Term Project

Applied Machine Learning

Group 12

Ahmed Al-Wasefy (300389391)

Ibrahim Elshenhapy (300389386)

Mohamed Gabr (300389919)

Problem's Overview

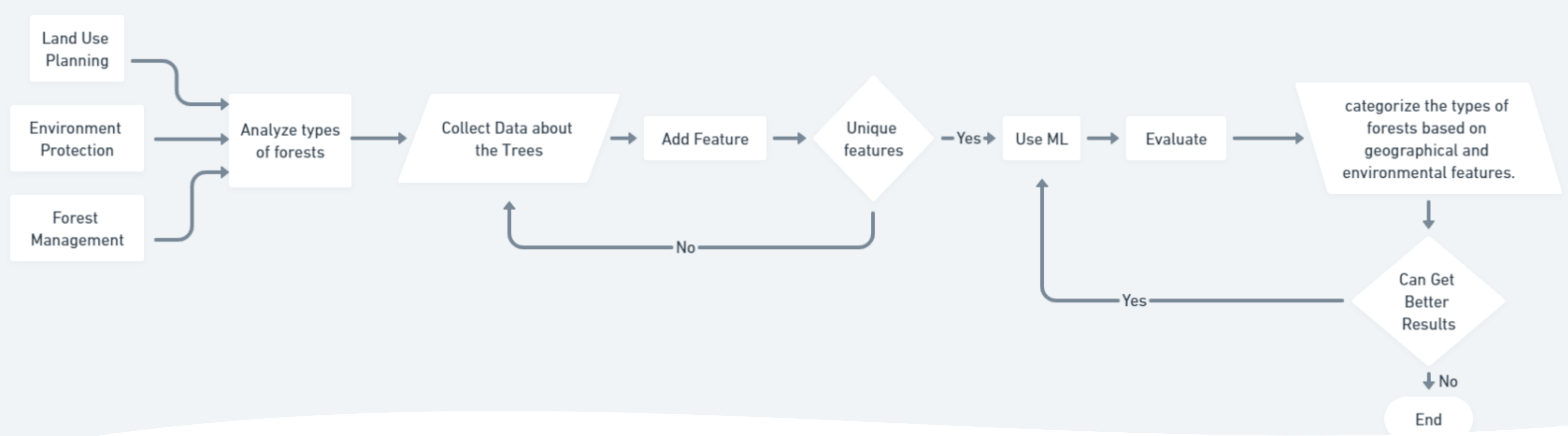
In The Forest of northern Colorado

Problem’s Overview

Land Use Planning

Environment Protection

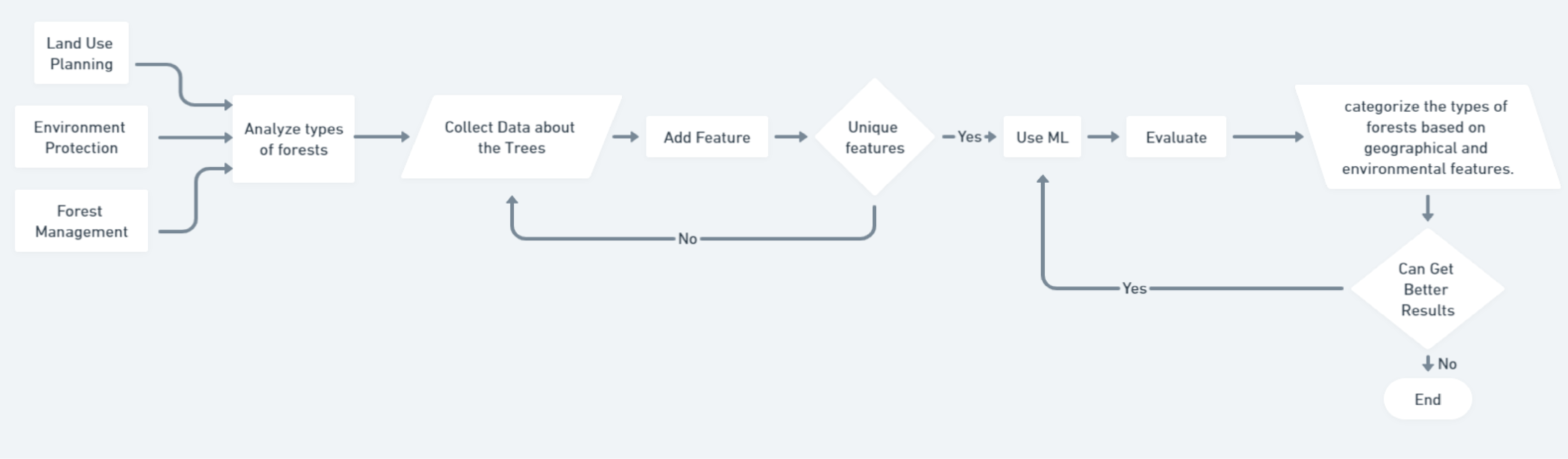
Forest Management



Problem's Overview

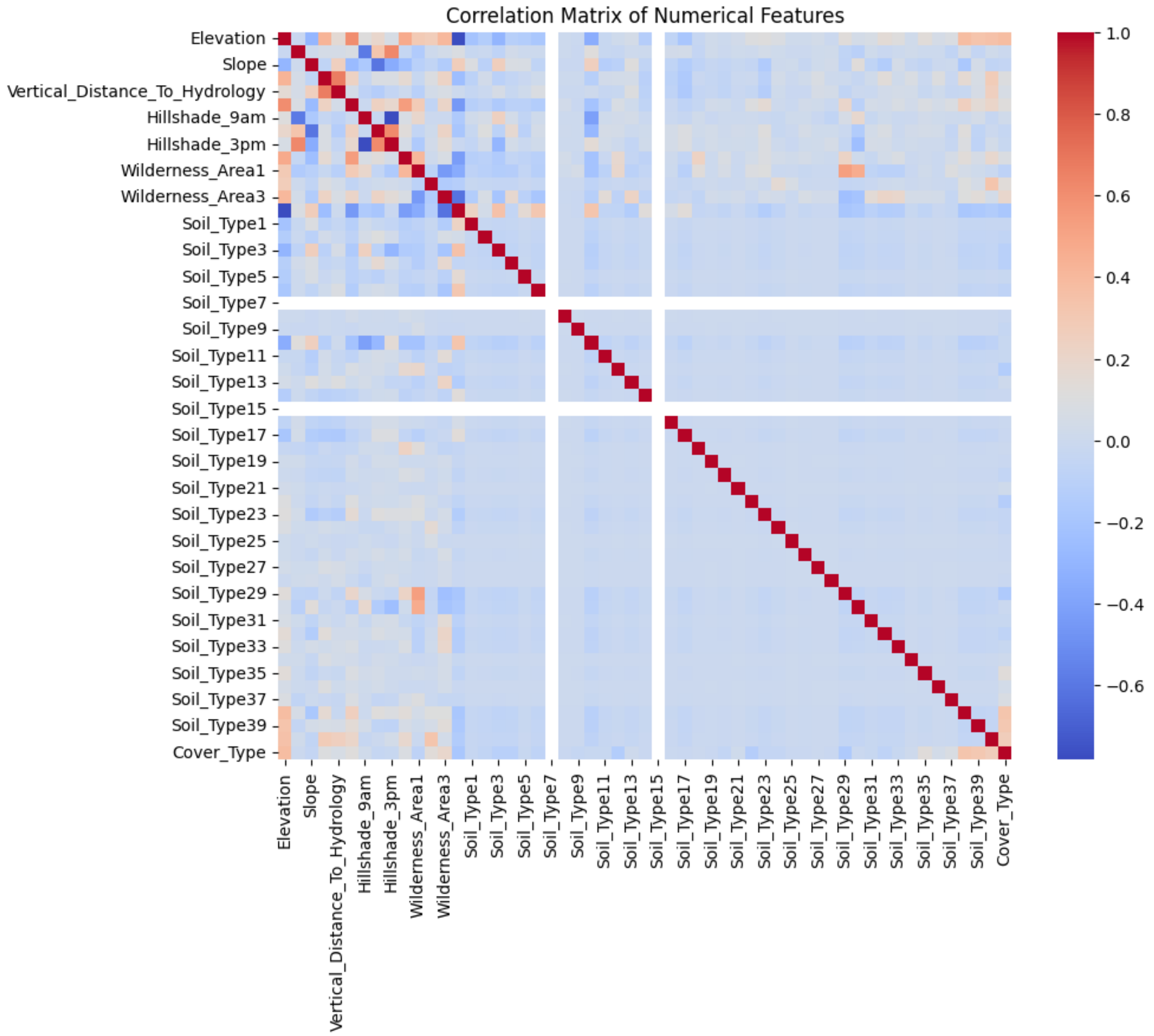
- Elevation
- Aspect
- Slope
- Distance to Hydrology (Horizontal and Vertical)
- Distance to Roadways
- Hillshade (Different time)
- Wilderness Area
- Soil Type
- Cover Type

