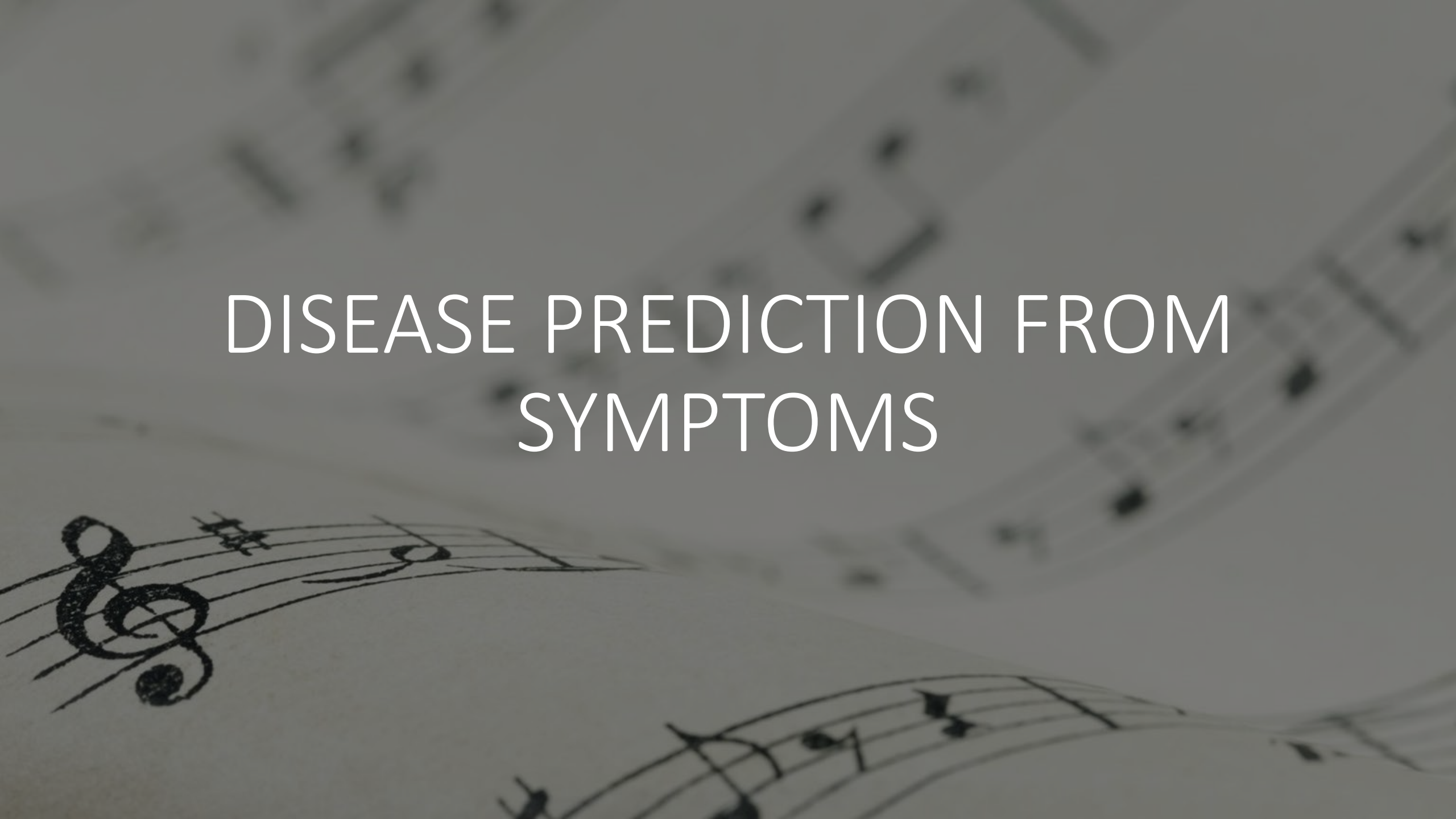

MACHINE LEARNING



DISEASE PREDICTION FROM SYMPTOMS

TEAM

Name	Roll Number
D.KRISHNA MURTHY	CB.EN.U4CSE21016
MEENAKSHI S	CB.EN.U4CSE21035
PICHERI LIKITHA	CB.EN.U4CSE21044
RAGALA TEJDEEP	CB.EN.U4CSE21046

Introduction

- People are currently suffering from a variety of diseases.
- Many people are unsure if the symptoms they are experiencing are indicative of a certain disease, and hence they are unable to take the required safeguards.
- Disease prediction of a human means predicting the probability of a patient's disease after examining the combinations of the patient's symptoms.

Motivation/significance

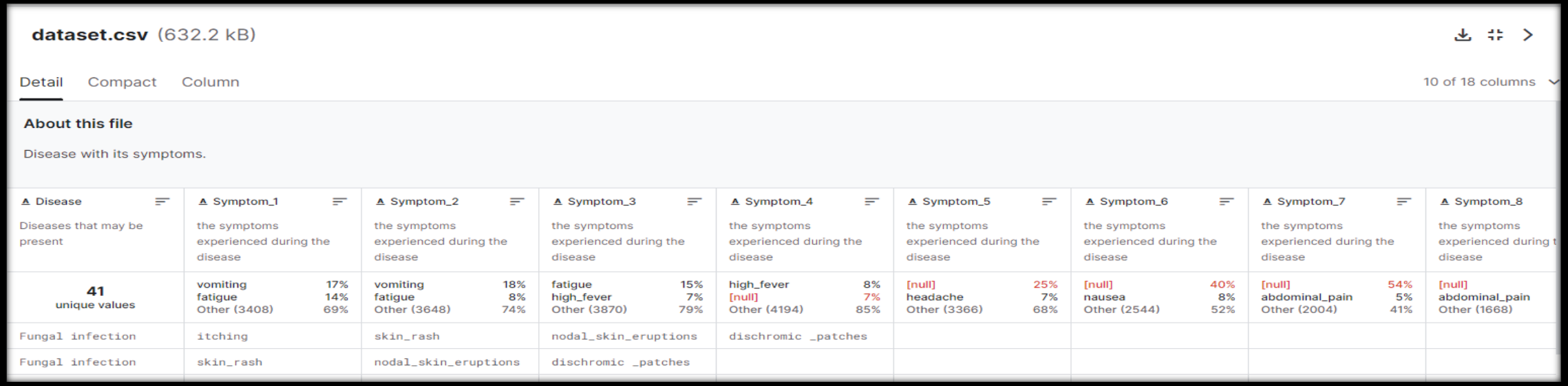
- The major goal of this project is to greatly aid physicians in predicting and diagnosing diseases at an early stage, as well as to design a model that will assist people in identifying diseases at an early stage so that risk may be decreased
- This analysis in the medical industry would lead to a streamlined and expedited treatment of patients

Dataset Considered

DATASET DESCRIPTION

- We used the below dataset from Kaggle.
- There are columns containing diseases, their symptoms and their weights.

- https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset/data?select=dataset.csv
  https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset/data?select=Symptom-severity.csv



dataset.csv (632.2 kB)

Detail | Compact | Column | 10 of 18 columns

About this file

Disease with its symptoms.

Disease	Symptom_1	Symptom_2	Symptom_3	Symptom_4	Symptom_5	Symptom_6	Symptom_7	Symptom_8
Diseases that may be present	the symptoms experienced during the disease	the symptoms experienced during the disease	the symptoms experienced during the disease	the symptoms experienced during the disease	the symptoms experienced during the disease	the symptoms experienced during the disease	the symptoms experienced during the disease	the symptoms experienced during t disease
41 unique values	vomiting 17% fatigue 14% Other (3408) 69%	vomiting 18% fatigue 8% Other (3648) 74%	fatigue 15% high_fever 7% Other (3870) 79%	high_fever 8% [null] 7% Other (4194) 85%	[null] 25% headache 7% Other (3366) 68%	[null] 40% nausea 8% Other (2544) 52%	[null] 54% abdominal_pain 5% Other (2004) 41%	[null] abdominal_pain Other (1668)
Fungal infection	itching	skin_rash	nodal_skin_eruptions	dischromic _patches				
Fungal infection	skin_rash	nodal_skin_eruptions	dischromic _patches					

Dataset description

- We used two csv files for predicting the disease. One csv file contains the symptoms and the disease columns and the next csv file contains the weights of the respective symptom.
- Disease csv file contains 18 features , in which 17 columns are the symptoms and the output is the disease of the respective symptoms
- The weight csv file contains the weight of the respective symptoms which is used for filling NAN symptoms values during encoding

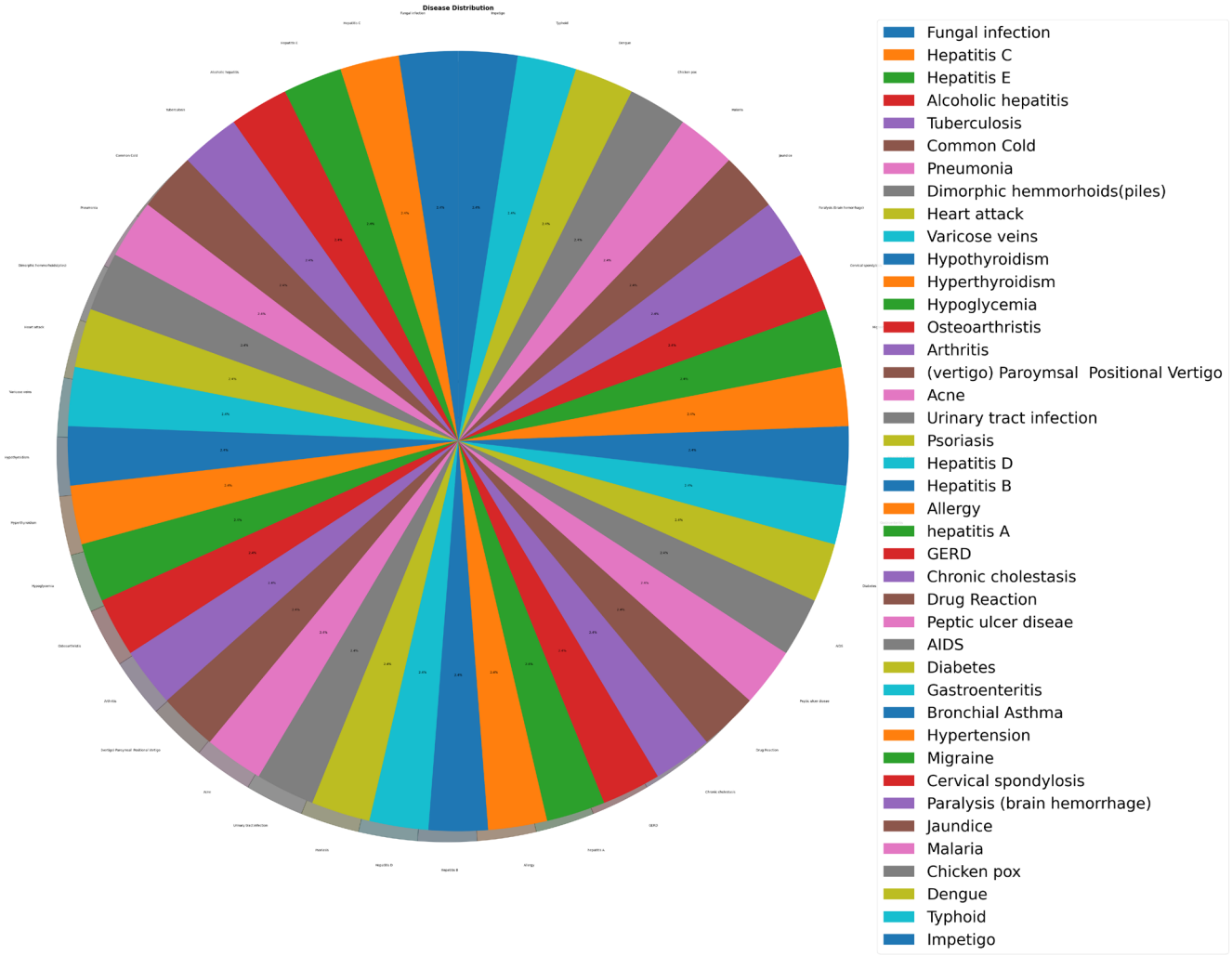
DATA PREPROCESSING

- After collecting that data, as that data is raw data we have to make it suitable for training our machine learning model. By using some python libraries like NumPy, and pandas, we have made that data suitable for machine learning models

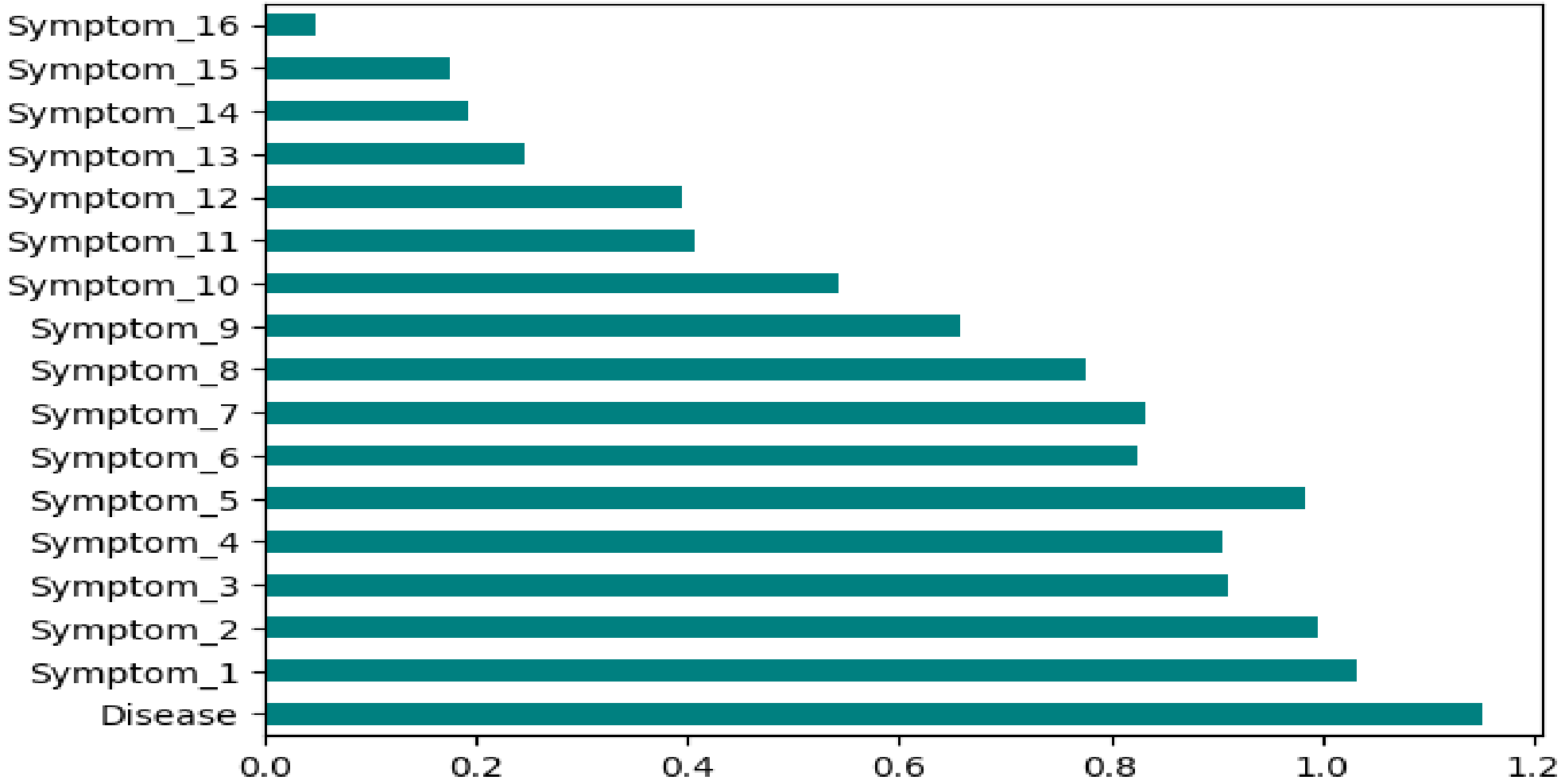
- The collected data are preprocessed for the availability of missing values in most of the structured data. Hence, it is essential to fill out the missed data or remove or modify them to enhance the quality of the data set. The preprocessing step also eliminates the commas, punctuations, and white spaces. Once the preprocessing of data has been completed, it is then subjected to feature extraction followed by disease prediction
- The numerical features are scaled using StandardScaler and MinMaxScaler to ensure that they have the same scale and to improve the performance of the model.

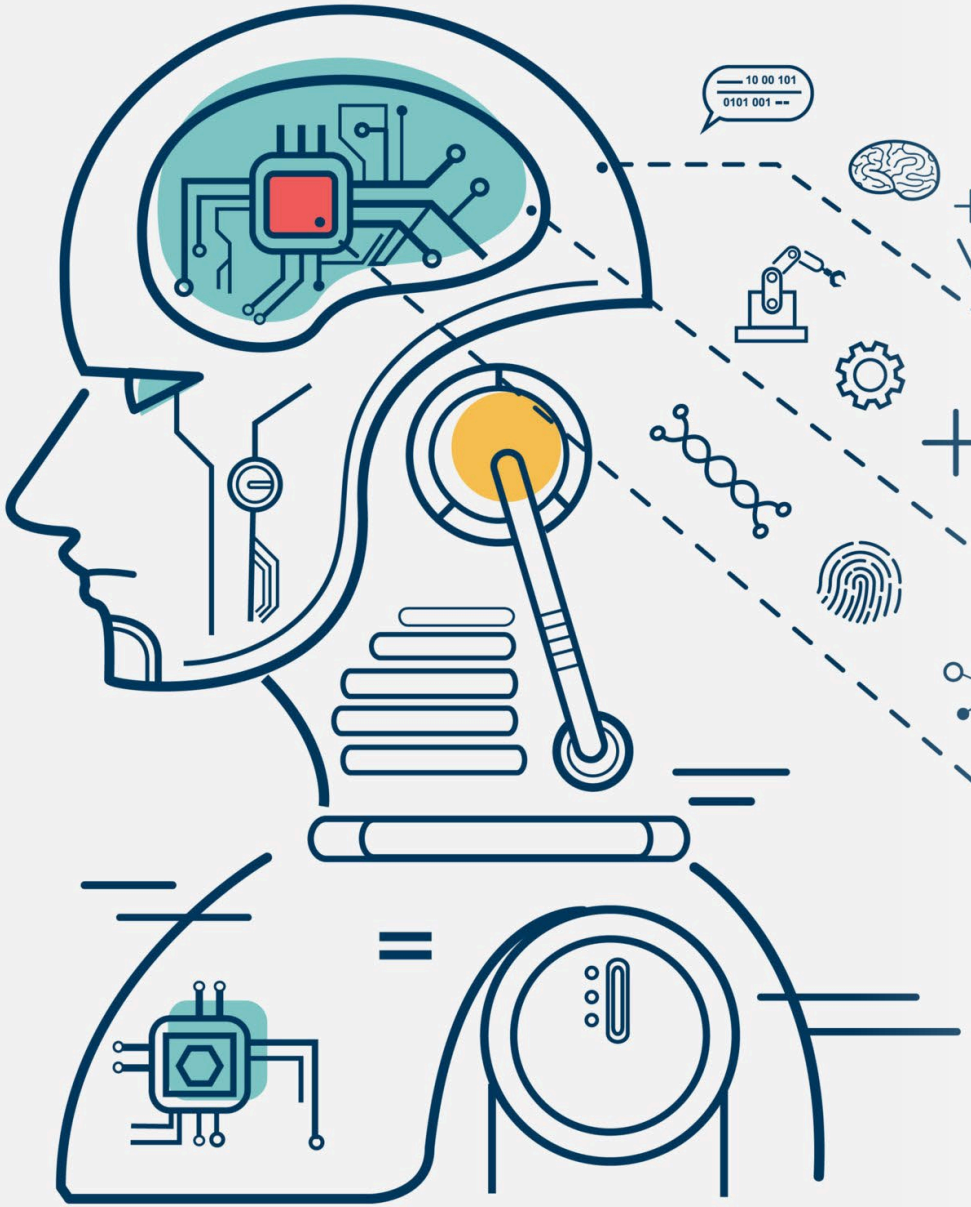

Some Visualizations

Visualization of Diseases



FEATURE SELECTION



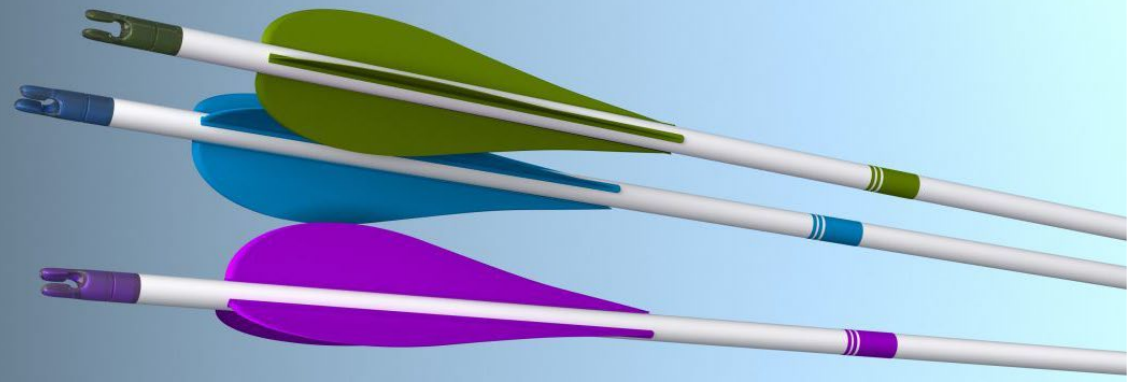


Machine Learning Algorithms Used

- Random Forest Classifier
- K-Nearest Neighbour Classifier
- Decision Tree Classifier
- SVM
- LOGISTIC REGRESSION
- LINEAR REGRESSION
- RANDOM FOREST
- ENSEMBLE LEARNING
- NAÏVE BAYES

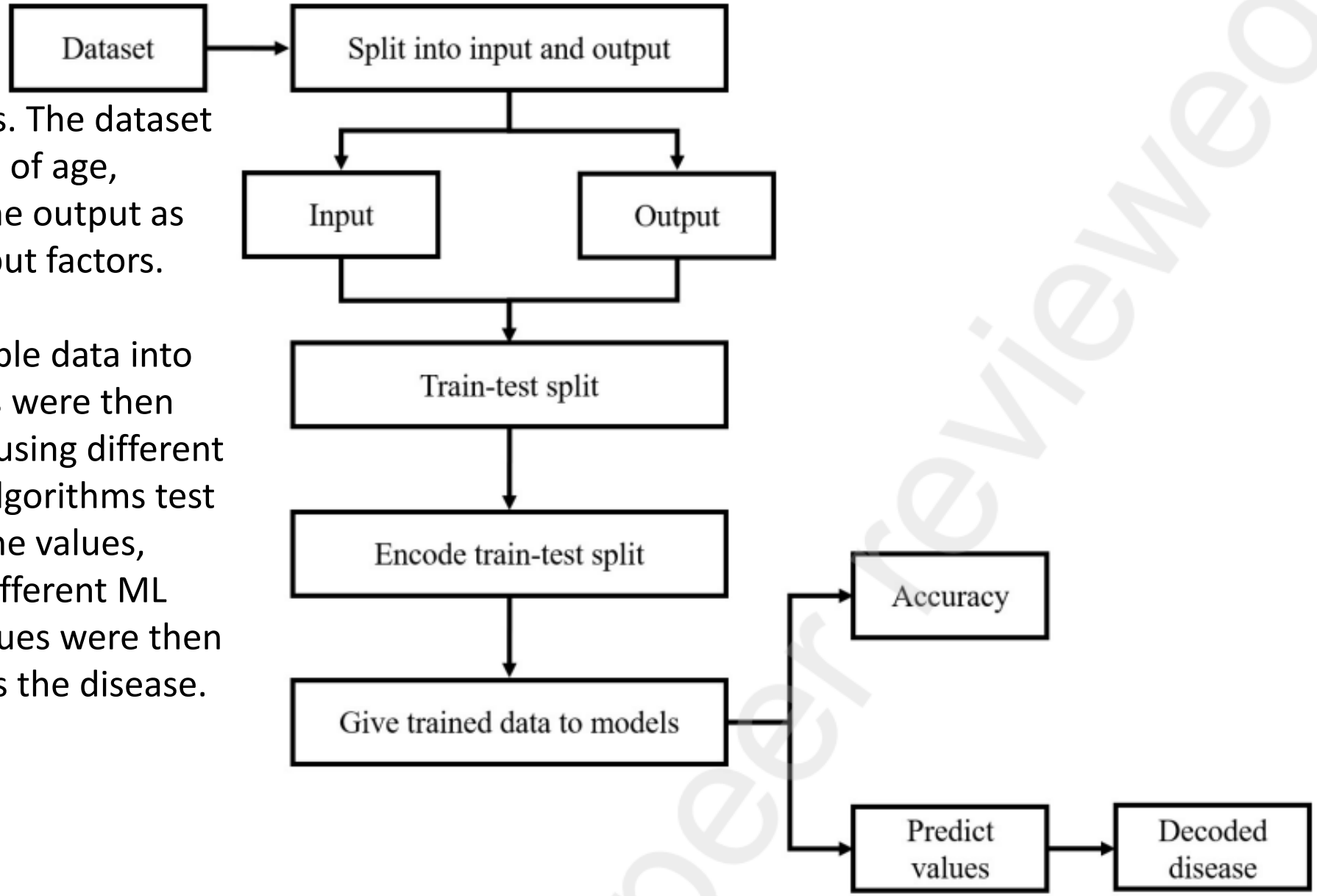
Performance Metrics Considered

- Accuracy
- Precision
- Recall
- ROC-AUC Score
- ROC Graphs
- Confusion Matrix



Functioning of the ML models. The dataset was split into input consisting of age, gender, and symptoms and the output as the diseases based on the input factors.

