Datesets :

1-Pima Indians Diabetes Database

- **Number of Classes**: There are two classes in this dataset:
  - 0: Represents individuals without diabetes
  - 1: Represents individuals with diabetes
- **Labels**: The dataset includes the following attributes (features) along with the class label:

1. Number of times pregnant

2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skinfold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

- **Total Number of Samples**: The dataset contains 768 samples/instances.

2-Boston Housing

- **Number of Classes**: This dataset is typically used for regression tasks, so it doesn't have distinct classes like classification datasets. Instead, the target variable is usually the median value of owner-occupied homes in $1000s (MEDV), which is a continuous variable representing house prices.
- **Labels**: The dataset includes the following features:

1. CRIM: Per capita crime rate by town
2. ZN: Proportion of residential land zoned for lots over 25,000 sq. ft.

.3 INDUS: Proportion of non-retail business acres per town

.4 CHAS: Charles River dummy variable (1 if tract bounds river; 0 otherwise)

.5 NOX: Nitric oxides concentration (parts per 10 million)

.6 RM: Average number of rooms per dwelling

.7 AGE: Proportion of owner-occupied units built prior to 1940

.8 DIS: Weighted distances to five Boston employment centers

.9 RAD: Index of accessibility to radial highways

.10 TAX: Full-value property tax rate per $10,000

PTRATIO: Pupil-teacher ratio by town .11

B: 1000(Bk - 0.63)^2 where Bk is the proportion of Black residents by town .12

LSTAT: Percentage of lower status of the population .13

MEDV: Median value of owner-occupied homes in $1000s (target variable) .14

- **Total Number of Samples**: The dataset contains 506 samples/instances.
- **Size of Each Sample**: Each sample in this dataset consists of 13 input features (attributes) and the target

variable (median value of owner-occupied homes).

3-Raisin binary classification

- **Number of Classes**: There are two classes in this dataset:
- 0: Indicates non-defective raisins
- 1: Indicates defective raisins
- **Labels**: The dataset includes the following attributes along with the class label:

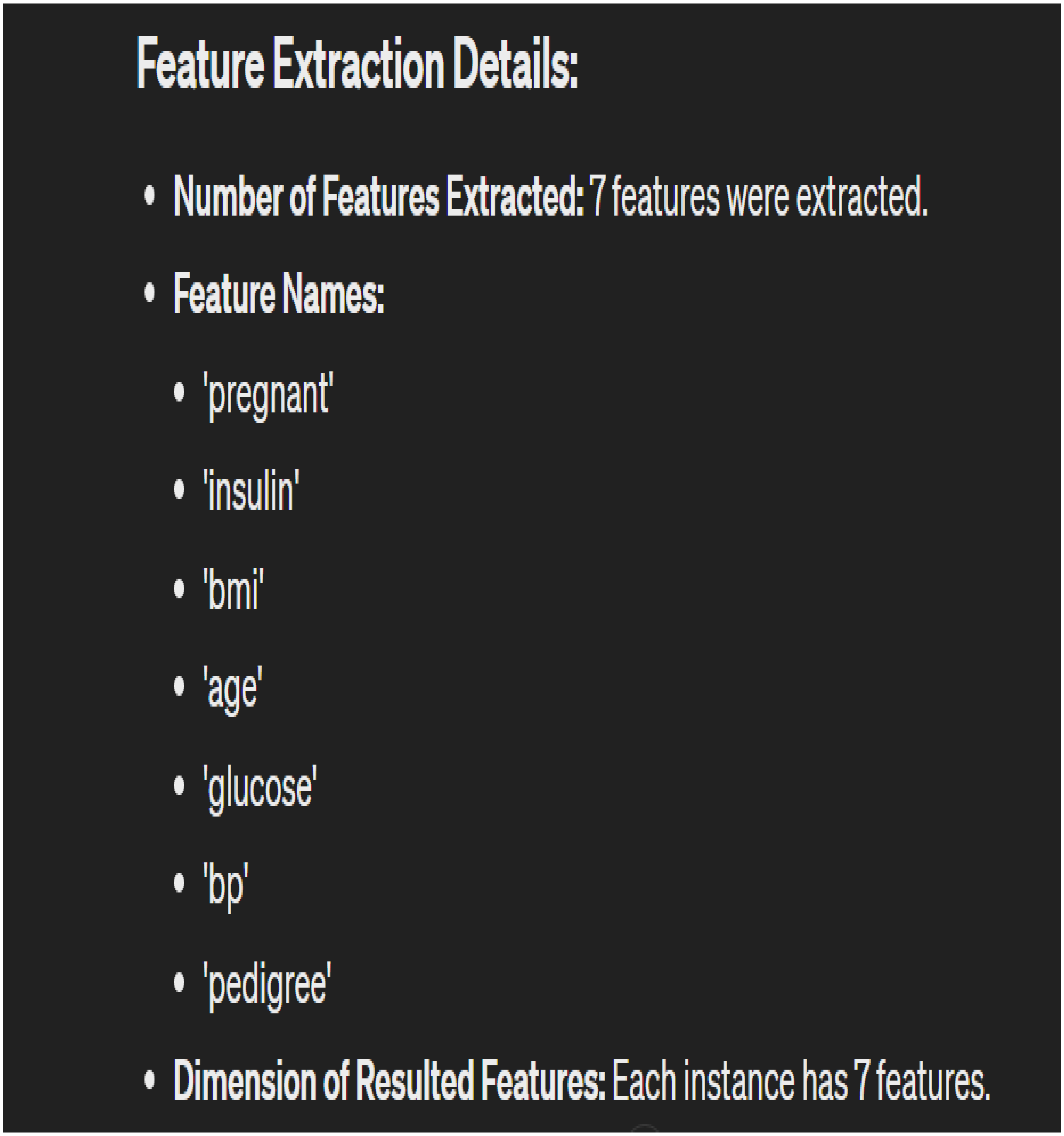
1. **Area** - The area of the raisin
2. **Perimeter** - The perimeter of the raisin
3. **MajorAxisLength** - The major axis length of the raisin

4. **MinorAxisLength** - The minor axis length of the raisin
5. **AspectRation** - The aspect ratio of the raisin
6. **Eccentricity** - The eccentricity of the raisin
7. **ConvexArea** - The convex area of the raisin
8. **EquivDiameter** - The equivalent diameter of the raisin
9. **Extent** - The extent of the raisin
10. **Solidity** - The solidity of the raisin
11. **Roundness** - The roundness of the raisin
12. **Compactness** - The compactness of the raisin

13. **ShapeFactor1** - Shape factor 1 of the raisin
14. **ShapeFactor2** - Shape factor 2 of the raisin
15. **ShapeFactor3** - Shape factor 3 of the raisin
16. **ShapeFactor4** - Shape factor 4 of the raisin
17. **Class** - Class variable (0 or 1)

- **Total Number of Samples**: The dataset contains a total of 106 samples/instances.
- **Size of Each Sample**: Each sample in this dataset contains 17 attributes including the class label.

Decision Tree Model:

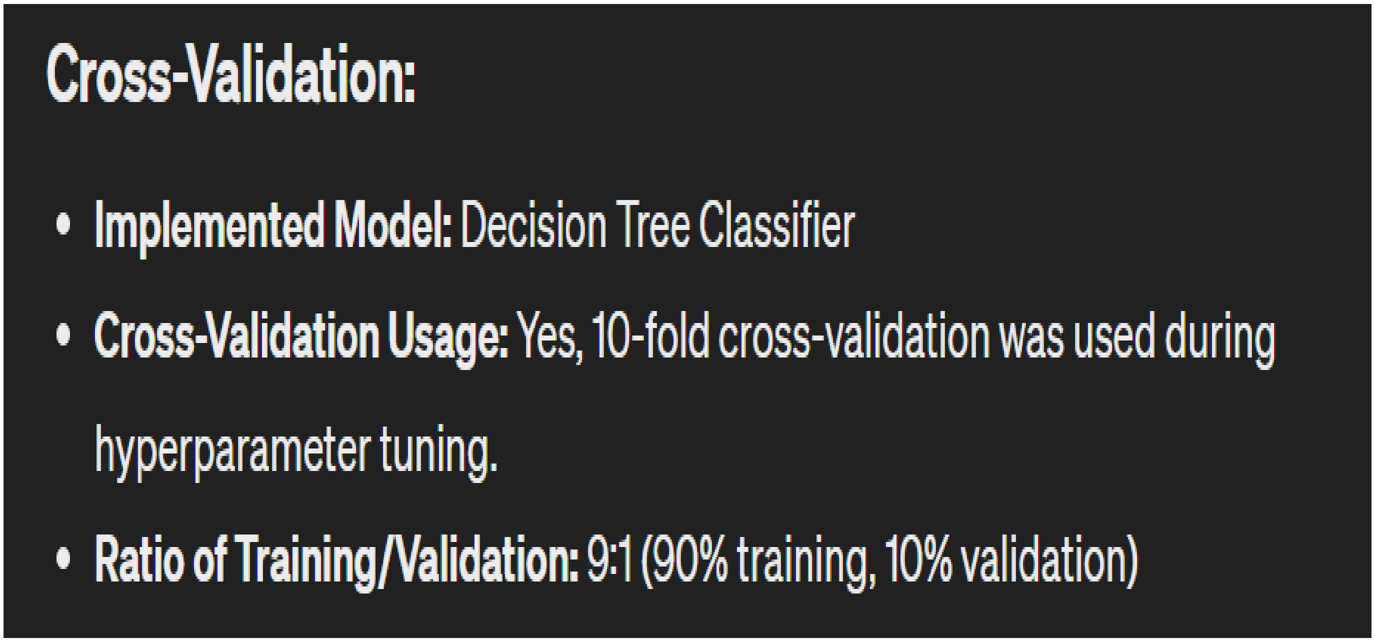


Feature Extraction Details:

- **Number of Features Extracted:** 7 features were extracted.
- **Feature Names:**
    - 'pregnant'
    - 'insulin'
    - 'bmi'
    - 'age'
    - 'glucose'
    - 'bp'
    - 'pedigree'
- **Dimension of Resulted Features:** Each instance has 7 features.