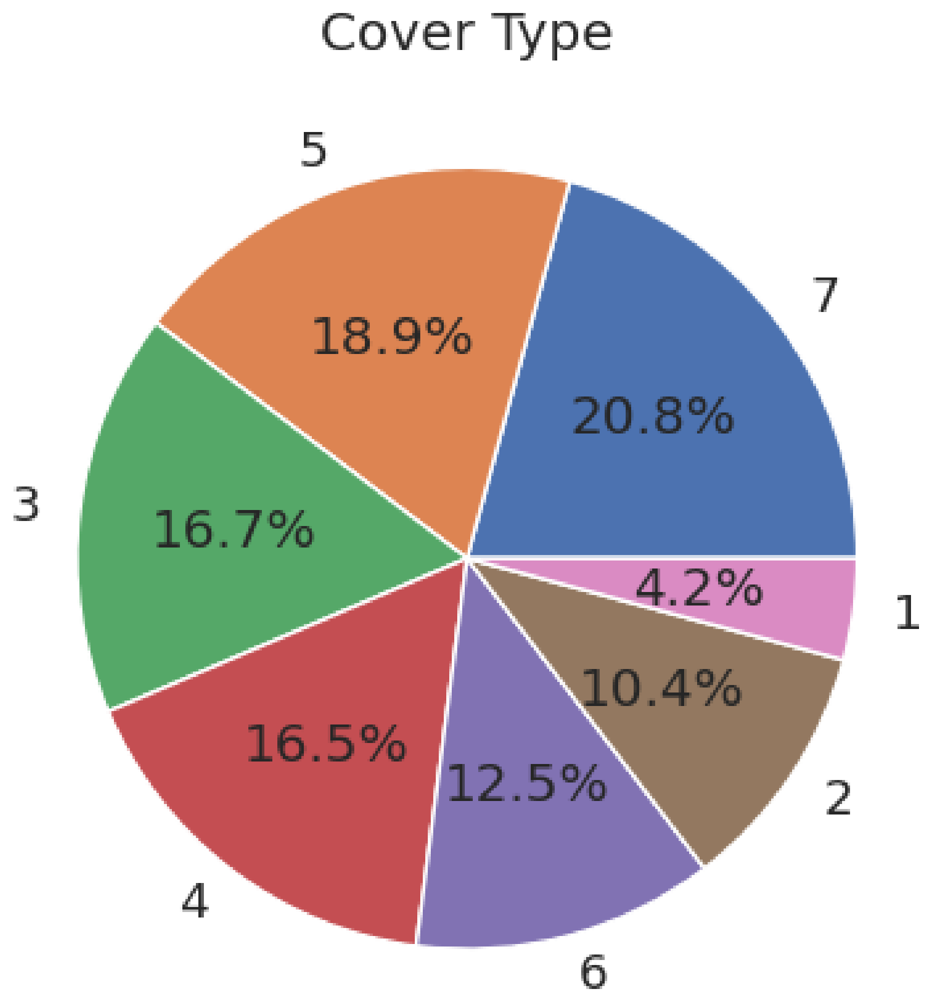
Problem's Overview



New Algorithms Developed

Dataset's overview (EDA)