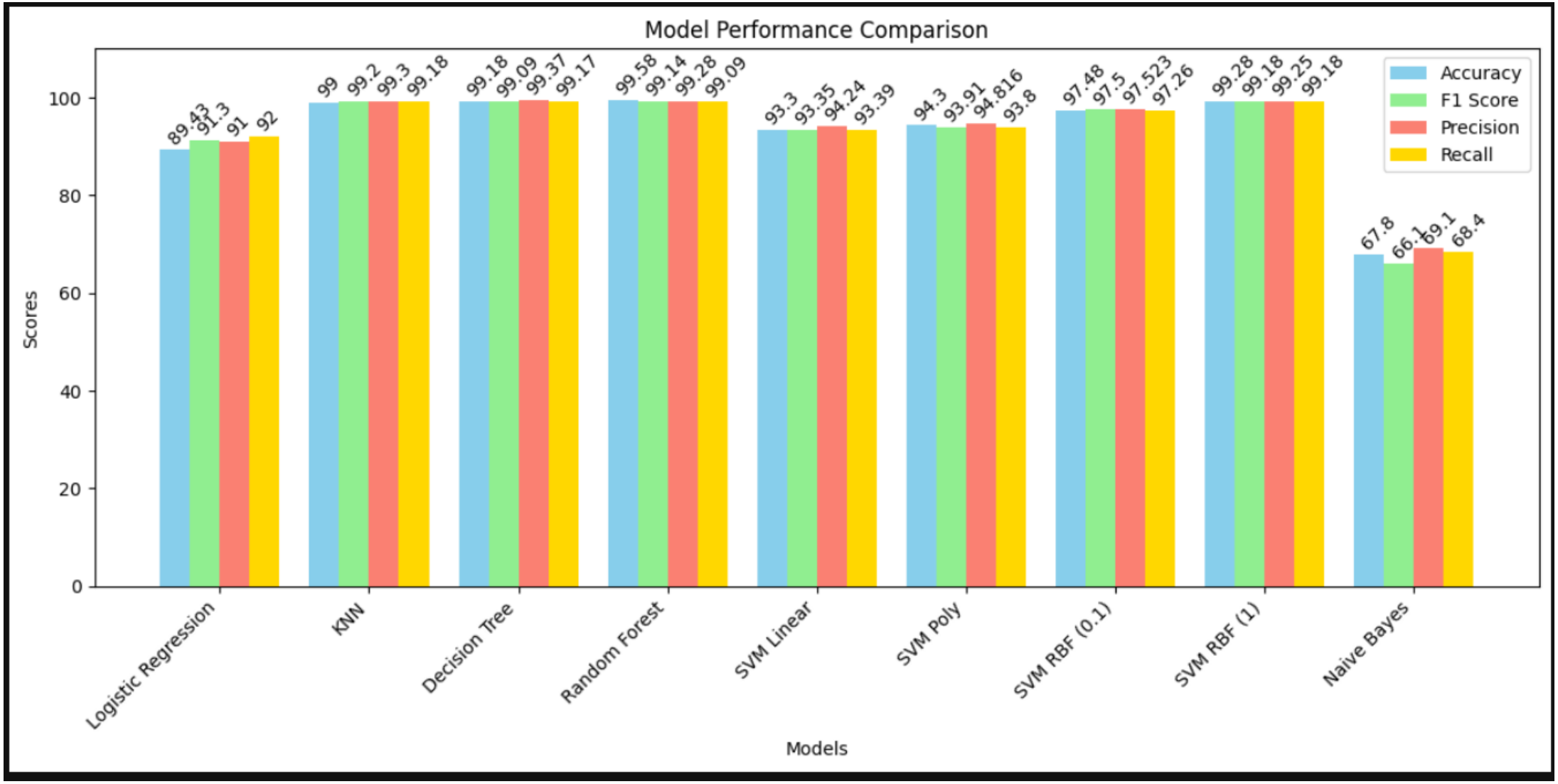
We randomly split the available data into train and test sets. These sets were then encoded and further trained using different algorithms. After which the algorithms test the training set and predict the values, resulting in the accuracy of different ML algorithms. The predicted values were then decoded to give the output as the disease.



Comparision Of Performance metrics of different models

S.NO	MODEL	ACCURACY	F1 SCORE	PRECISION	RECALL
1.	LOGISTIC REGRESSION	89.43%	91.3	91	92
2.	KNN(K=4)	99%	99.2	99.3	99.18
3.	DECISION TREE	99.18%	99.09	99.37	99.17
4.	RANDOM FOREST	99.58%	99.14	99.28	99.09
5.	SVM-LINEAR	93.3%	93.35	94.24	93.39
6.	SVM-POLYNOMIAL (DEGREE=3)	94.3%	93.91	94.816	93.8
7.	SVM-RBF KERNAL (GAMMA=0.1)	97.48%	97.5	97.523	97.26
8.	SVM-RBF KERNAL (GAMMA=1)	99.28%	99.18	99.25	99.18
9.	NAIVE BAYES	67.8%	66.1	69.1	68.4

- COMPARATIVE ANALYSIS

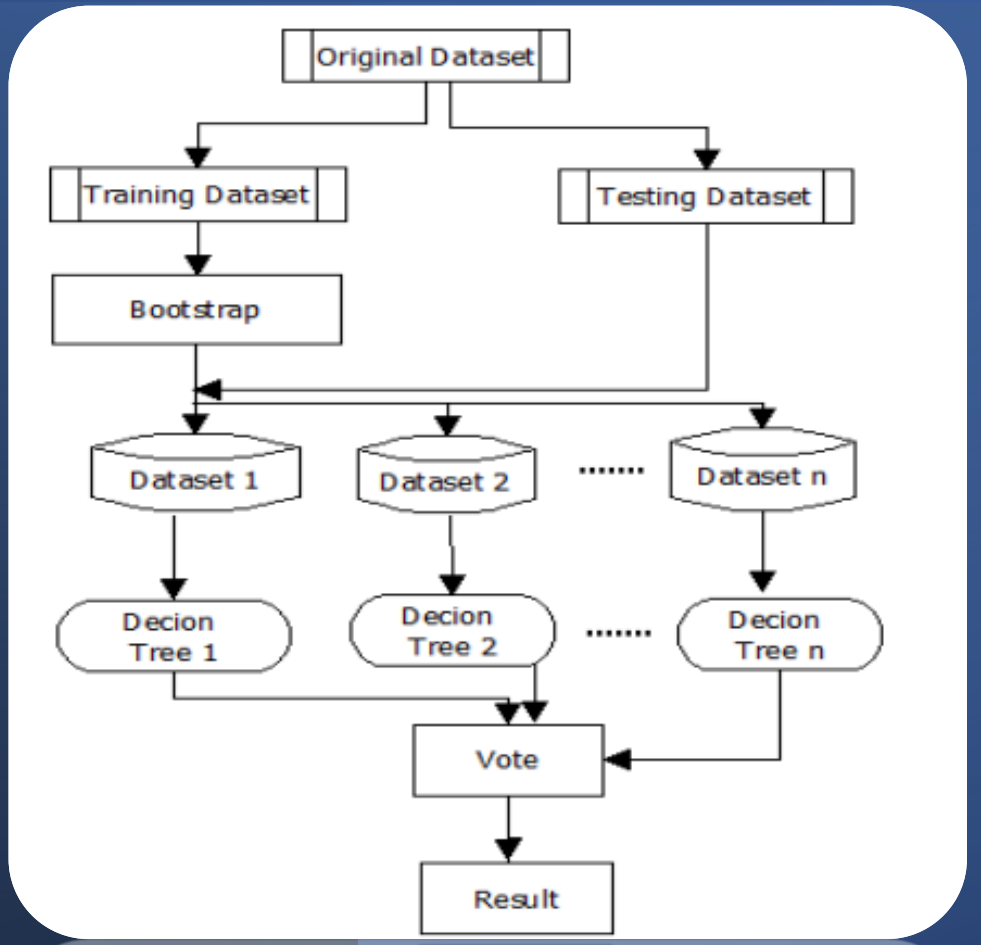


INFERENCE

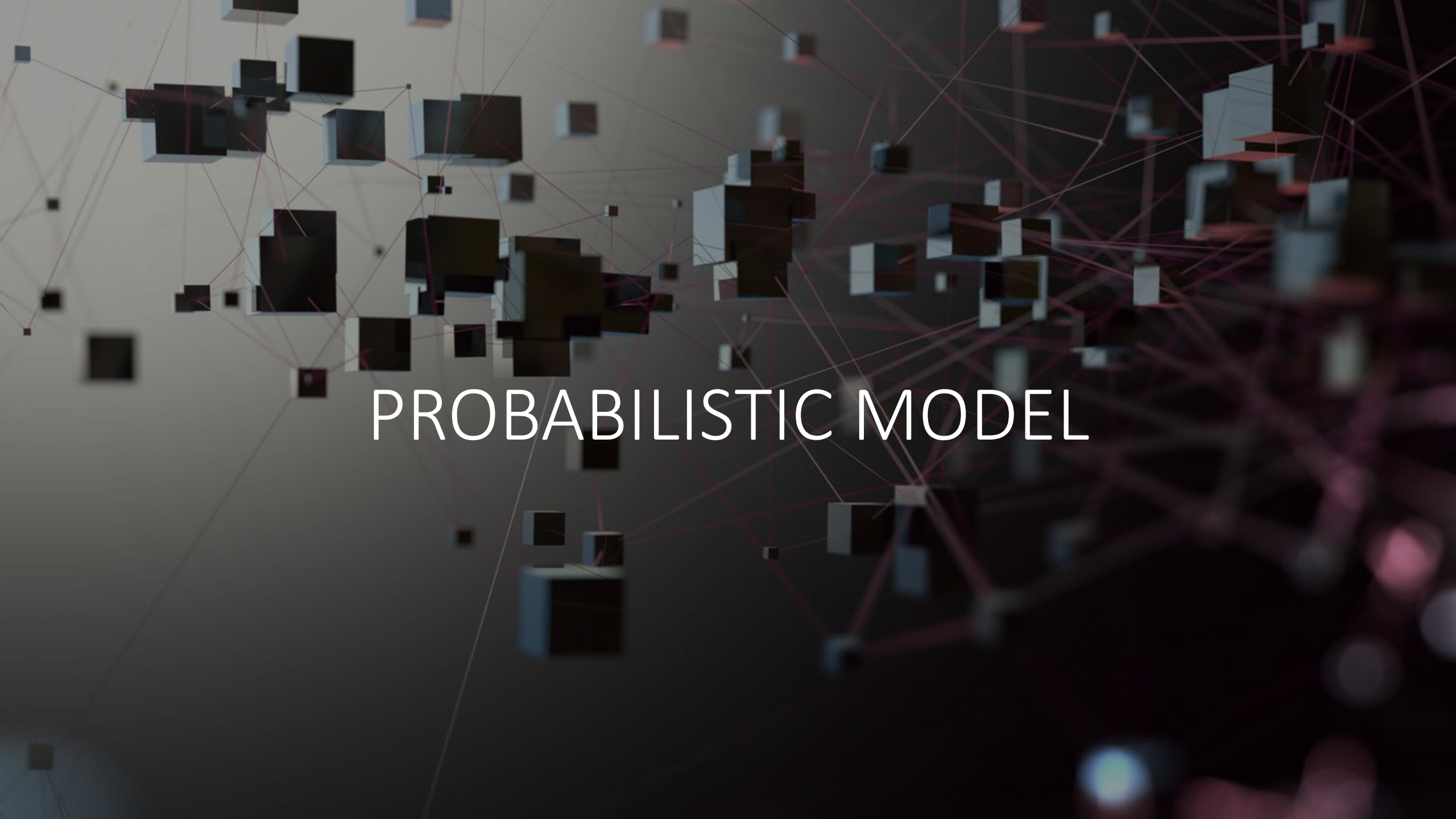
Different machine learning models were used to examine the prediction of disease for available input dataset.
The table shows the comparative analysis of the accuracy of the training models. From the earlier necessities, Random Forest Algorithm was best with a model accuracy of 99.58%.
Following the Random Forest is the Decision Tree with an accuracy of 99.18%.
The SVM and KNN models were also very close.
However, the suggested model, that is Random forest model, yields the most accurate result, 99.5% as compared to earlier methods.
Among all the models, we gained the highest accuracy for the Random Forest model of 99.58 %.

Advantage of Random Forest:

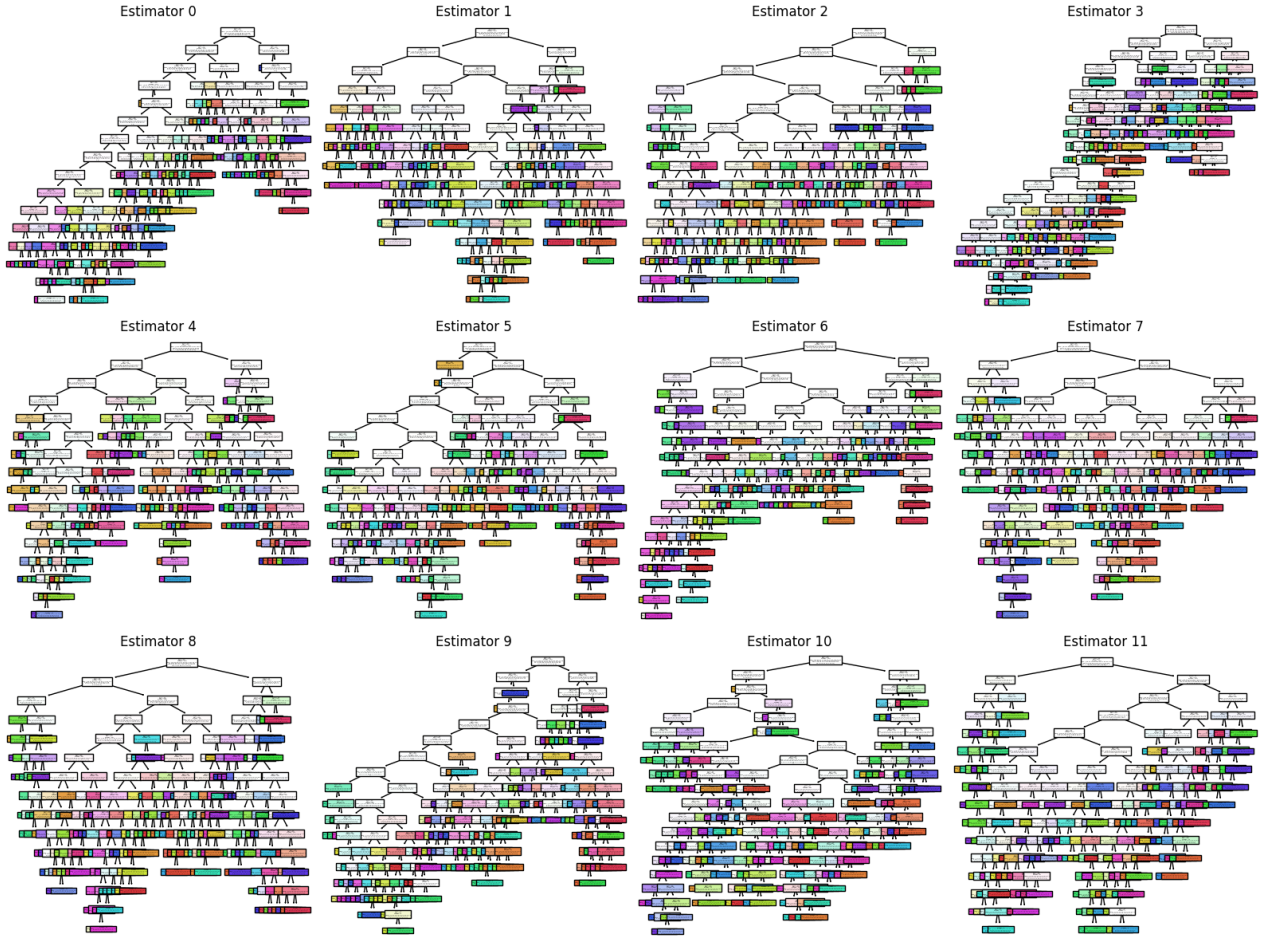
- Different machine learning models were used to examine the prediction of disease for available input dataset.
- While training the model, the decision forests that are formed while concluding are pruned as soon as they encounter a weak symptom or a symptom that does not occur in a location.
- Thus, Random Forest Algorithm minimizes the cost whilst predicting a more realistic model.

