**84.34%**

Table 8

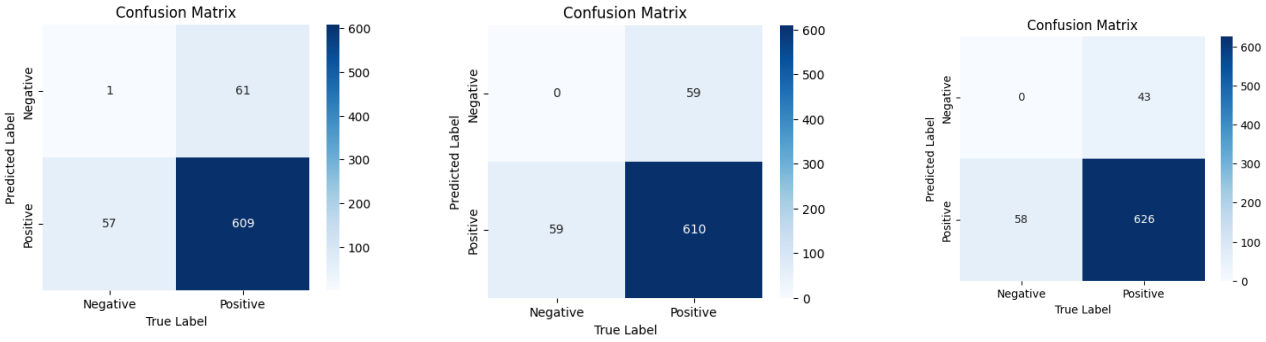
Fold 1



Fold 2



Fold 3 **Fold 4** **Fold 5**



Classification Report for Class "1":

Fold	Precision	Recall	F1 score	Support
CV1	0.91	.90	0.91	670
CV2	0.91	0.93	0.92	670
CV3	1.00	1.00	1.00	670
CV4	0.91	0.91	0.91	669
CV5	0.92	0.94	0.93	669

Table 9

Classification Report for Class "0":

Fold	Precision	Recall	F1 score	Support
CV1	0.01	0.02	0.02	58
CV2	0.00	0.00	0.00	58
CV3	0.02	0.02	0.02	58
CV4	0.00	0.00	0.00	59
CV5	0.00	0.00	0.00	58

Table 10

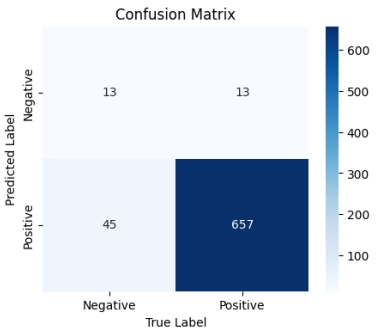
K nearest neighbors

Accuracy Table:

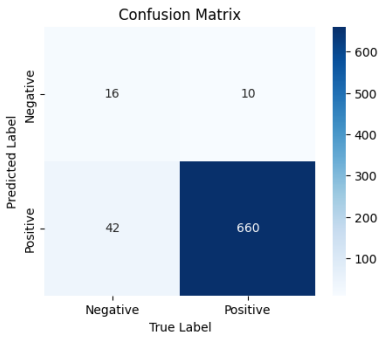
Fold	Accuracy
CV1	92.03%
CV2	92.86%
CV3	93.27%
CV4	92.86%
CV5	93.67%
Average	**92.94%**

Table 11

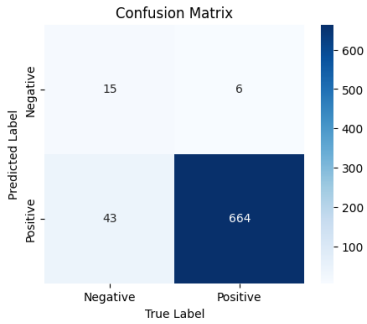
Fold 1



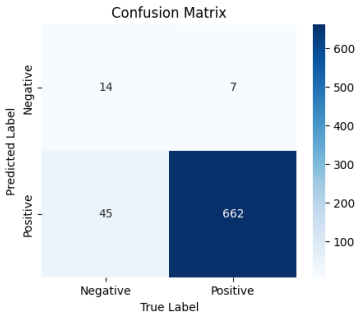
Fold 2



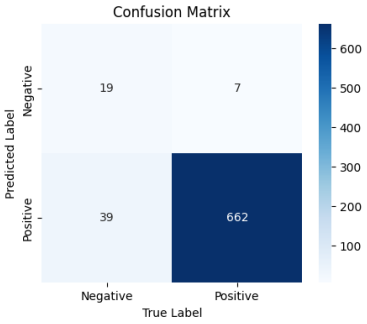
Fold 3



Fold 4



Fold 5



Classification Report for Class "1":

Fold	Precision	Recall	F1 score	Support
CV1	0.94	0.98	0.96	670
CV2	0.94	0.99	0.96	670
CV3	0.94	0.99	0.96	670
CV4	0.94	0.99	0.96	669
CV5	0.94	0.99	0.97	669

Table 12

Classification Report for Class "0":

Fold	Precision	Recall	F1 score	Support
CV1	0.50	0.22	0.31	58
CV2	0.62	0.28	0.38	58
CV3	0.71	0.26	0.38	58
CV4	0.67	0.24	0.35	59
CV5	0.73	0.33	0.45	58

Table 13

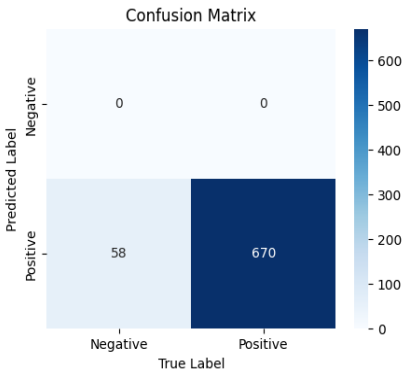
Naïve Bayes

Accuracy Table:

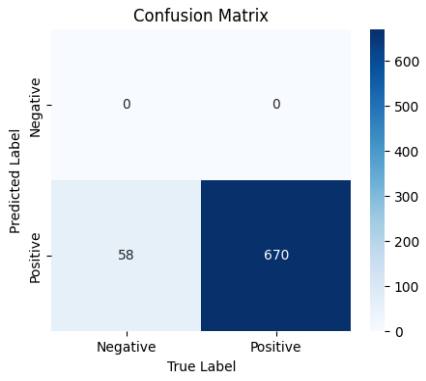
Fold	Accuracy
CV1	92.03%
CV2	92.03%
CV3	92.03%
CV4	91.90%
CV5	92.02%
Average	**92.00%**

Table 14

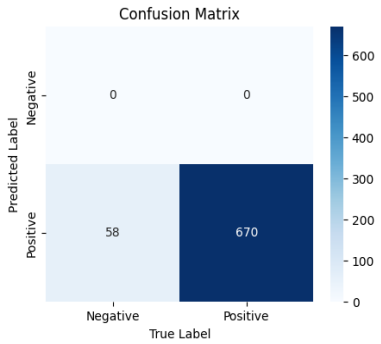
Fold 1



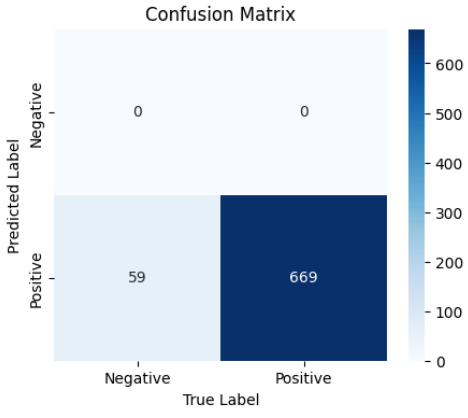
Fold 2



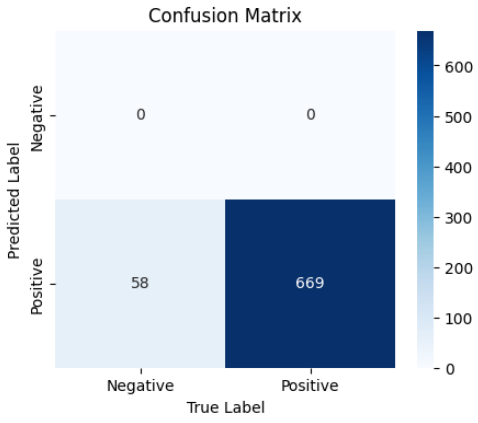
Fold 3



Fold 4



Fold 5



Classification Report for Class "1":

Fold	Precision	Recall	F1 score	Support
CV1	0.92	1.00	0.96	670
CV2	0.92	1.00	0.96	670
CV3	0.92	1.00	0.96	670
CV4	0.92	1.00	0.96	669
CV5	0.92	1.00	0.96	669

Table 15

Classification Report for Class "0":

Fold	Precision	Recall	F1 score	Support
CV1	0.00	0.00	0.00	58
CV2	0.00	0.00	0.00	58
CV3	0.00	0.00	0.00	58
CV4	0.00	0.00	0.00	59
CV5	0.00	0.00	0.00	58

Table 16

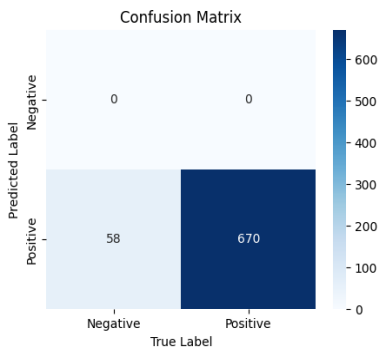
Artificial Neural Network

Accuracy Table:

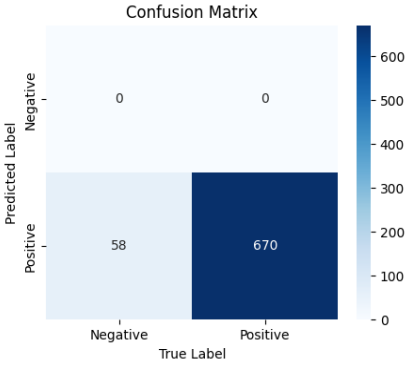
Fold	Accuracy
CV1	92.03%
CV2	92.03%
CV3	92.03%
CV4	92.99%
CV5	92.26%
Average	**92.47%**

Table 17

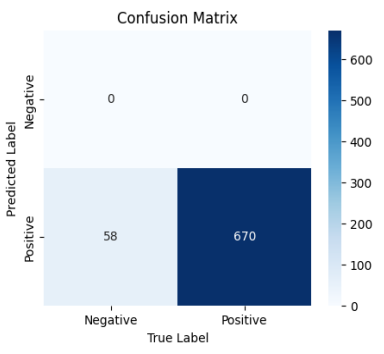
Fold 1



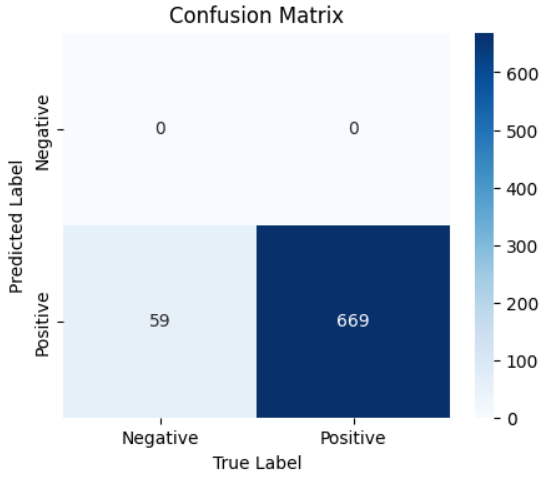
Fold 2



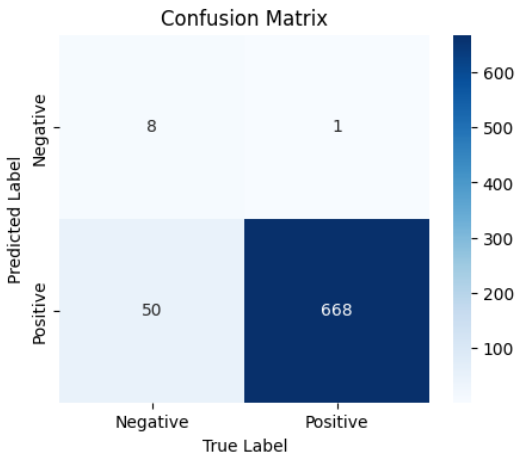
Fold 3



Fold 4



Fold 5



Classification Report for Class "1":

Fold	Precision	Recall	F1 score	Support
CV1	0.92	1.00	0.96	670
CV2	0.92	1.00	0.96	670
CV3	0.92	1.00	0.96	670
CV4	0.92	1.00	0.96	669
CV5	0.93	1.00	0.96	669

Table 18

Classification Report for Class "0":