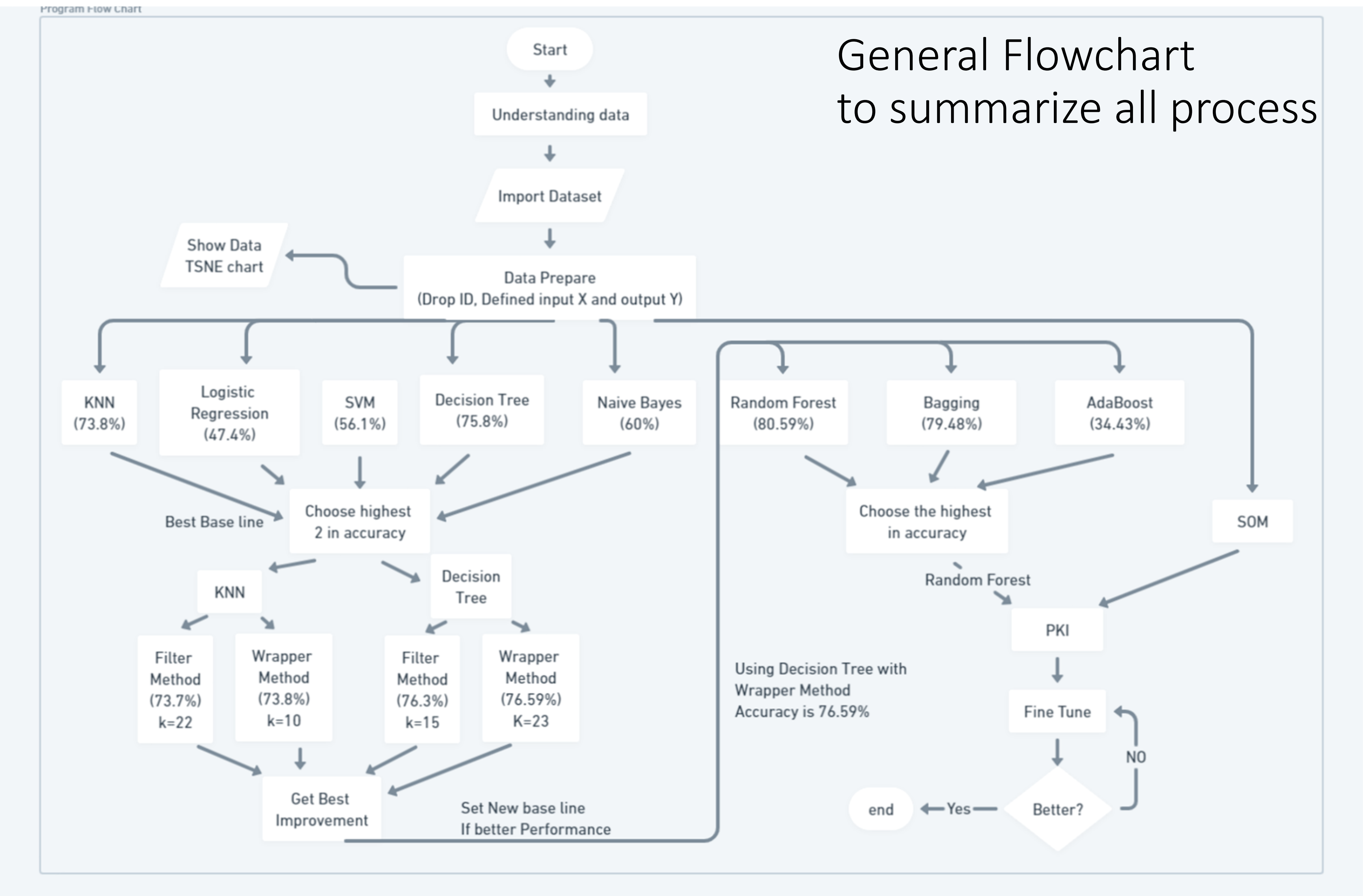




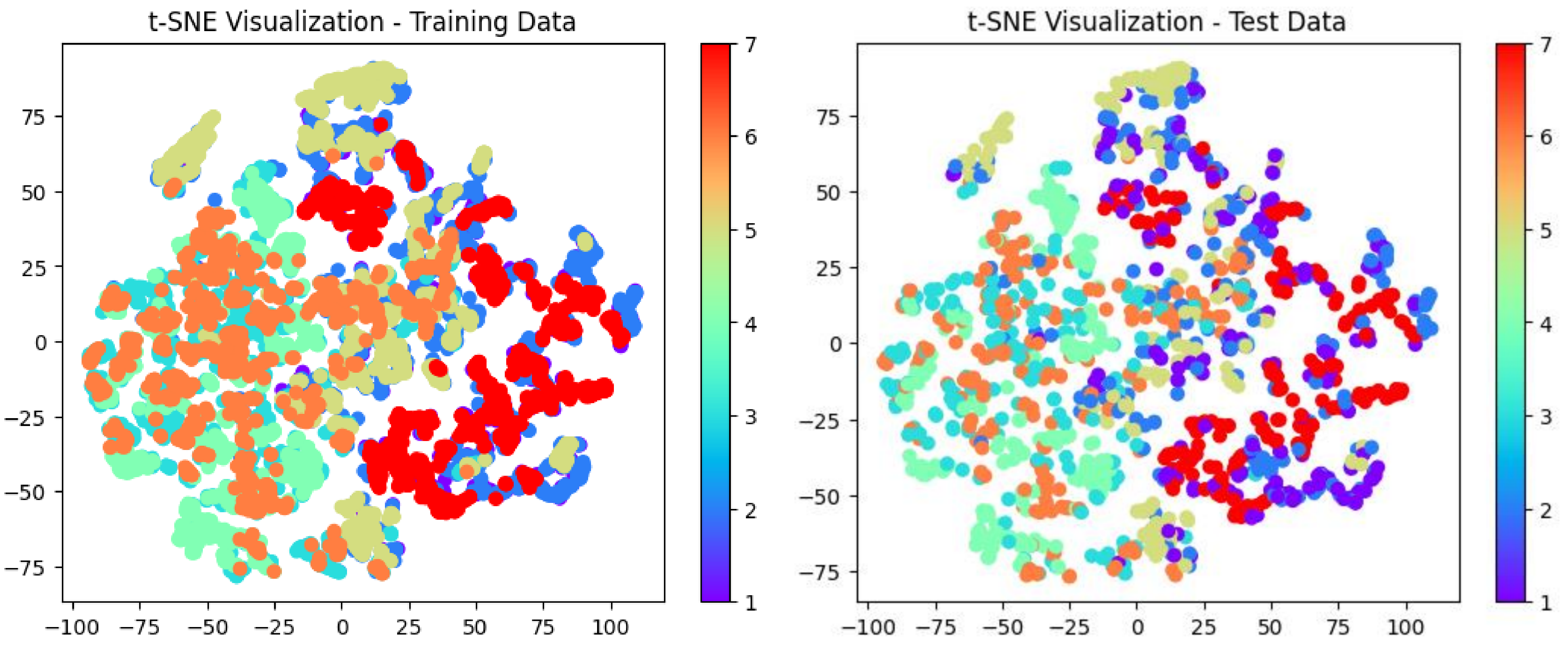
Dataset’s overview (EDA)