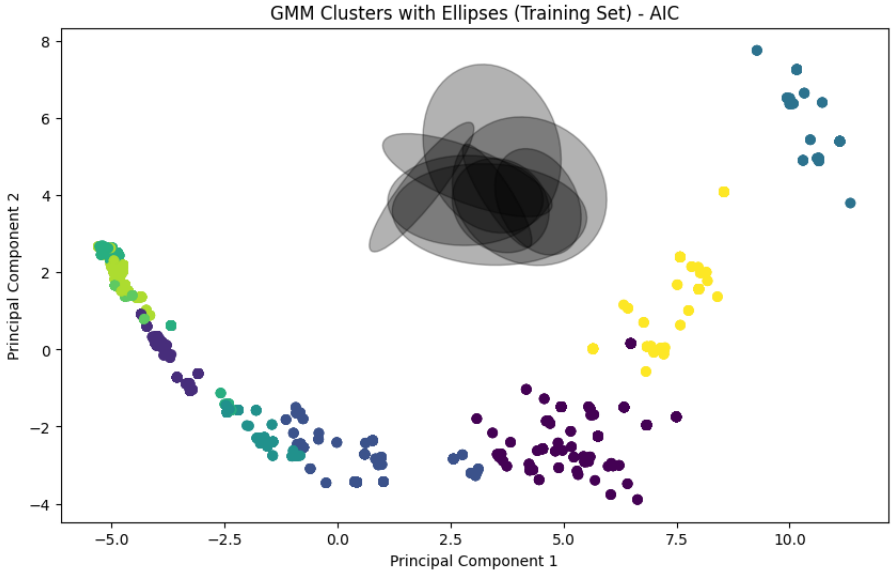


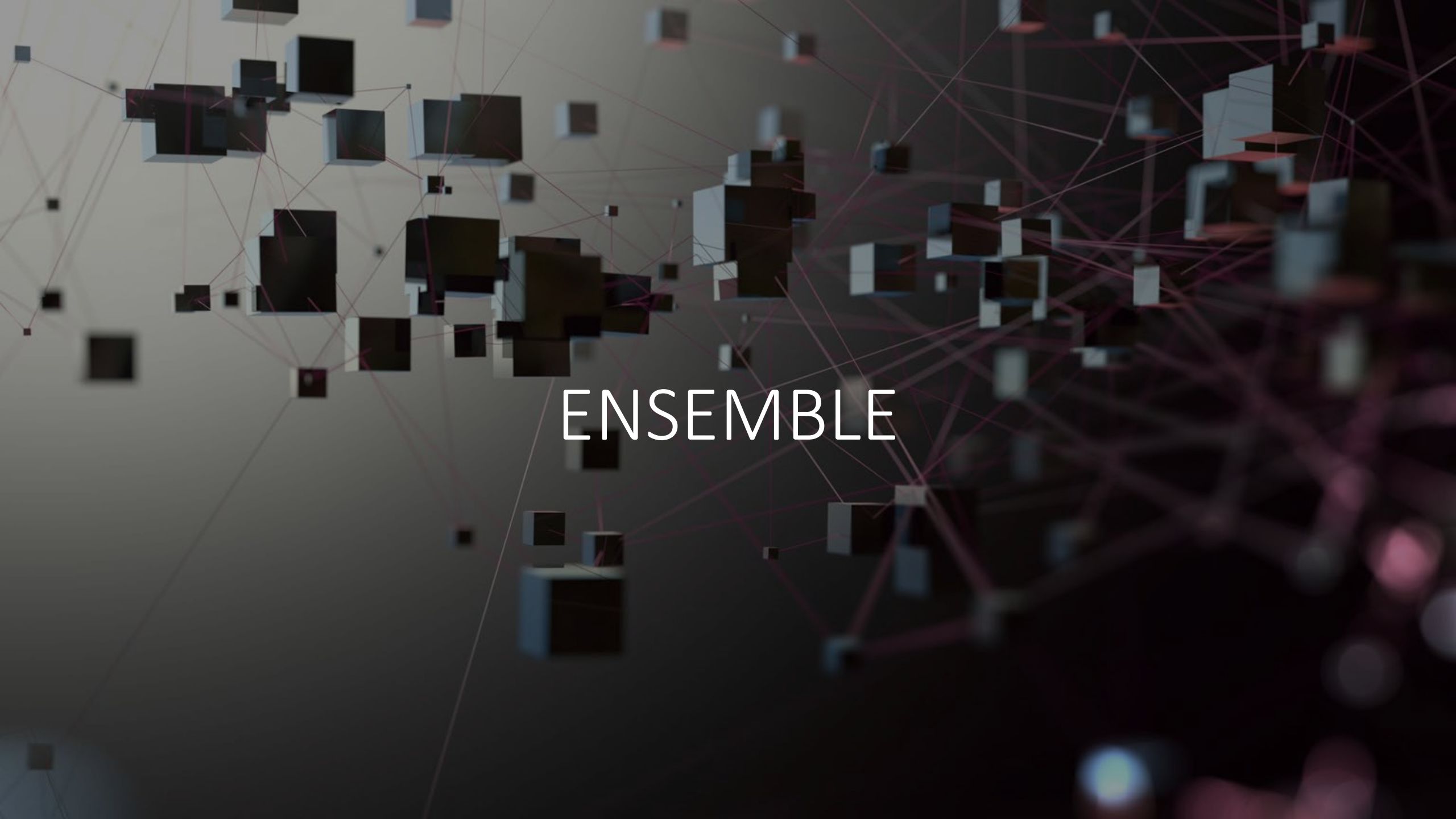
REVIEW 2

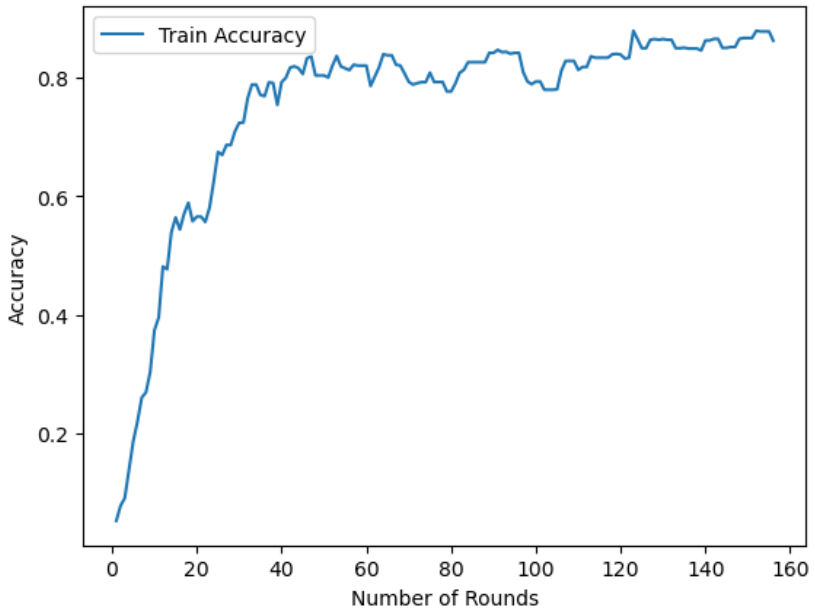

PROBABILISTIC MODEL

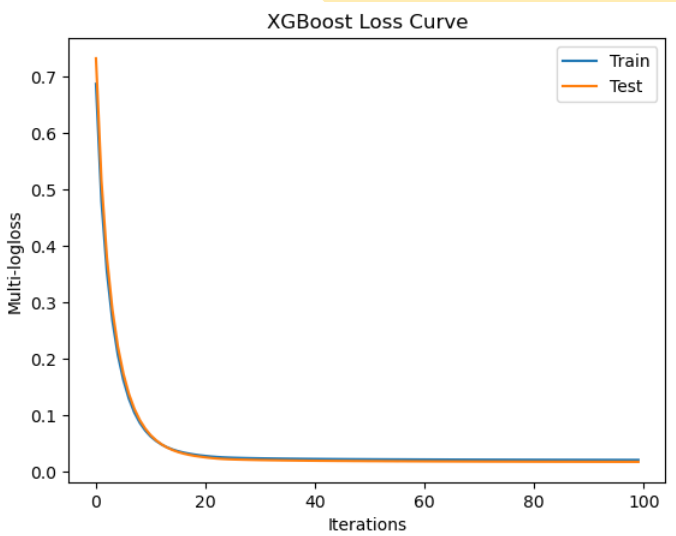
Random forest

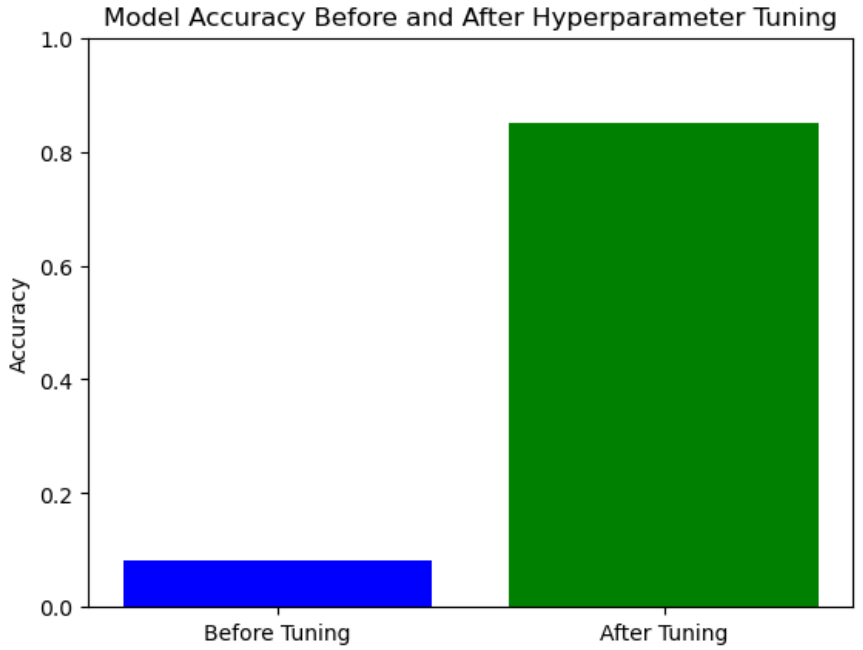


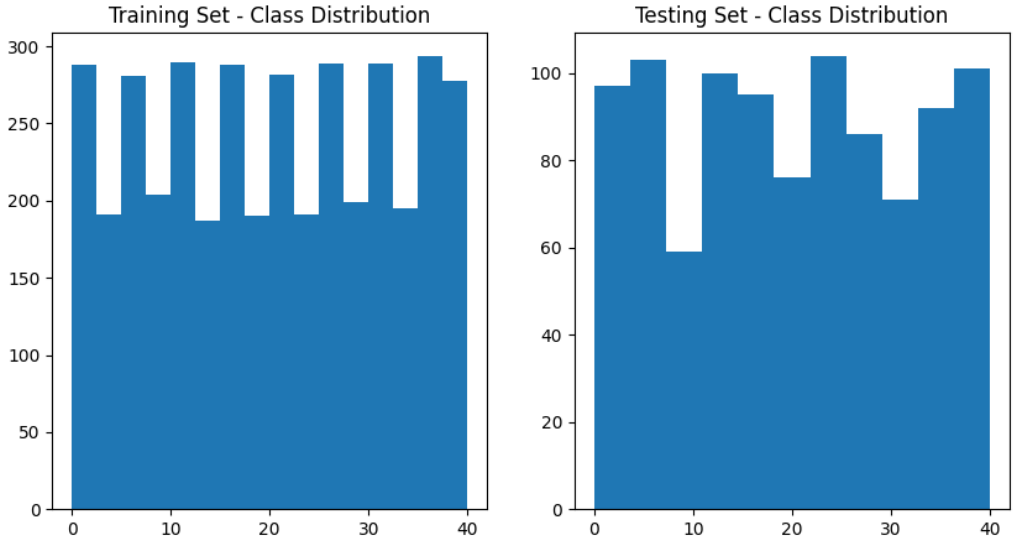


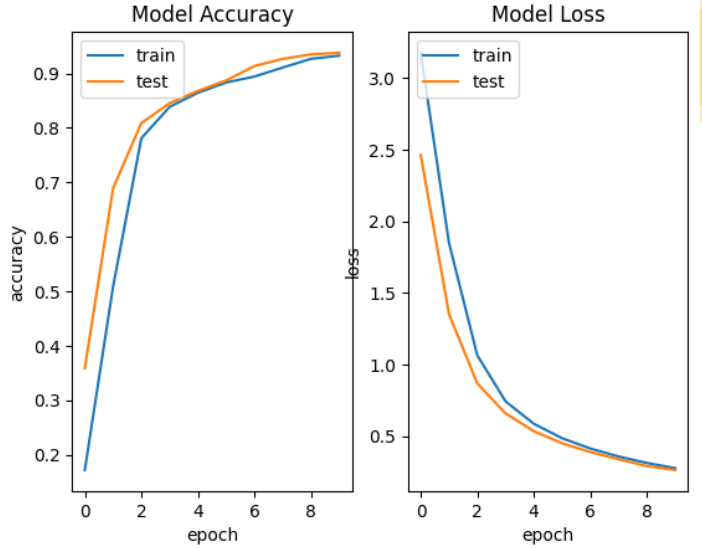

ENSEMBLE


Train Accuracy
Accuracy
Number of Rounds
0.2
0.4
0.6
0.8
0
20
40
60
80
100
120
140
160


XGBoost Loss Curve
Train
Test
Multi-logloss
Iterations
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
20
40
60
80
100


Model Accuracy Before and After Hyperparameter Tuning
Accuracy
Before Tuning
After Tuning
0.0
0.2
0.4
0.6
0.8
1.0



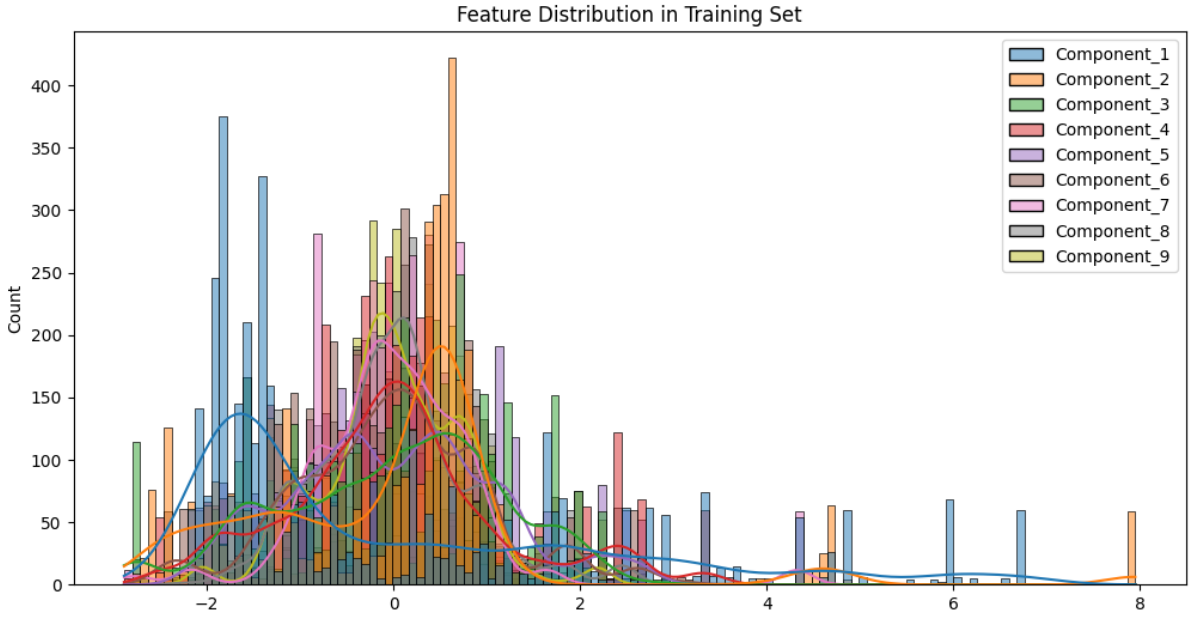


dense_24_input	input:	[(None, 9)]
InputLayer	output:	[(None, 9)]

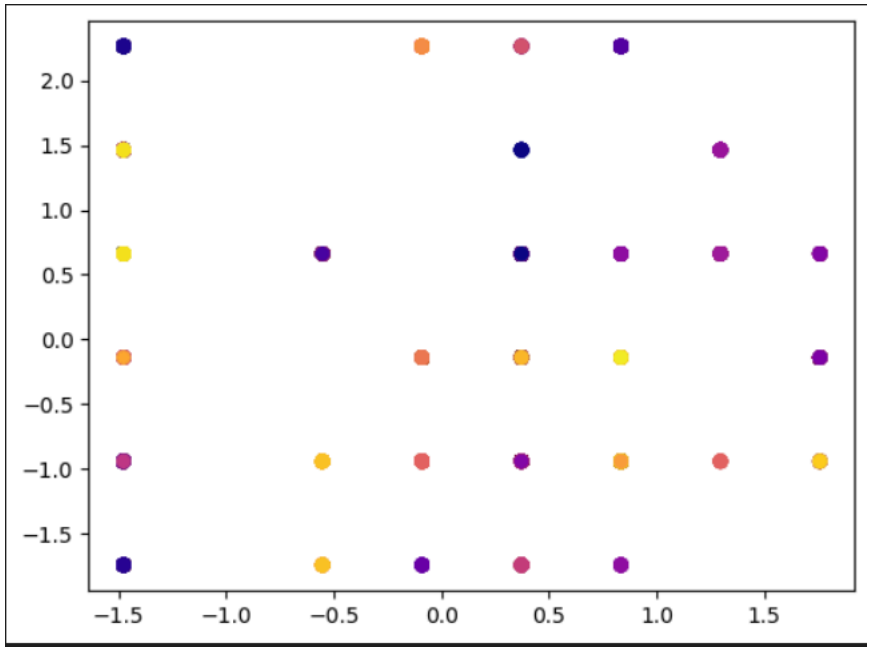
dense_24	input:	(None, 9)
Dense	output:	(None, 32)

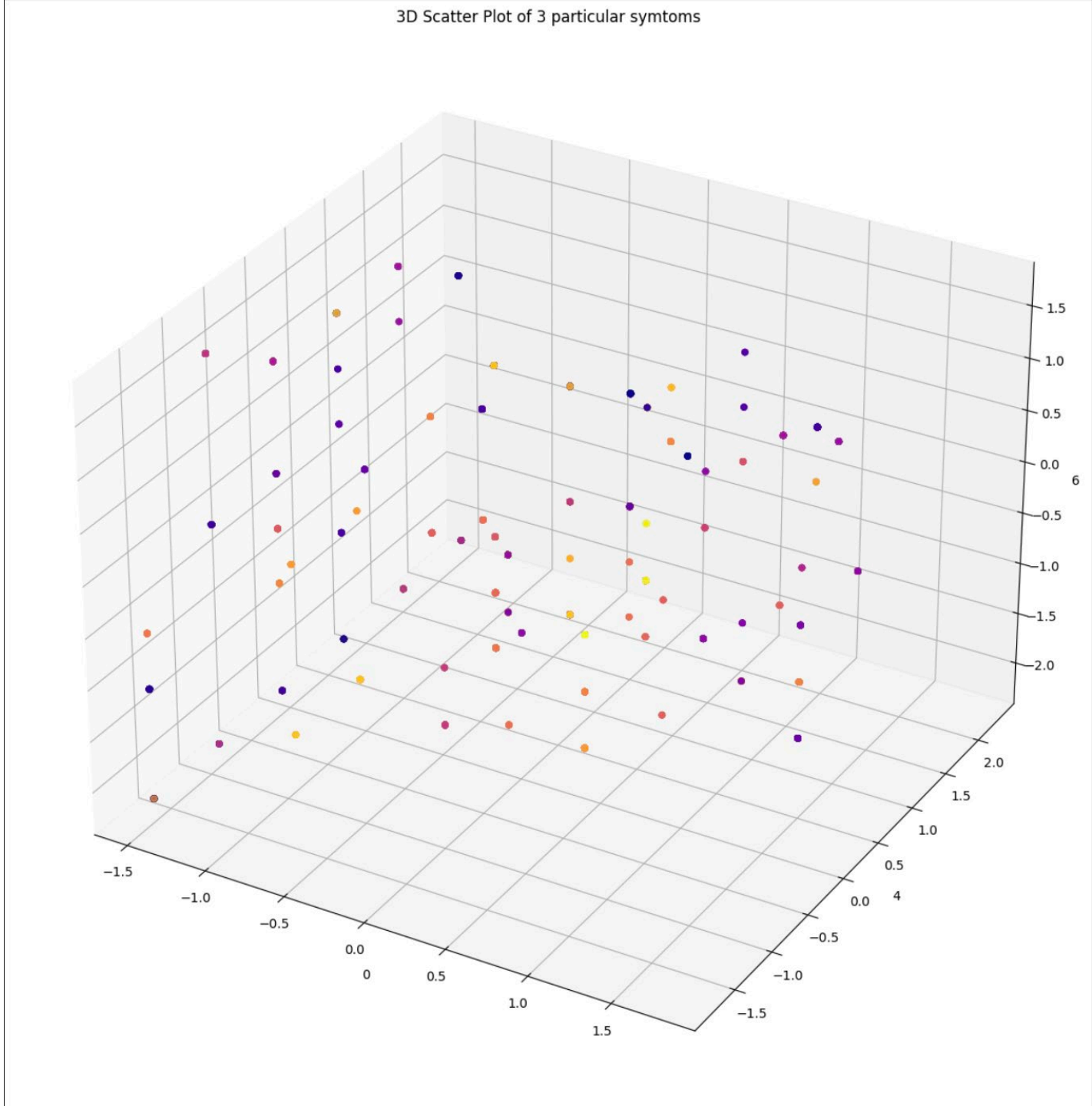
dense_25	input:	(None, 32)
Dense	output:	(None, 16)

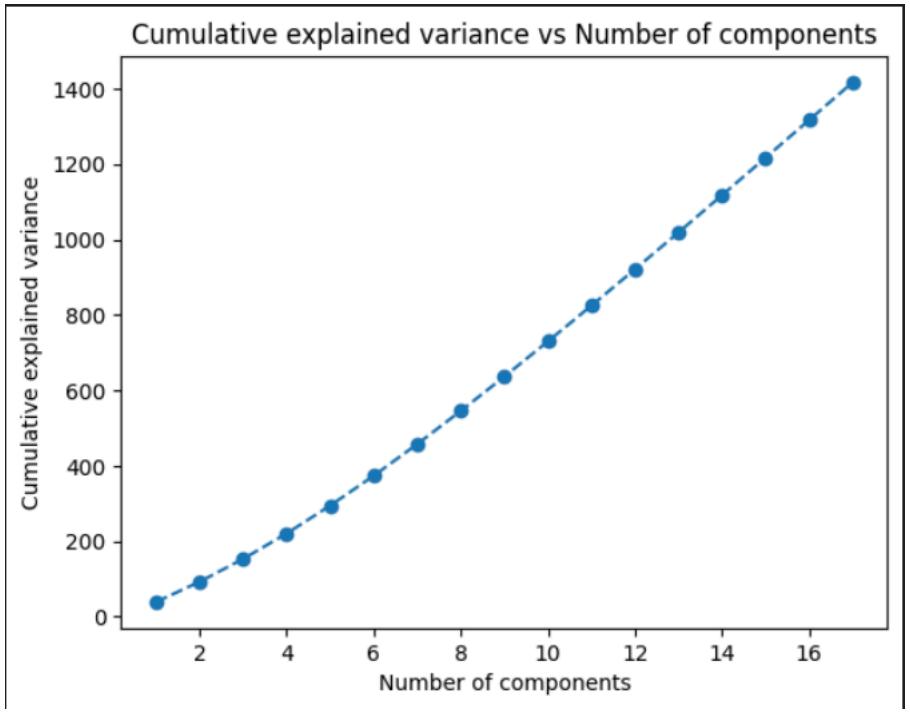
dense_26	input:	(None, 16)
Dense	output:	(None, 41)