Fold	Precision	Recall	F1 score	Support
CV1	0.00	0.00	0.00	58
CV2	0.00	0.00	0.00	58
CV3	0.00	0.00	0.00	58
CV4	0.00	0.00	0.00	59
CV5	0.89	0.14	0.24	58

Table 19

Comparison:

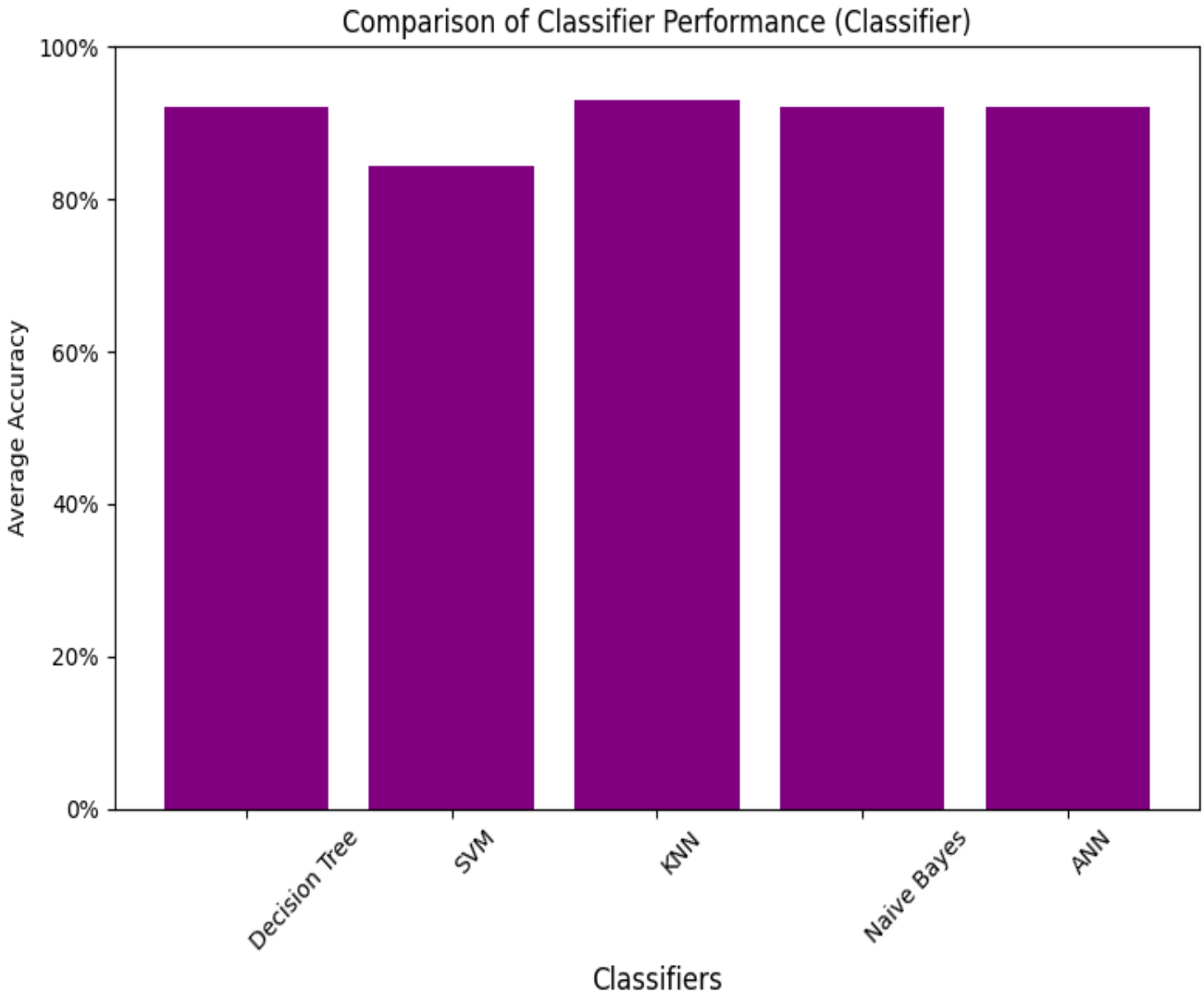


Figure 4

K-means Dataset

Dataset Overview: This dataset is created by the clustered class using K-means clustering and turned into supervised dataset.

Dataset Splitting: Using 5 folds splits k-means dataset where train set contains 2911 data and test set contains 728 data but last fold contains 727 data. Each fold maintains the class ratio which is 92.3% for class "1" and 7.7% for class "0".

Distribution of Result Class:

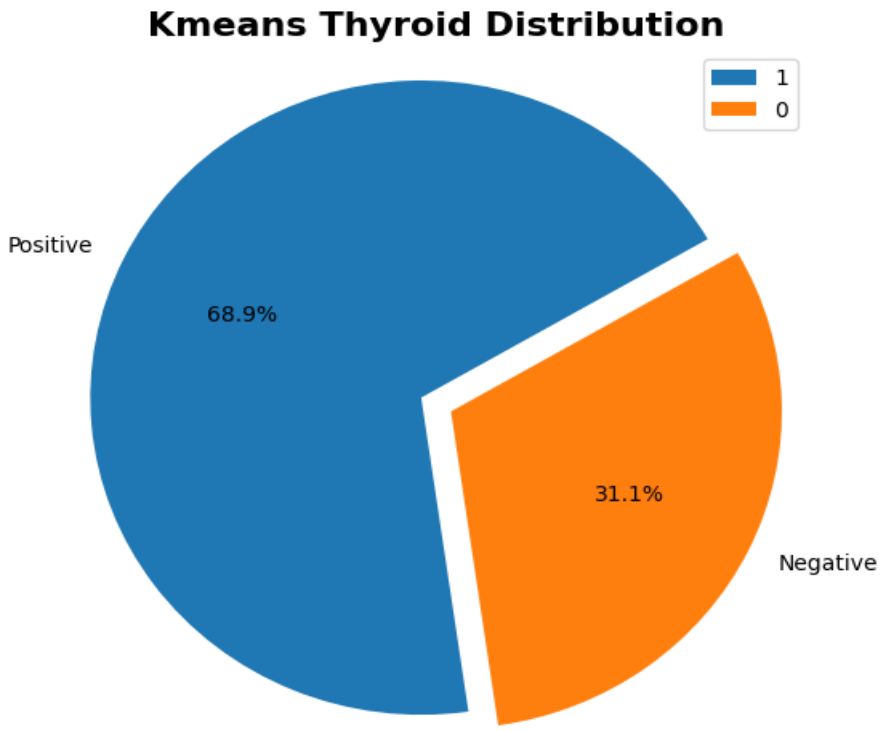


Figure 5

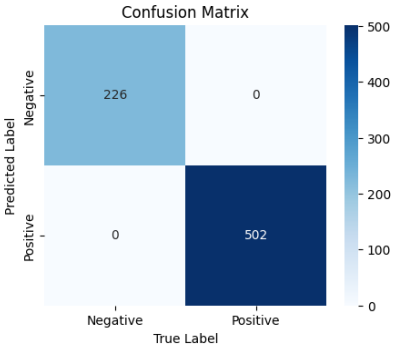
Decision Tree

Accuracy Table:

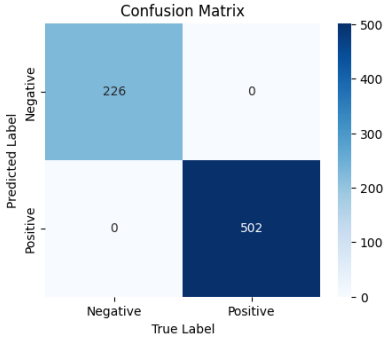
Fold	Accuracy
CV1	100.00%
CV2	100.00%
CV3	100.00%
CV4	100.00%
CV5	100.00%
Average	**100.00%**

Table 20

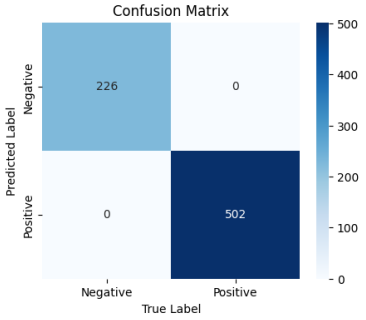
Fold 1



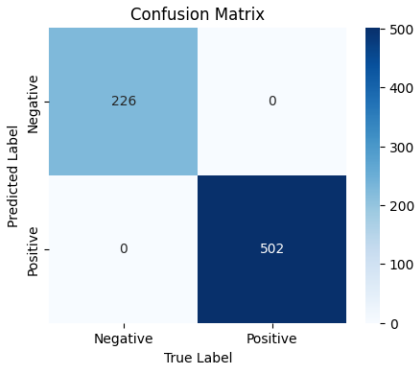
Fold 2



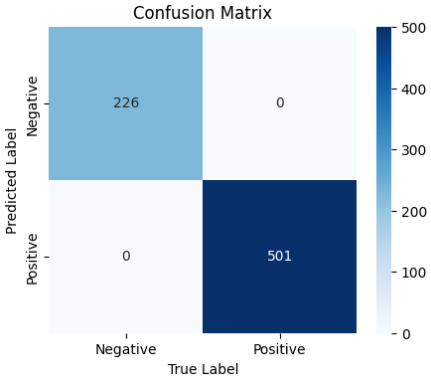
Fold 3



Fold 4



Fold 5



Classification Report for Class "1":

Fold	Precision	Recall	F1 score	Support
CV1	1.00	1.00	1.00	502
CV2	1.00	1.00	1.00	502
CV3	1.00	1.00	1.00	502
CV4	1.00	1.00	1.00	502
CV5	1.00	1.00	1.00	501

Table 21

Classification Report for Class "0":

Fold	Precision	Recall	F1 score	Support
CV1	1.00	1.00	1.00	226
CV2	1.00	1.00	1.00	226
CV3	1.00	1.00	1.00	226
CV4	1.00	1.00	1.00	226
CV5	1.00	1.00	1.00	226

Table 22

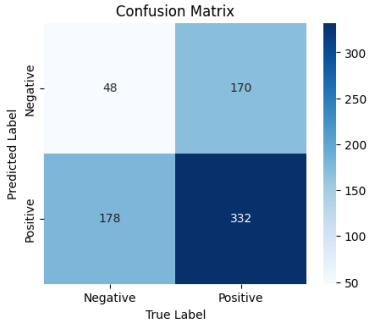
Support Vector Machine

Accuracy Table:

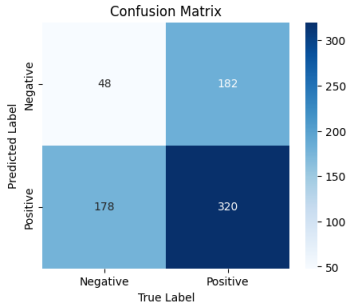
Fold	Accuracy
CV1	52.20%
CV2	50.55%
CV3	51.51%
CV4	53.02%
CV5	53.09%
Average	**52.08%**

Table 23

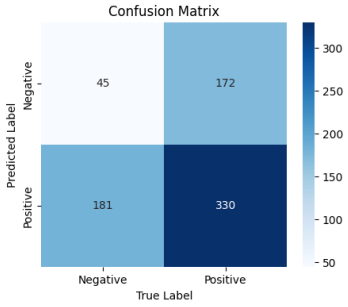
Fold 1



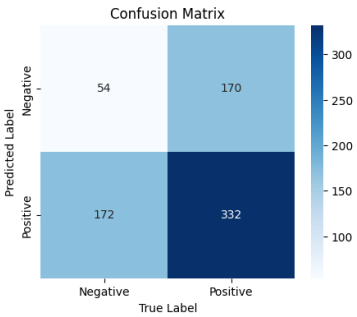
Fold 2



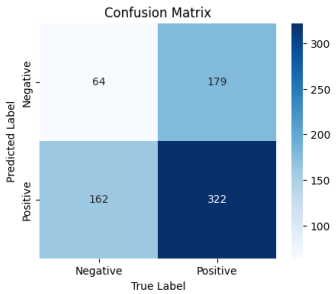
Fold 3



Fold 4



Fold 5



Classification Report for Class "1":

Fold	Precision	Recall	F1 score	Support
CV1	0.65	0.66	0.66	502
CV2	0.64	0.64	0.64	502
CV3	0.65	0.66	0.65	502
CV4	0.66	0.66	0.66	502
CV5	0.67	0.64	0.65	501

Table 24

Classification Report for Class "0":

Fold	Precision	Recall	F1 score	Support
CV1	0.22	0.21	0.22	226
CV2	0.21	0.21	0.21	226
CV3	0.21	0.20	0.20	226
CV4	0.24	0.24	0.24	226
CV5	0.26	0.28	0.27	226

Table 25

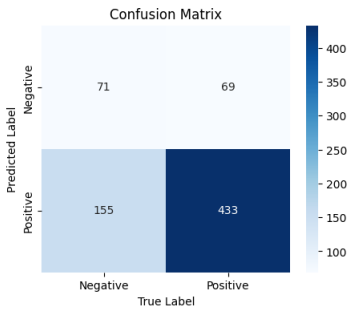
K nearest neighbors

Accuracy Table:

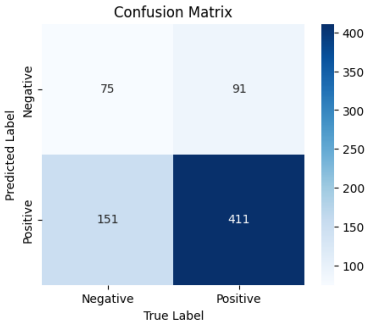
Fold	Accuracy
CV1	69.23%
CV2	66.76%
CV3	64.29%
CV4	63.60%
CV5	66.57%
Average	**66.09%**

Table 26

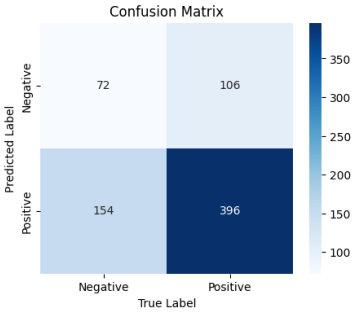
Fold 1



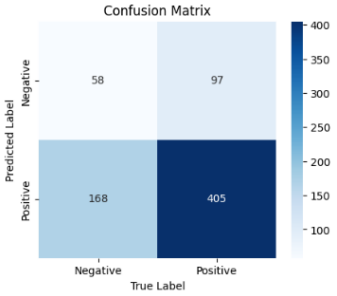
Fold 2



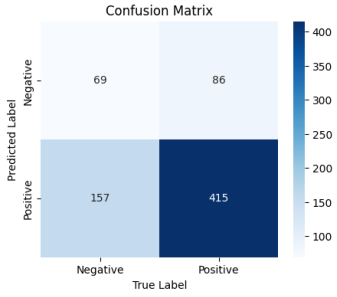
Fold 3



Fold 4



Fold 5



Classification Report for Class "1":

Fold	Precision	Recall	F1 score	Support
CV1	0.74	0.86	0.79	502
CV2	0.73	0.82	0.77	502
CV3	0.72	0.79	0.75	502
CV4	0.71	0.81	0.75	502
CV5	0.73	0.83	0.77	501

Table 27

Classification Report for Class "0":

Fold	Precision	Recall	F1 score	Support
CV1	0.51	0.31	0.39	226
CV2	0.45	0.33	0.38	226
CV3	0.40	0.32	0.36	226
CV4	0.37	0.26	0.30	226
CV5	0.45	0.31	0.36	226

Table 28

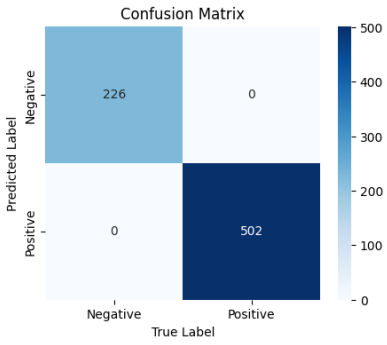
Naïve Bayes

Accuracy Table:

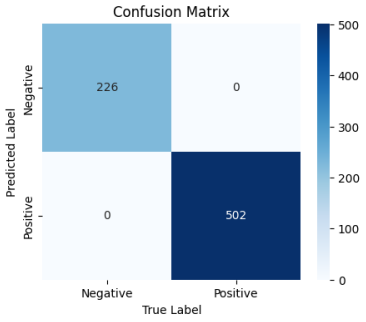
Fold	Accuracy
CV1	100.00%
CV2	100.00%
CV3	100.00%
CV4	100.00%
CV5	100.00%
Average	**100.00%**

Table 29

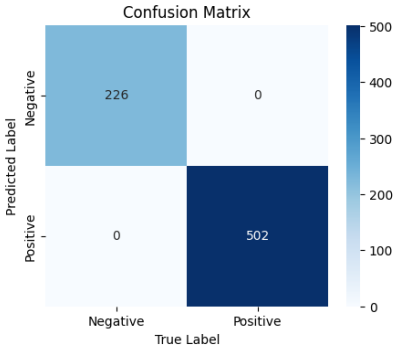
Fold 1



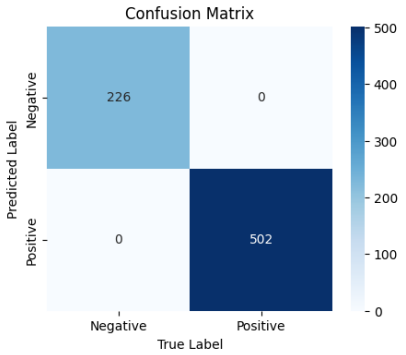
Fold 2



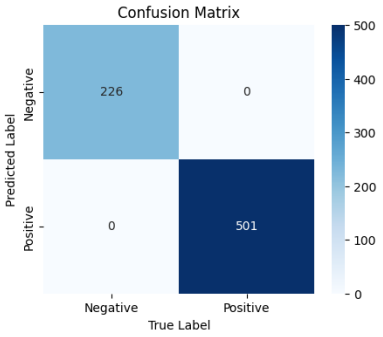
Fold 3



Fold 4



Fold 5



Classification Report for Class "1":

Fold	Precision	Recall	F1 score	Support
CV1	1.00	1.00	1.00	502
CV2	1.00	1.00	1.00	502
CV3	1.00	1.00	1.00	502
CV4	1.00	1.00	1.00	502
CV5	1.00	1.00	1.00	501

Table 30

Classification Report for Class "0":

Fold	Precision	Recall	F1 score	Support
CV1	1.00	1.00	1.00	226
CV2	1.00	1.00	1.00	226
CV3	1.00	1.00	1.00	226
CV4	1.00	1.00	1.00	226
CV5	1.00	1.00	1.00	226

Table 31

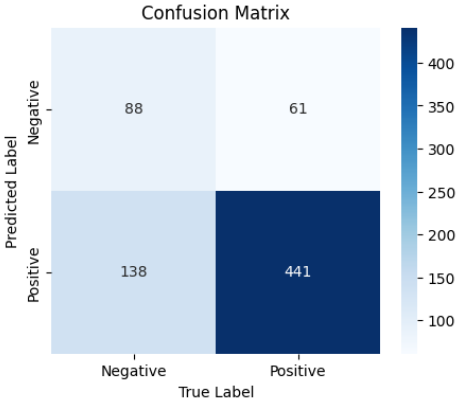
Artificial Neural Network

Accuracy Table:

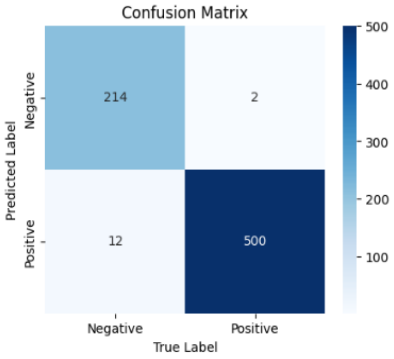
Fold	Accuracy
CV1	72.66%
CV2	98.08%
CV3	99.73%
CV4	100.00%
CV5	100.00%
Average	**94.09%**

Table 32

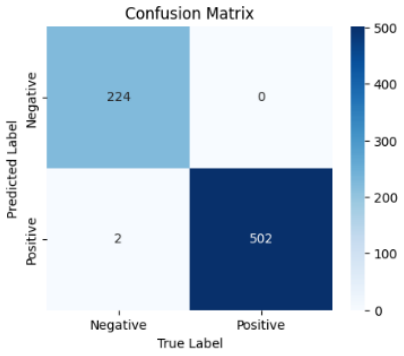
Fold 1



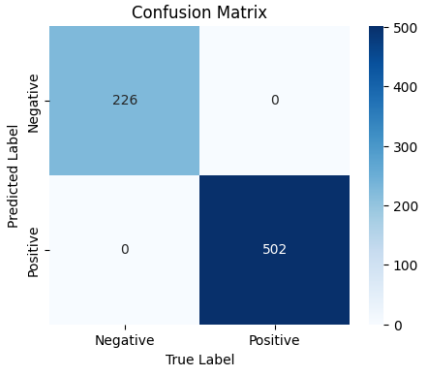
Fold 2



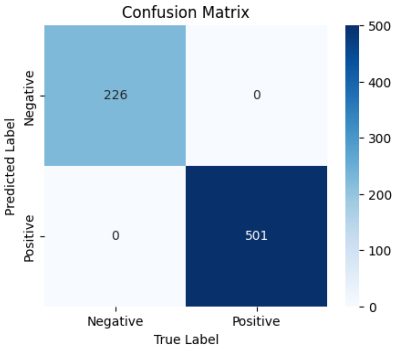
Fold 3



Fold 4



Fold 5



Classification Report for Class "1":

Fold	Precision	Recall	F1 score	Support
CV1	0.76	0.88	0.82	502
CV2	0.98	1.00	0.99	502
CV3	1.00	1.00	1.00	502
CV4	1.00	1.00	1.00	502
CV5	1.00	1.00	1.00	501

Table 33

Classification Report for Class "0":

Fold	Precision	Recall	F1 score	Support
CV1	0.59	0.49	0.47	226
CV2	0.99	0.95	0.97	226
CV3	1.00	0.99	1.00	226
CV4	1.00	1.00	1.00	226
CV5	1.00	1.00	1.00	226

Table 34

Comparison:

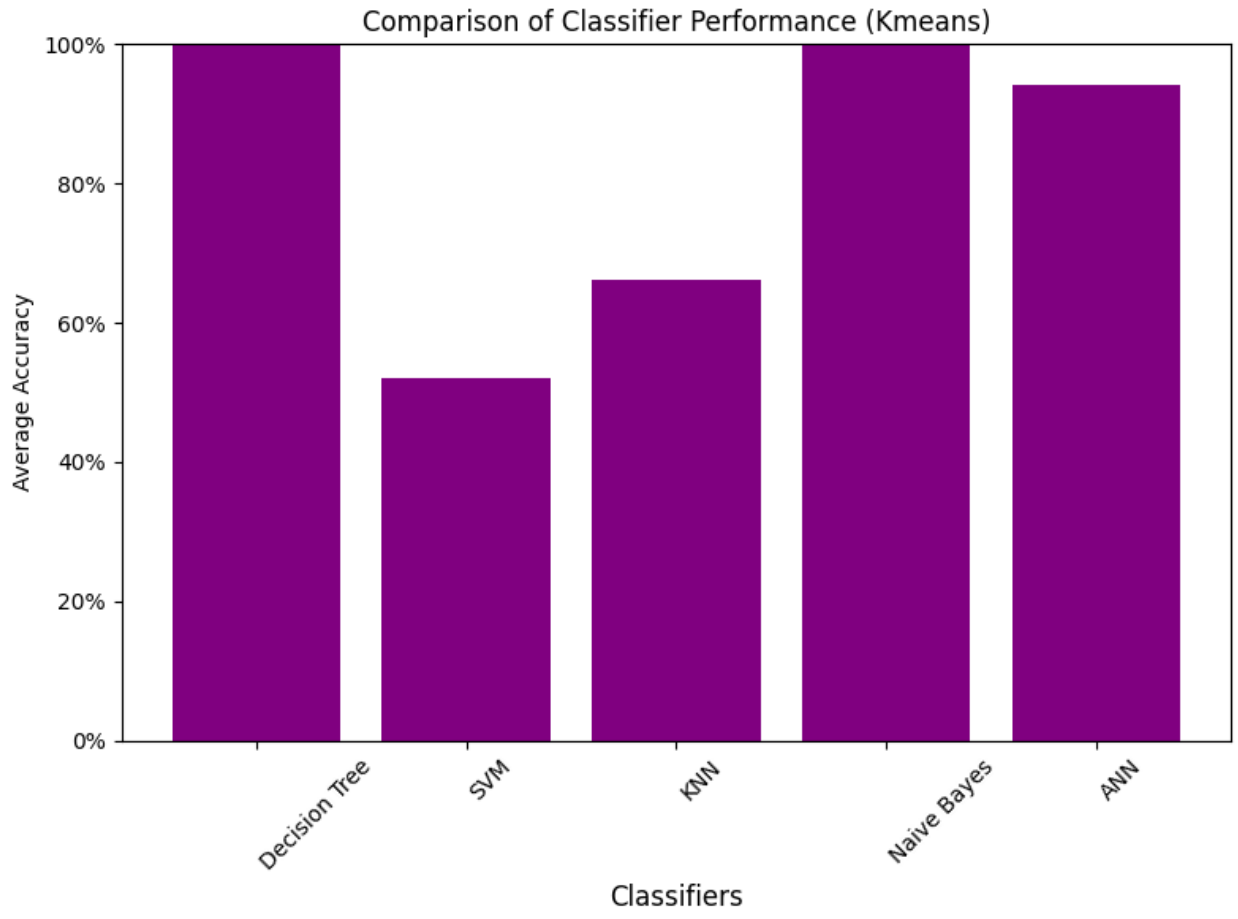


Figure 6

K-medoids Dataset

Dataset Overview: This dataset is created by the clustered class using K-medoids clustering and turned into supervised dataset.