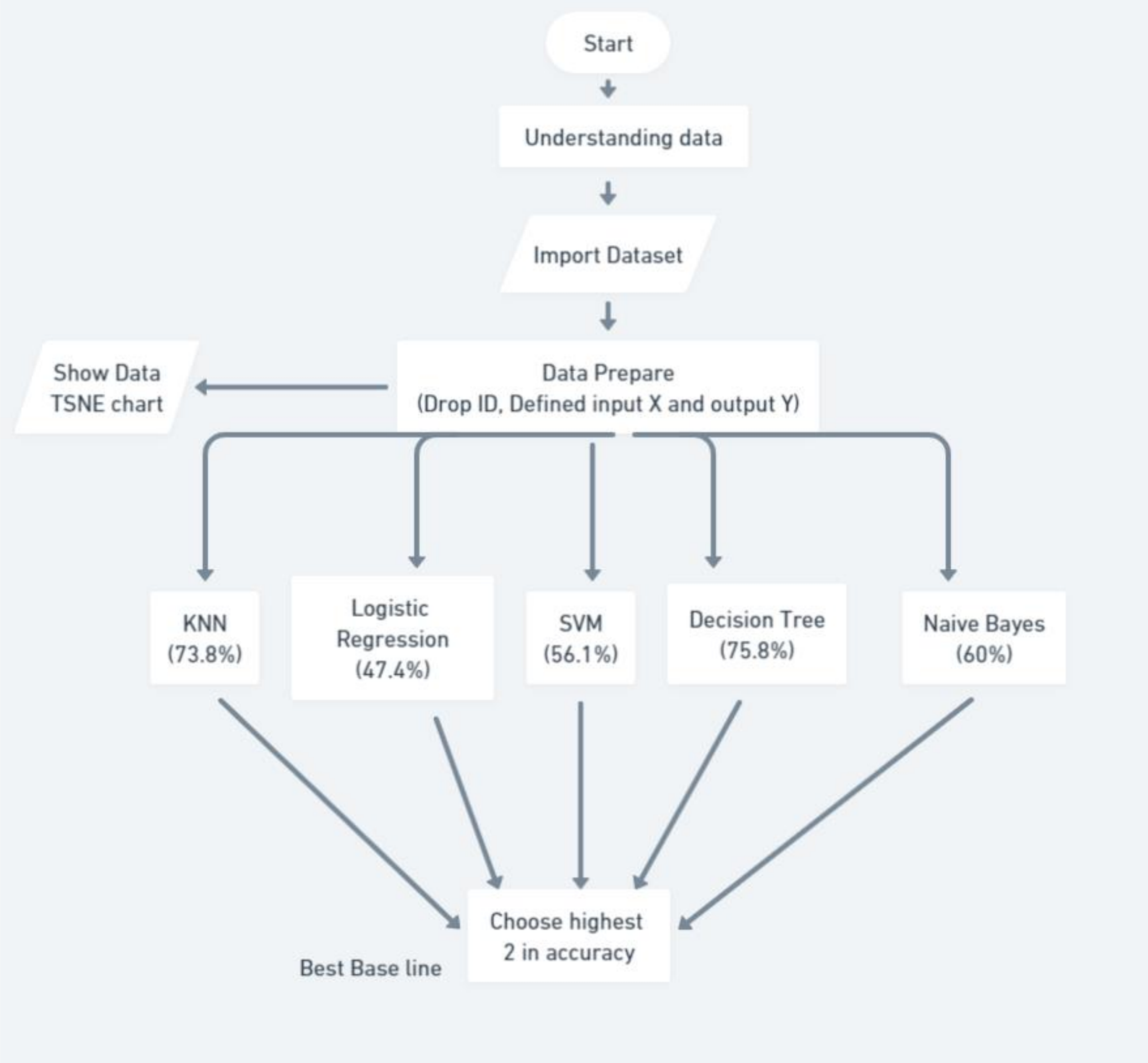
```
Data columns (total 55 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   Elevation                           8286 non-null   int64
 1   Aspect                              8286 non-null   int64
 2   Slope                               8286 non-null   int64
 3   Horizontal_Distance_To_Hydrology    8286 non-null   int64
 4   Vertical_Distance_To_Hydrology      8286 non-null   int64
 5   Horizontal_Distance_To_Roadways     8286 non-null   int64
 6   Hillshade_9am                       8286 non-null   int64
 7   Hillshade_Noon                      8286 non-null   int64
 8   Hillshade_3pm                       8286 non-null   int64
 9   Horizontal_Distance_To_Fire_Points  8286 non-null   int64
 10  Wilderness_Area1                    8286 non-null   int64
 11  Wilderness_Area2                    8286 non-null   int64
 12  Wilderness_Area3                    8286 non-null   int64
 13  Wilderness_Area4                    8286 non-null   int64
 14  Soil_Type1                          8286 non-null   int64
 15  Soil_Type2                          8286 non-null   int64
 16  Soil_Type3                          8286 non-null   int64
 17  Soil_Type4                          8286 non-null   int64
 18  Soil_Type5                          8286 non-null   int64
 19  Soil_Type6                          8286 non-null   int64
 20  Soil_Type7                          8286 non-null   int64
 21  Soil_Type8                          8286 non-null   int64
 22  Soil_Type9                          8286 non-null   int64
 23  Soil_Type10                         8286 non-null   int64
 24  Soil_Type11                         8286 non-null   int64
 25  Soil_Type12                         8286 non-null   int64
 26  Soil_Type13                         8286 non-null   int64
 27  Soil_Type14                         8286 non-null   int64
 28  Soil_Type15                         8286 non-null   int64
 29  Soil_Type16                         8286 non-null   int64
 30  Soil_Type17                         8286 non-null   int64
 31  Soil_Type18                         8286 non-null   int64
 32  Soil_Type19                         8286 non-null   int64
 33  Soil_Type20                         8286 non-null   int64
 34  Soil_Type21                         8286 non-null   int64
 35  Soil_Type22                         8286 non-null   int64
 36  Soil_Type23                         8286 non-null   int64
```

```
         Elevation       Aspect        Slope  \