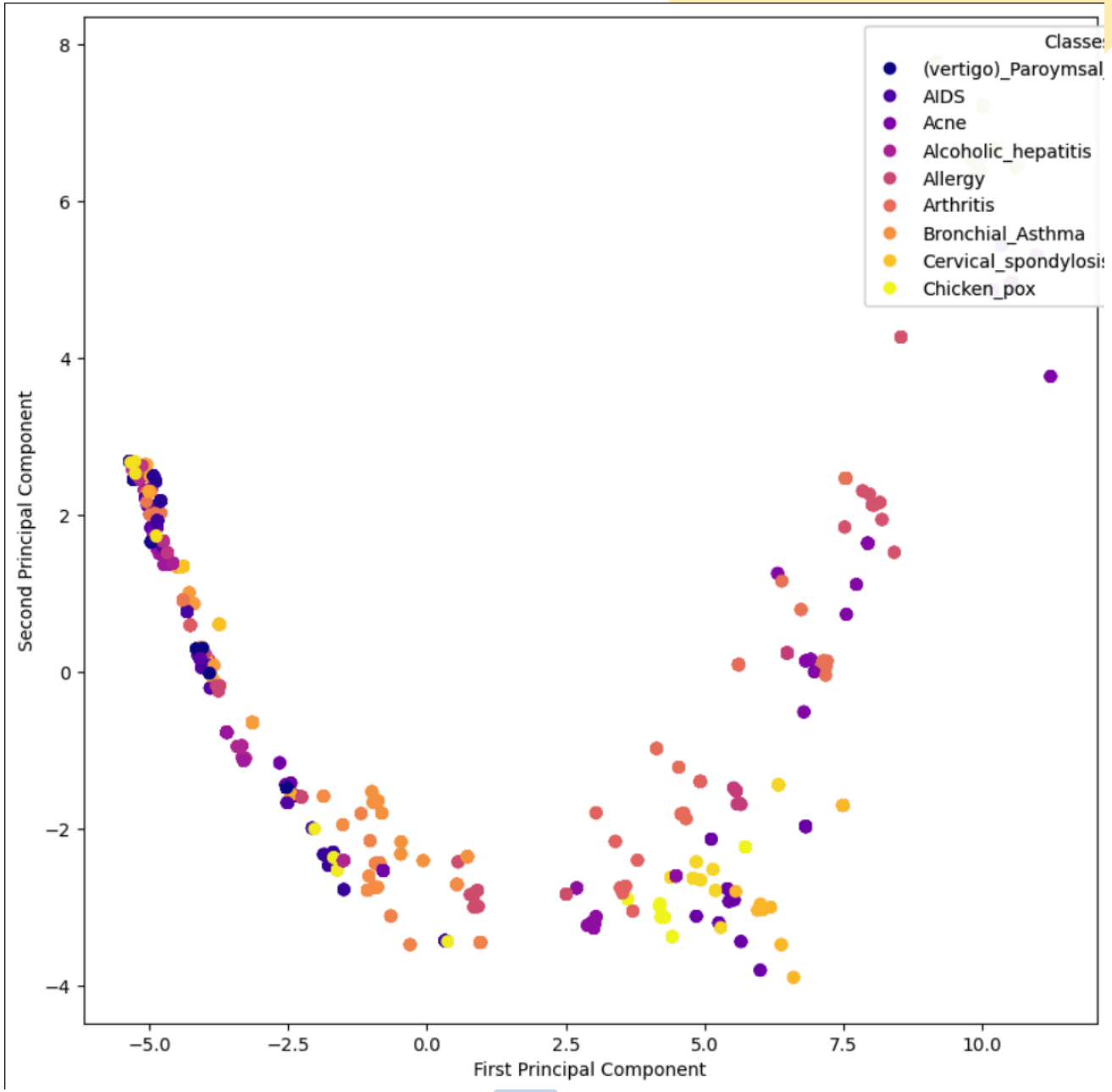


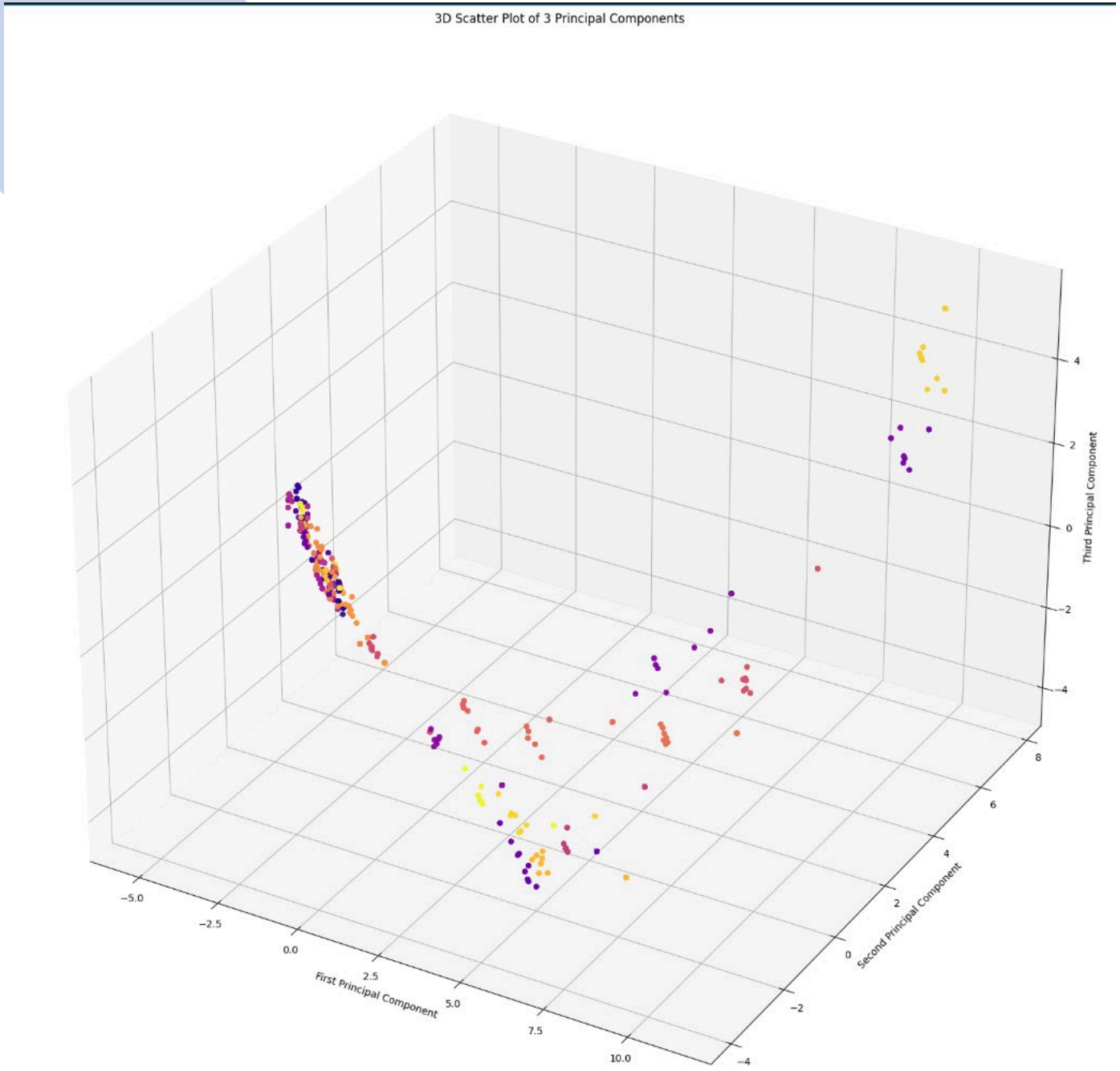

PCA

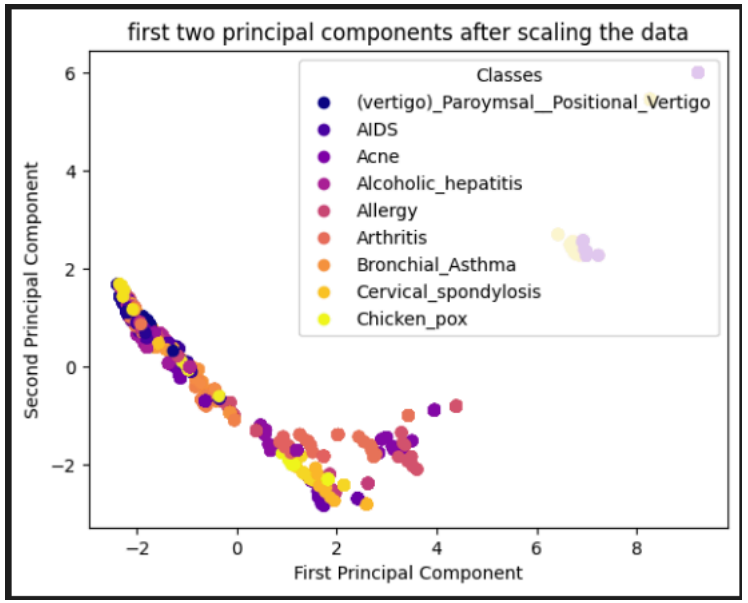

2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

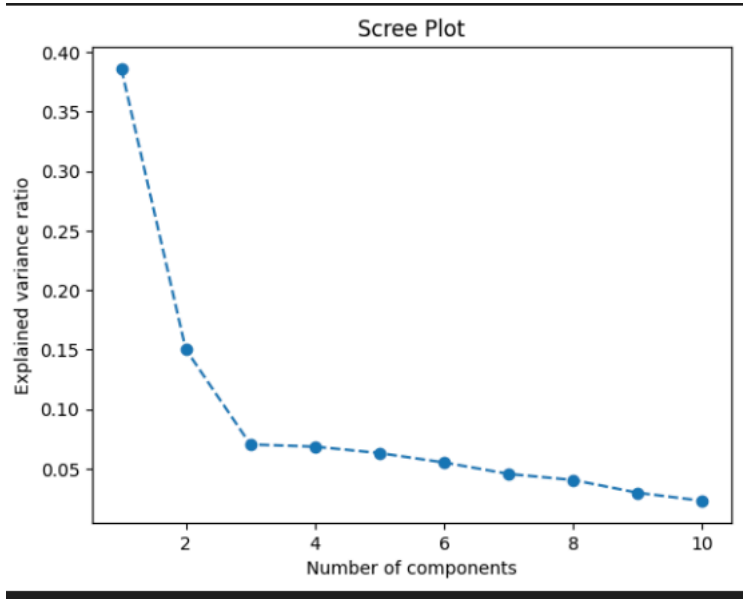

3D Scatter Plot of 3 particular symtoms
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
6
2.0
1.5
1.0
0.5
0.0
4
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0
0.5
1.0
1.5

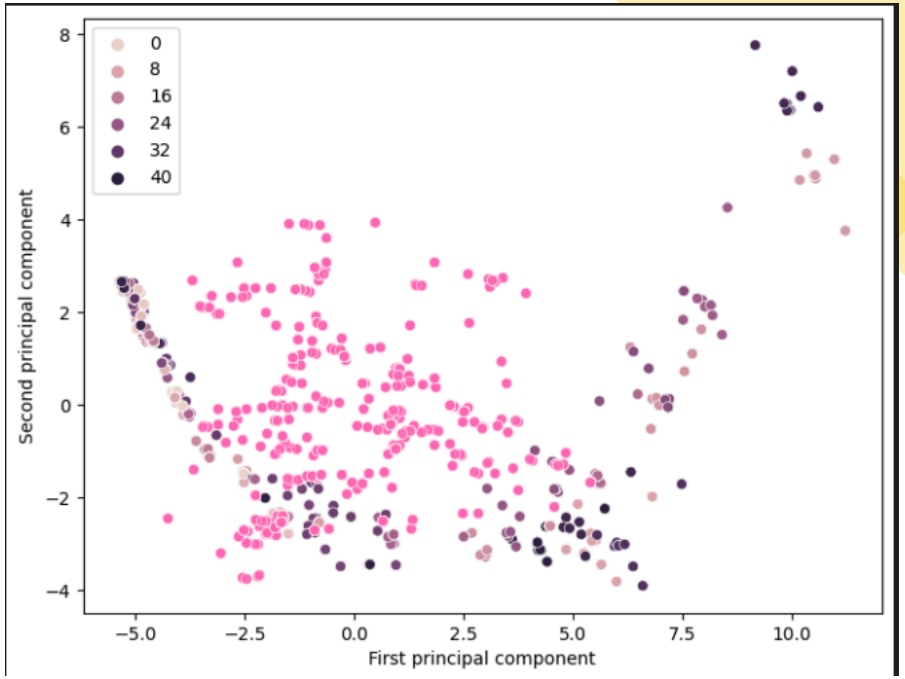

Cumulative explained variance vs Number of components
Cumulative explained variance
Number of components
0
200
400
600
800
1000
1200
1400
2
4
6
8
10
12
14
16

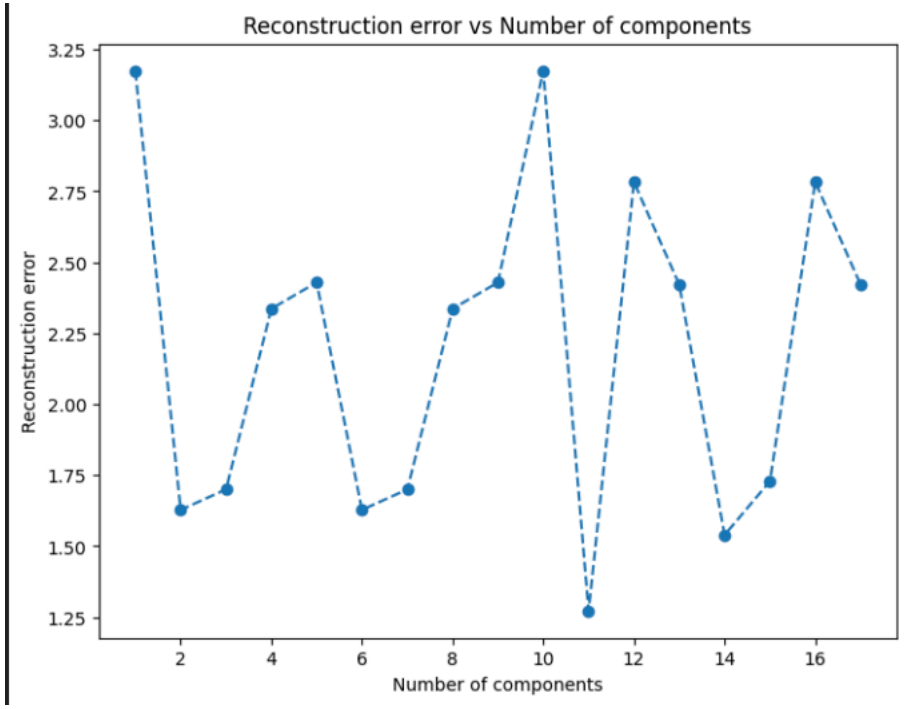

Classes
(vertigo)_Paroymsal
AIDS
Acne
Alcoholic_hepatitis
Allergy
Arthritis
Bronchial_Asthma
Cervical_spondylosis
Chicken_pox
Second Principal Component
First Principal Component
8
6
4
2
0
−2
−4
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

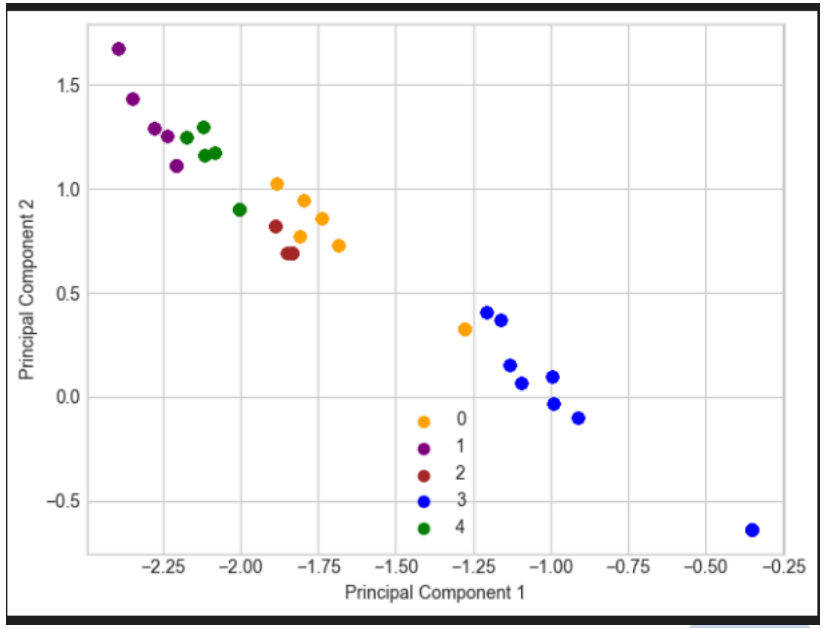

3D Scatter Plot of 3 Principal Components
First Principal Component
Second Principal Component
Third Principal Component

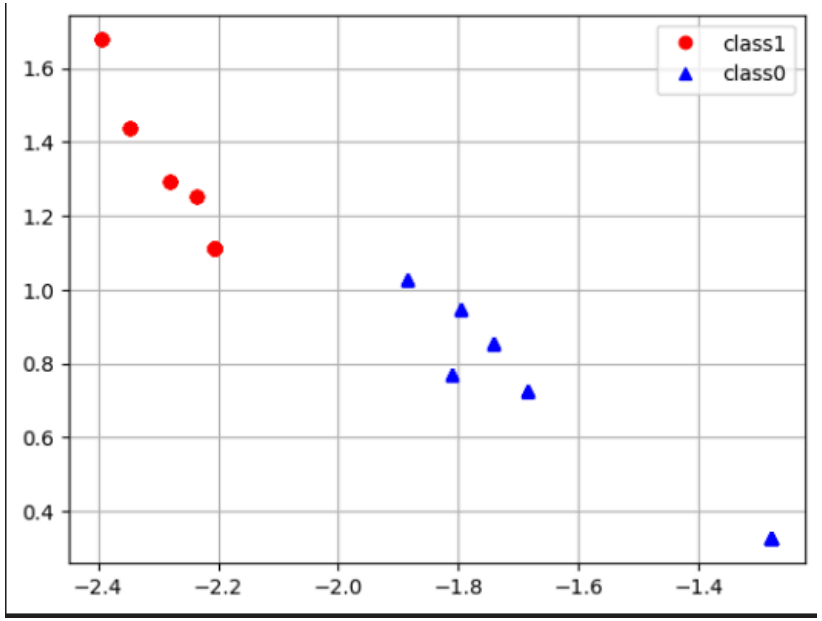

first two principal components after scaling the data
Classes
(vertigo)_Paroymsal__Positional_Vertigo
AIDS
Acne
Alcoholic_hepatitis
Allergy
Arthritis
Bronchial_Asthma
Cervical_spondylosis
Chicken_pox
Second Principal Component
First Principal Component

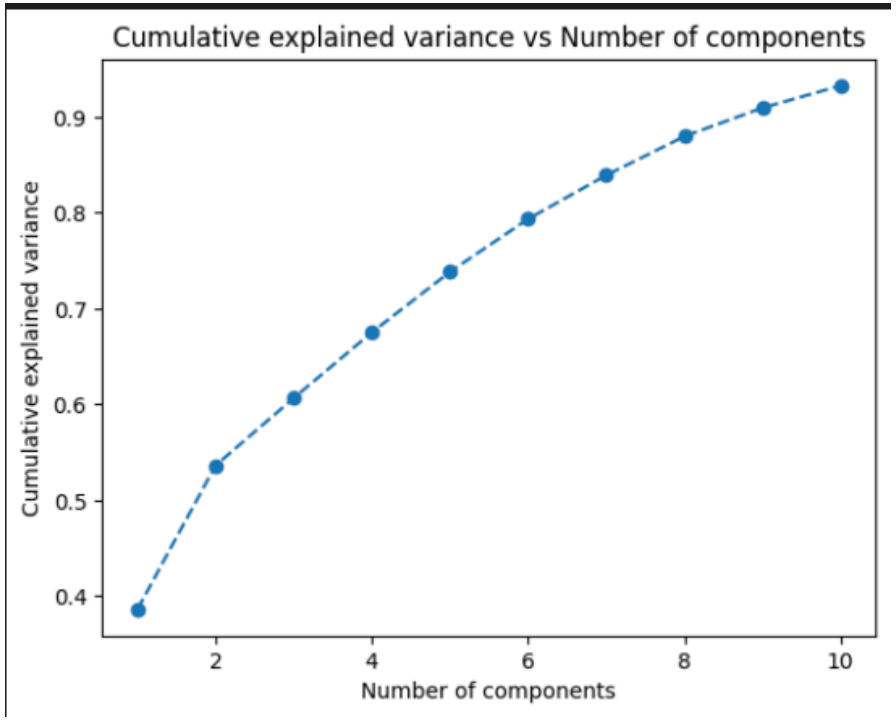

Scree Plot
Explained variance ratio
Number of components
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
2
4
6
8
10

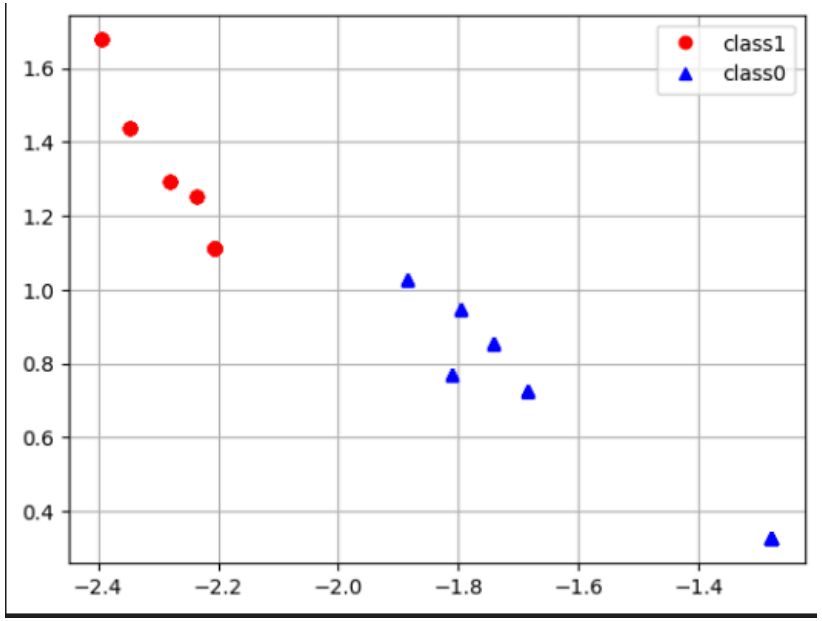

0
8
16
24
32
40
Second principal component
First principal component
8
6
4
2
0
−2
−4
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

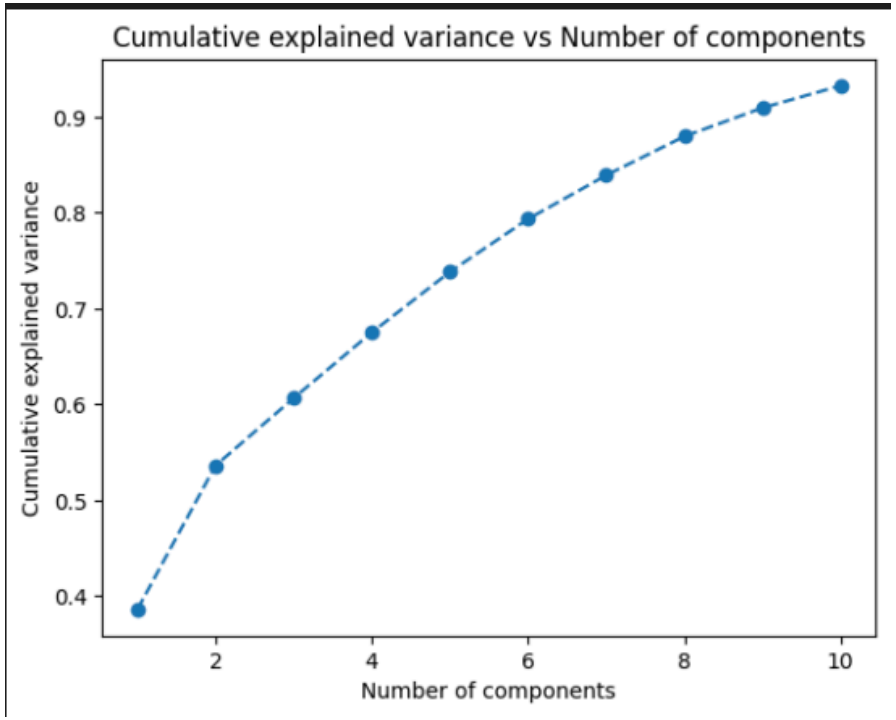

Reconstruction error vs Number of components
Reconstruction error
Number of components
3.25
3.00
2.75
2.50
2.25
2.00
1.75
1.50
1.25
2
4
6
8
10
12
14
16