Results Details (on Testing Data):

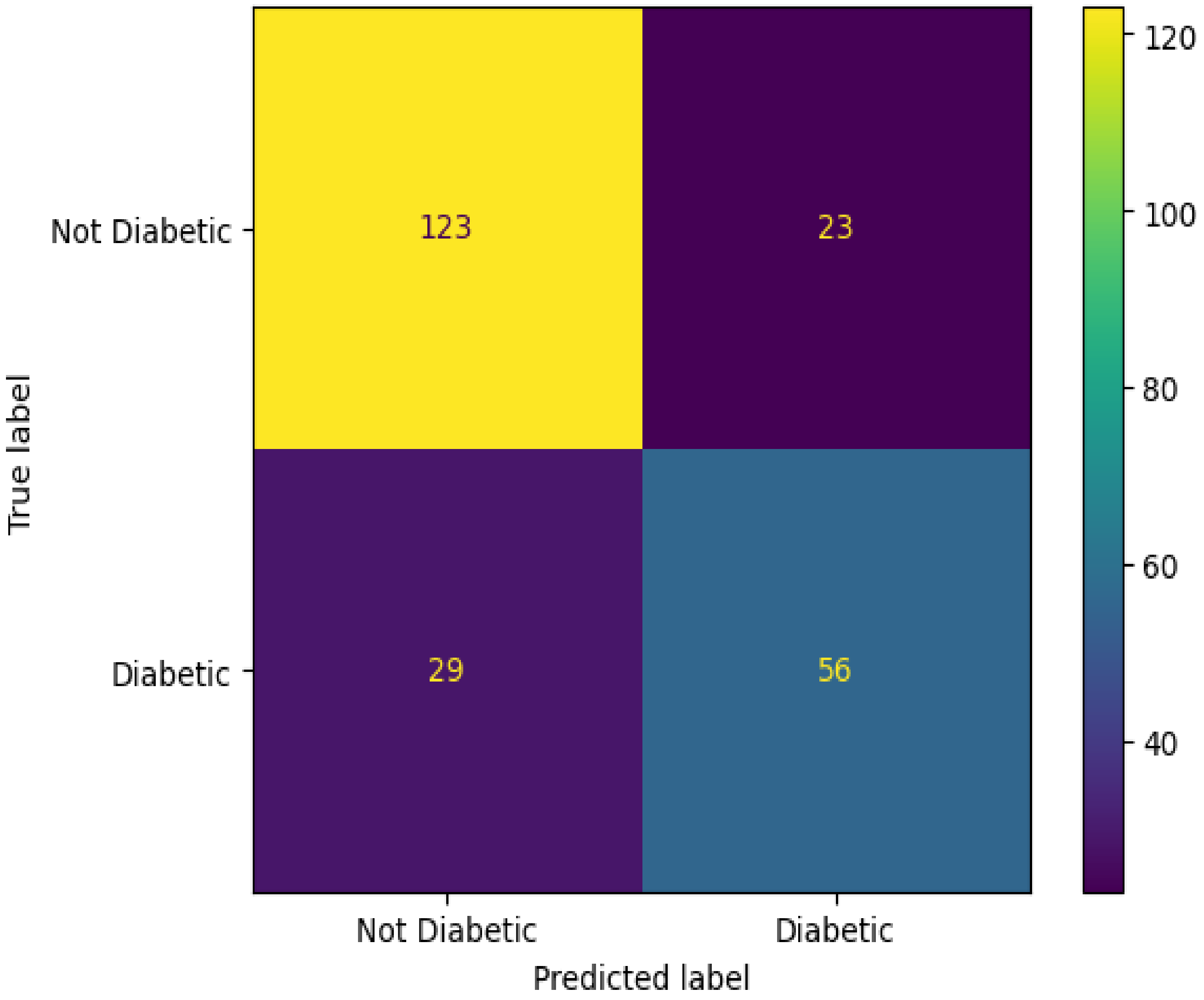
- **Accuracy with Hyperparameter Tuning:** 77%



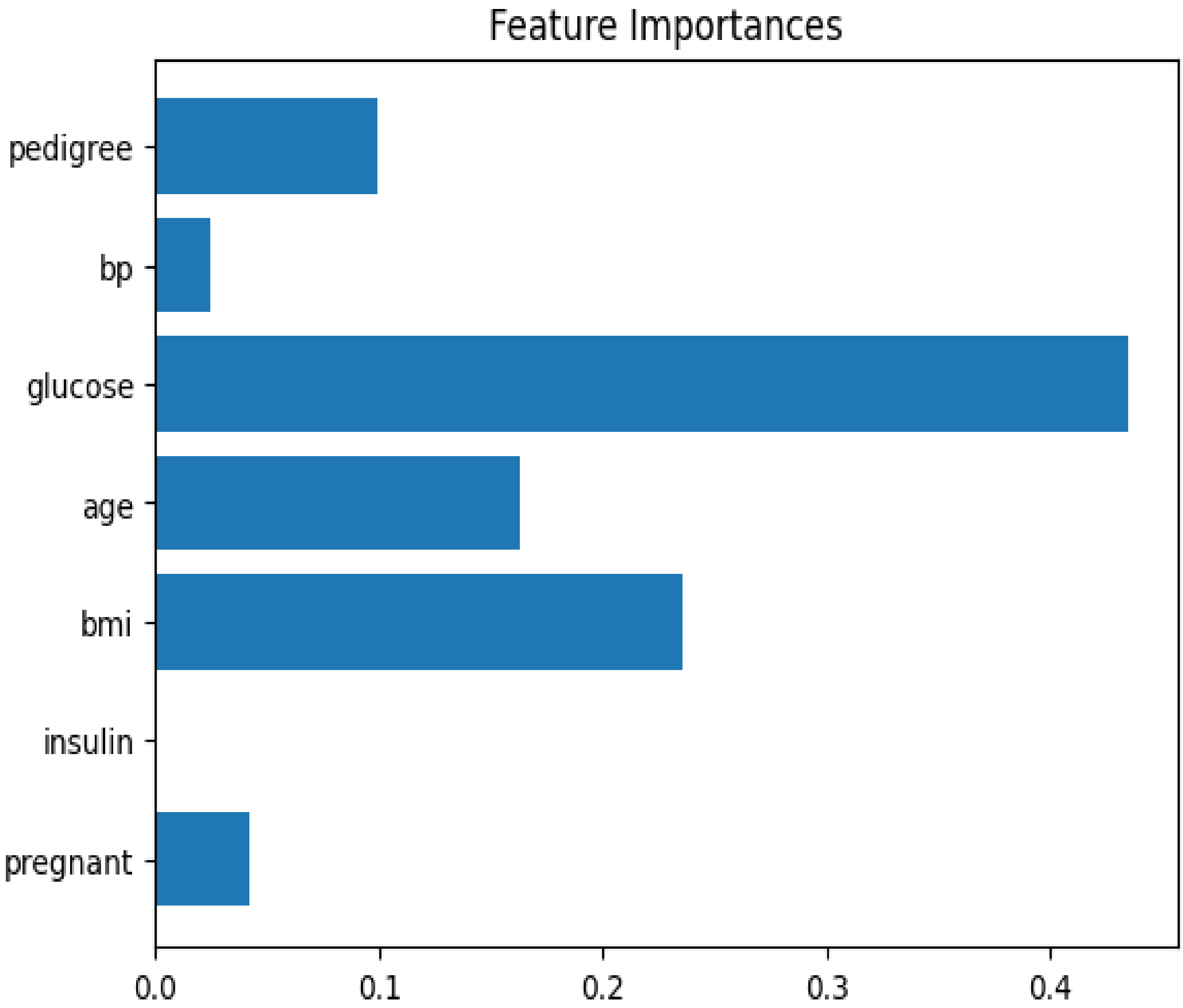
Cross-Validation:

- **Implemented Model:** Decision Tree Classifier
- **Cross-Validation Usage:** Yes, 10-fold cross-validation was used during hyperparameter tuning.
- **Ratio of Training/Validation:** 9:1 (90% training, 10% validation)