count  8286.000000  8286.000000  8286.000000
mean   2732.983104   155.366643    16.868694
std     432.906958   108.392758     8.514811
min    1863.000000     0.000000     0.000000
25%    2350.000000    66.000000    10.000000
50%    2720.500000   125.000000    16.000000
75%    3099.750000   252.000000    23.000000
max    3849.000000   360.000000    50.000000

       Horizontal_Distance_To_Hydrology  Vertical_Distance_To_Hydrology  \
count                       8286.000000                     8286.000000
mean                         225.249698                       53.233888
std                          213.670866                       62.890107
min                            0.000000                     -134.000000
25%                           60.000000                        5.000000
50%                          175.000000                       34.000000
75%                          323.000000                       84.000000
max                         1343.000000                      547.000000

       Horizontal_Distance_To_Roadways  Hillshade_9am  Hillshade_Noon  \
count                      8286.000000    8286.000000     8286.000000
mean                       1629.840574     213.732682      218.534999
std                        1259.714393      30.675904       23.142959
min                           0.000000      58.000000       99.000000
25%                         726.000000     197.000000      206.000000
50%                        1273.000000     221.000000      222.000000
75%                        2155.000000     237.000000      235.000000
max                        6508.000000     254.000000      254.000000
```


Program Flow Chart
General Flowchart
to summarize all process
Start
Understanding data
Import Dataset
Show Data
TSNE chart
Data Prepare
(Drop ID, Defined input X and output Y)
KNN
(73.8%)
Logistic
Regression
(47.4%)
SVM
(56.1%)
Decision Tree
(75.8%)
Naive Bayes
(60%)
Random Forest
(80.59%)
Bagging
(79.48%)
AdaBoost
(34.43%)
Best Base line
Choose highest
2 in accuracy
Choose the highest
in accuracy
SOM
KNN
Decision
Tree
Random Forest
PKI
Filter
Method
(73.7%)
k=22
Wrapper
Method
(73.8%)
k=10
Filter
Method
(76.3%)
k=15
Wrapper
Method
(76.59%)
K=23
Using Decision Tree with
Wrapper Method
Accuracy is 76.59%
Fine Tune
NO
Get Best
Improvement
Set New base line
If better Performance
end
Yes
Better?

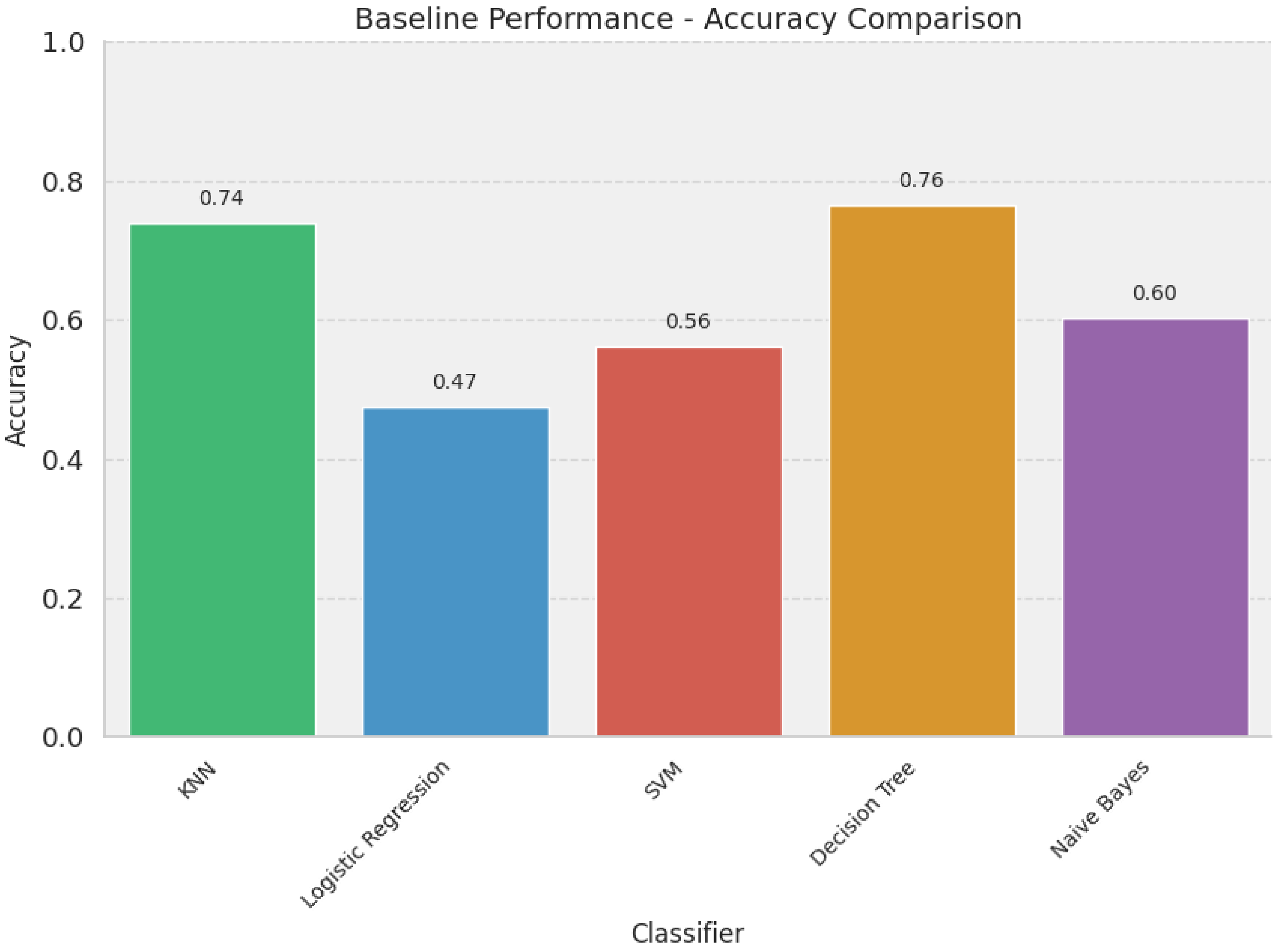
Visualize the training and test set to understand problem nature