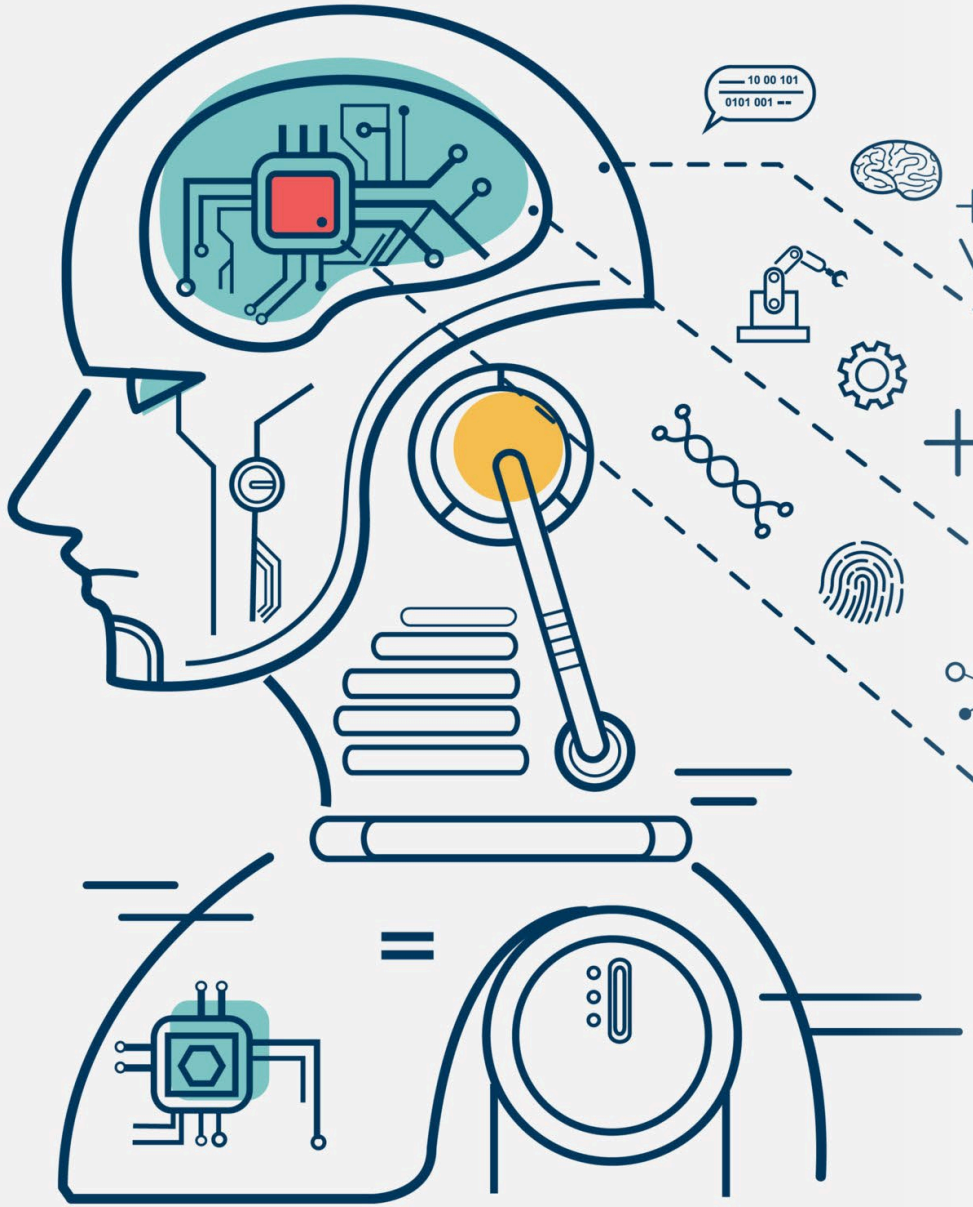

Principal Component 2
Principal Component 1
0
1
2
3
4
1.5
1.0
0.5
0.0
−0.5
−2.25
−2.00
−1.75
−1.50
−1.25
−1.00
−0.75
−0.50
−0.25


class1
class0
1.6
1.4
1.2
1.0
0.8
0.6
0.4
−2.4
−2.2
−2.0
−1.8
−1.6
−1.4


Cumulative explained variance vs Number of components
Cumulative explained variance
0.9
0.8
0.7
0.6
0.5
0.4
2
4
6
8
10
Number of components


class1
class0
1.6
1.4
1.2
1.0
0.8
0.6
0.4
−2.4
−2.2
−2.0
−1.8
−1.6
−1.4

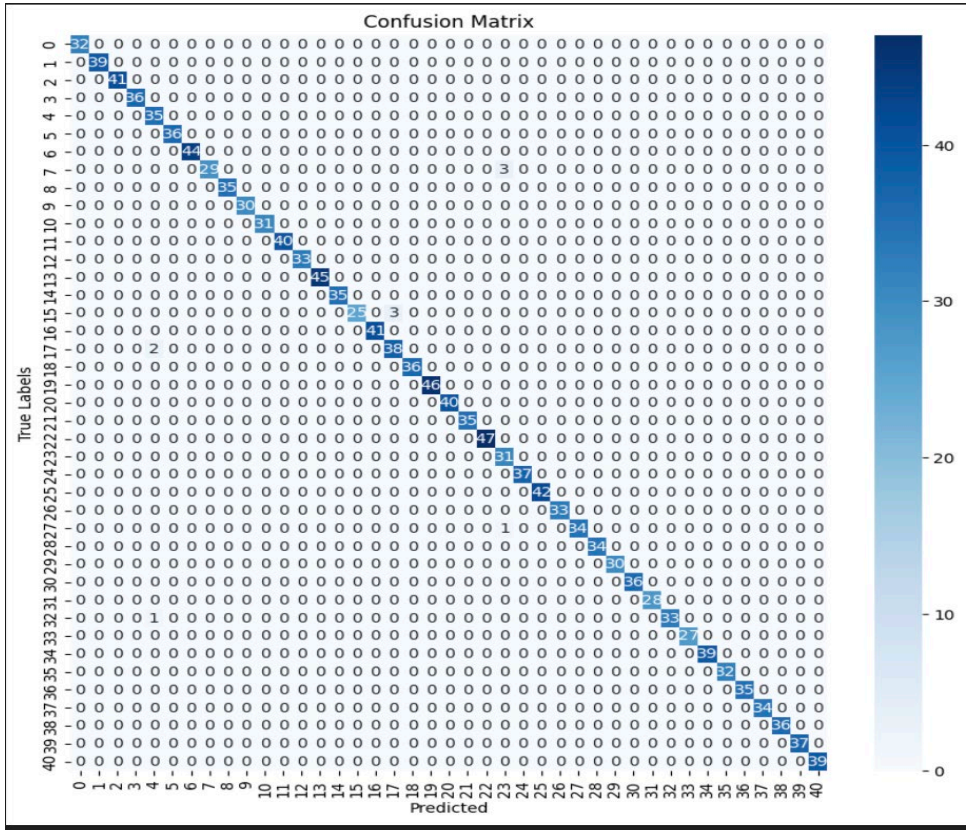

Cumulative explained variance vs Number of components
Cumulative explained variance
0.9
0.8
0.7
0.6
0.5
0.4
2
4
6
8
10
Number of components

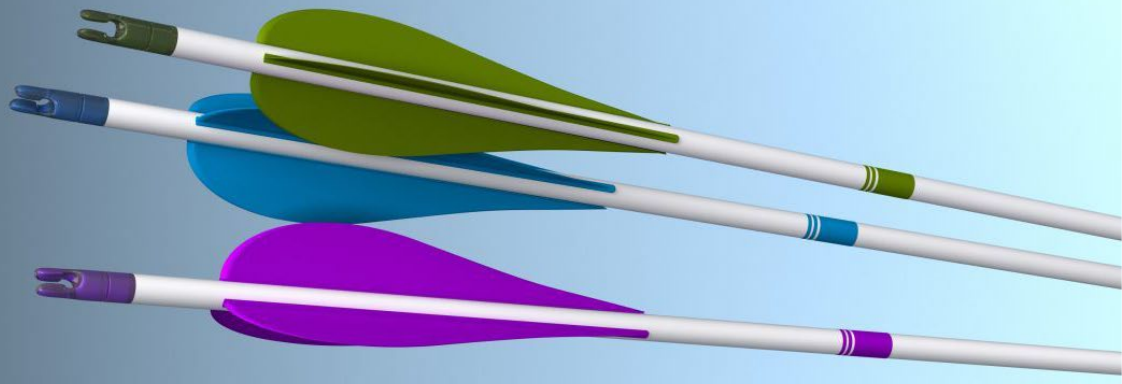


Machine Learning Algorithms Used

- Random Forest Classifier
- K-Nearest Neighbour Classifier
- Decision Tree Classifier

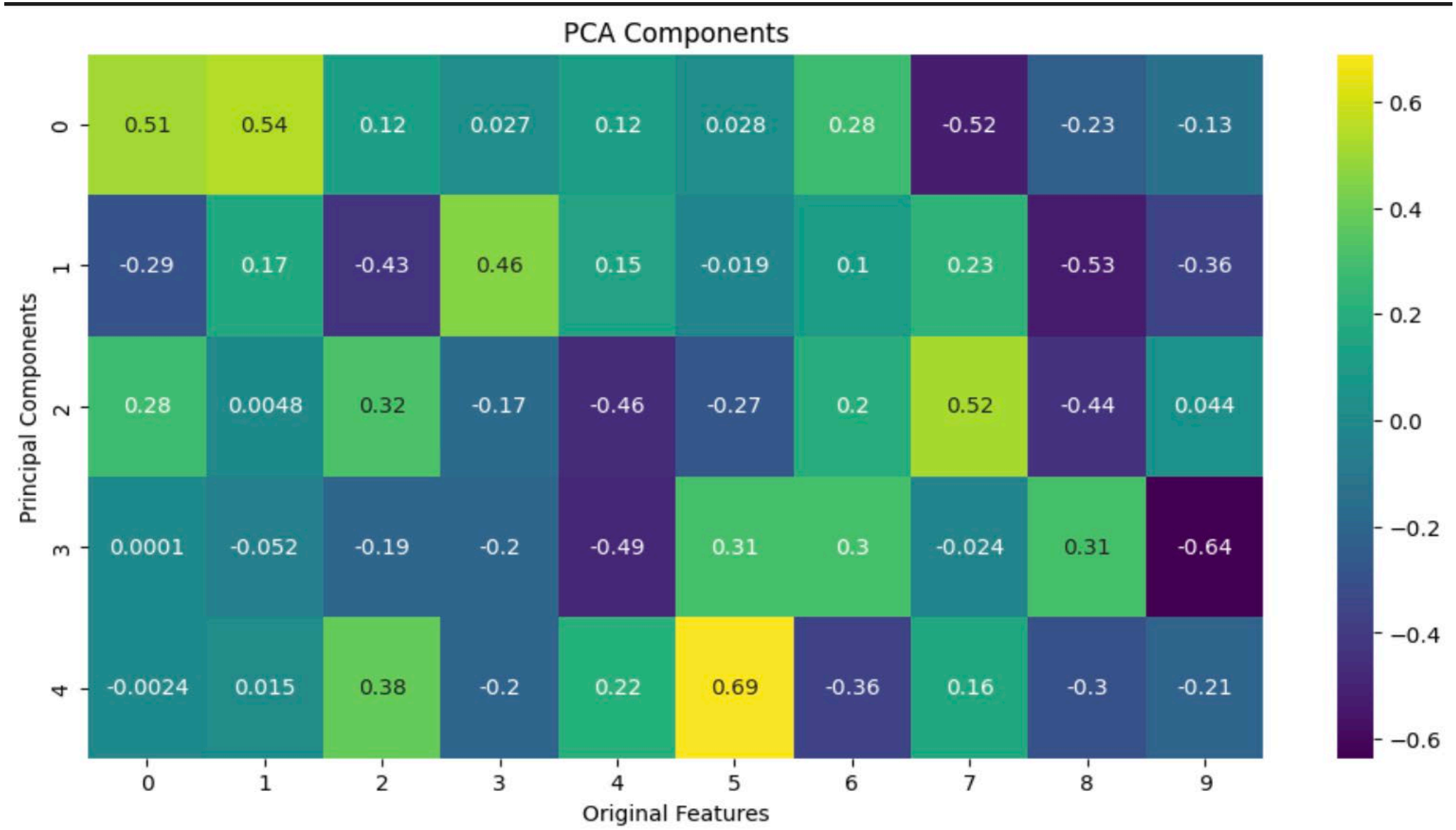
SVM

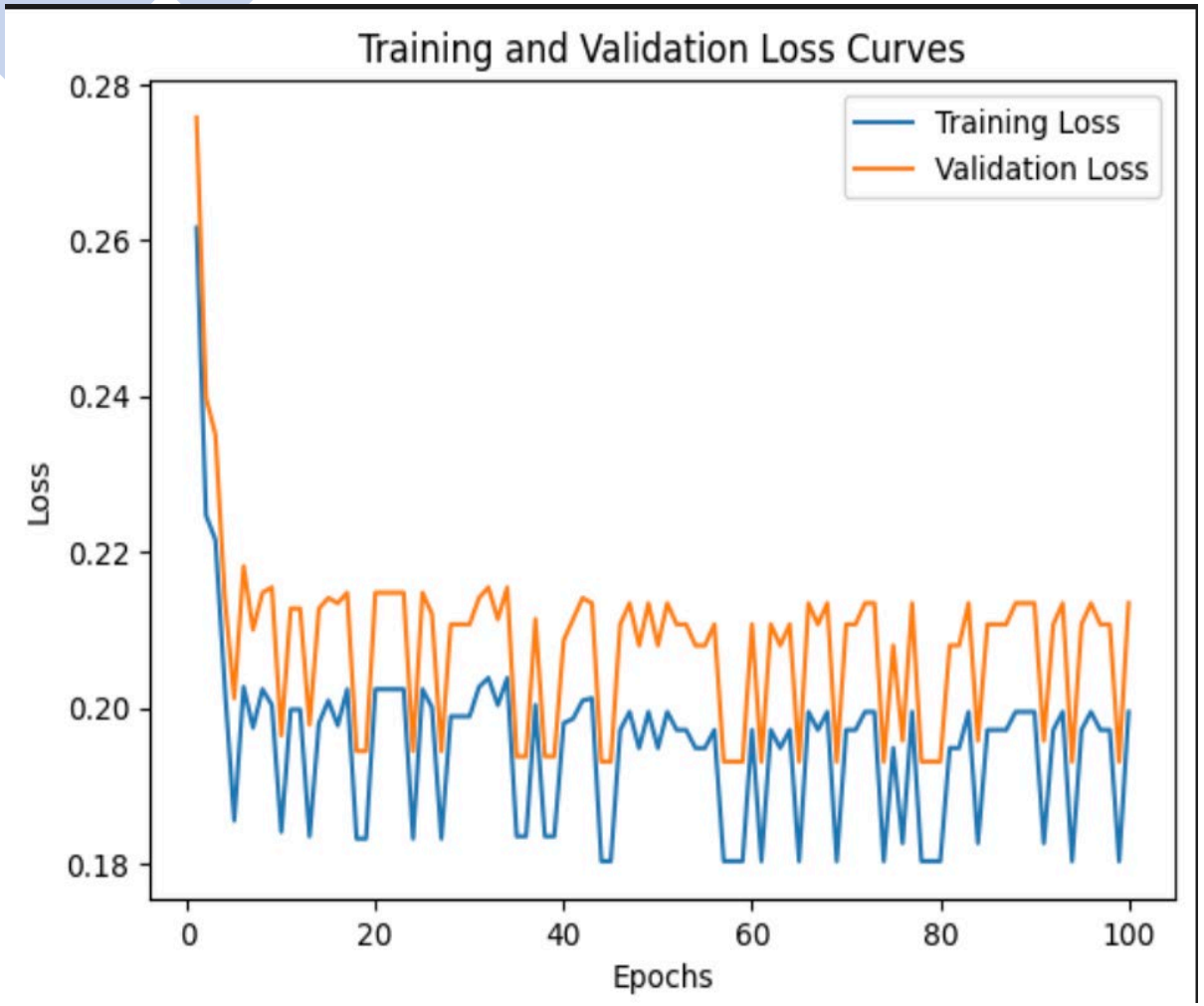


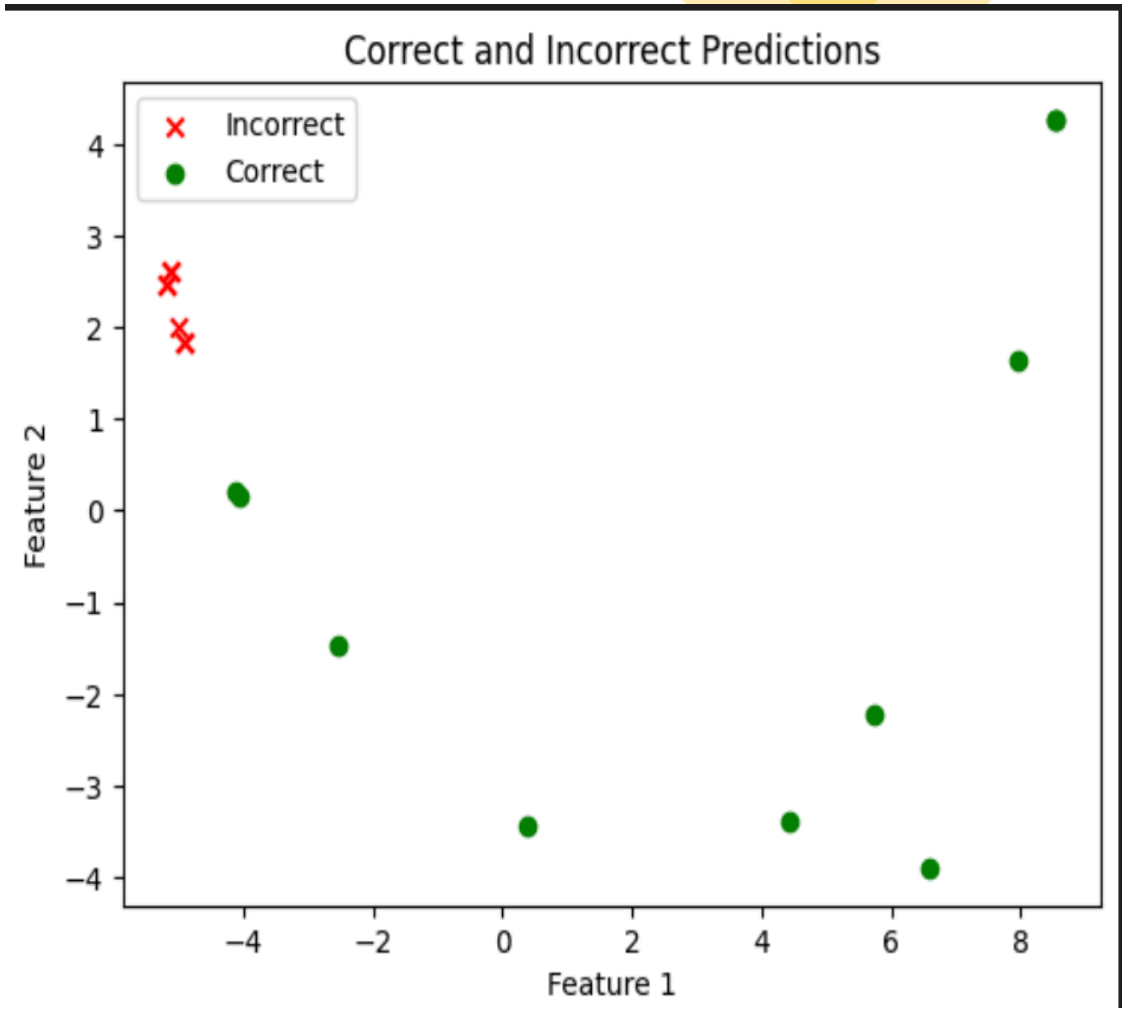


SVM

Metrics	
Accuracy	0.9932249322493225
Precision	0.9937446854296159
Recall	0.9932249322493225


PCA Components
Principal Components
Original Features
0
1
2
3
4
0.51 0.54 0.12 0.027 0.12 0.028 0.28 -0.52 -0.23 -0.13
-0.29 0.17 -0.43 0.46 0.15 -0.019 0.1 0.23 -0.53 -0.36
0.28 0.0048 0.32 -0.17 -0.46 -0.27 0.2 0.52 -0.44 0.044
0.0001 -0.052 -0.19 -0.2 -0.49 0.31 0.3 -0.024 0.31 -0.64
-0.0024 0.015 0.38 -0.2 0.22 0.69 -0.36 0.16 -0.3 -0.21
0 1 2 3 4 5 6 7 8 9
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

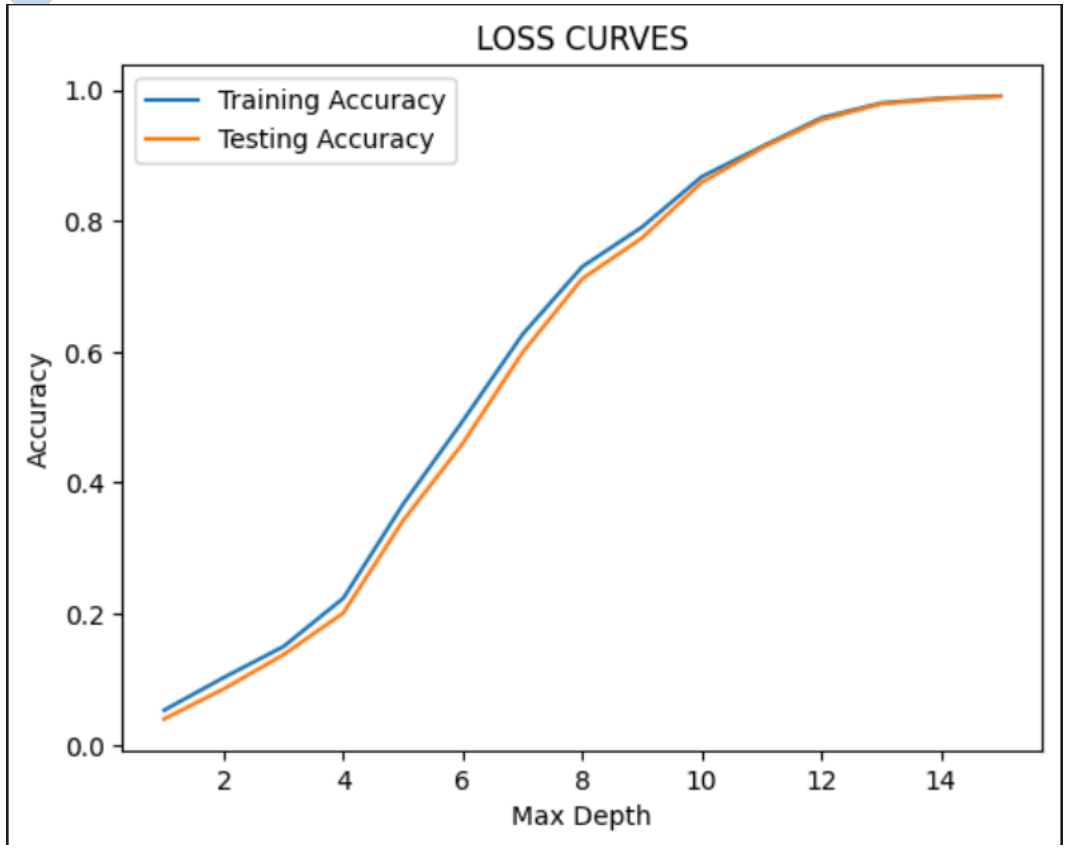

Training and Validation Loss Curves
Training Loss
Validation Loss
Loss
Epochs
0.28
0.26
0.24
0.22
0.20
0.18
0
20
40
60
80
100