Dataset Splitting: Using 5 folds splits k-means dataset where train set contains 2911 data and test set contains 728 data but last fold contains 727 data. Each fold maintains the class ratio which is 54.1% for class "1" and 45.9% for class "0".

Distribution of Result Class:

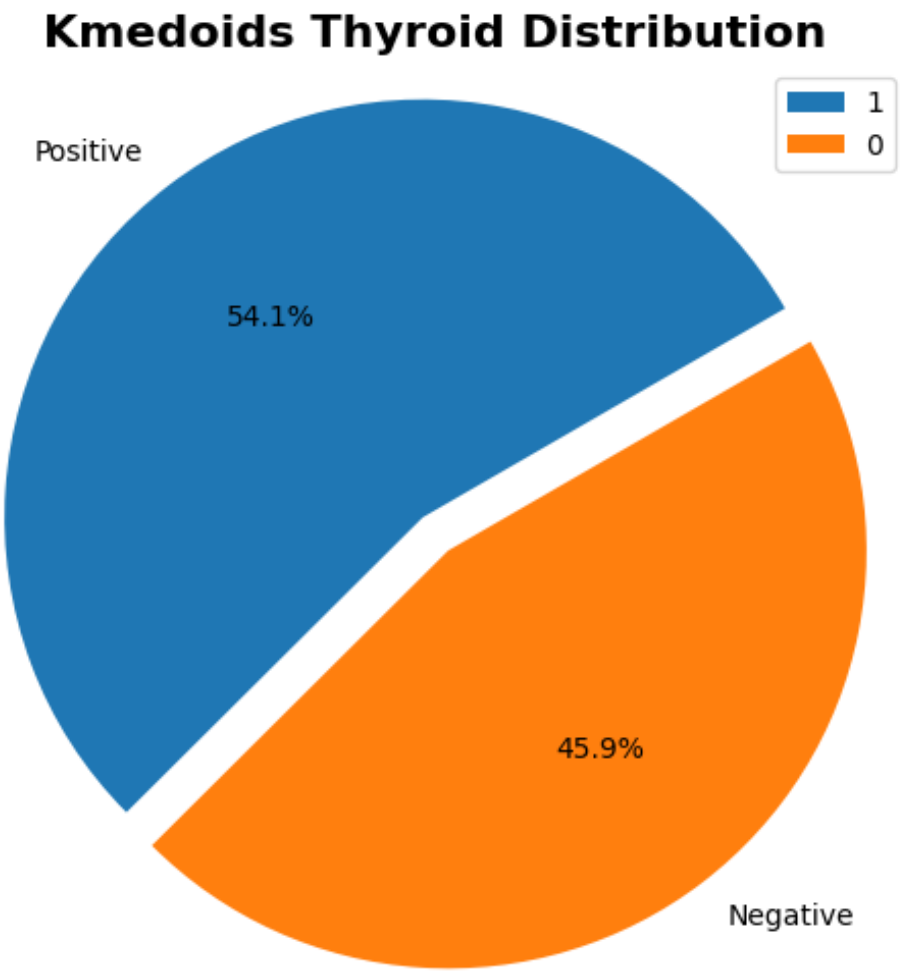


Figure 7

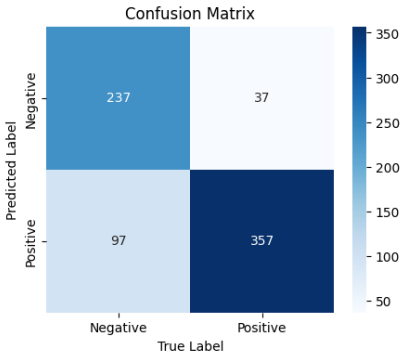
Decision Tree

Accuracy Table:

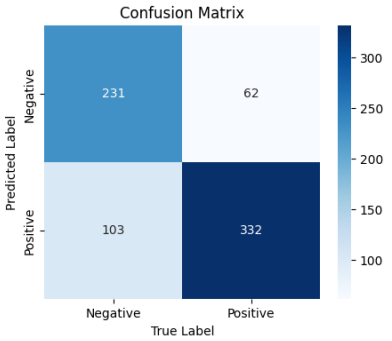
Fold	Accuracy
CV1	81.59%
CV2	77.34%
CV3	80.22%
CV4	82.28%
CV5	85.42%
Average	**81.37%**

Table 35

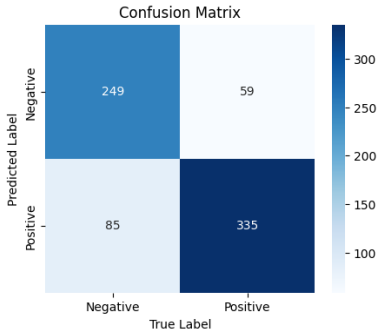
Fold 1



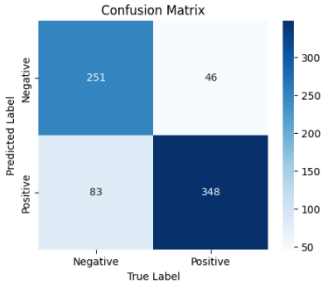
Fold 2



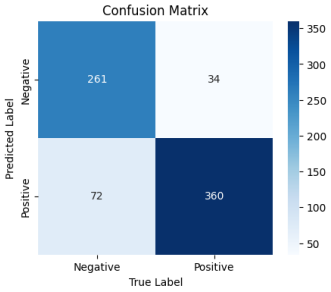
Fold 3



Fold 4



Fold 5



Classification Report for Class "1":

Fold	Precision	Recall	F1 score	Support
CV1	0.79	0.91	0.84	394
CV2	0.76	0.84	0.80	394
CV3	0.80	0.85	0.82	394
CV4	0.81	0.88	0.84	394
CV5	0.83	0.91	0.87	394

Table 36

Classification Report for Class "0":

Fold	Precision	Recall	F1 score	Support
CV1	0.76	0.91	0.84	334
CV2	0.79	0.69	0.74	334
CV3	0.81	0.75	0.78	334
CV4	0.85	0.75	0.80	334
CV5	0.88	0.78	0.83	333

Table 37

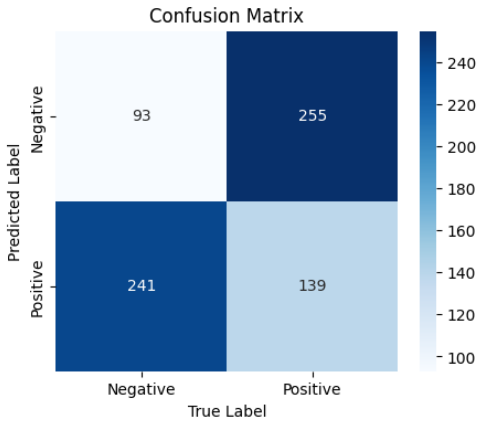
Support Vector Machine

Accuracy Table:

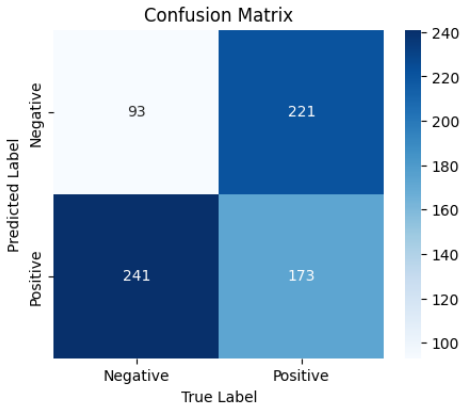
Fold	Accuracy
CV1	31.87%
CV2	36.54%
CV3	35.03%
CV4	33.65%
CV5	78.27%
Average	**43.07%**

Table 38

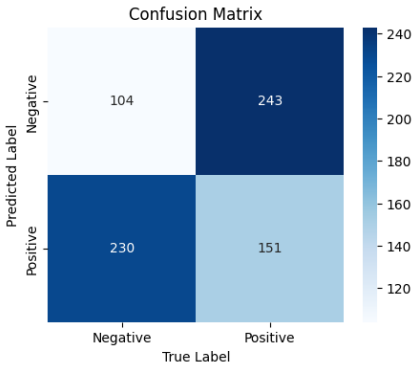
Fold 1



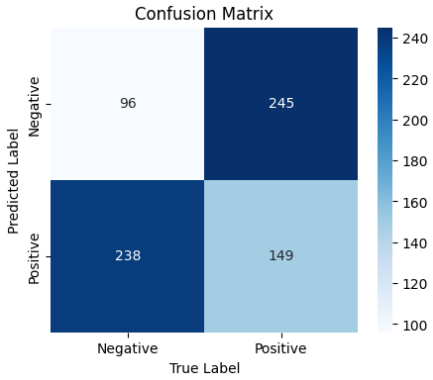
Fold 2



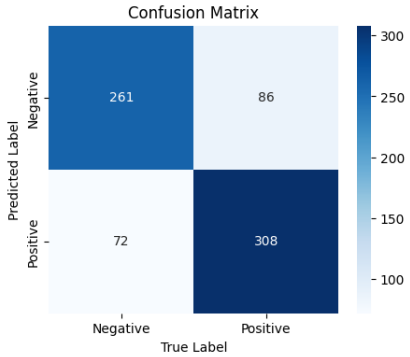
Fold 3



Fold 4



Fold 5



Classification Report for Class "1":

Fold	Precision	Recall	F1 score	Support
CV1	0.37	0.35	0.36	394
CV2	0.42	0.44	0.43	394
CV3	0.40	0.38	0.39	394
CV4	0.39	0.38	0.38	394
CV5	0.81	0.78	0.80	394

Table 39

Classification Report for Class "0":

Fold	Precision	Recall	F1 score	Support
CV1	0.27	0.28	0.27	334
CV2	0.30	0.28	0.29	334
CV3	0.30	0.31	0.31	334
CV4	0.28	0.29	0.28	334
CV5	0.75	0.78	0.77	333

Table 40

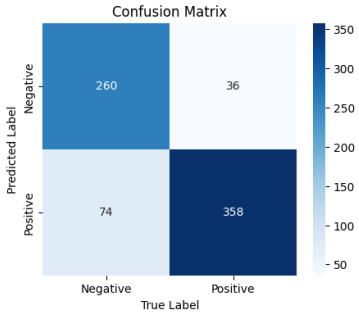
K nearest neighbors

Accuracy Table:

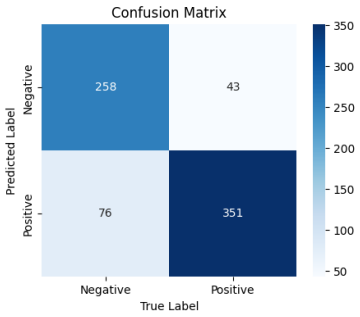
Fold	Accuracy
CV1	84.89%
CV2	83.65%
CV3	71%
CV4	86.81%
CV5	87.21%
Average	**85.66%**

Table 41

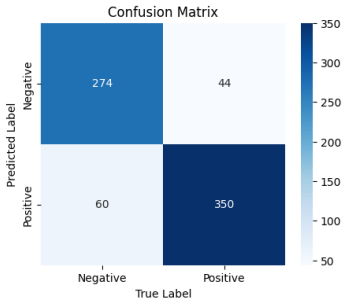
Fold 1



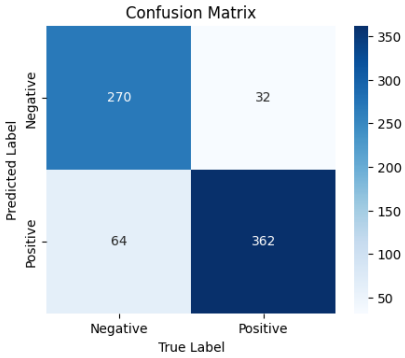
Fold 2



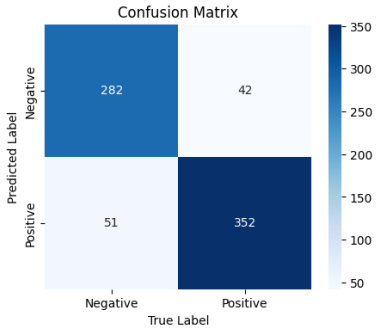
Fold 3



Fold 4



Fold 5



Classification Report for Class "1":

Fold	Precision	Recall	F1 score	Support
CV1	0.83	0.91	0.77	394
CV2	0.82	0.89	0.86	394
CV3	0.85	0.89	0.87	394
CV4	0.85	0.92	0.88	394
CV5	0.87	0.89	0.86	394

Table 42

Classification Report for Class "0":