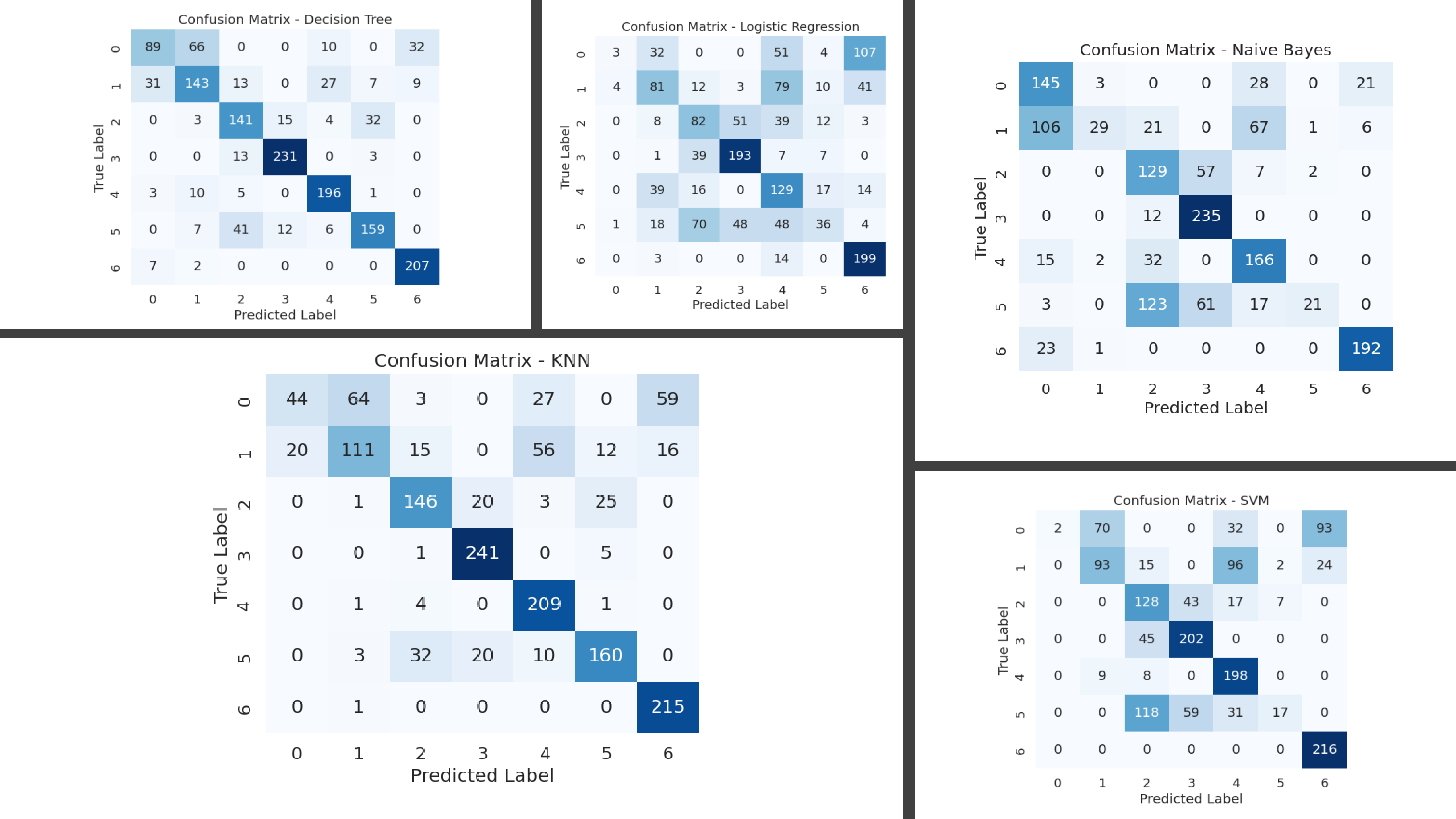


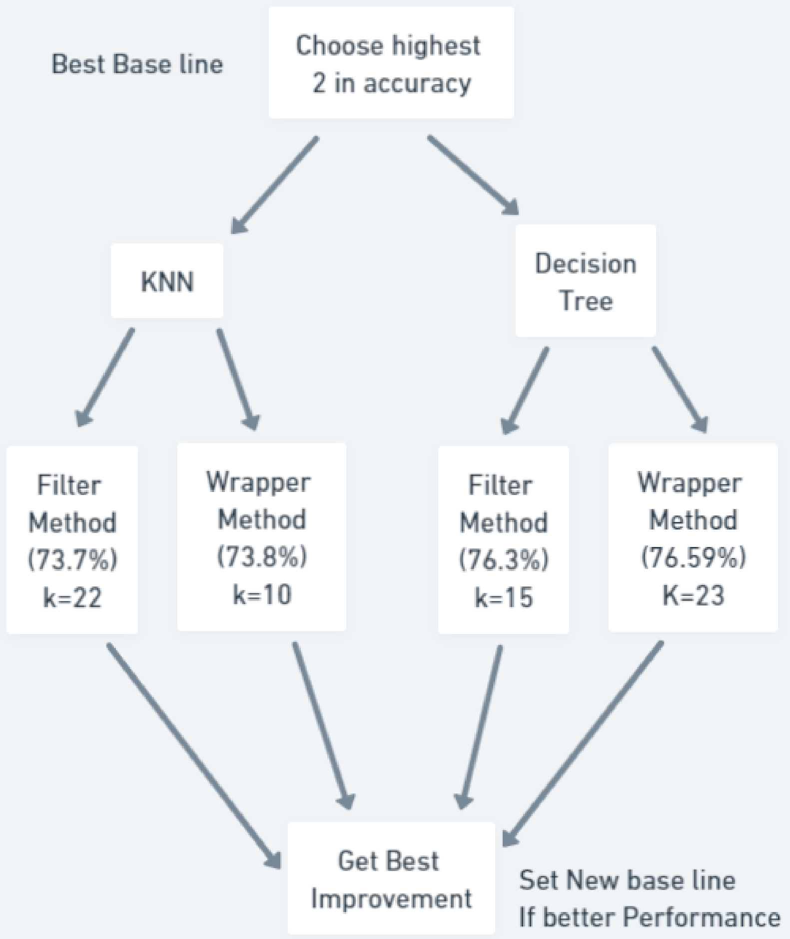


- Flow Chart for Question 1

Obtain a baseline performance



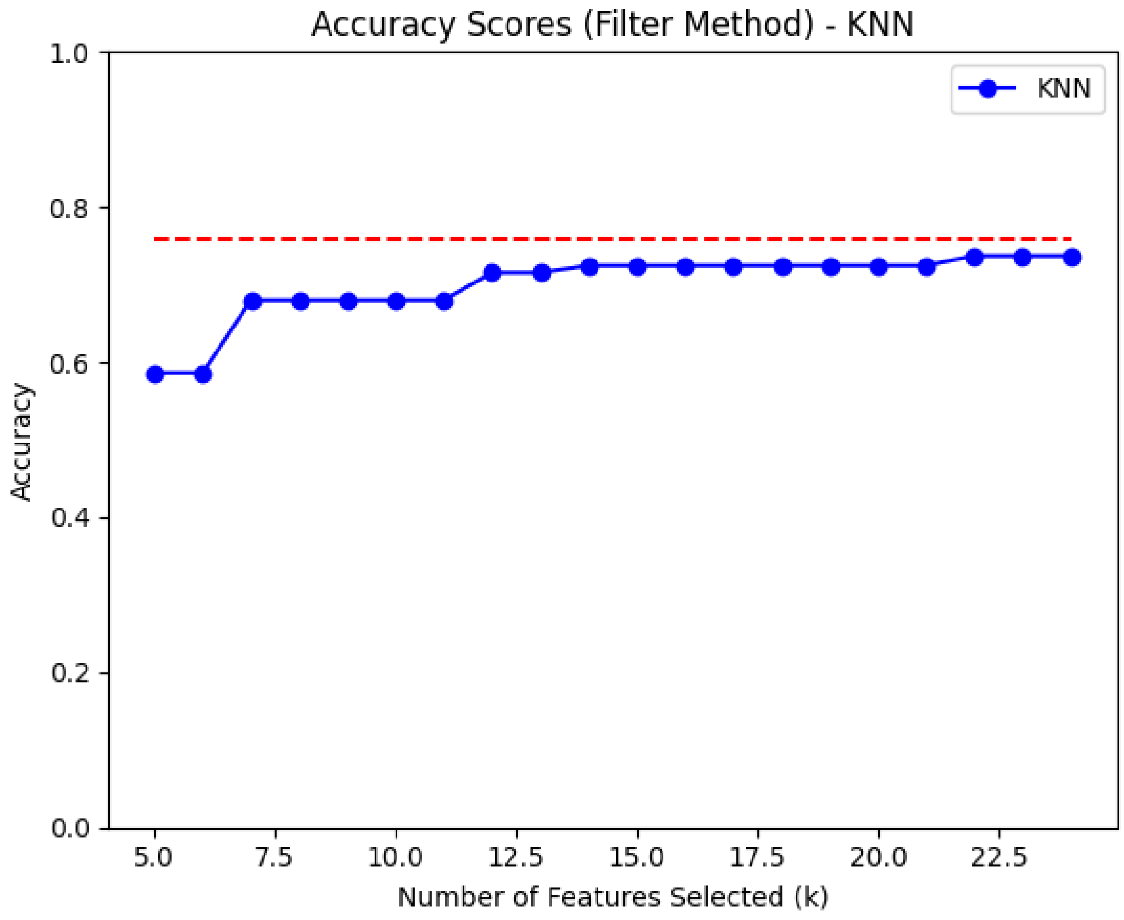

Confusion Matrix - Decision Tree
89 66 0 0 10 0 32
31 143 13 0 27 7 9
0 3 141 15 4 32 0
0 0 13 231 0 3 0
3 10 5 0 196 1 0
0 7 41 12 6 159 0
7 2 0 0 0 0 207
True Label
Predicted Label
Confusion Matrix - Logistic Regression
3 32 0 0 51 4 107
4 81 12 3 79 10 41
0 8 82 51 39 12 3
0 1 39 193 7 7 0
0 39 16 0 129 17 14
1 18 70 48 48 36 4
0 3 0 0 14 0 199
True Label
Predicted Label
Confusion Matrix - Naive Bayes
145 3 0 0 28 0 21
106 29 21 0 67 1 6
0 0 129 57 7 2 0
0 0 12 235 0 0 0
15 2 32 0 166 0 0
3 0 123 61 17 21 0
23 1 0 0 0 0 192
True Label
Predicted Label
Confusion Matrix - KNN
44 64 3 0 27 0 59
20 111 15 0 56 12 16
0 1 146 20 3 25 0
0 0 1 241 0 5 0
0 1 4 0 209 1 0
0 3 32 20 10 160 0
0 1 0 0 0 0 215
True Label
Predicted Label
Confusion Matrix - SVM
2 70 0 0 32 0 93
0 93 15 0 96 2 24
0 0 128 43 17 7 0
0 0 45 202 0 0 0
0 9 8 0 198 0 0
0 0 118 59 31 17 0
0 0 0 0 0 0 216
True Label
Predicted Label