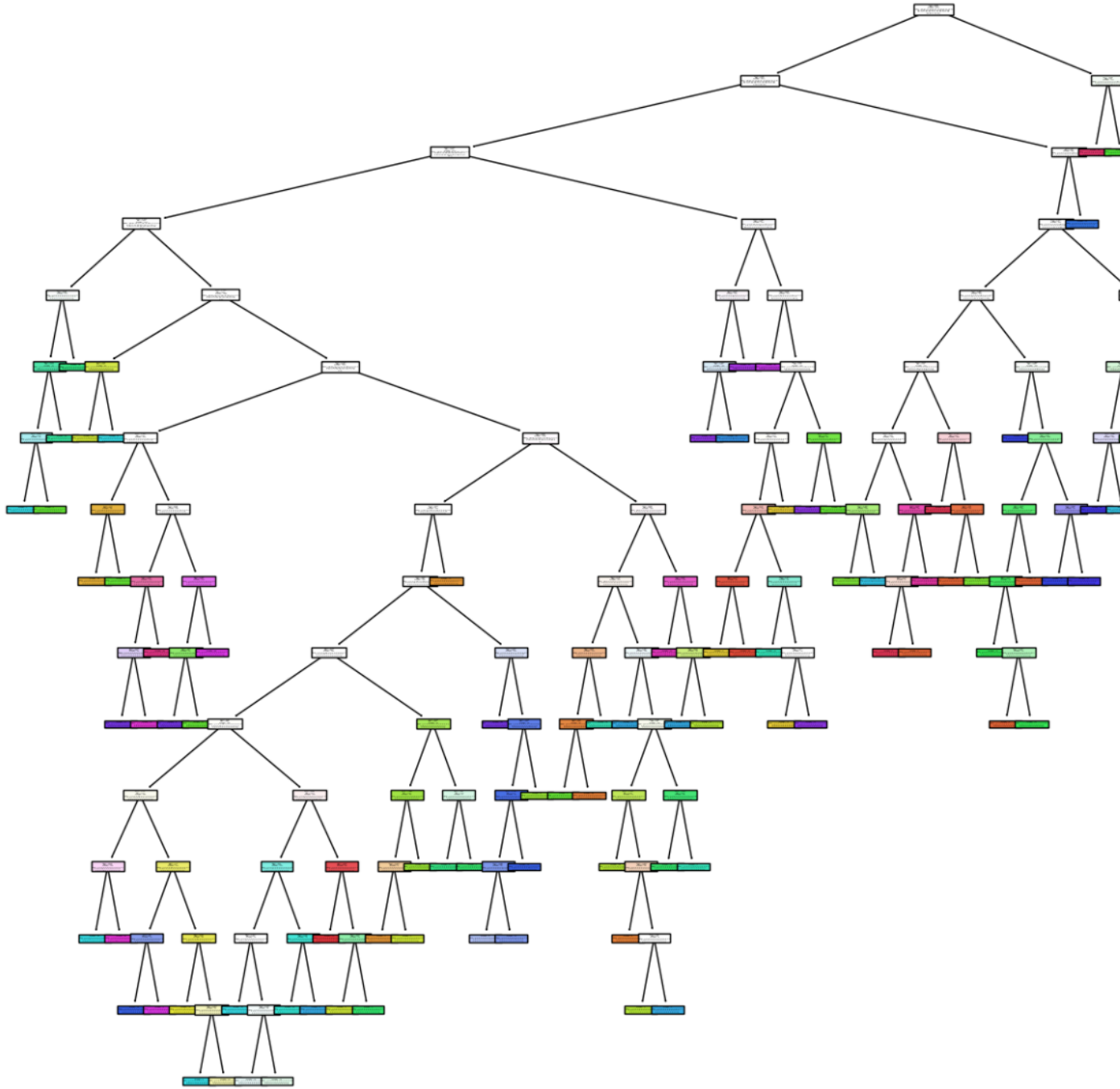

Correct and Incorrect Predictions
Incorrect
Correct
Feature 2
Feature 1
4
3
2
1
0
−1
−2
−3
−4
−4
−2
0
2
4
6
8

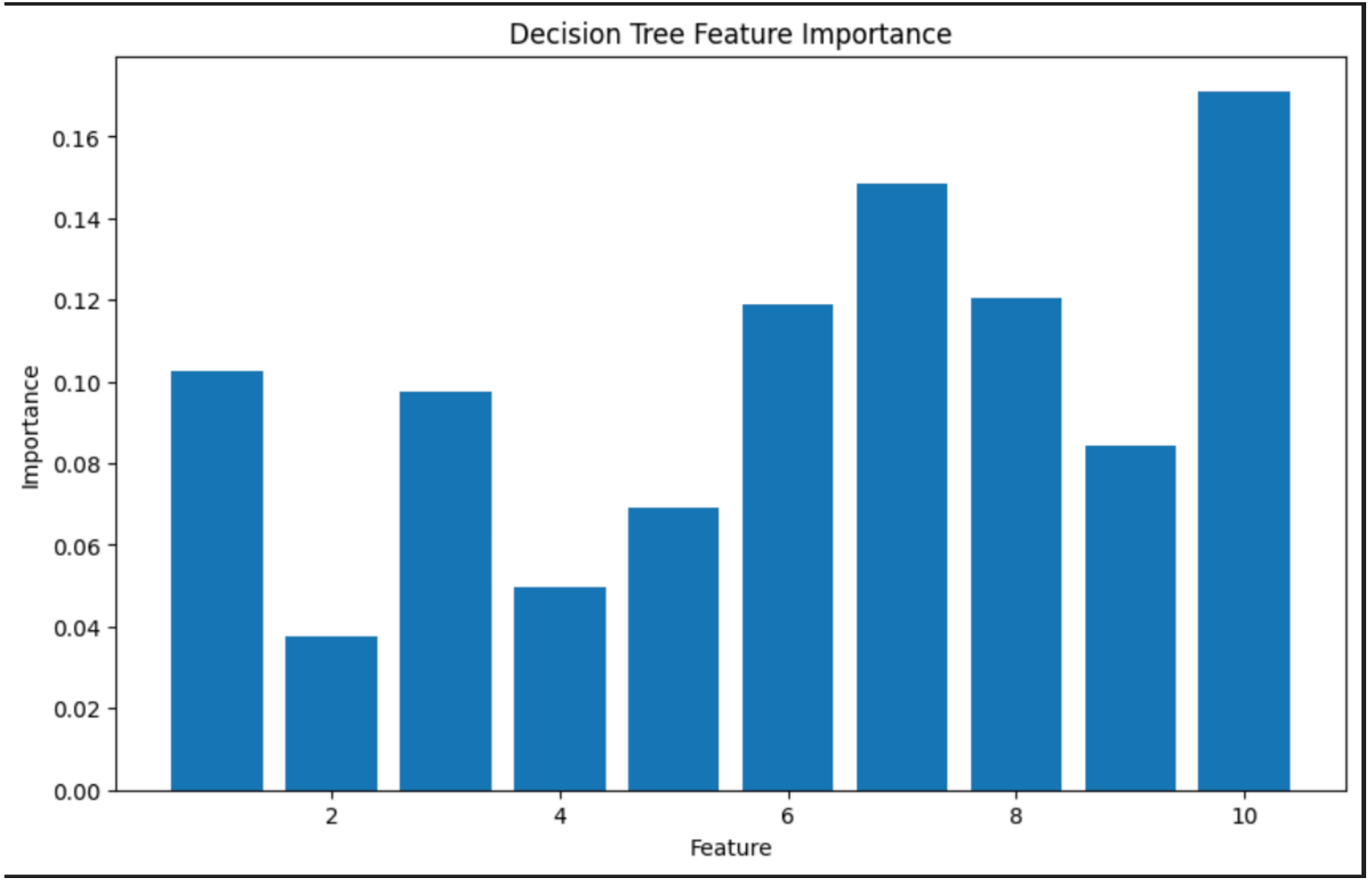
DECISION TREE

Decision Tree

Metrics	Accuracy
Accuracy	0.9932249322493225
Precision	0.9937446854296159
Recall	0.9932249322493225


LOSS CURVES
Training Accuracy
Testing Accuracy
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
2
4
6
8
10
12
14
Max Depth

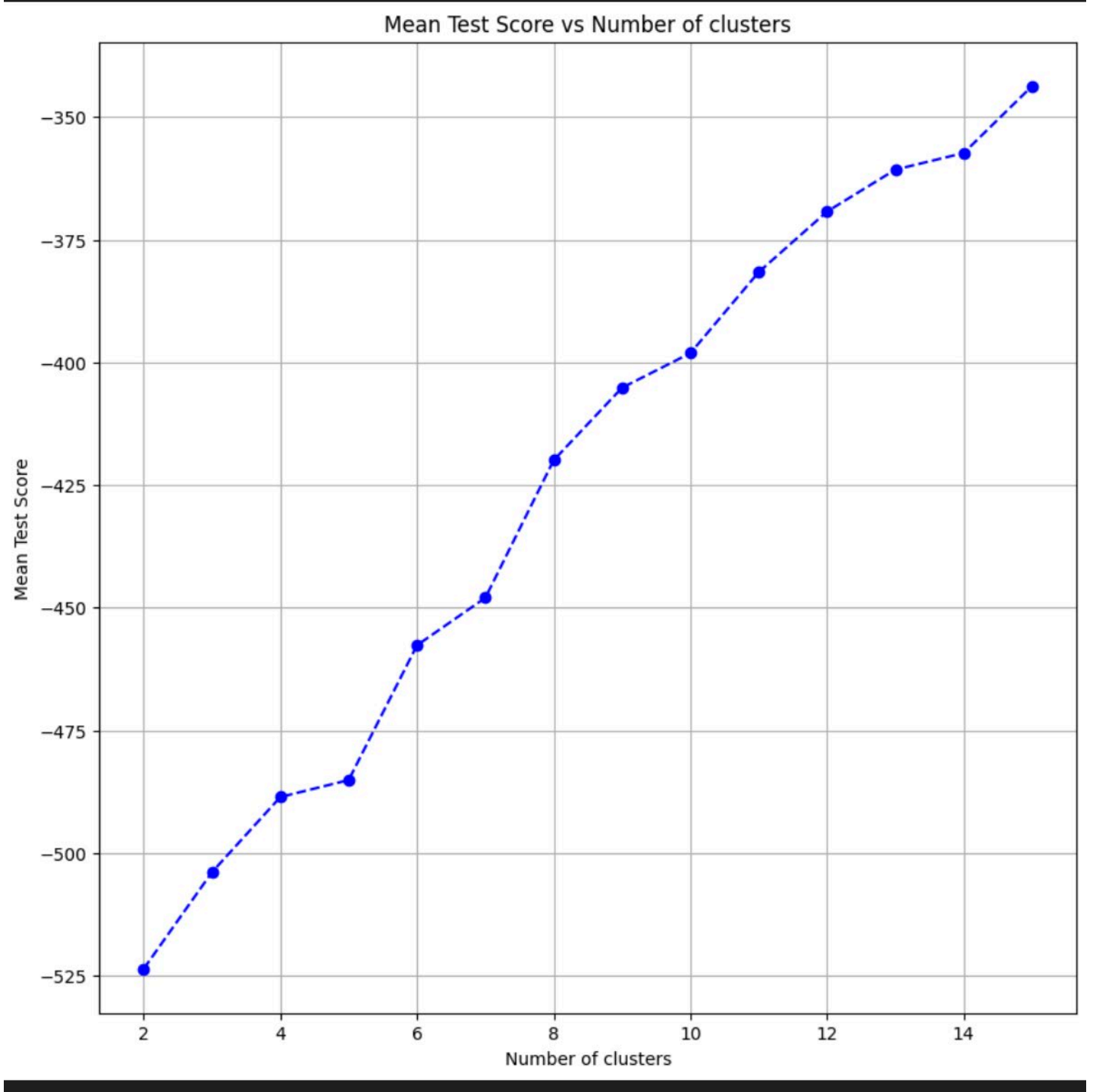



Decision Tree Feature Importance
Importance
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
2
4
6
8
10
Feature

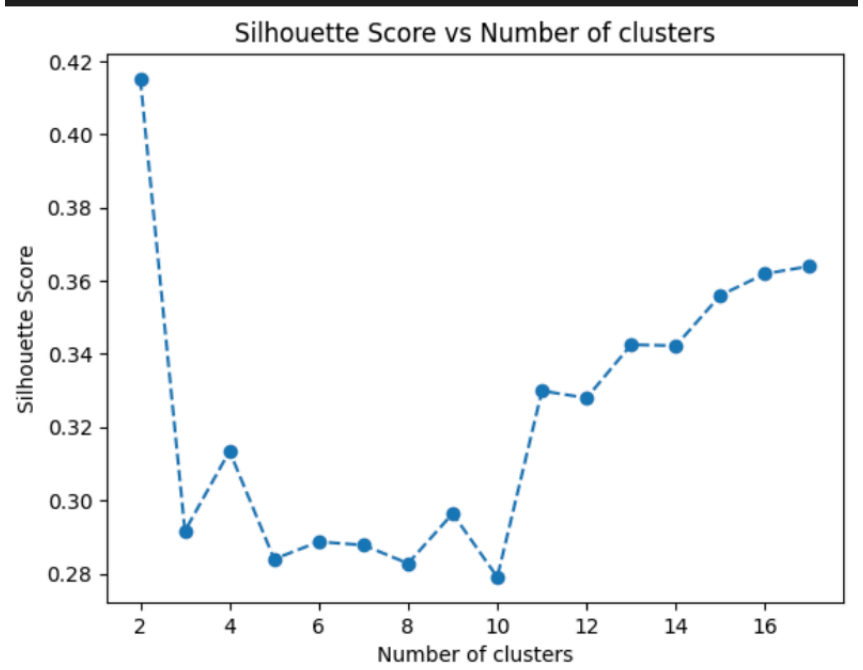
CLUSTERING

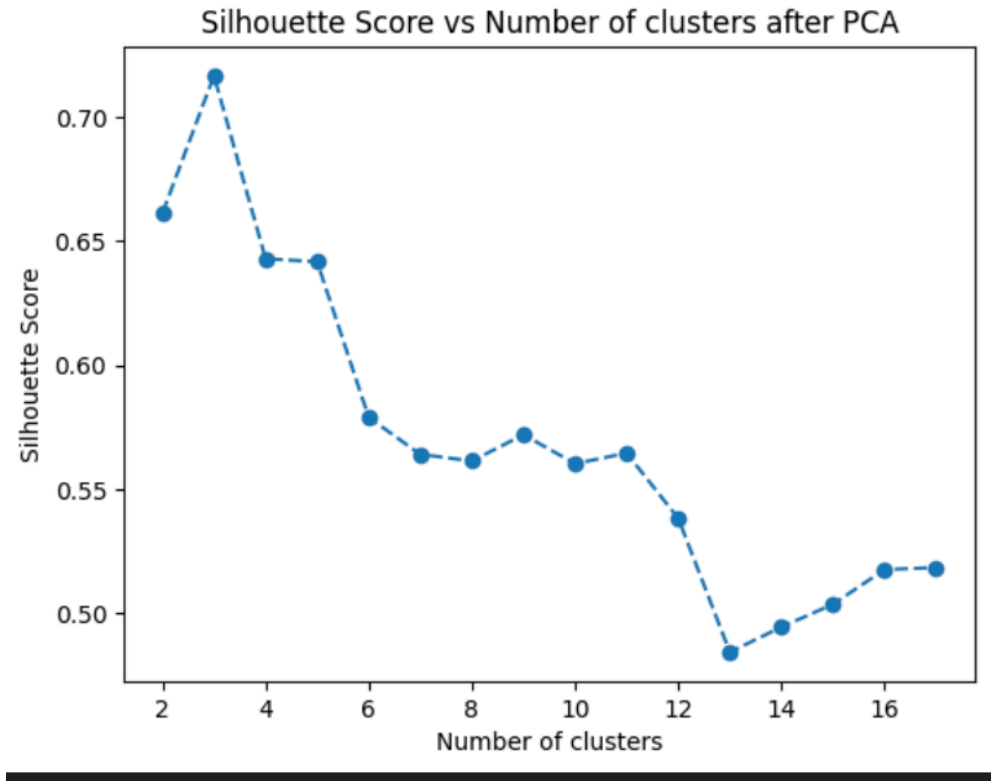
- K-Means Clustering
- Agglomarative Clustering