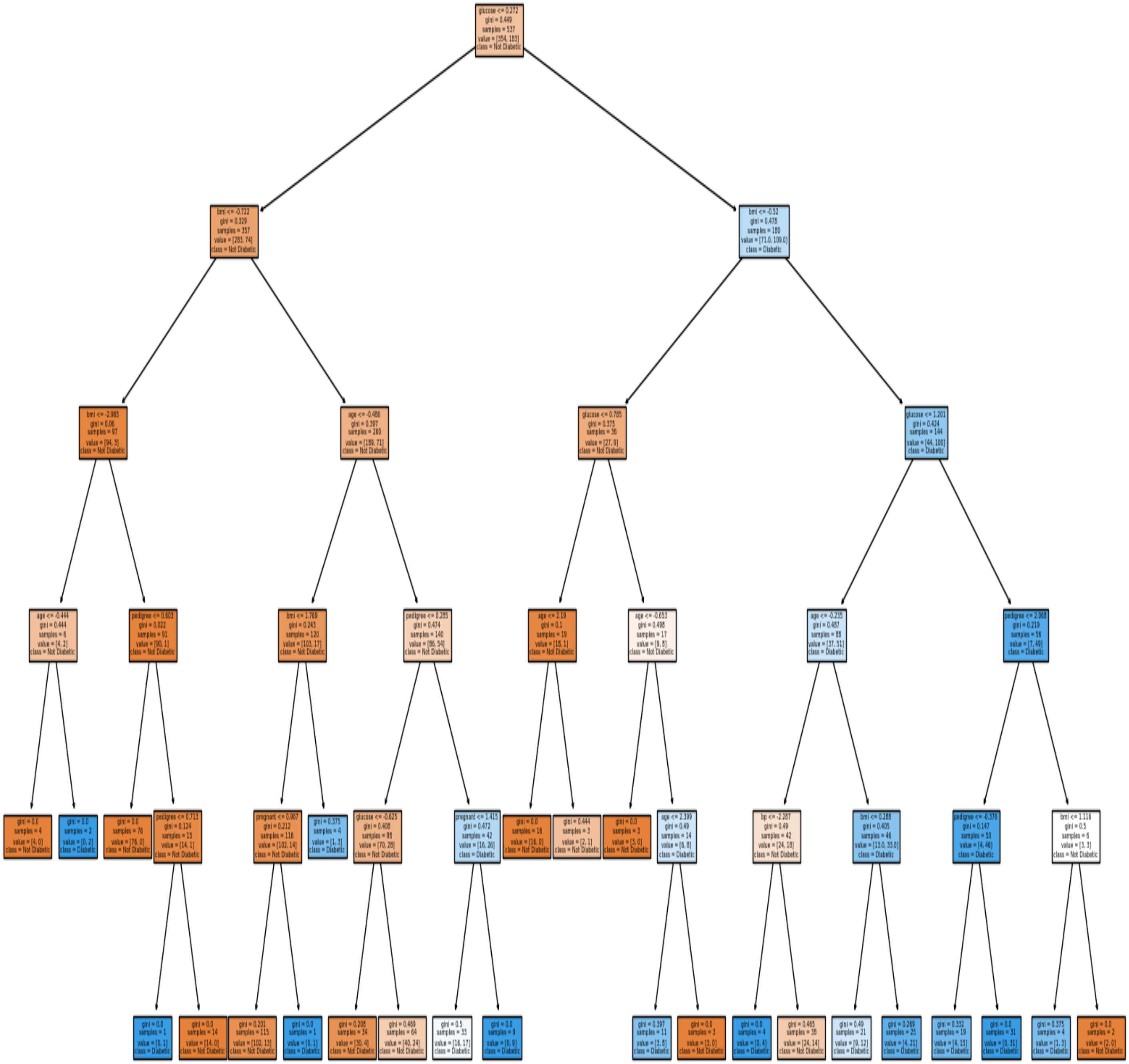
Confusion Matrix:



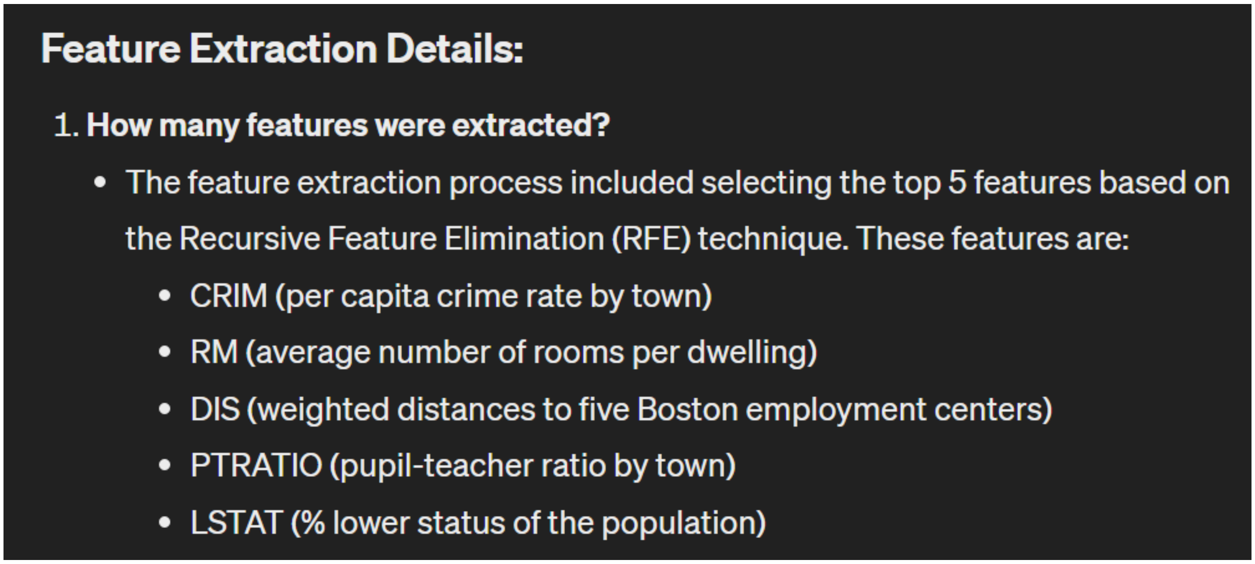
Feature Importance:



Decision tree:

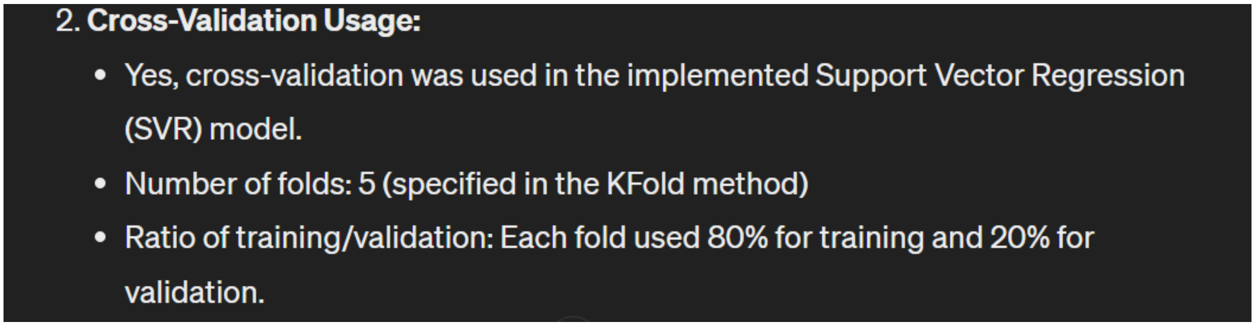

glucose <= 0.272 gini = 0.449 samples = 537 value = [354, 183] class = Not Diabetic
bmi <= -0.722 gini = 0.329 samples = 357 value = [283, 74] class = Not Diabetic
bmi <= -0.52 gini = 0.478 samples = 180 value = [71.0, 109.0] class = Diabetic
bmi <= -2.965 gini = 0.06 samples = 97 value = [94, 3] class = Not Diabetic
age <= -0.486 gini = 0.397 samples = 260 value = [189, 71] class = Not Diabetic
glucose <= 0.785 gini = 0.375 samples = 36 value = [27, 9] class = Not Diabetic
glucose <= 1.201 gini = 0.424 samples = 144 value = [44, 100] class = Diabetic
age <= -0.444 gini = 0.444 samples = 6 value = [4, 2] class = Not Diabetic
pedigree <= 0.603 gini = 0.022 samples = 91 value = [90, 1] class = Not Diabetic
bmi <= 1.769 gini = 0.243 samples = 120 value = [103, 17] class = Not Diabetic
pedigree <= 0.285 gini = 0.474 samples = 140 value = [86, 54] class = Not Diabetic
age <= 2.19 gini = 0.1 samples = 19 value = [18, 1] class = Not Diabetic
age <= -0.653 gini = 0.498 samples = 17 value = [9, 8] class = Not Diabetic
age <= -0.235 gini = 0.487 samples = 88 value = [37, 51] class = Diabetic
pedigree <= 2.068 gini = 0.219 samples = 56 value = [7, 49] class = Diabetic
gini = 0.0 samples = 4 value = [4, 0] class = Not Diabetic
gini = 0.0 samples = 2 value = [0, 2] class = Diabetic
gini = 0.0 samples = 76 value = [76, 0] class = Not Diabetic
pedigree <= 0.713 gini = 0.124 samples = 15 value = [14, 1] class = Not Diabetic
pregnant <= 0.967 gini = 0.212 samples = 116 value = [102, 14] class = Not Diabetic
gini = 0.375 samples = 4 value = [1, 3] class = Diabetic
glucose <= -0.625 gini = 0.408 samples = 98 value = [70, 28] class = Not Diabetic
pregnant <= 1.415 gini = 0.472 samples = 42 value = [16, 26] class = Diabetic
gini = 0.0 samples = 16 value = [16, 0] class = Not Diabetic
gini = 0.444 samples = 3 value = [2, 1] class = Not Diabetic
gini = 0.0 samples = 3 value = [3, 0] class = Not Diabetic
age <= 2.399 gini = 0.49 samples = 14 value = [6, 8] class = Diabetic
bp <= -2.287 gini = 0.49 samples = 42 value = [24, 18] class = Not Diabetic
bmi <= 0.288 gini = 0.405 samples = 46 value = [13.0, 33.0] class = Diabetic
pedigree <= -0.376 gini = 0.147 samples = 50 value = [4, 46] class = Diabetic
bmi <= 1.116 gini = 0.5 samples = 6 value = [3, 3] class = Not Diabetic
gini = 0.0 samples = 1 value = [0, 1] class = Diabetic
gini = 0.0 samples = 14 value = [14, 0] class = Not Diabetic
gini = 0.201 samples = 115 value = [102, 13] class = Not Diabetic
gini = 0.0 samples = 1 value = [0, 1] class = Diabetic
gini = 0.208 samples = 34 value = [30, 4] class = Not Diabetic
gini = 0.469 samples = 64 value = [40, 24] class = Not Diabetic
gini = 0.5 samples = 33 value = [16, 17] class = Diabetic
gini = 0.0 samples = 9 value = [0, 9] class = Diabetic
gini = 0.397 samples = 11 value = [3, 8] class = Diabetic
gini = 0.0 samples = 3 value = [3, 0] class = Not Diabetic
gini = 0.0 samples = 4 value = [0, 4] class = Diabetic
gini = 0.465 samples = 38 value = [24, 14] class = Not Diabetic
gini = 0.49 samples = 21 value = [9, 12] class = Diabetic
gini = 0.269 samples = 25 value = [4, 21] class = Diabetic
gini = 0.332 samples = 19 value = [4, 15] class = Diabetic
gini = 0.0 samples = 31 value = [0, 31] class = Diabetic
gini = 0.375 samples = 4 value = [1, 3] class = Diabetic
gini = 0.0 samples = 2 value = [2, 0] class = Not Diabetic

SVR Model :

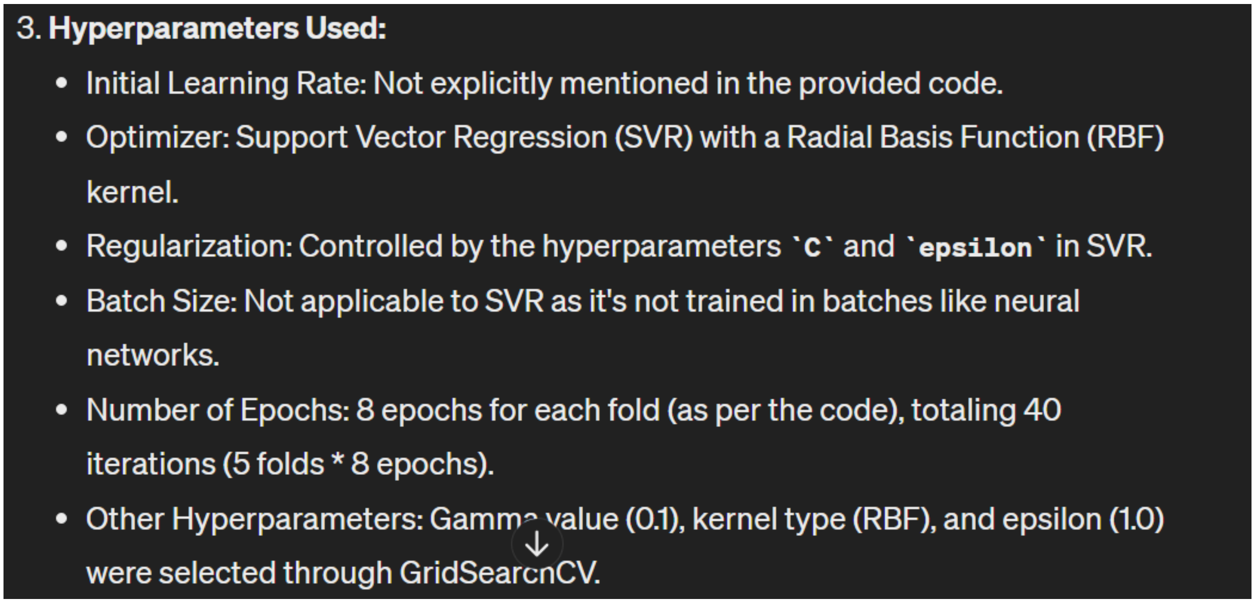


Feature Extraction Details:

1. **How many features were extracted?**
   - The feature extraction process included selecting the top 5 features based on the Recursive Feature Elimination (RFE) technique. These features are:
     - CRIM (per capita crime rate by town)
     - RM (average number of rooms per dwelling)
     - DIS (weighted distances to five Boston employment centers)
     - PTRATIO (pupil-teacher ratio by town)
     - LSTAT (% lower status of the population)