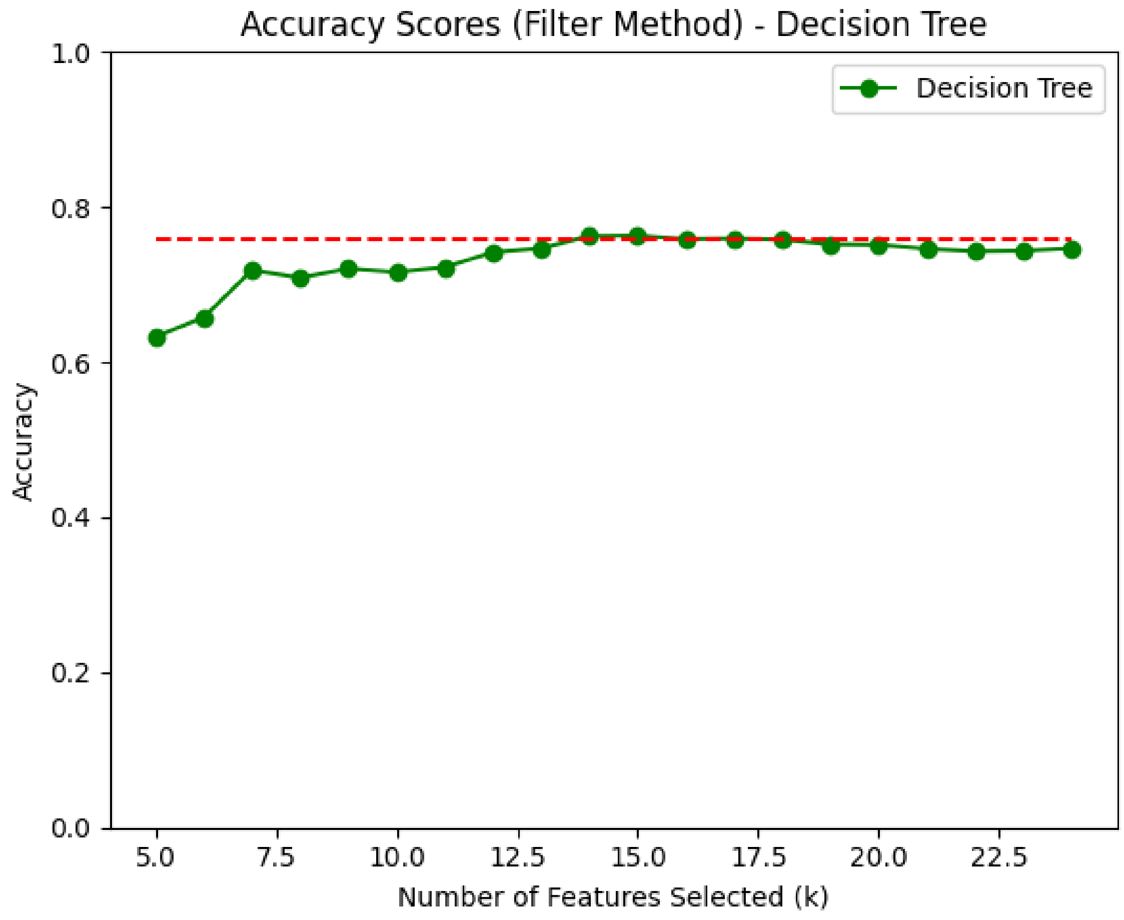


- Flow Chart for Question 2

First Improvement strategy Feature Selection

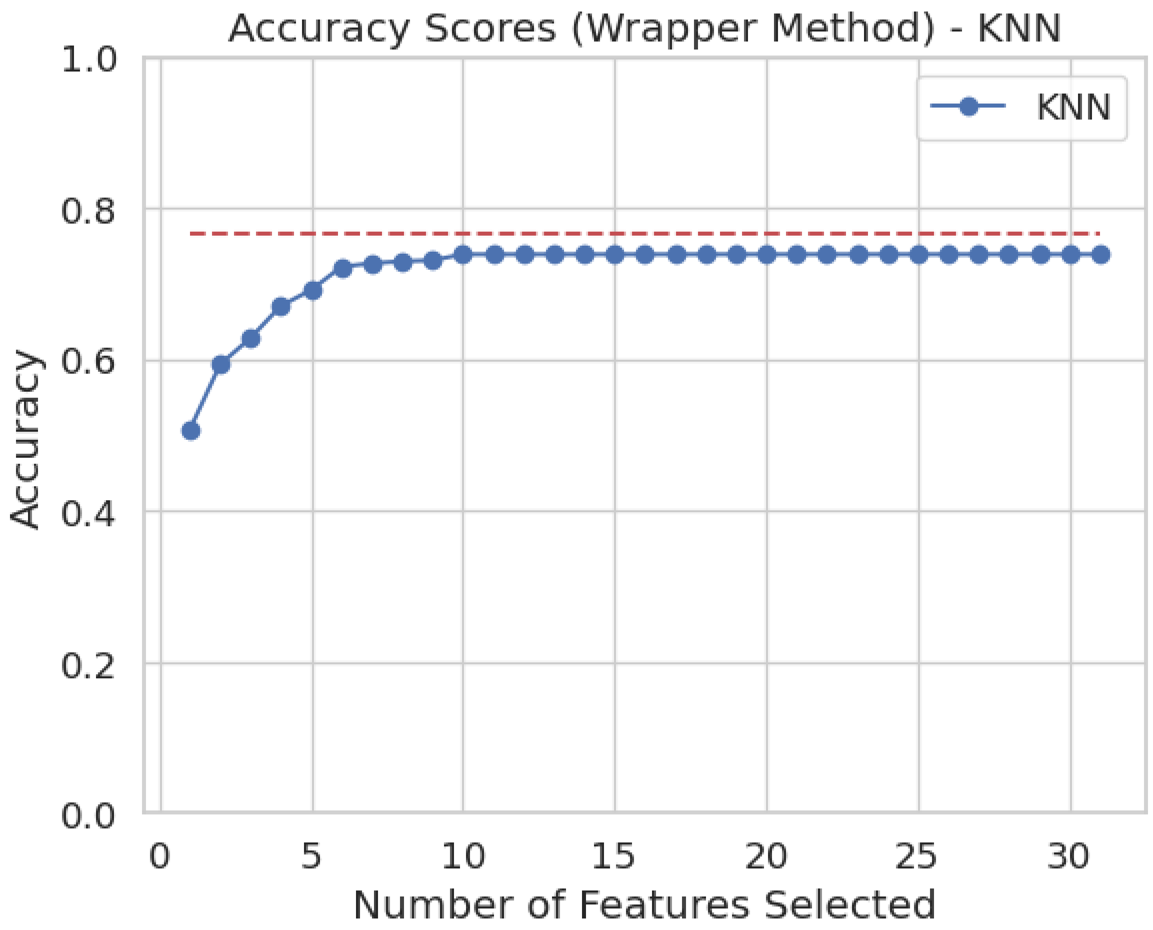
K= (5 - 25)