Metrics for K-Means

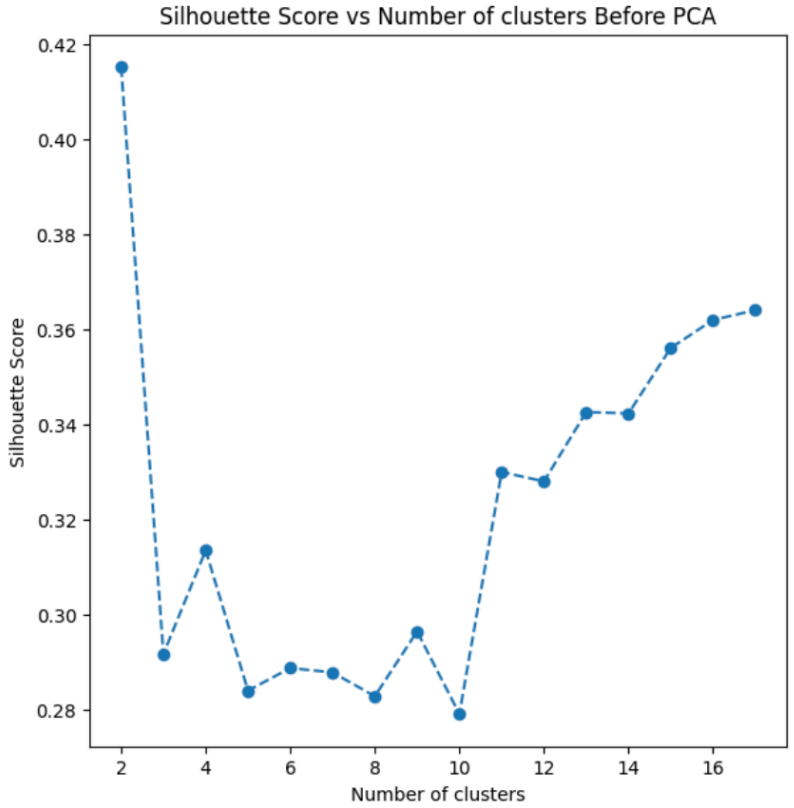


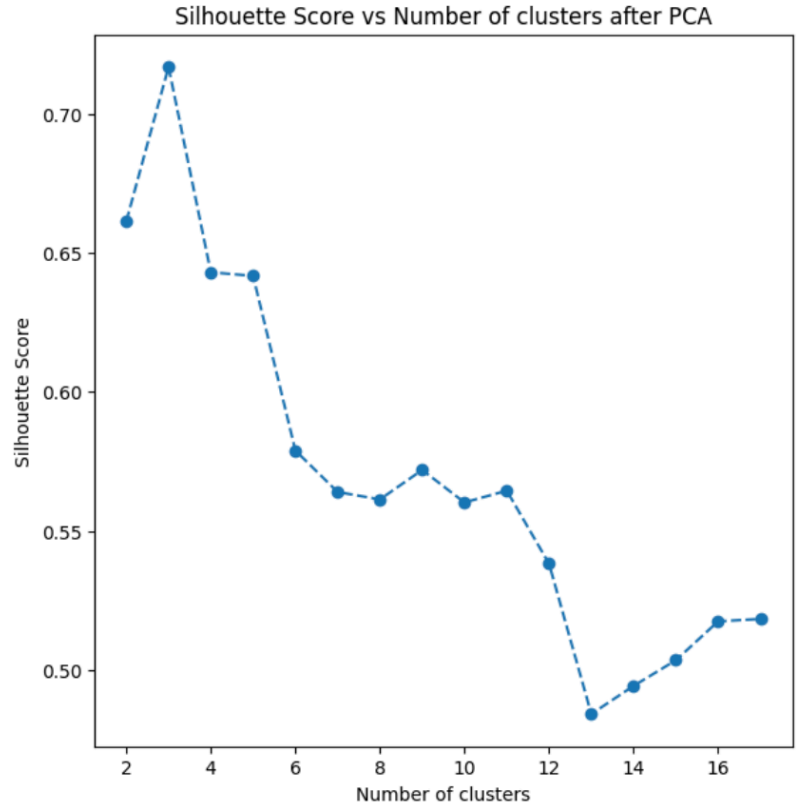
Silhouette Score



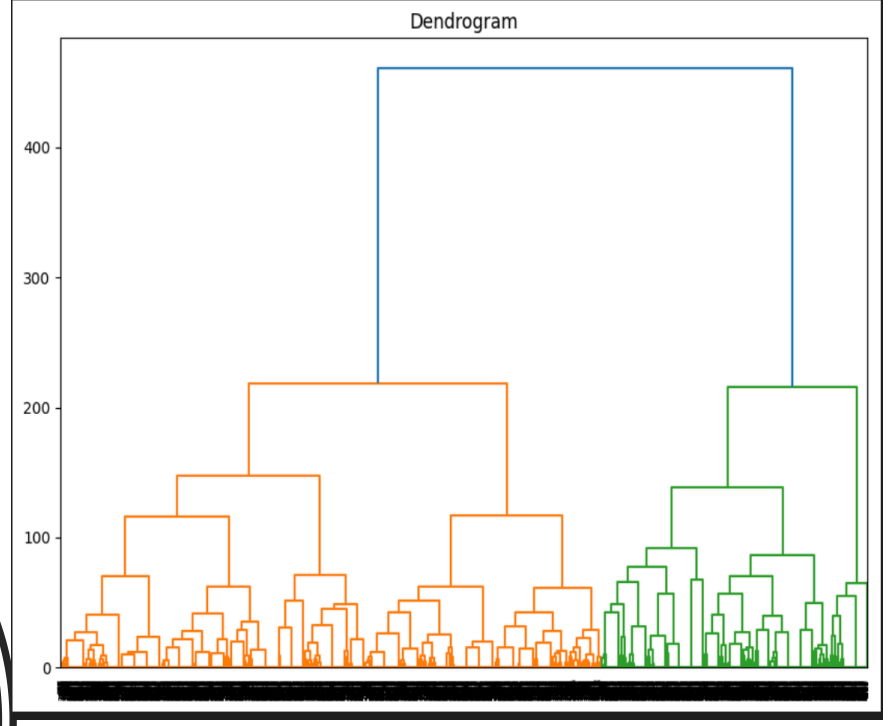


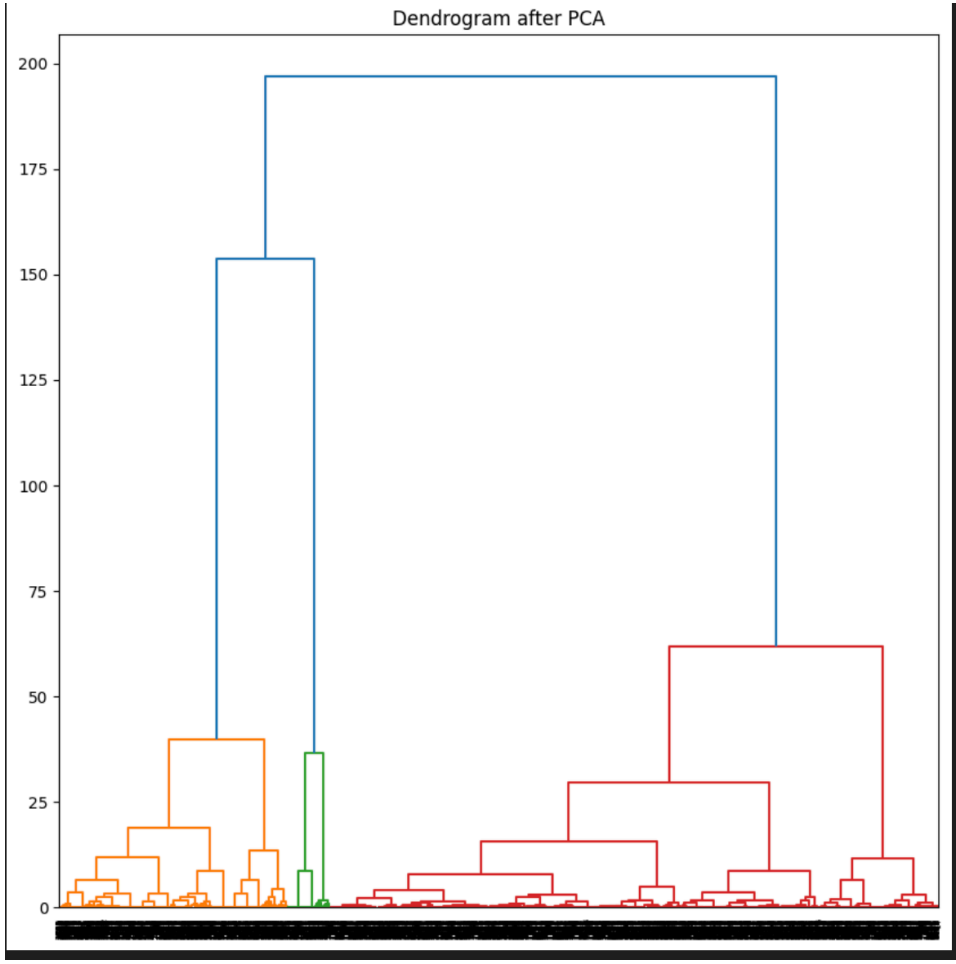
Silhouette Score Before Vs After PCA



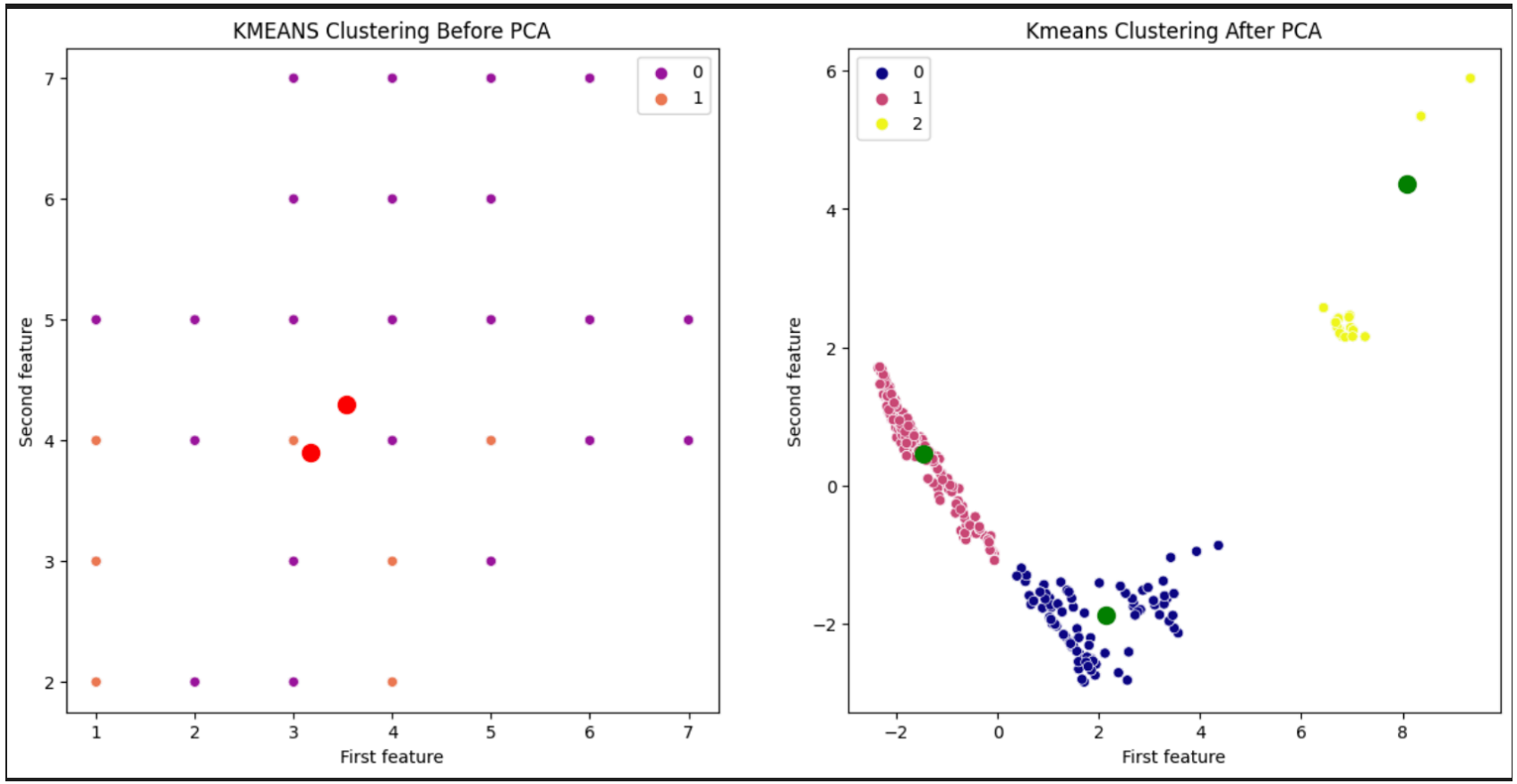


Dendrogram

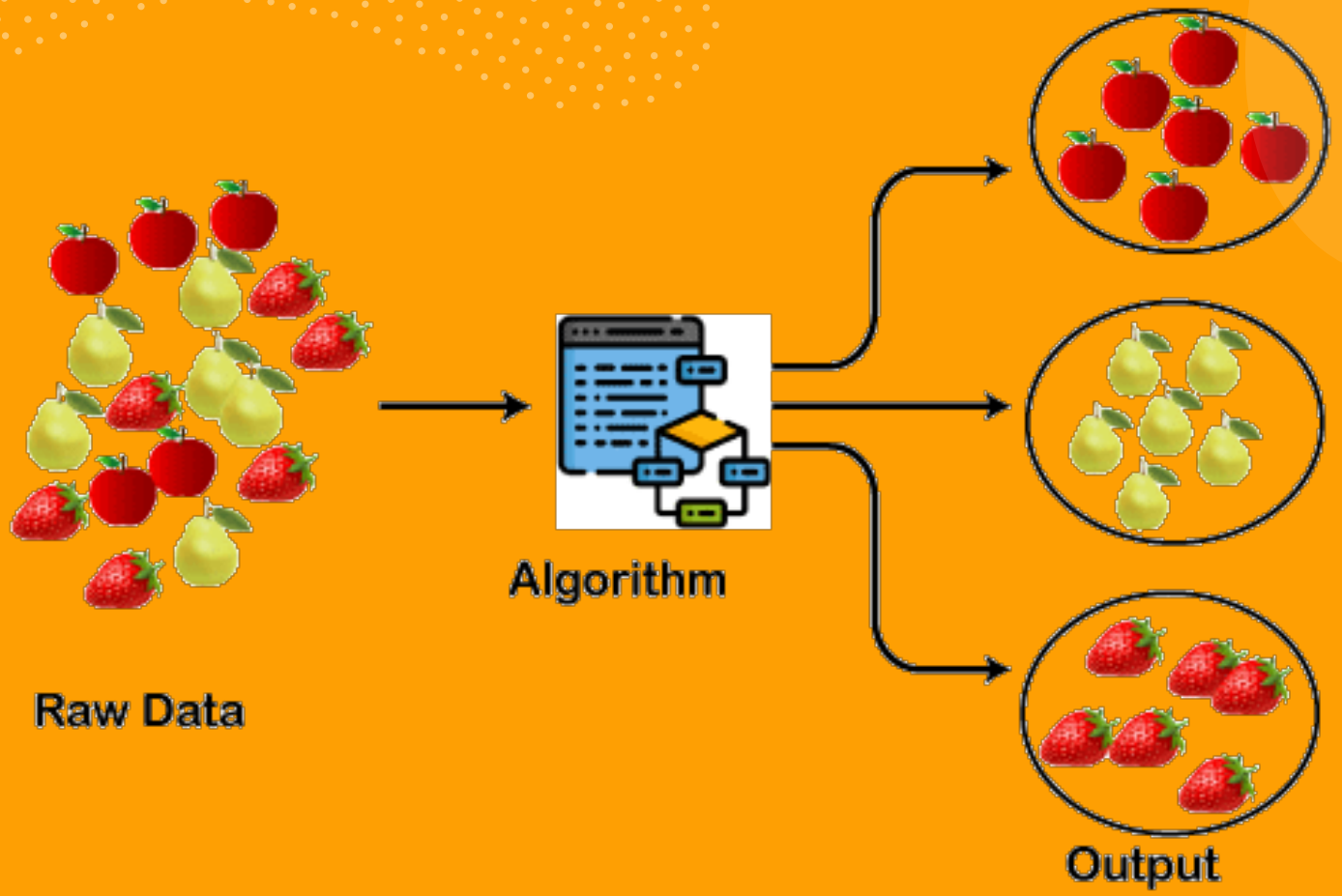

Dendrogram
400
300
200
100
0

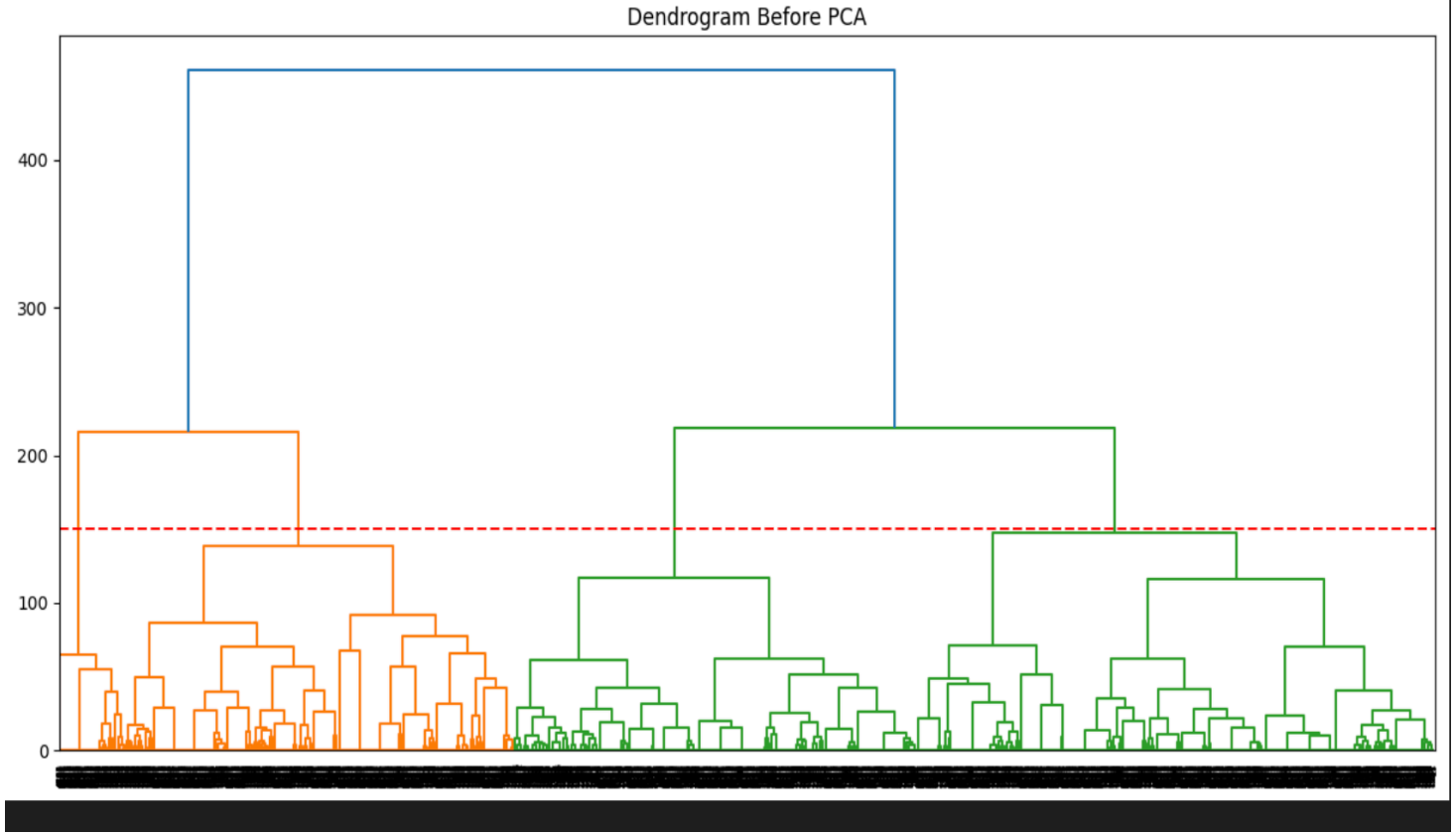

Dendrogram after PCA
200
175
150
125
100
75
50
25
0

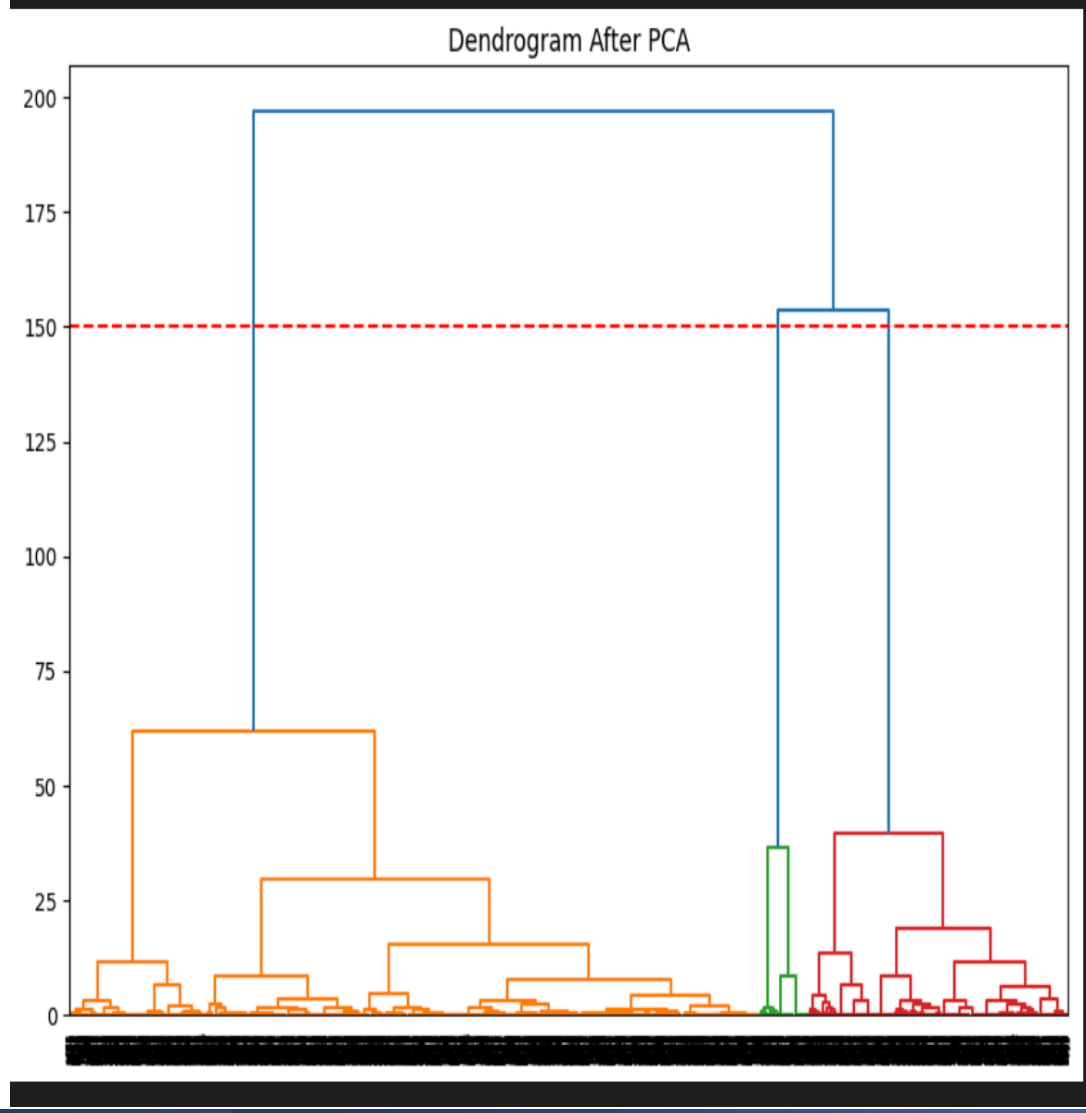
K-MEANS Before Vs After PCA



Hierarchical clustering

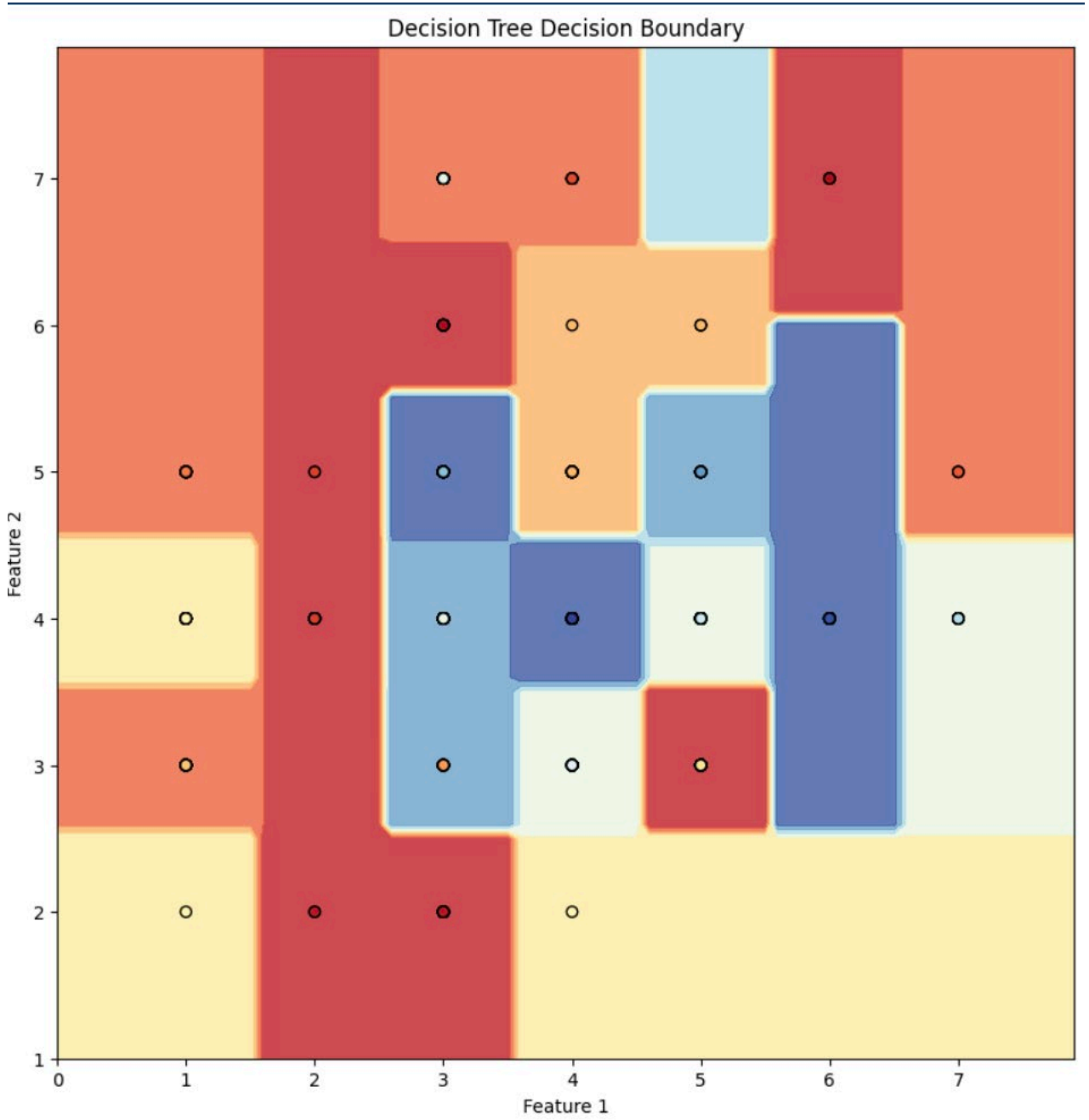




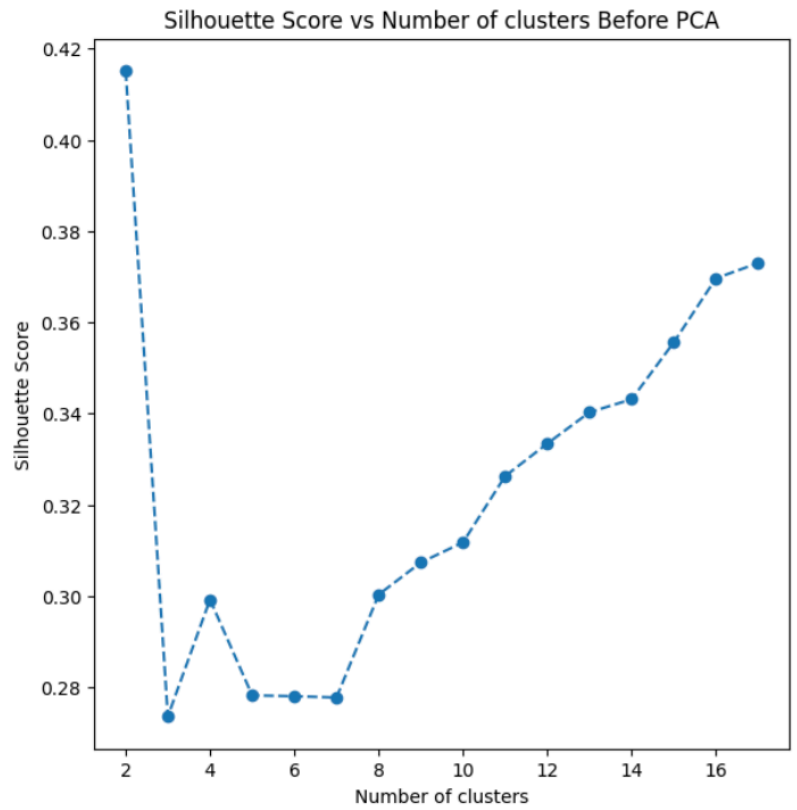


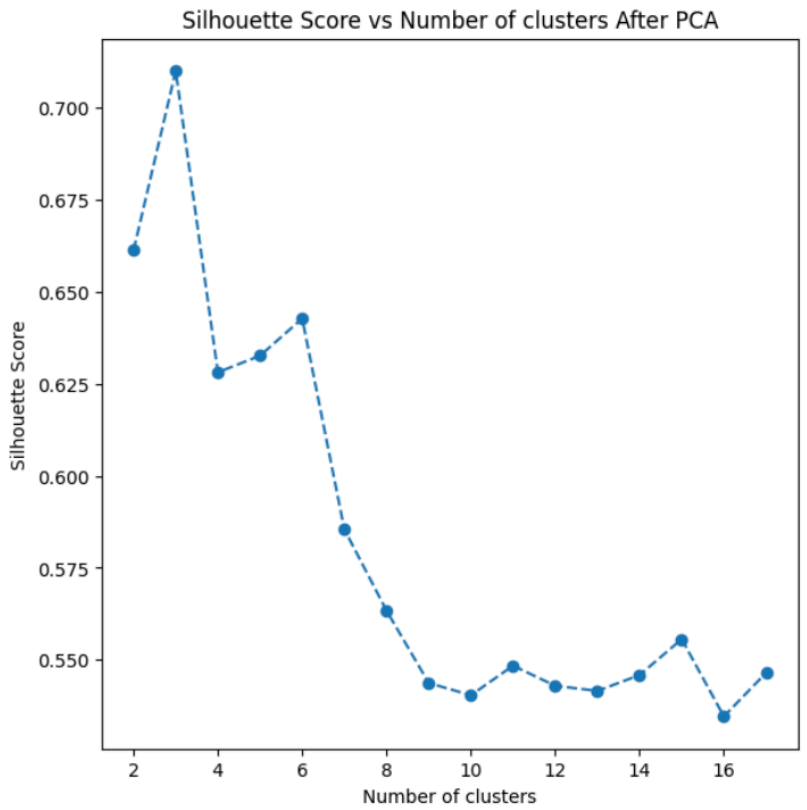
Hierarchial Clustering

Decision Tree Decision Boundary



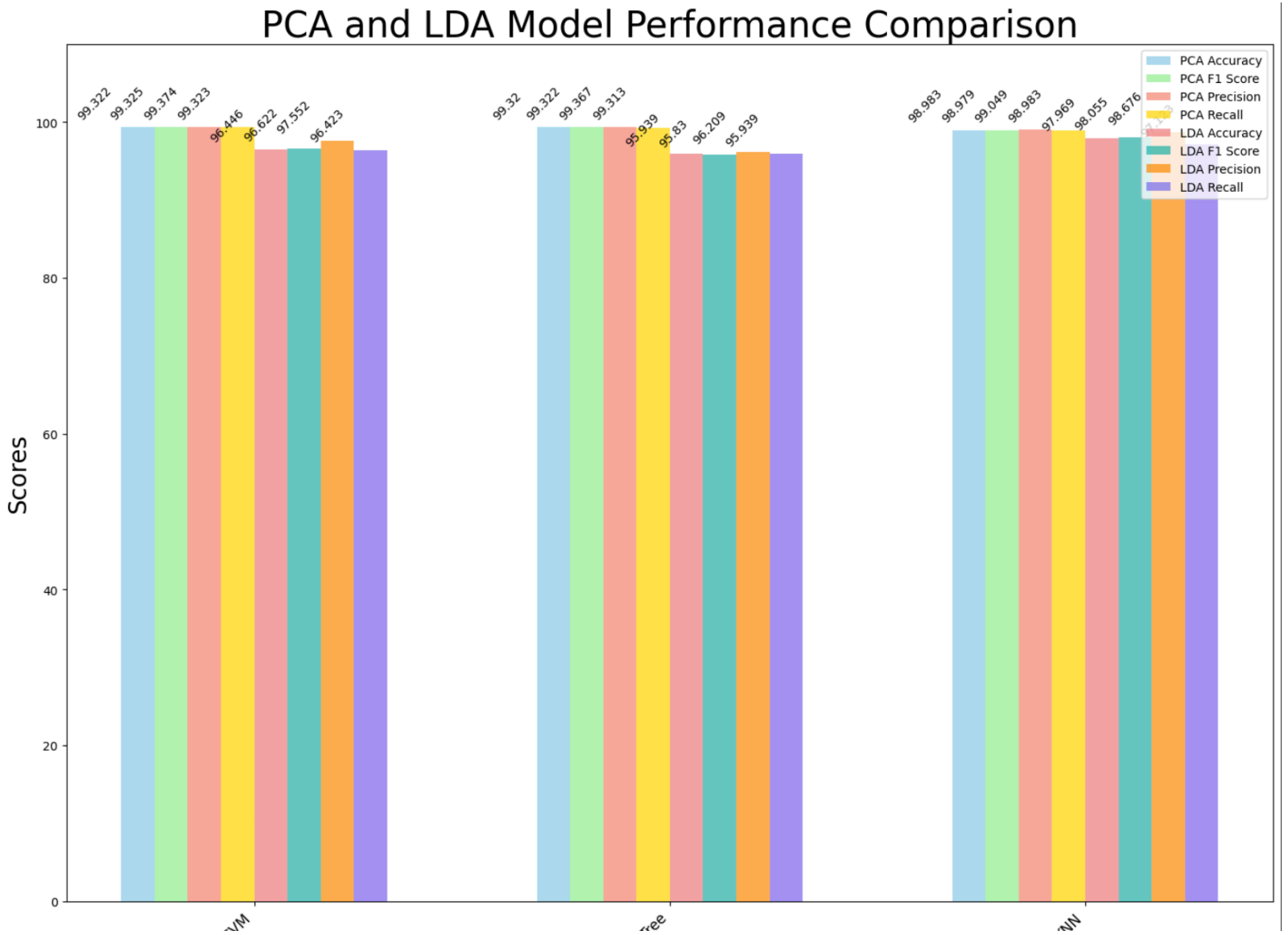
Agglomerative Clustering





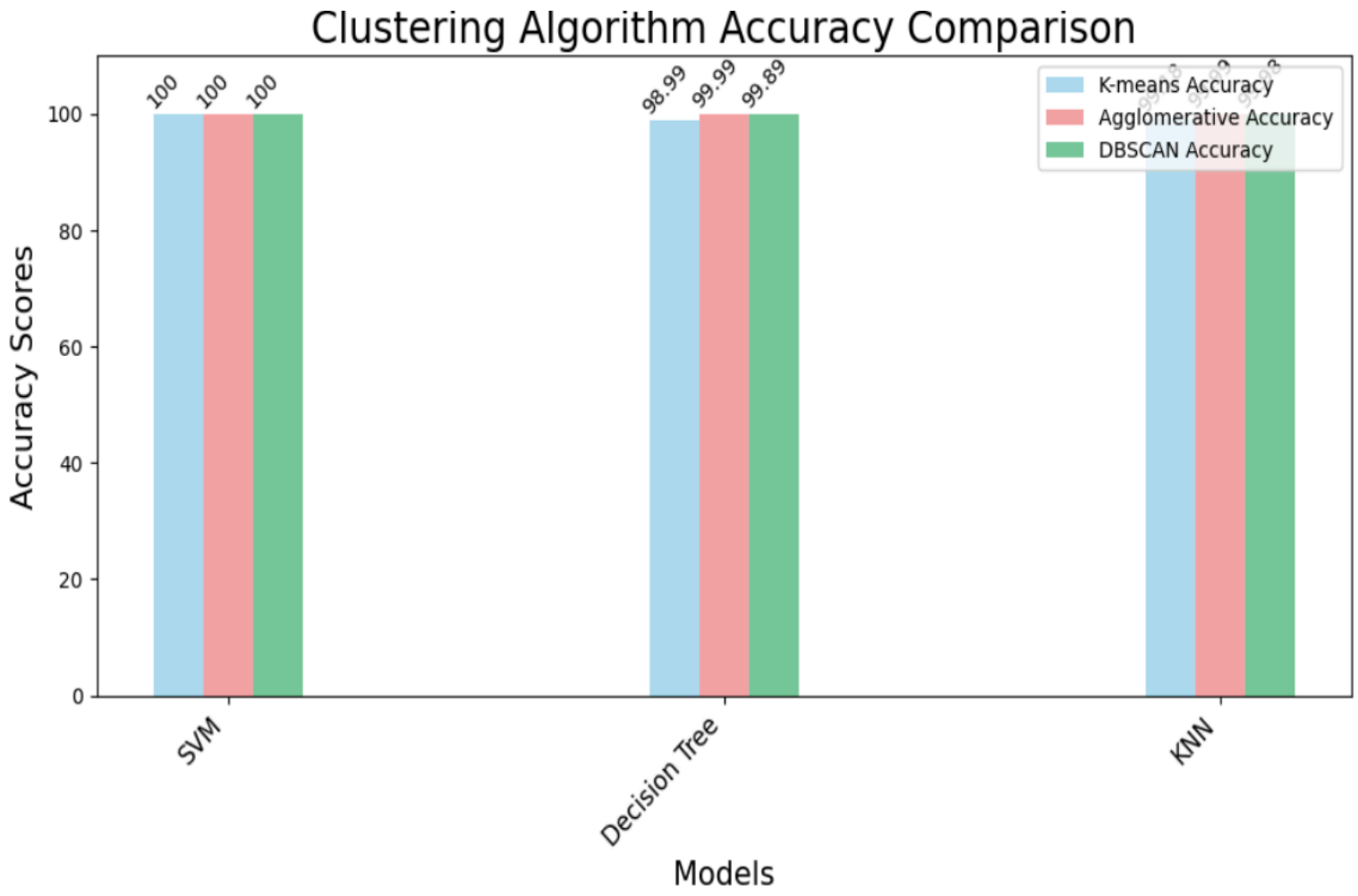
S.NO	MODEL	ACCURACY	F1 SCORE	PRECISION	RECALL
1	KNN-PCA	99%	99.2	99.3	99.18
2	DECISION TREE-PCA	99.18%	99.09	99.37	99.17
3	RANDOM FOREST	99.58%	99.14	99.28	99.09
4	SVM-Kernel-PCA				
5	DECISION-TREE-clustering	100%	93.91	94.816	93.8
6	SVM-RBF KERNAL clustering		97.5	97.523	97.26
7	KNN-CLUSTERING	99.28%	99.18	99.25	99.18

Performance Analysis for PCA AND LDA

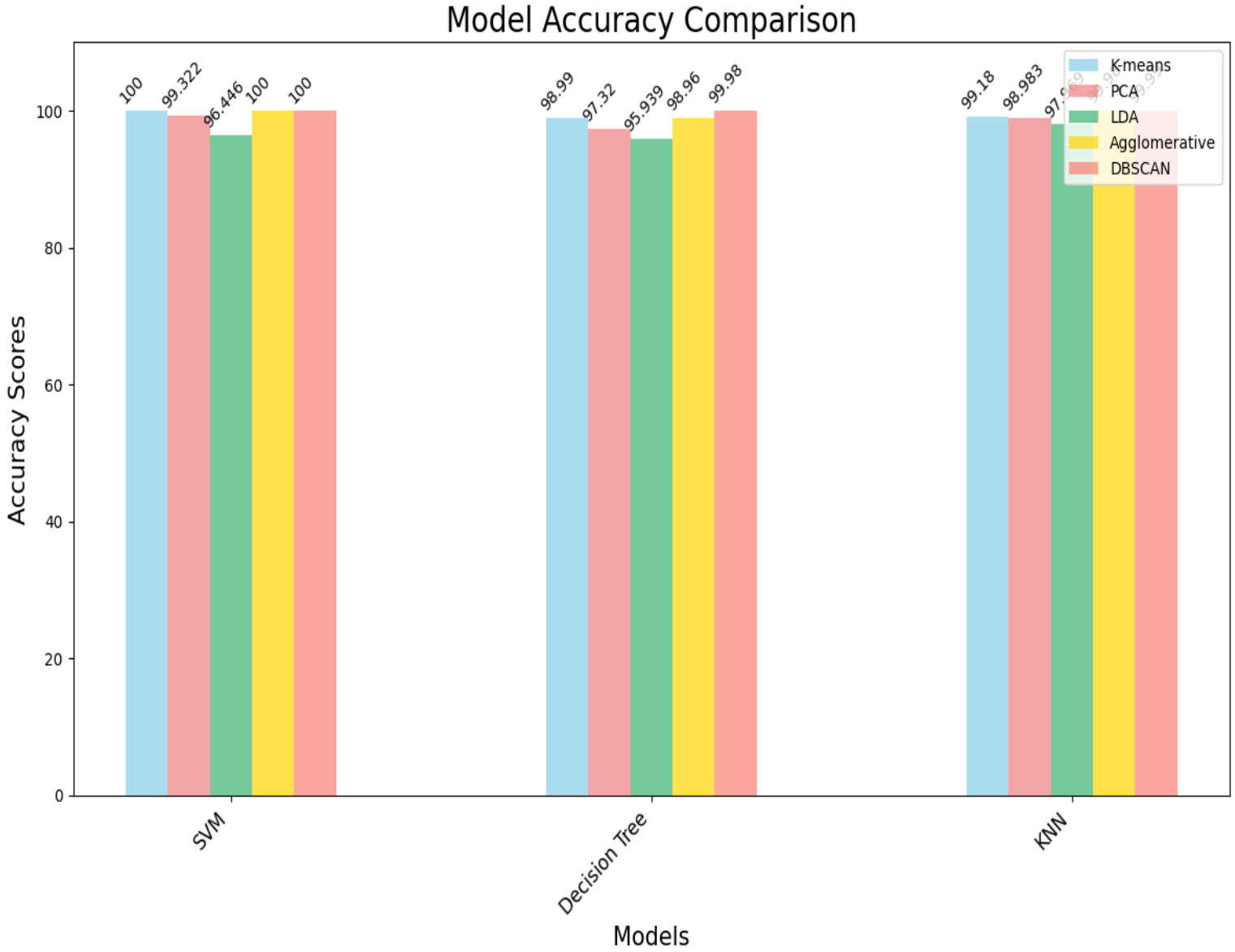


PERFORMANCE ANALYSIS
FOR K MEANS

ACCURARCY ANALYSIS FOR CLUSTERING ALGORITHMS



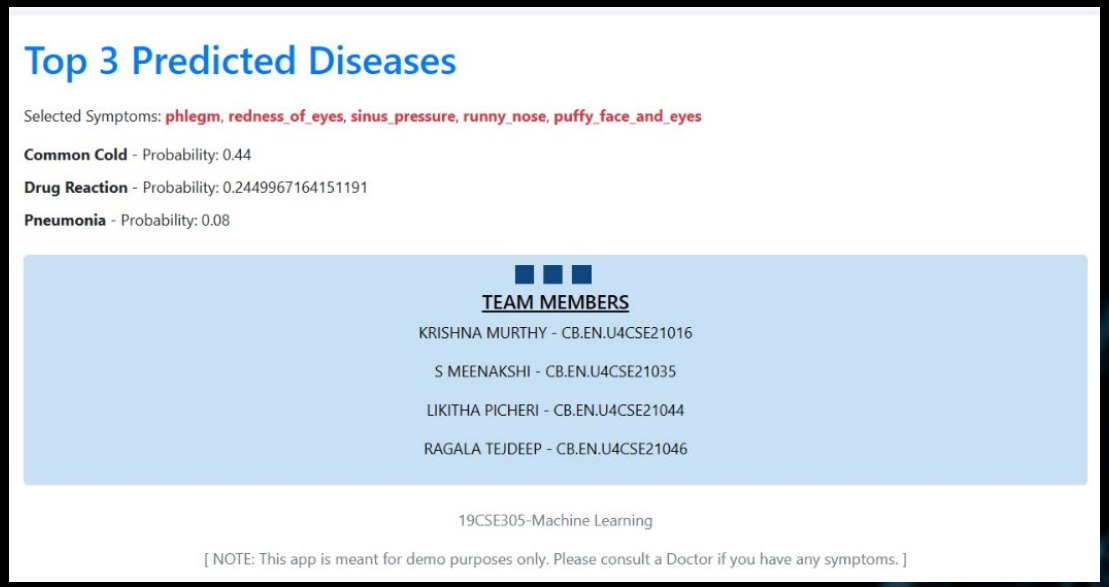
ACCURARCY COMPARISION OF ALL MODELS

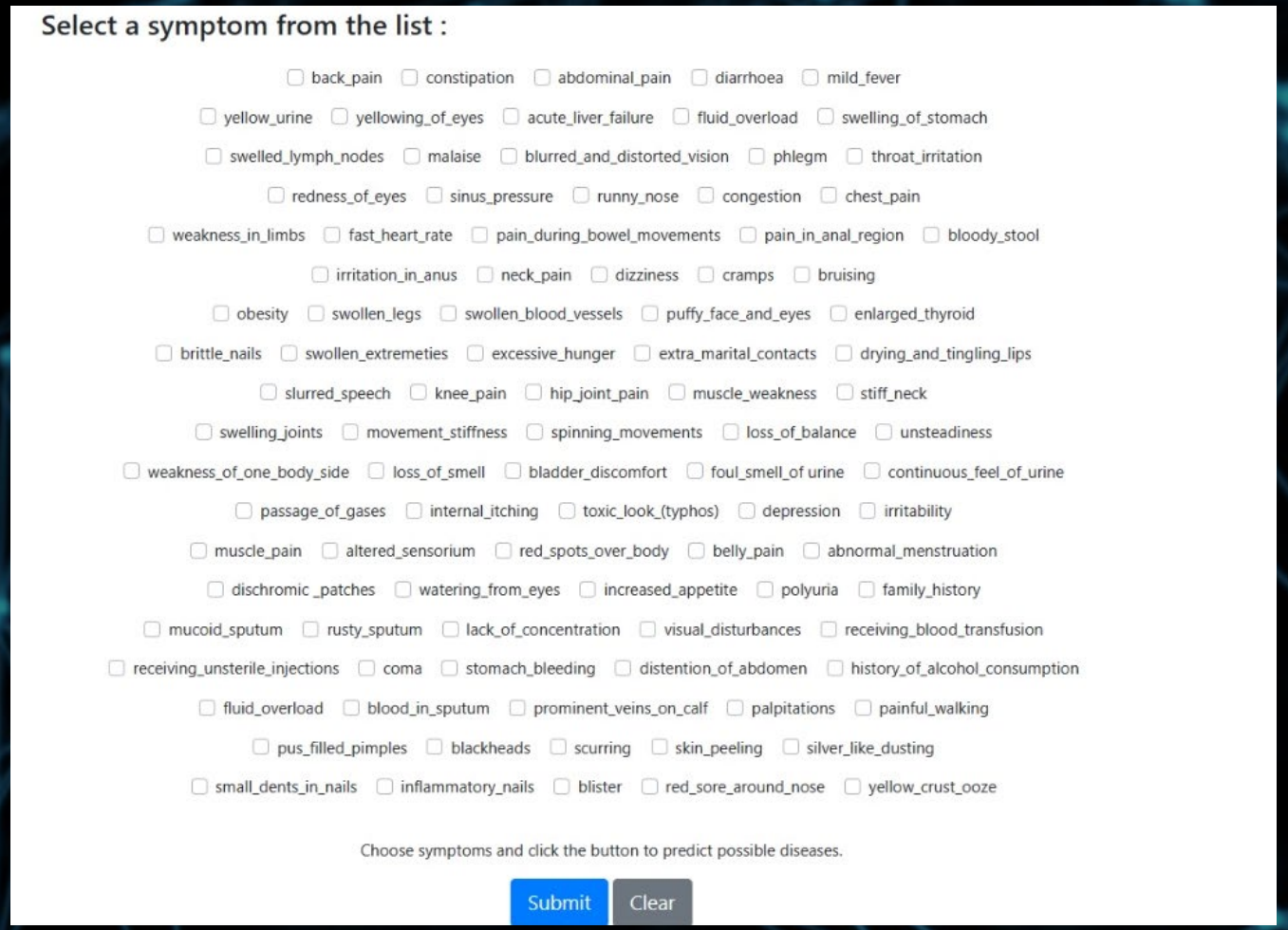


INFERENCE

In conclusion, the Support Vector Machine (SVM) model consistently demonstrated superior performance across clustering and Principal Component Analysis (PCA) tasks. Its accuracy, precision, recall, and F1 score outperformed alternative models, including Decision Trees and K-Nearest Neighbors, indicating its robustness in capturing complex patterns within the data. The SVM's exceptional performance suggests its suitability for both clustering and dimensionality reduction tasks, highlighting its versatility as a reliable and effective machine learning algorithm for the given dataset. These findings are integral to the model selection process, providing a strong basis for recommending SVM as the preferred choice in diverse analytical scenarios