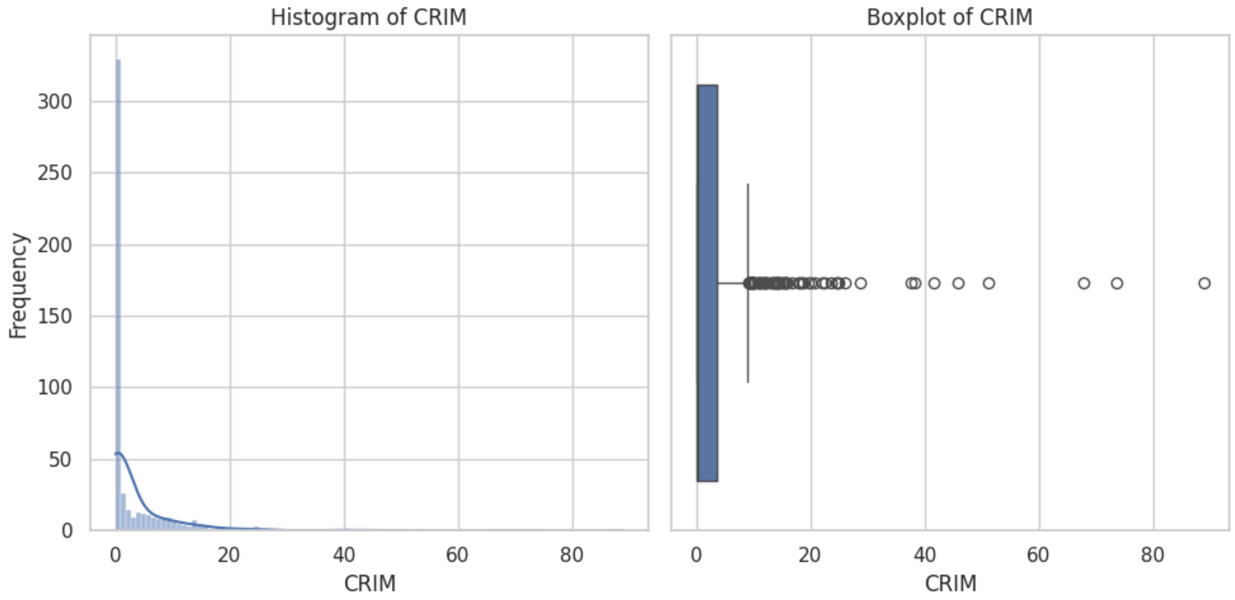


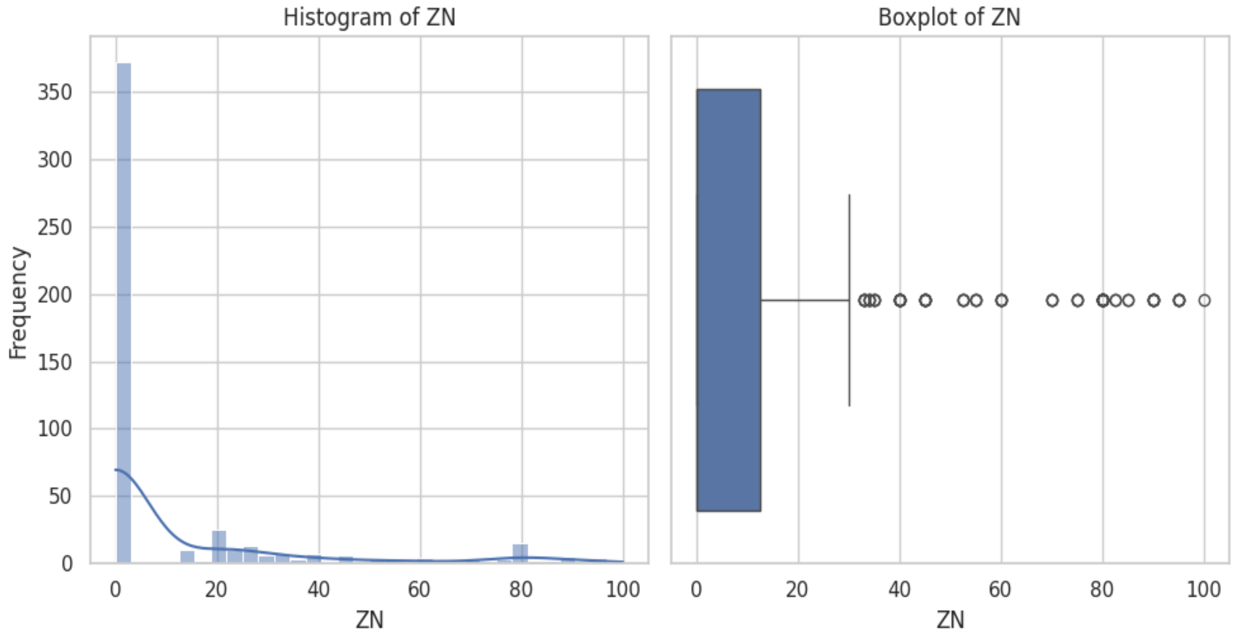
2. **Cross-Validation Usage:**
   - Yes, cross-validation was used in the implemented Support Vector Regression (SVR) model.
   - Number of folds: 5 (specified in the KFold method)
   - Ratio of training/validation: Each fold used 80% for training and 20% for validation.

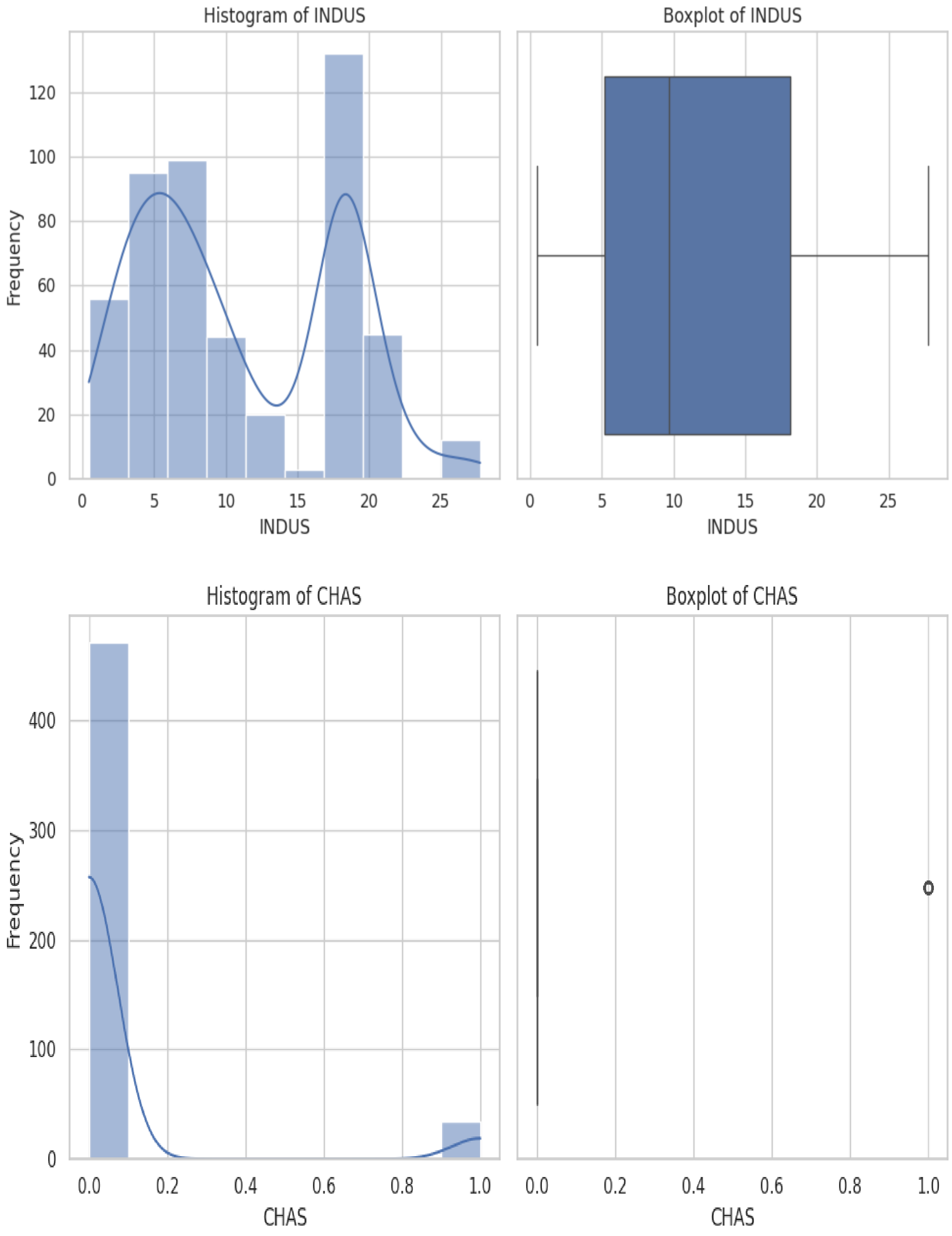


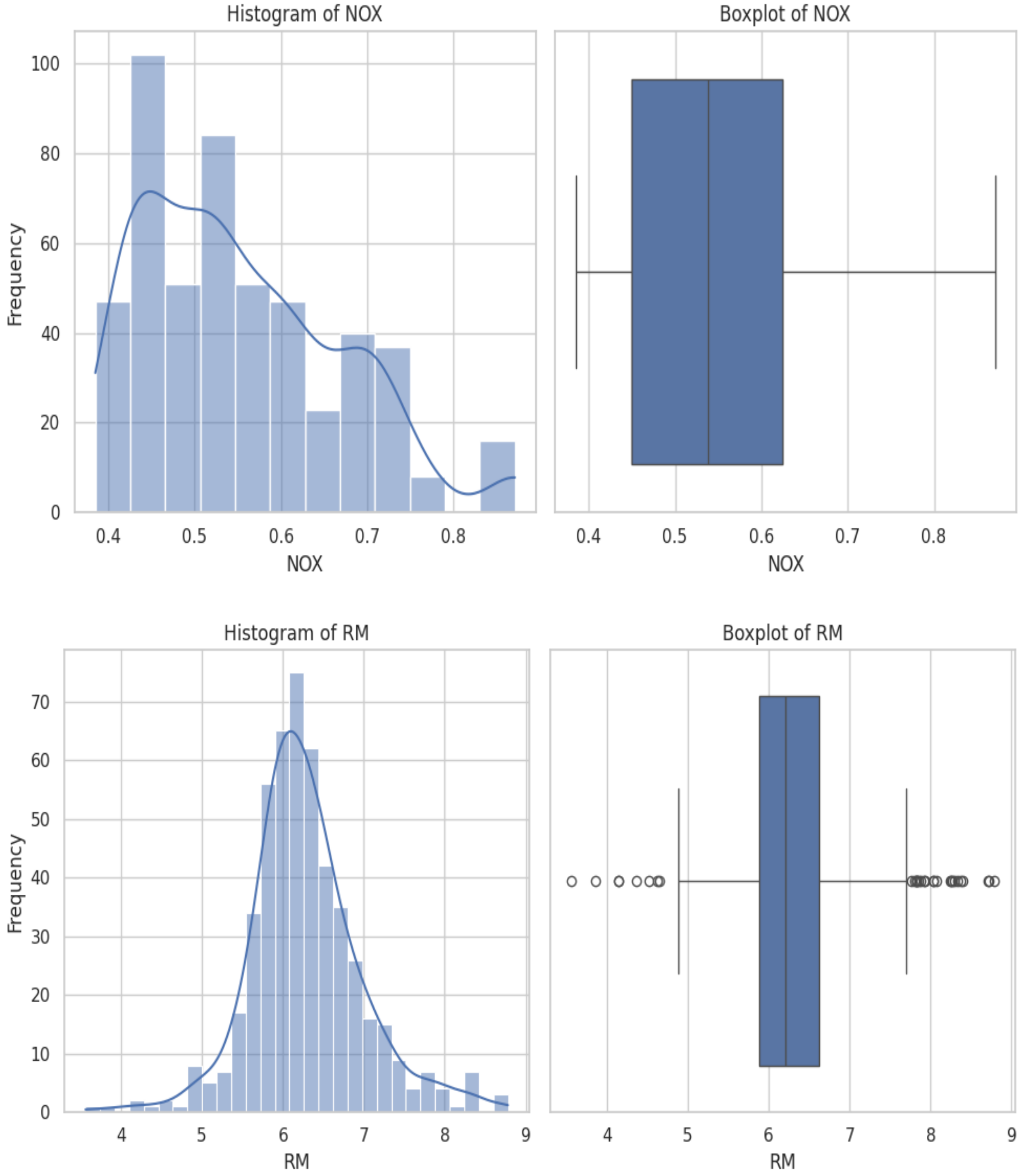
3. **Hyperparameters Used:**
   - Initial Learning Rate: Not explicitly mentioned in the provided code.
   - Optimizer: Support Vector Regression (SVR) with a Radial Basis Function (RBF) kernel.
   - Regularization: Controlled by the hyperparameters `C` and `epsilon` in SVR.
   - Batch Size: Not applicable to SVR as it's not trained in batches like neural networks.
   - Number of Epochs: 8 epochs for each fold (as per the code), totaling 40 iterations (5 folds * 8 epochs).
   - Other Hyperparameters: Gamma value (0.1), kernel type (RBF), and epsilon (1.0) were selected through GridSearchCV.