Fold	Precision	Recall	F1 score	Support
CV1	0.86	0.77	0.81	334
CV2	0.88	0.78	0.83	334
CV3	0.86	0.82	0.84	334
CV4	0.89	0.81	0.85	334
CV5	0.87	0.85	0.86	333

Table 43

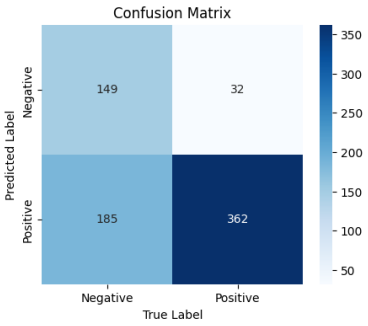
Naïve Bayes

Accuracy Table:

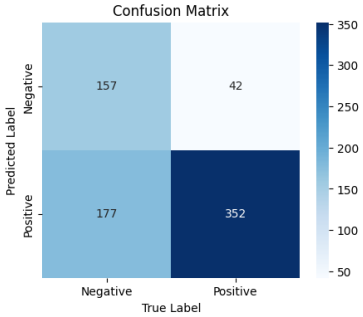
Fold	Accuracy
CV1	70.19%
CV2	69.92%
CV3	67.58%
CV4	70.33%
CV5	68.50%
Average	**69.30%**

Table 44

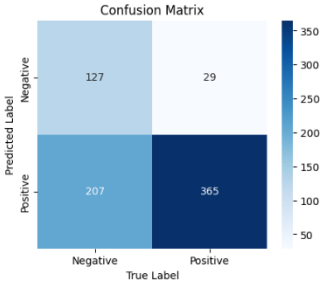
Fold 1



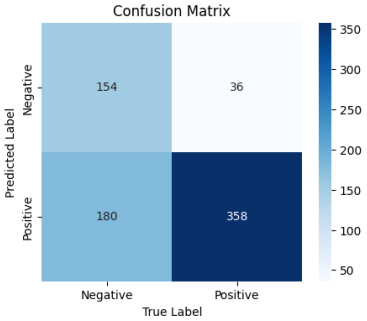
Fold 2



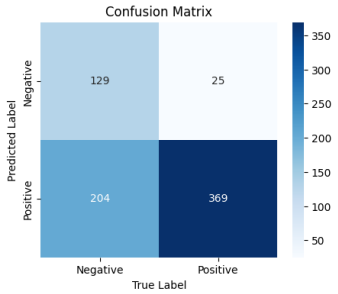
Fold 3



Fold 4



Fold 5



Classification Report for Class "1":

Fold	Precision	Recall	F1 score	Support
CV1	0.66	0.92	0.77	394
CV2	0.79	0.47	0.76	394
CV3	0.80	0.38	0.92	394
CV4	0.67	0.91	0.77	394
CV5	0.66	0.45	0.77	394

Table 45

Classification Report for Class "0":

Fold	Precision	Recall	F1 score	Support
CV1	0.82	0.45	0.58	334
CV2	0.79	0.47	0.59	334
CV3	0.81	0.38	0.52	334
CV4	0.81	0.46	0.59	334
CV5	0.84	0.39	0.53	333

Table 46

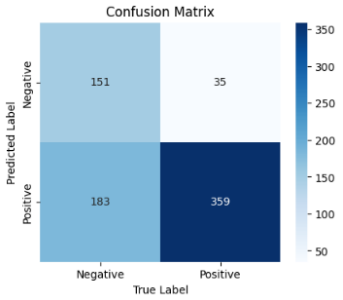
Artificial Neural Network

Accuracy Table:

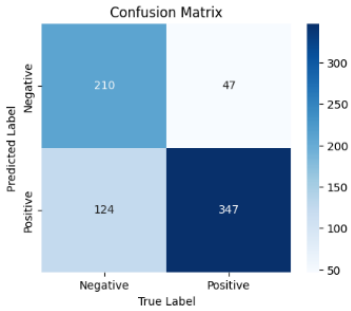
Fold	Accuracy
CV1	70.06%
CV2	76.08%
CV3	86.73%
CV4	86.00%
CV5	88.00%
Average	**80.96%**

Table 47

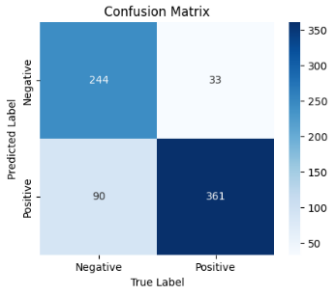
Fold 1



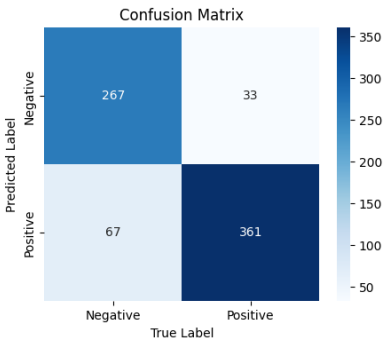
Fold 2



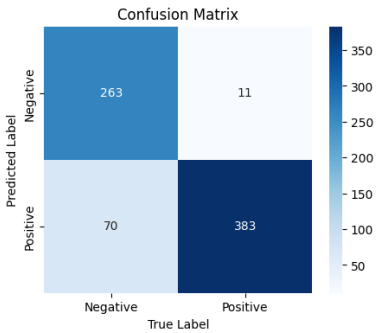
Fold 3



Fold 4



Fold 5



Classification Report for Class "1":

Fold	Precision	Recall	F1 score	Support
CV1	0.66	0.91	0.77	394
CV2	0.74	0.88	0.80	394
CV3	0.80	0.92	0.85	394
CV4	0.84	0.92	0.88	394
CV5	0.85	0.97	0.90	394

Table 48

Classification Report for Class "0":

Fold	Precision	Recall	F1 score	Support
CV1	0.81	0.45	0.58	334
CV2	0.82	0.63	0.71	334
CV3	0.88	0.73	0.80	334
CV4	0.81	0.80	0.84	334
CV5	0.96	0.79	0.87	333

Table 49

Comparison:

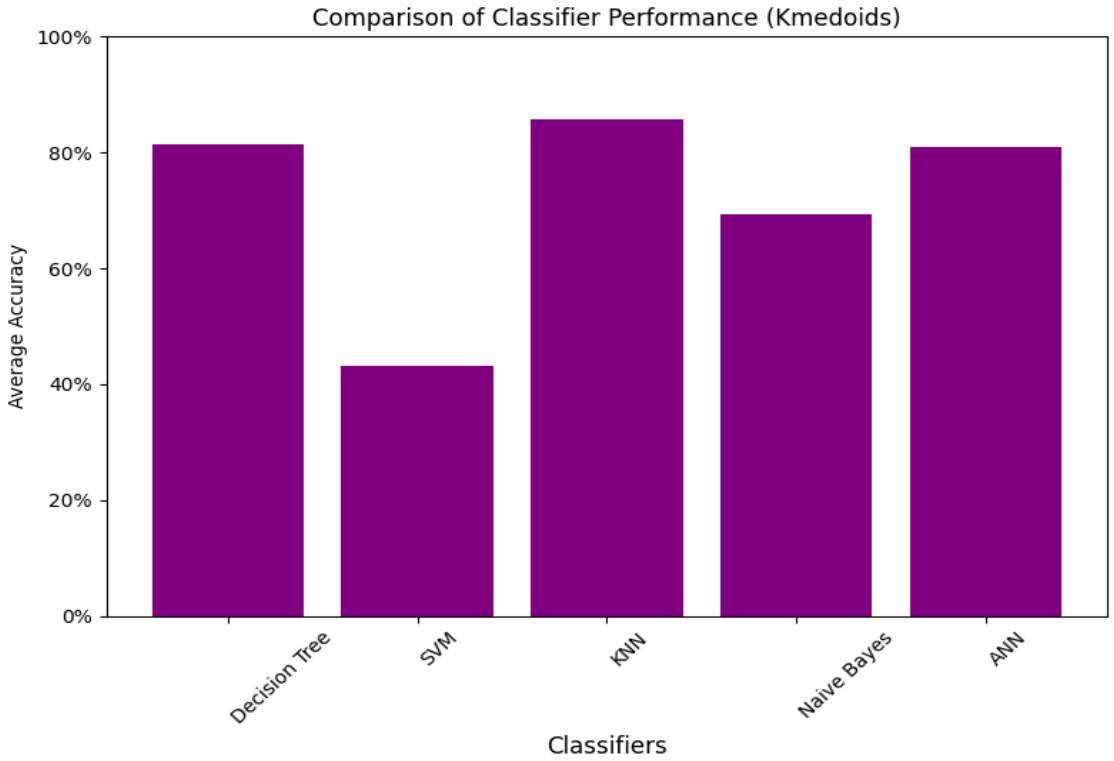


Figure 8

DBSCAN Dataset

Dataset Overview: This dataset is created by the clustered class using DBSCAN clustering and turned into supervised dataset.

Dataset Splitting: Using 5 folds splits k-means dataset where train set contains 2358 data and test set contains 590 data but last fold contains 727 data. Each fold maintains the class ratio which is 67.5% for class "1" and 32.5% for class "0".

Distribution of Result Class:

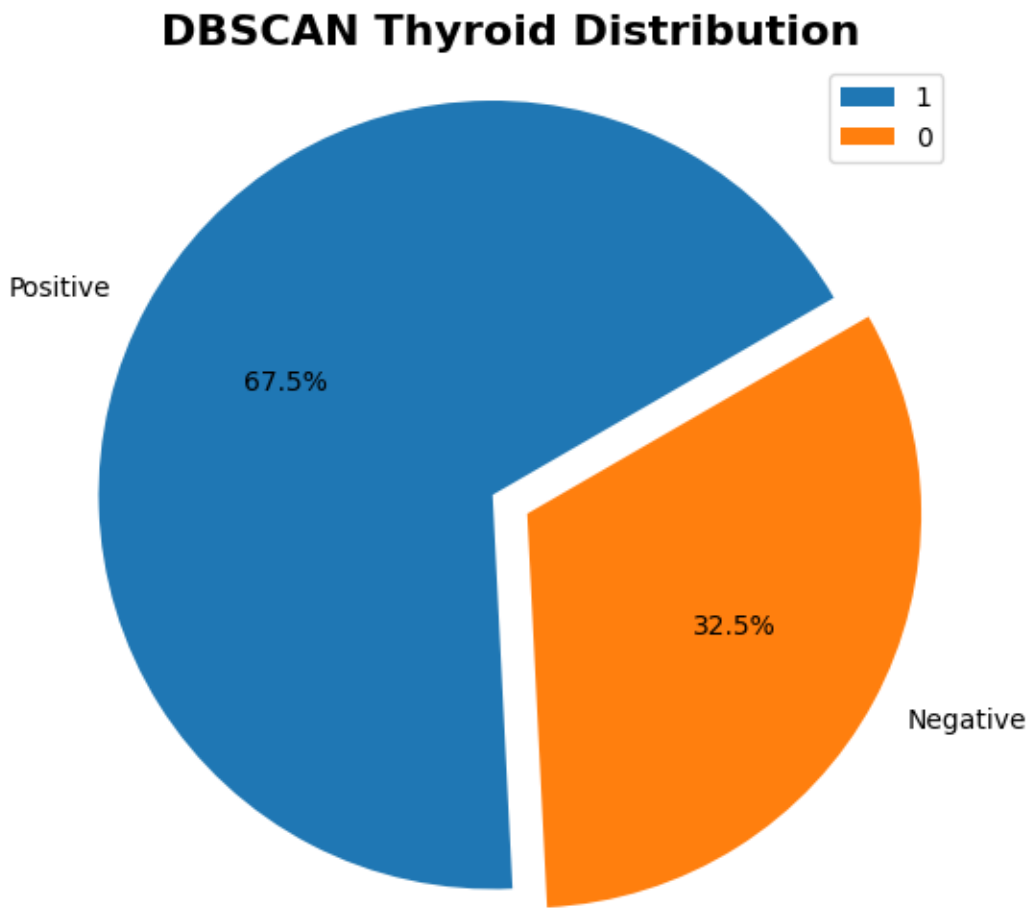


Figure 9

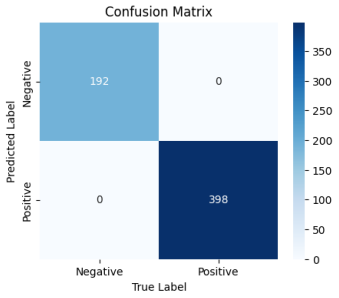
Decision Tree

Accuracy Table:

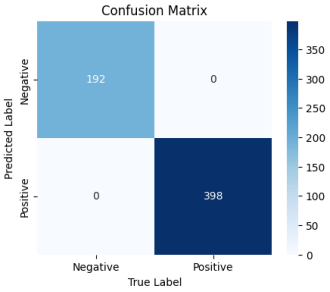
Fold	Accuracy
CV1	100.00%
CV2	100.00%
CV3	100.00%
CV4	100.00%
CV5	100.00%
Average	**100.00%**