GUI


Top 3 Predicted Diseases
Selected Symptoms: phlegm, redness_of_eyes, sinus_pressure, runny_nose, puffy_face_and_eyes
Common Cold - Probability: 0.44
Drug Reaction - Probability: 0.2449967164151191
Pneumonia - Probability: 0.08
TEAM MEMBERS
KRISHNA MURTHY - CB.EN.U4CSE21016
S MEENAKSHI - CB.EN.U4CSE21035
LIKITHA PICHERI - CB.EN.U4CSE21044
RAGALA TEJDEEP - CB.EN.U4CSE21046
19CSE305-Machine Learning
[NOTE: This app is meant for demo purposes only. Please consult a Doctor if you have any symptoms.]


Select a symptom from the list :
back_pain constipation abdominal_pain diarrhoea mild_fever
yellow_urine yellowing_of_eyes acute_liver_failure fluid_overload swelling_of_stomach
swelled_lymph_nodes malaise blurred_and_distorted_vision phlegm throat_irritation
redness_of_eyes sinus_pressure runny_nose congestion chest_pain
weakness_in_limbs fast_heart_rate pain_during_bowel_movements pain_in_anal_region bloody_stool
irritation_in_anus neck_pain dizziness cramps bruising
obesity swollen_legs swollen_blood_vessels puffy_face_and_eyes enlarged_thyroid
brittle_nails swollen_extremeties excessive_hunger extra_marital_contacts drying_and_tingling_lips
slurred_speech knee_pain hip_joint_pain muscle_weakness stiff_neck
swelling_joints movement_stiffness spinning_movements loss_of_balance unsteadiness
weakness_of_one_body_side loss_of_smell bladder_discomfort foul_smell_of urine continuous_feel_of_urine
passage_of_gases internal_itching toxic_look_(typhos) depression irritability
muscle_pain altered_sensorium red_spots_over_body belly_pain abnormal_menstruation
dischromic _patches watering_from_eyes increased_appetite polyuria family_history
mucoid_sputum rusty_sputum lack_of_concentration visual_disturbances receiving_blood_transfusion
receiving_unsterile_injections coma stomach_bleeding distention_of_abdomen history_of_alcohol_consumption
fluid_overload blood_in_sputum prominent_veins_on_calf palpitations painful_walking
pus_filled_pimples blackheads scurring skin_peeling silver_like_dusting
small_dents_in_nails inflammatory_nails blister red_sore_around_nose yellow_crust_ooze
Choose symptoms and click the button to predict possible diseases.
Submit
Clear