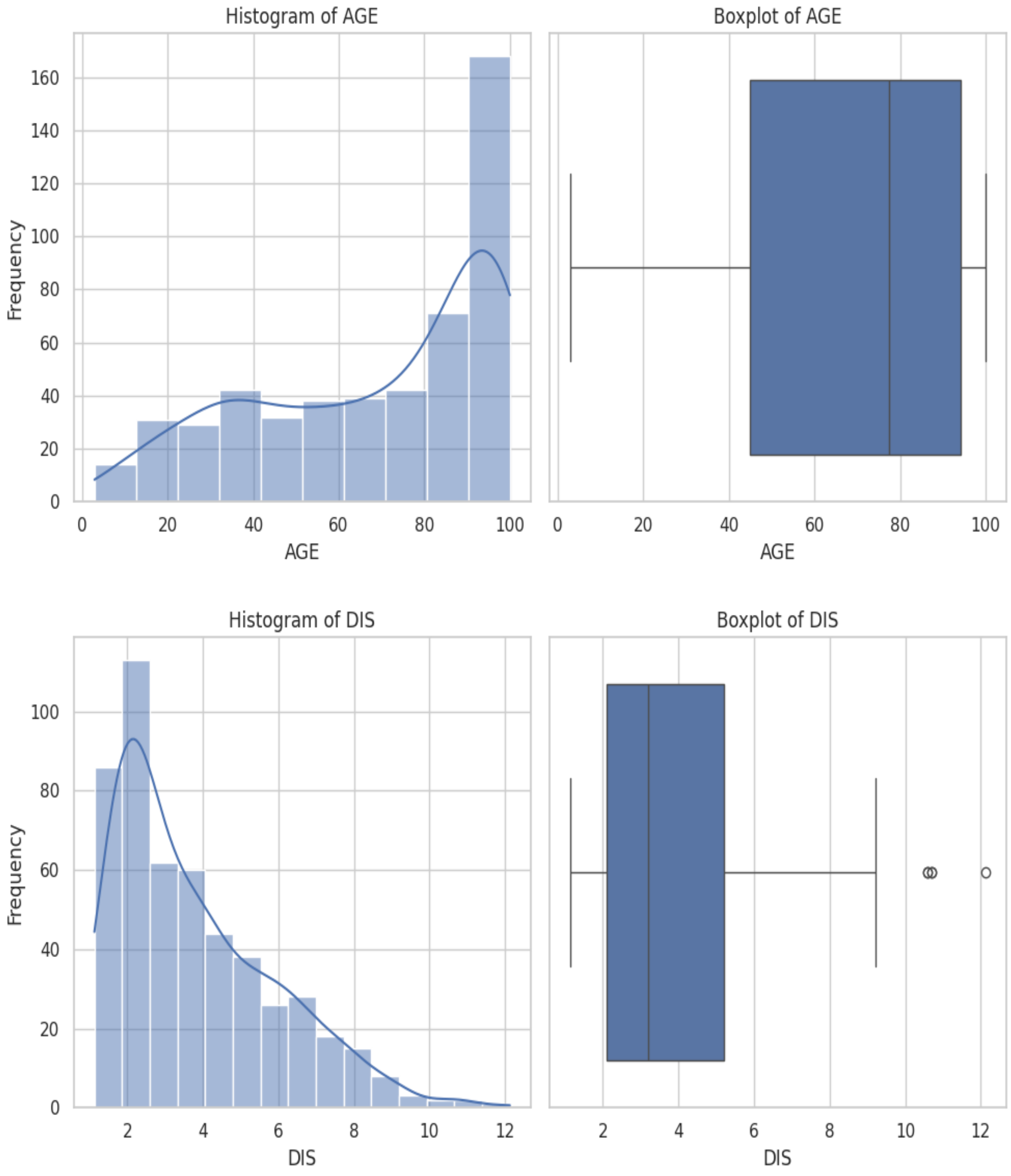
histogram & boxplot to check for outliers visually:

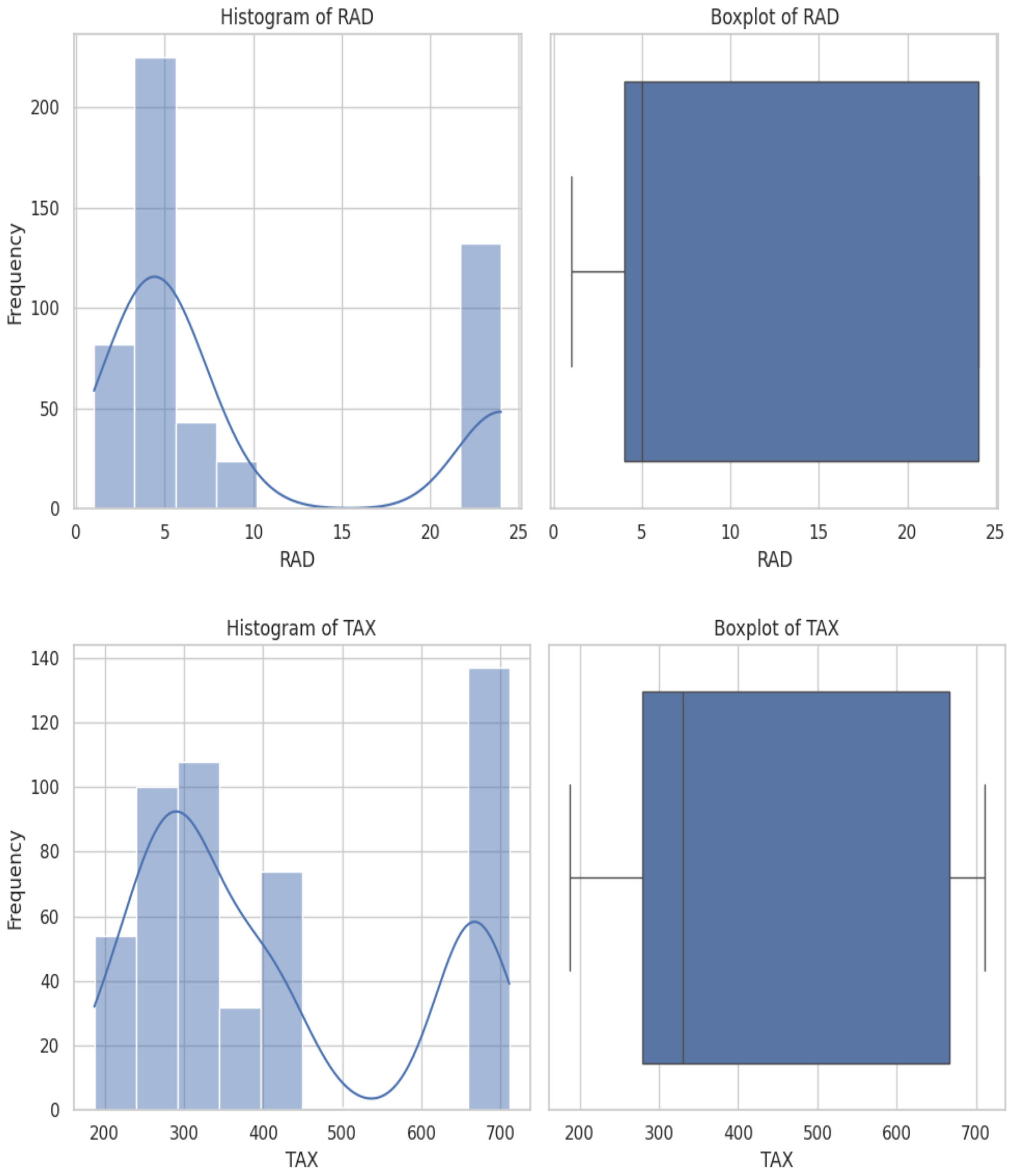


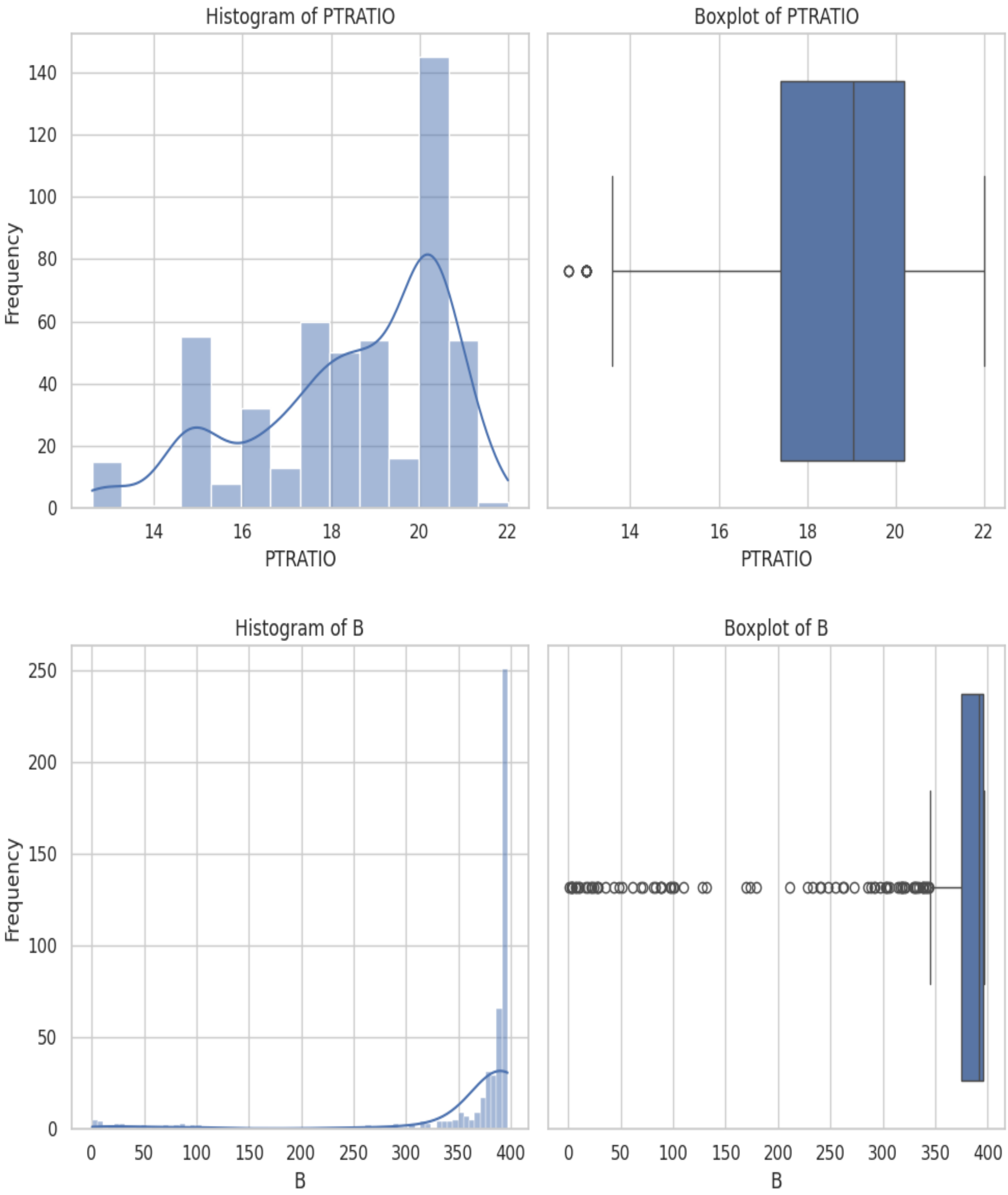


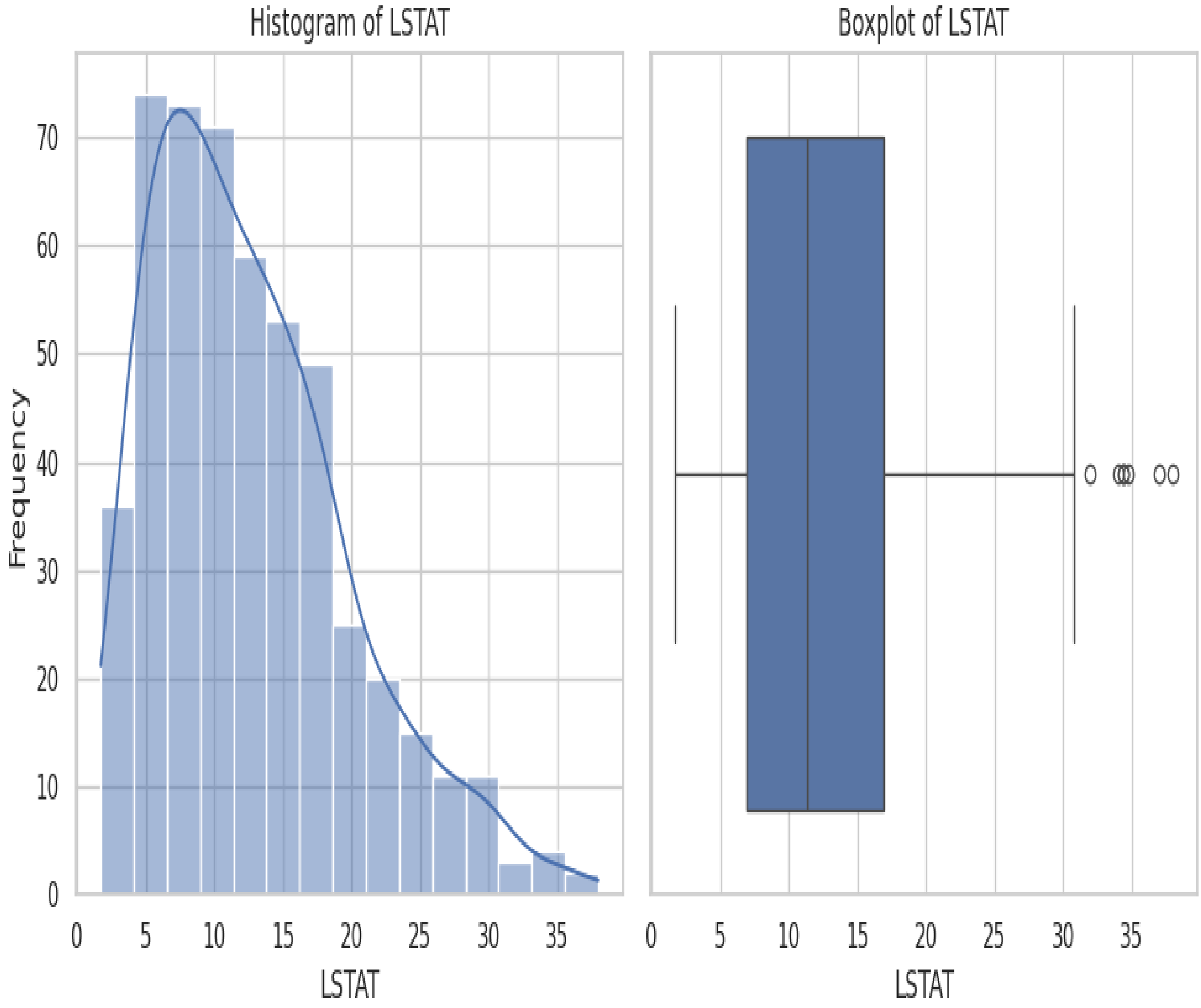

Histogram of INDUS
Frequency
0
20
40
60
80
100
120
0
5
10
15
20
25
INDUS
Boxplot of INDUS
0
5
10
15
20
25
INDUS
Histogram of CHAS
Frequency
0
100
200
300
400
0.0
0.2
0.4
0.6
0.8
1.0
CHAS
Boxplot of CHAS
0.0
0.2
0.4
0.6
0.8
1.0
CHAS


Histogram of NOX
Boxplot of NOX
Frequency
0
20
40
60
80
100
0.4
0.5
0.6
0.7
0.8
NOX
0.4
0.5
0.6
0.7
0.8
NOX
Histogram of RM
Boxplot of RM
Frequency
0
10
20
30
40
50
60
70
4
5
6
7
8
9
RM
4
5
6
7
8
9
RM


Histogram of AGE
Frequency
0
20
40
60
80
100
120
140
160
0
20
40
60
80
100
AGE
Boxplot of AGE
0
20
40
60
80
100
AGE
Histogram of DIS
Frequency
0
20
40
60
80
100
2
4
6
8
10
12
DIS
Boxplot of DIS
2
4
6
8
10
12
DIS


Histogram of RAD
Frequency
0
50
100
150
200
0
5
10
15
20
25
RAD
Boxplot of RAD
0
5
10
15
20
25
RAD
Histogram of TAX
Frequency
0
20
40
60
80
100
120
140
200
300
400
500
600
700
TAX
Boxplot of TAX
200
300
400
500
600
700
TAX


Histogram of PTRATIO
Frequency
0
20
40
60
80
100
120
140
14
16
18
20
22
PTRATIO
Boxplot of PTRATIO
14
16
18
20
22
PTRATIO
Histogram of B
Frequency
0
50
100
150
200
250
0
50
100
150
200
250
300
350
400
B
Boxplot of B
0
50
100
150
200
250
300
350
400
B


Histogram of LSTAT
Boxplot of LSTAT
Frequency
0
10
20
30
40
50
60
70
0
5
10
15
20
25
30
35
LSTAT
0
5
10
15
20
25
30
35
LSTAT

NN Model:

Features extracted details:

Area (not used) .1

Gives the number of pixels within the boundaries of the raisin.

MajorAxisLength .2

Gives the length of the main axis, which is the longest line that can be drawn on the raisin.

MinorAxisLength .3

Gives the length of the small axis, which is the shortest line that can be drawn on the raisin.

Eccentricity .4

It gives a measure of the eccentricity of the ellipse, which has the same moments as raisins.