Table 50

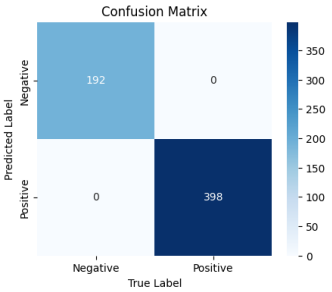
Fold 1



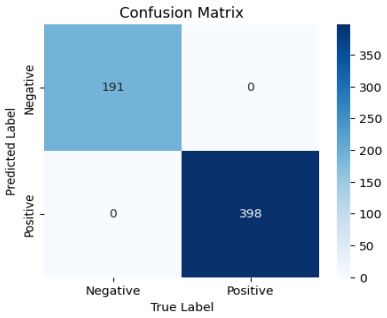
Fold 2



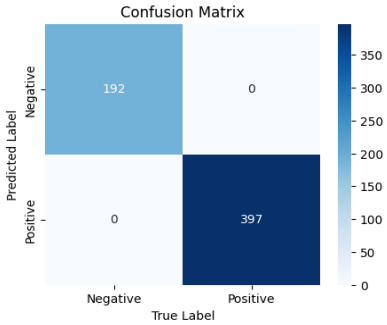
Fold 3



Fold 4



Fold 5



Classification Report for Class "1":

Fold	Precision	Recall	F1 score	Support
CV1	1.00	1.00	1.00	398
CV2	1.00	1.00	1.00	398
CV3	1.00	1.00	1.00	398
CV4	1.00	1.00	1.00	398
CV5	1.00	1.00	1.00	397

Table 51

Classification Report for Class "0":

Fold	Precision	Recall	F1 score	Support
CV1	1.00	1.00	1.00	192
CV2	1.00	1.00	1.00	192
CV3	1.00	1.00	1.00	192
CV4	1.00	1.00	1.00	191
CV5	1.00	1.00	1.00	192

Table 52

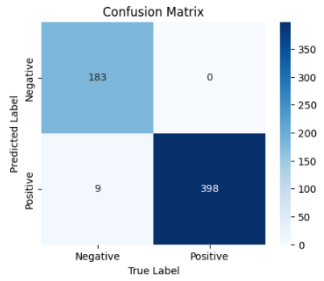
Support Vector Machine

Accuracy Table:

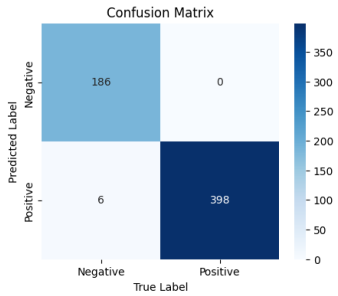
Fold	Accuracy
CV1	98.47%
CV2	98.98%
CV3	98.64%
CV4	99.15%
CV5	99.15%
Average	**98.88%**

Table 53

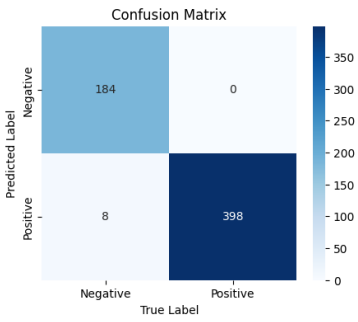
Fold 1



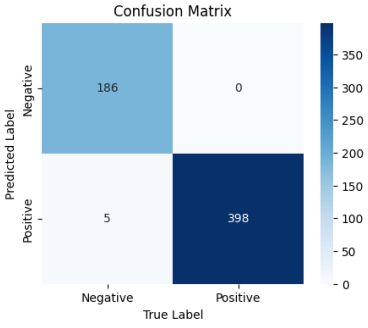
Fold 2



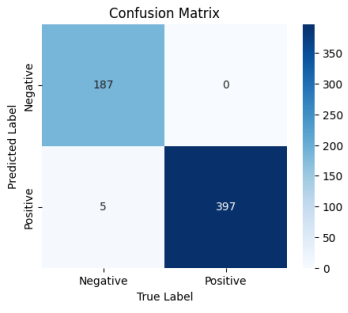
Fold 3



Fold 4



Fold 5



Classification Report for Class "1":

Fold	Precision	Recall	F1 score	Support
CV1	0.98	1.00	0.97	398
CV2	0.99	1.00	0.99	398
CV3	0.98	1.00	0.99	398
CV4	0.99	1.00	0.99	398
CV5	0.99	1.00	0.99	397

Table 54

Classification Report for Class "0":

Fold	Precision	Recall	F1 score	Support
CV1	1.00	0.95	0.98	192
CV2	1.00	0.97	0.98	192
CV3	1.00	0.96	0.98	192
CV4	1.00	0.97	0.99	191
CV5	1.00	0.97	0.98	192

Table 55

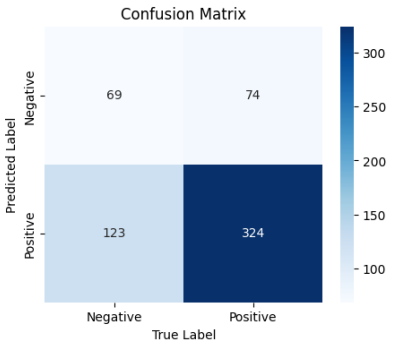
K nearest neighbors

Accuracy Table:

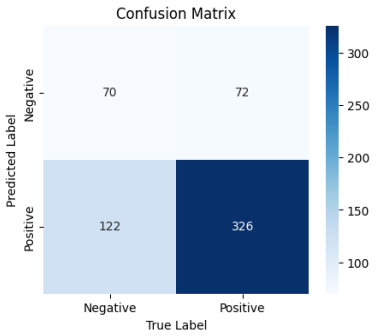
Fold	Accuracy
CV1	66.61%
CV2	67.12%
CV3	65.08%
CV4	61.63%
CV5	66.21%
Average	**65.33%**

Table 56

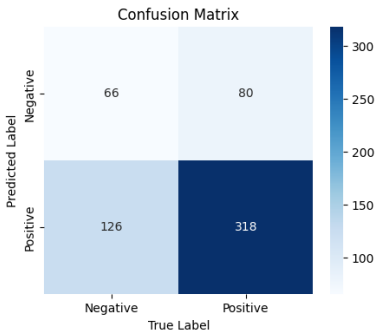
Fold 1



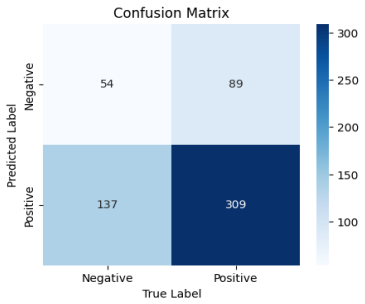
Fold 2



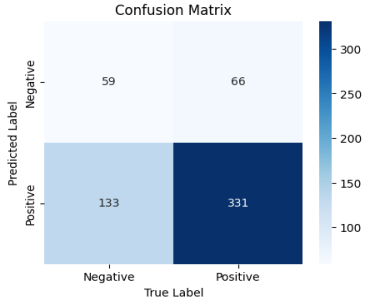
Fold 3



Fold 4



Fold 5



Classification Report for Class "1":

Fold	Precision	Recall	F1 score	Support
CV1	0.72	0.81	0.77	398
CV2	0.73	0.82	0.77	398
CV3	0.72	0.81	0.76	398
CV4	0.69	0.78	0.73	398
CV5	0.71	0.83	0.77	397

Table 57

Classification Report for Class "0":

Fold	Precision	Recall	F1 score	Support
CV1	0.48	0.36	0.41	192
CV2	0.49	0.36	0.42	192
CV3	0.38	0.28	0.32	192
CV4	0.48	0.36	0.41	191
CV5	0.47	0.31	0.37	192

Table 58

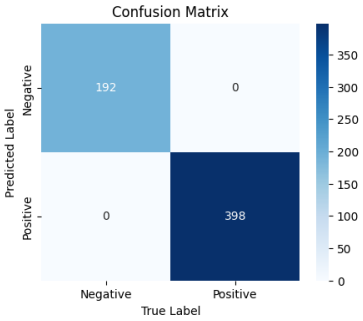
Naïve Bayes

Accuracy Table:

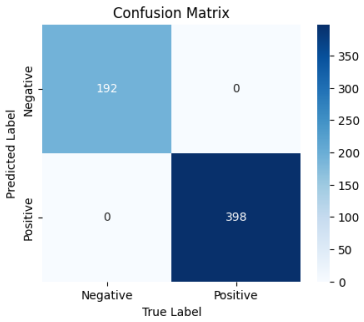
Fold	Accuracy
CV1	100.00%
CV2	100.00%
CV3	100.00%
CV4	100.00%
CV5	100.00%
Average	**100.00%**

Table 59

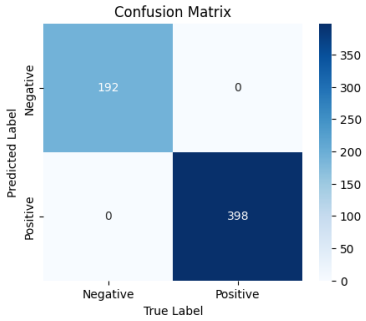
Fold 1



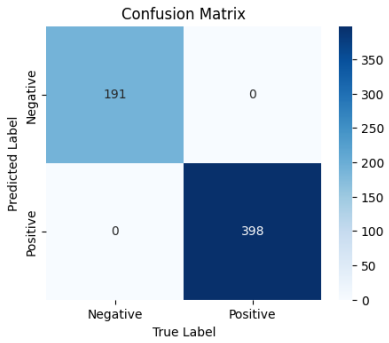
Fold 2



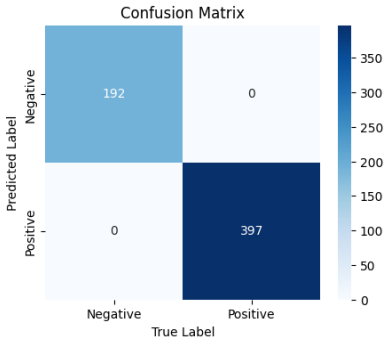
Fold 3



Fold 4



Fold 5



Classification Report for Class "1":

Fold	Precision	Recall	F1 score	Support
CV1	1.00	1.00	1.00	398
CV2	1.00	1.00	1.00	398
CV3	1.00	1.00	1.00	398
CV4	1.00	1.00	1.00	398
CV5	1.00	1.00	1.00	397

Table 60

Classification Report for Class "0":

Fold	Precision	Recall	F1 score	Support
CV1	1.00	1.00	1.00	192
CV2	1.00	1.00	1.00	192
CV3	1.00	1.00	1.00	192
CV4	1.00	1.00	1.00	191
CV5	1.00	1.00	1.00	192

Table 61

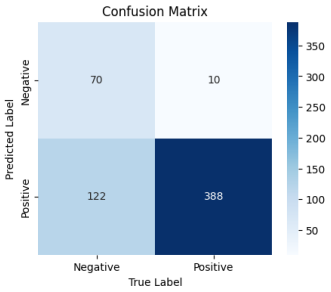
Artificial Neural Network

Accuracy Table:

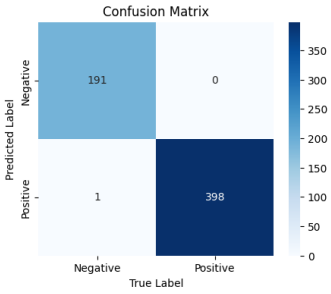
Fold	Accuracy
CV1	77.63%
CV2	99.83%
CV3	100.00%
CV4	100.00%
CV5	100.00%
Average	**95.49%**

Table 62

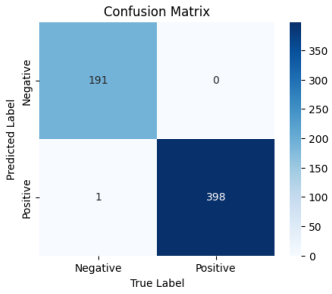
Fold 1



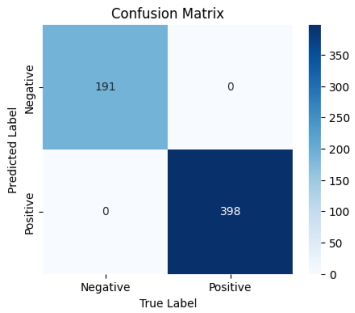
Fold 2



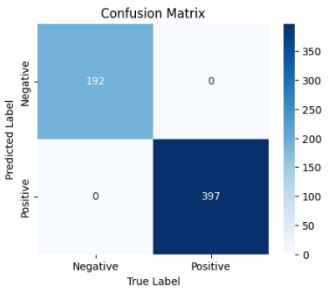
Fold 3



Fold 4



Fold 5



Classification Report for Class "1":