ConvexArea (not used) .5

Gives the number of pixels of the smallest convex shell of the region formed by the raisin.

Extent .6

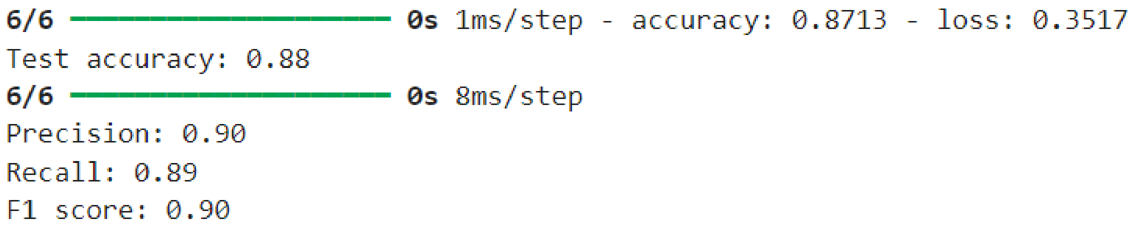
Gives the ratio of the region formed by the raisin to the total pixels in the bounding box.

Perimeter .7

It measures the environment by calculating the distance between the boundaries of the raisin and the pixels

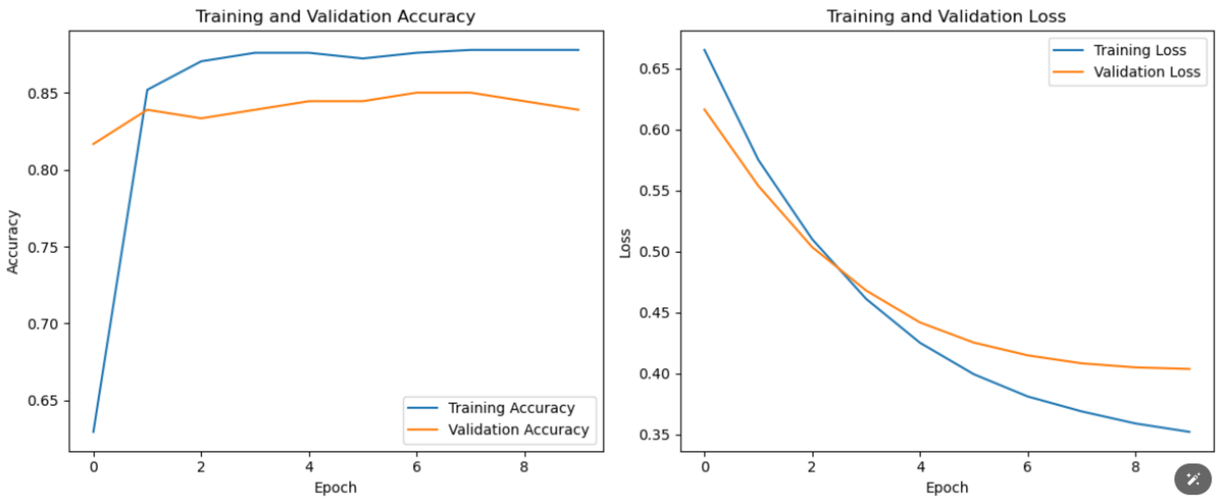
Results :

Accuracy:-

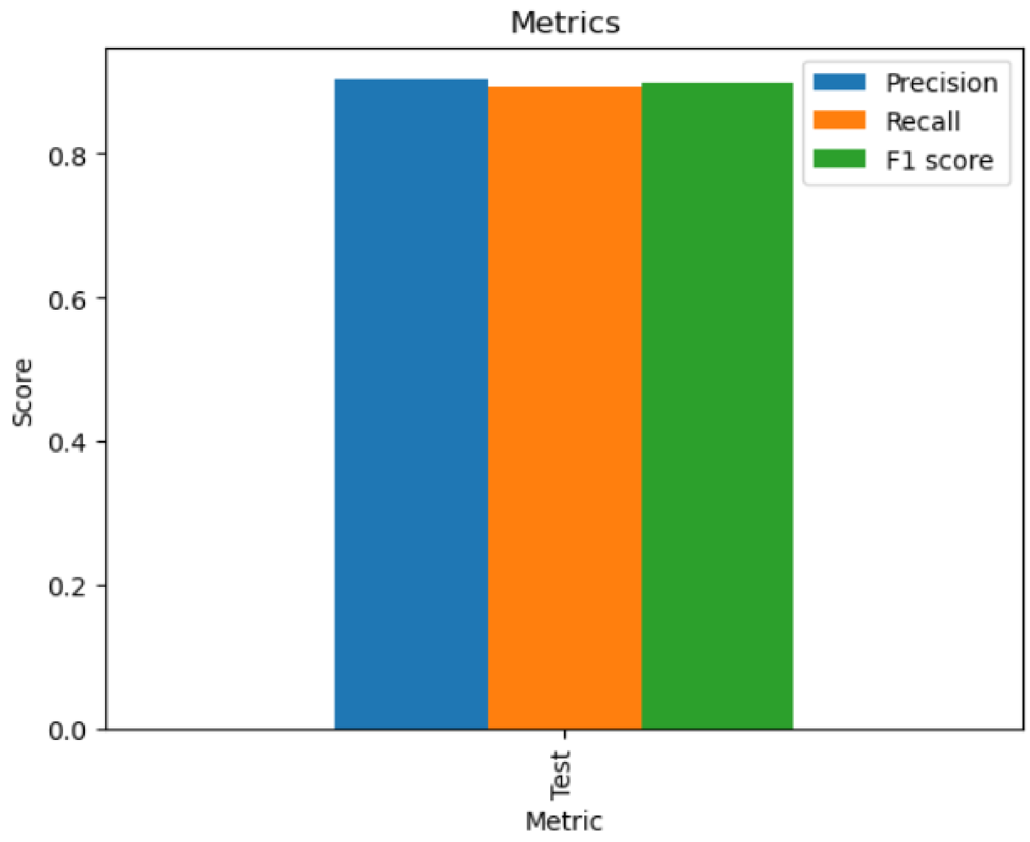


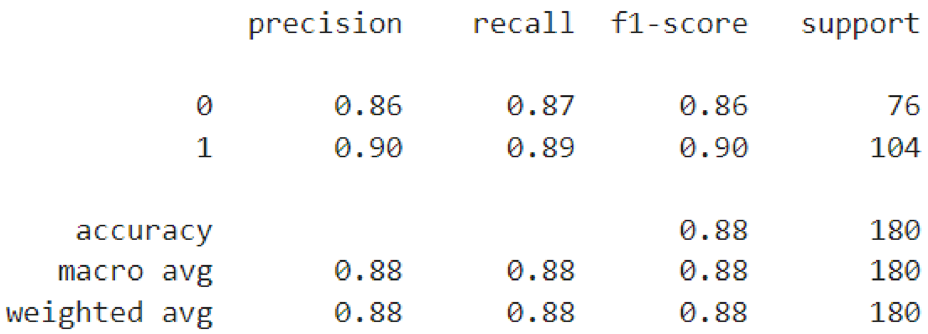
```
6/6 ━━━━━━━━━━━━━━━━━━━━ 0s 1ms/step - accuracy: 0.8713 - loss: 0.3517
Test accuracy: 0.88
6/6 ━━━━━━━━━━━━━━━━━━━━ 0s 8ms/step
Precision: 0.90
Recall: 0.89
F1 score: 0.90
```

Training vs validation:-



Precision, Recall and F1 Score:-





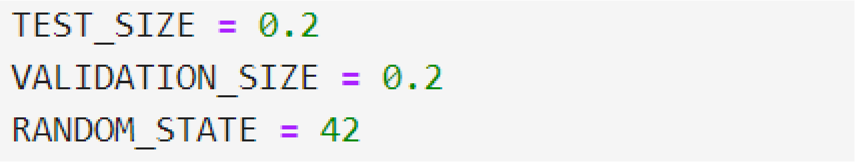
```
              precision    recall  f1-score   support

           0       0.86      0.87      0.86        76
           1       0.90      0.89      0.90       104

    accuracy                           0.88       180
   macro avg       0.88      0.88      0.88       180
weighted avg       0.88      0.88      0.88       180
```

Validation Size :-

yes validation was used.



```
TEST_SIZE = 0.2
VALIDATION_SIZE = 0.2
RANDOM_STATE = 42
```

Hyperparameters:-

several hyperparameters are used:

1. TEST_SIZE: This hyperparameter determines the proportion of the dataset to include in the test split when splitting the data into training, validation, and test sets. It is set to 0.2, meaning 20% of the data will be used for testing.
2. VALIDATION_SIZE: This hyperparameter determines the proportion of the dataset to include in the validation split when splitting the data into training, validation, and test sets. It is also set to 0.2, indicating that 20% of the data will be used for validation.
3. RANDOM_STATE: This hyperparameter is used for randomizing the dataset splitting. It ensures reproducibility of results by fixing the random state. It is set to 42.
4. Number of Epochs: In the model training part, the number of epochs determines how many times the

learning algorithm will work through the entire training dataset. Here it's set to 10.

Batch Size: The batch size specifies the number of samples that will be propagated through the network. Here it's set to 32. .5