Fold	Precision	Recall	F1 score	Support
CV1	0.76	0.97	0.85	398
CV2	1.00	1.00	1.00	398
CV3	1.00	1.00	1.00	398
CV4	1.00	1.00	1.00	398
CV5	1.00	1.00	1.00	397

Table 63

Classification Report for Class "0":

Fold	Precision	Recall	F1 score	Support
CV1	0.88	0.36	0.51	192
CV2	1.00	0.99	1.00	192
CV3	1.00	1.00	1.00	192
CV4	1.00	1.00	1.00	191
CV5	1.00	1.00	1.00	192

Table 64

Comparison:

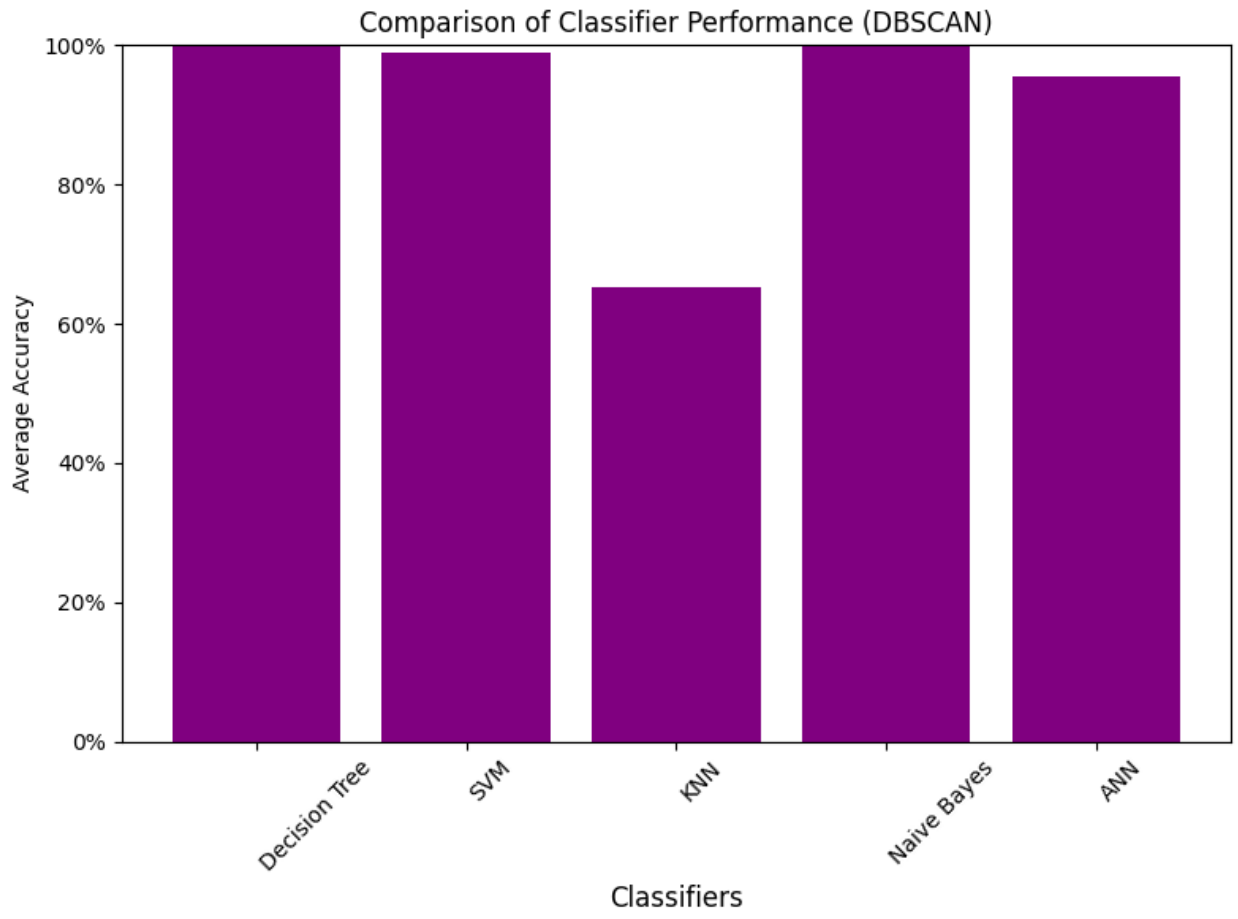


Figure 10

Hierarchical Dataset

Dataset Overview: This dataset is created by the clustered class using Hierarchical clustering and turned into supervised dataset.

Dataset Splitting: Using 5 folds splits k-means dataset where train set contains 2911 data and test set contains 728 data but last fold contains 727 data. Each fold maintains the class ratio which is 71.1% for class "1" and 28.9% for class "0".

Distribution of Result Class:

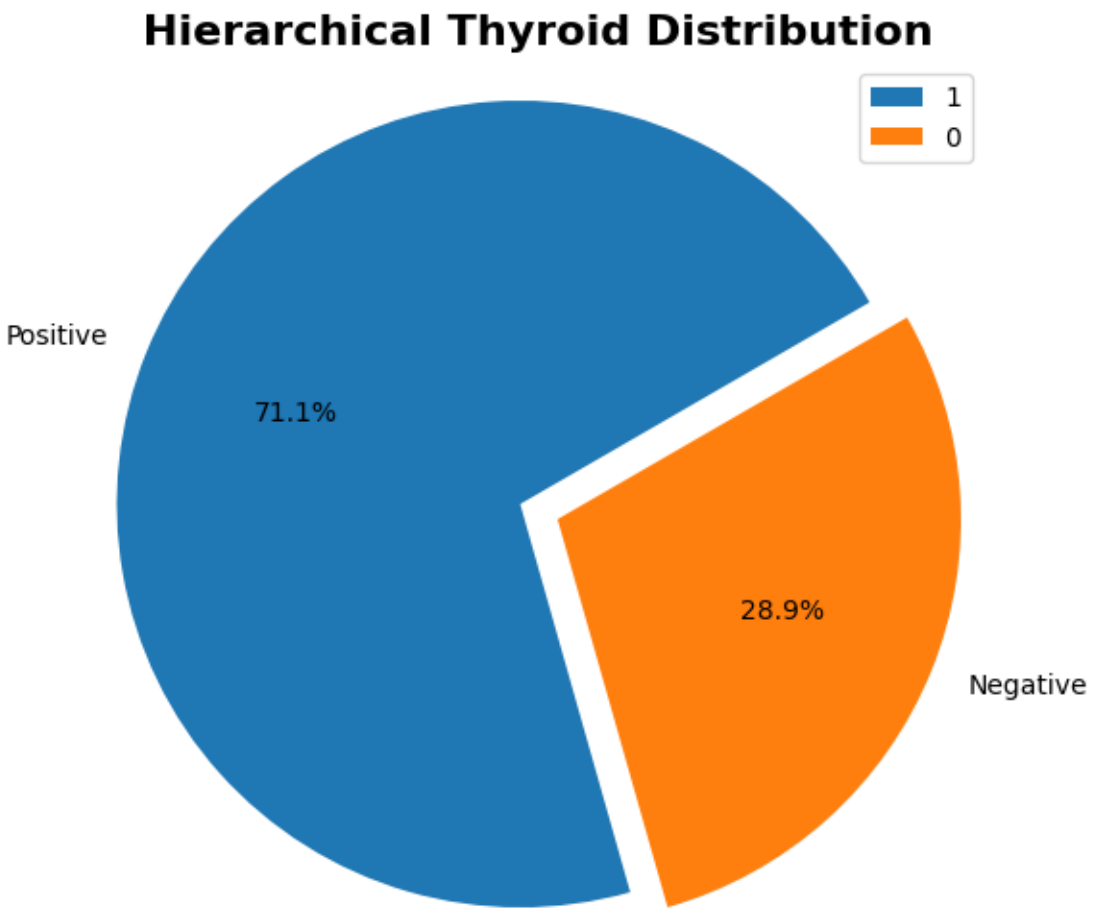


Figure 11

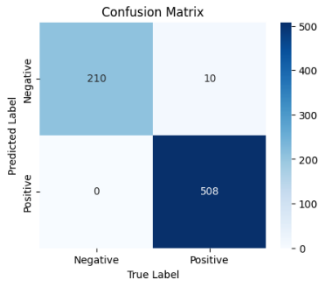
Decision Tree

Accuracy Table:

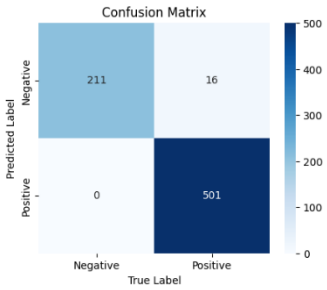
Fold	Accuracy
CV1	98.63%
CV2	97.80%
CV3	98.35%
CV4	97.80%
CV5	96.84%
Average	**97.88%**

Table 65

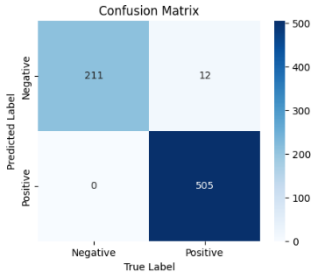
Fold 1



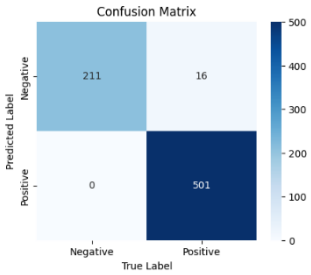
Fold 2



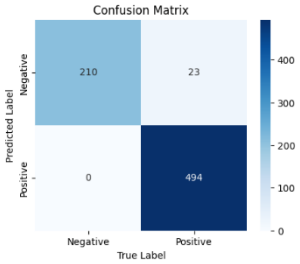
Fold 3



Fold 4



Fold 5



Classification Report for Class "1":

Fold	Precision	Recall	F1 score	Support
CV1	1.00	0.98	0.99	518
CV2	1.00	0.97	0.98	517
CV3	1.00	0.98	0.99	517
CV4	1.00	0.97	0.98	517
CV5	1.00	0.96	0.98	517

Table 66

Classification Report for Class "0":

Fold	Precision	Recall	F1 score	Support
CV1	0.95	1.00	0.98	210
CV2	0.93	1.00	0.96	211
CV3	0.95	1.00	0.97	211
CV4	0.93	1.00	0.96	211
CV5	0.90	1.00	0.95	210

Table 67

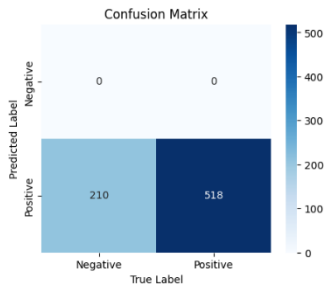
Support Vector Machine

Accuracy Table:

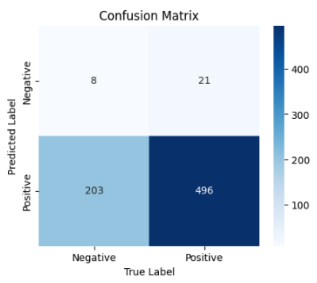
Fold	Accuracy
CV1	71.15%
CV2	71.02%
CV3	71.02%
CV4	71.02%
CV5	70.11%
Average	**70.71%**

Table 68

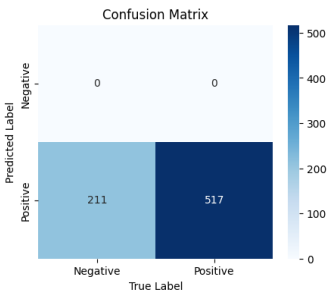
Fold 1



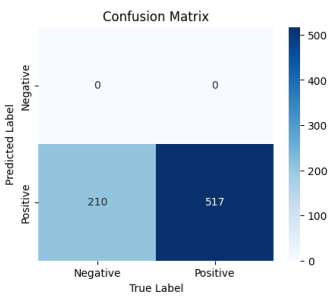
Fold 2



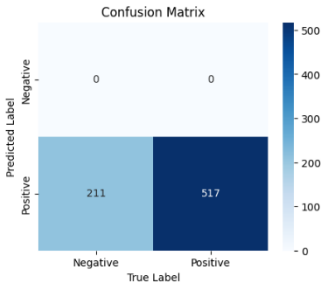
Fold 3



Fold 4



Fold 5



Classification Report for Class “1”:

Fold	Precision	Recall	F1 score	Support
CV1	0.71	1.00	0.83	518
CV2	0.71	0.96	0.82	517
CV3	0.71	1.00	0.83	517
CV4	0.71	1.00	0.83	517
CV5	0.71	1.00	0.83	517

Table 69

Classification Report for Class “0”:

Fold	Precision	Recall	F1 score	Support
CV1	0.00	0.00	0.00	210
CV2	0.28	0.04	0.07	211
CV3	0.00	0.00	0.00	211
CV4	0.00	0.00	0.00	211
CV5	0.00	0.00	0.00	210

Table 70

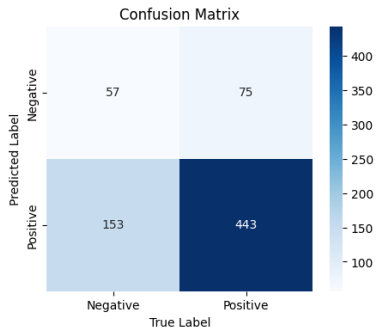
K nearest neighbors

Accuracy Table:

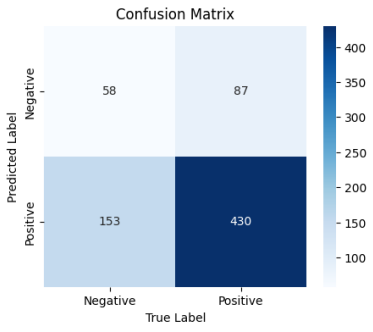
Fold	Accuracy
CV1	68.68%
CV2	67.03%
CV3	66.07%
CV4	68.41%
CV5	68.23%
Average	**67.78%**

Table 71

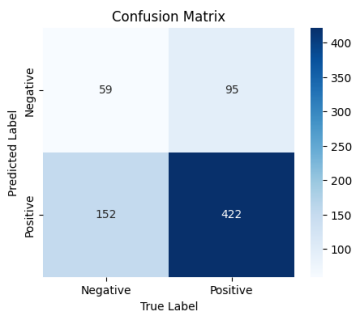
Fold 1



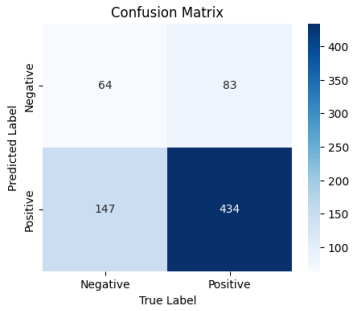
Fold 2



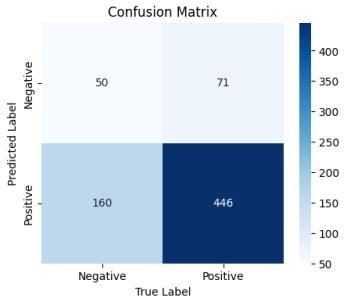
Fold 3



Fold 4



Fold 5



Classification Report for Class "1":

Fold	Precision	Recall	F1 score	Support
CV1	0.74	0.86	0.80	518
CV2	0.74	0.83	0.78	517
CV3	0.74	0.82	0.77	517
CV4	0.75	0.84	0.79	517
CV5	0.74	0.86	0.79	517

Table 72

Classification Report for Class "0":

Fold	Precision	Recall	F1 score	Support
CV1	0.43	0.27	0.33	210
CV2	0.40	0.27	0.33	211
CV3	0.38	0.28	0.32	211
CV4	0.44	0.30	0.36	211
CV5	0.41	0.24	0.30	210

Table 73

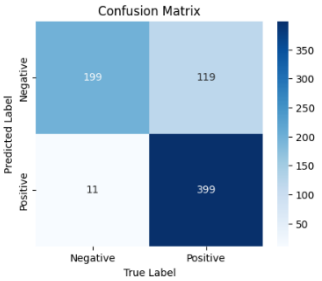
Naïve Bayes

Accuracy Table:

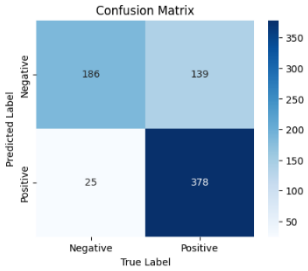
Fold	Accuracy
CV1	82.14%
CV2	77.47%
CV3	84.07%
CV4	79.95%
CV5	86.93%
Average	**82.11%**

Table 74

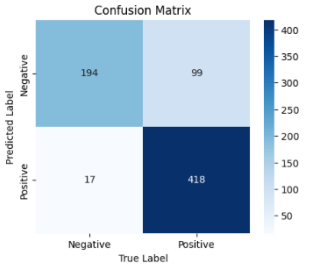
Fold 1



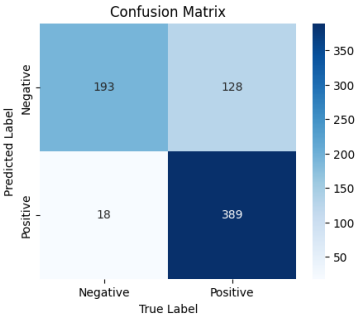
Fold 2



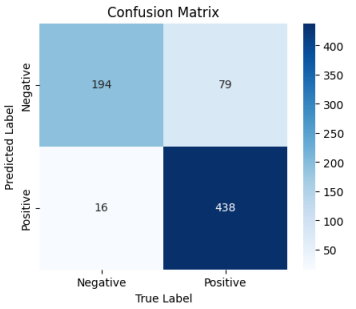
Fold 3



Fold 4



Fold 5



Classification Report for Class "1":

Fold	Precision	Recall	F1 score	Support
CV1	0.97	0.77	0.86	518
CV2	0.94	0.73	0.82	517
CV3	0.96	0.81	0.88	517
CV4	0.96	0.75	0.84	517
CV5	0.96	0.85	0.90	517

Table 75

Classification Report for Class "0":

Fold	Precision	Recall	F1 score	Support
CV1	0.63	0.95	0.75	210
CV2	0.57	0.88	0.69	211
CV3	0.66	0.92	0.77	211
CV4	0.60	0.91	0.73	211
CV5	0.71	0.92	0.80	210

Table 76

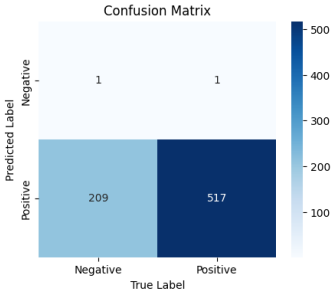
Artificial Neural Network

Accuracy Table:

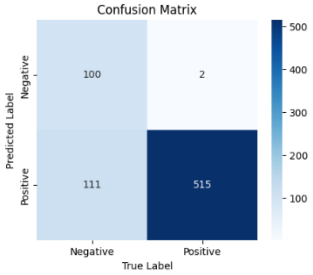
Fold	Accuracy
CV1	71.15%
CV2	84.48%
CV3	98.63%
CV4	99.86%
CV5	100.00%
Average	**90.82%**

Table 77

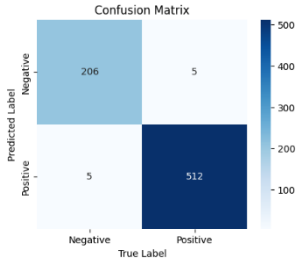
Fold 1



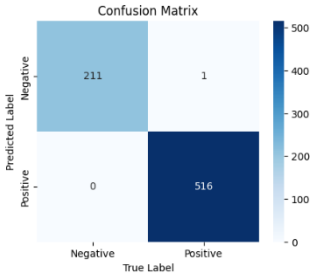
Fold 2



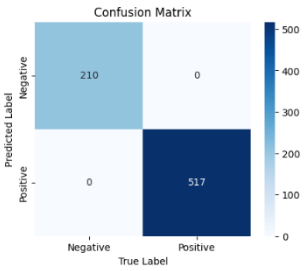
Fold 3



Fold 4



Fold 5



Classification Report for Class "1":

Fold	Precision	Recall	F1 score	Support
CV1	0.71	1.00	0.83	518
CV2	0.82	1.00	0.90	517
CV3	0.99	0.99	0.99	517
CV4	1.00	1.00	1.00	517
CV5	1.00	1.00	1.00	517

Table 78

Classification Report for Class "0":

Fold	Precision	Recall	F1 score	Support
CV1	0.50	0.00	0.01	210
CV2	0.98	0.47	0.64	211
CV3	0.98	0.98	0.98	211
CV4	1.00	1.00	1.00	211
CV5	1.00	1.00	1.00	210

Table 79

Comparison:

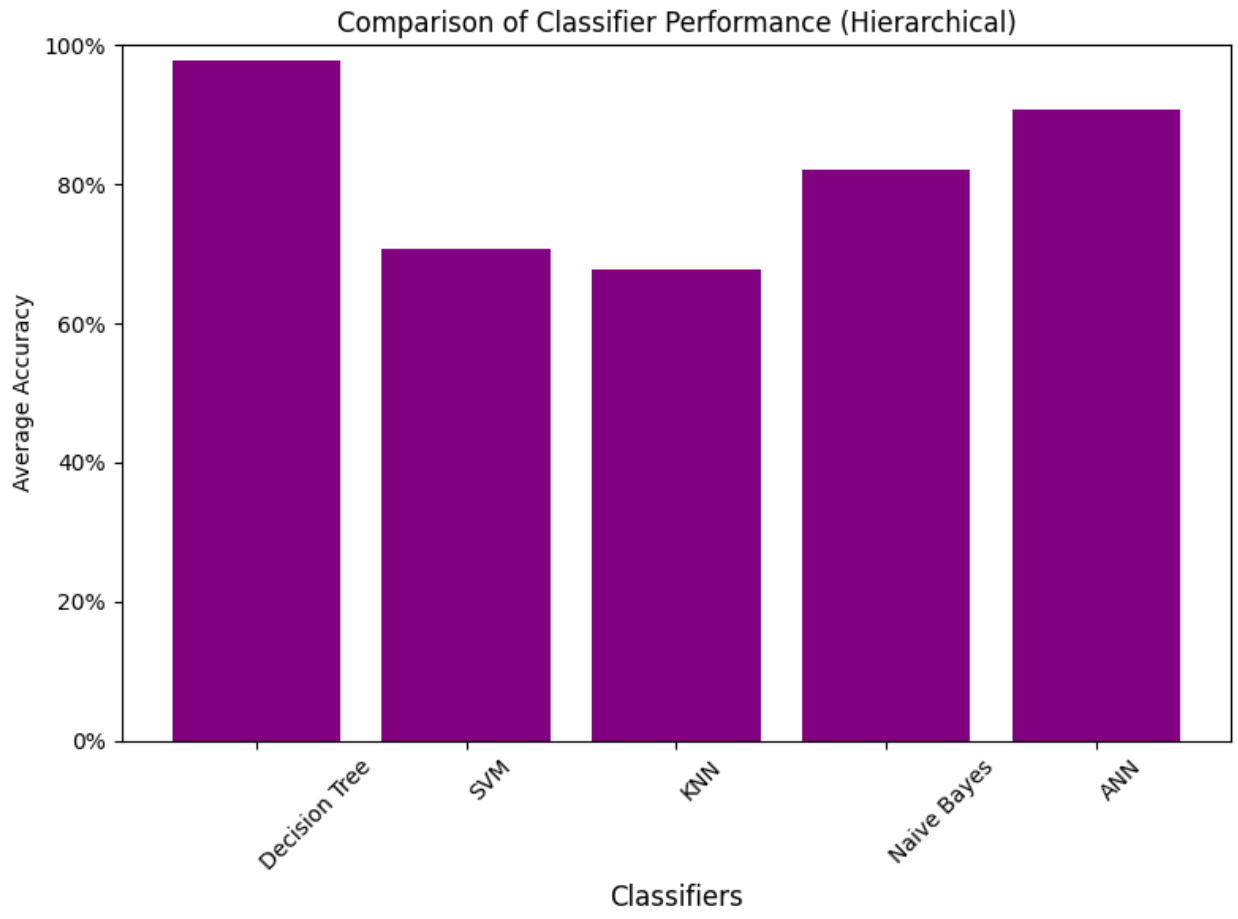


Figure 12

PERFORMANCE COMPARISON

The overall comparison of the cross-validation process on each dataset is shown below:

Accuracy Table:

	Original	K-means	K-medoids	DBSCAN	Hierarchical
DT	92.00%	100.00%	81.37%	100.00%	97.88%
SVM	84.34%	58.08%	48.07%	98.88%	67.71%
KNN	92.94%	66.09%	85.66%	65.33%	67.68%
NB	92.00%	100.00%	69.30%	100.00%	87.11%
ANN	92.20%	94.09%	80.96%	95.49%	90.82%

Table 80

We can see that the accuracy in the original dataset was good in all classifier but the precision for 0 class is not at all good. This is because of the minority of 0 class.

Again, SVM classifier performed comparatively better for DBSCAN as this dataset was free of outliers and had less data point. So, we can say DBSCAN clustering did a good job for improving performance and others for improving class precision.

CONCLUSION

Cross-validation is a powerful technique for evaluating and validating machine learning models. It helps to prevent overfitting and provides a more accurate estimate of the model's performance. By following some best practices, we can get the most out of cross-validation and improve the quality of our models. We finally got an idea of the impact of cross validation and could detect sum difference. In conclusion, mastering the implementation of cross-validation empowers practitioners to build models that not only perform well on training data but also generalize effectively to new and unseen instances, making them robust and reliable in real-world applications.