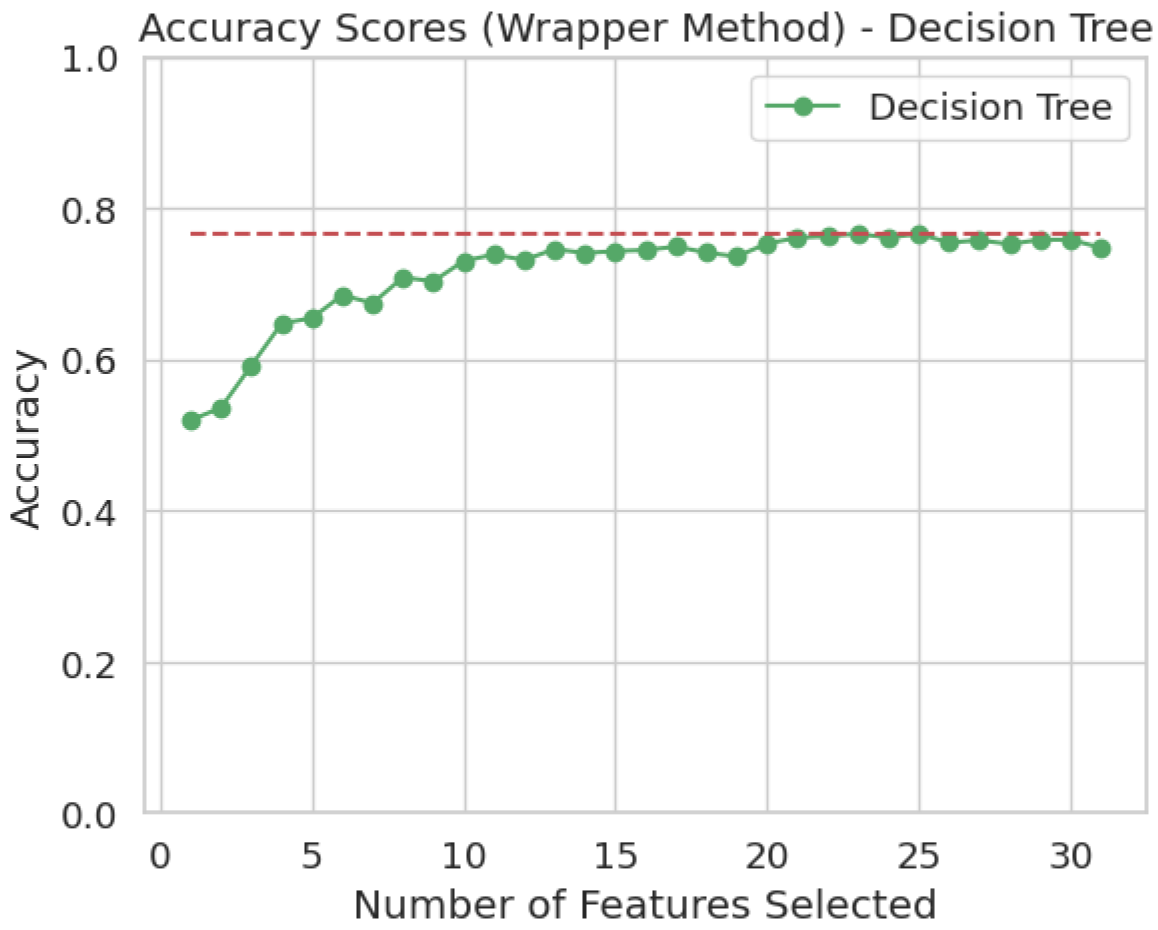




First Improvement strategy Feature Selection

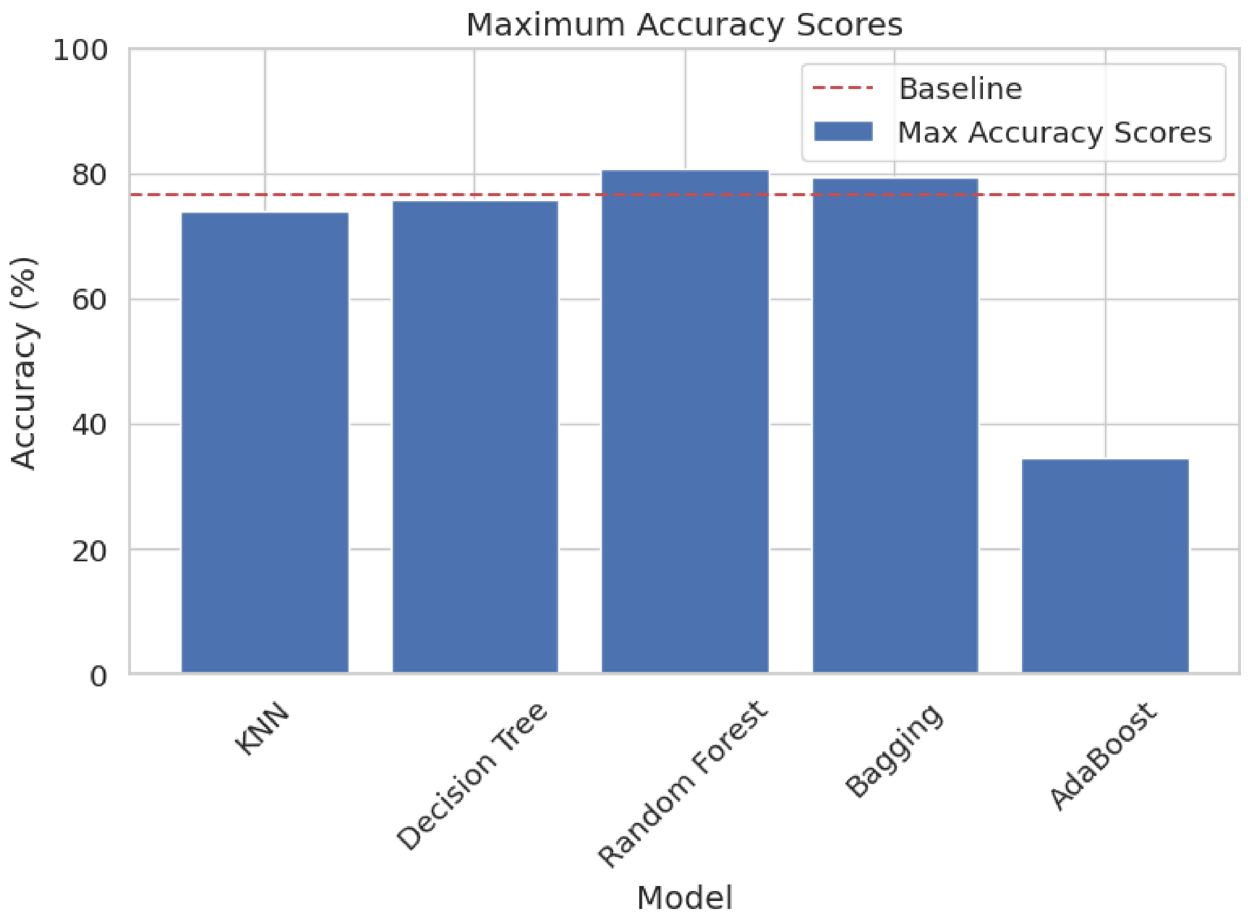
K= (5 - 25)

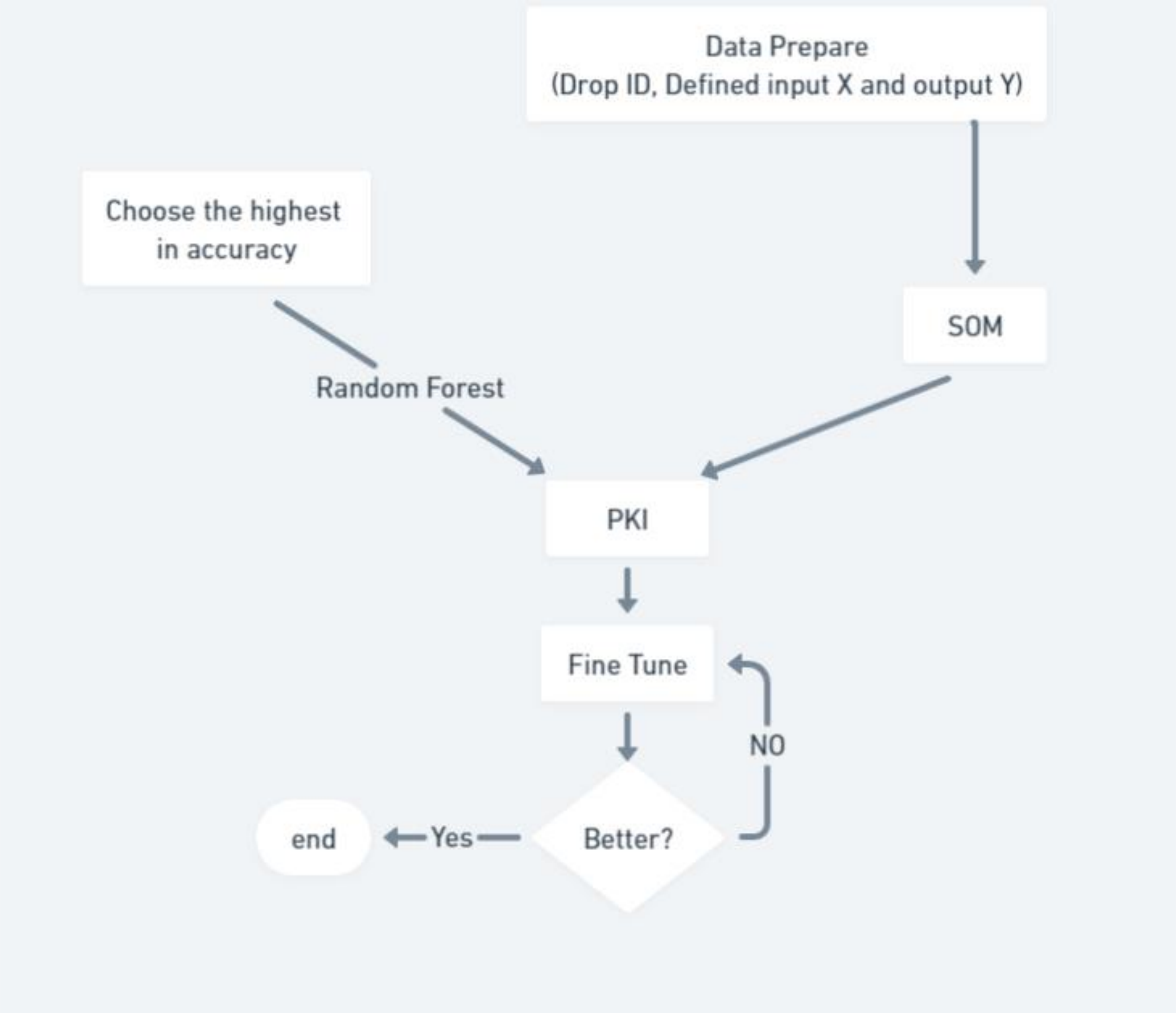




- Flow Chart for Question 3

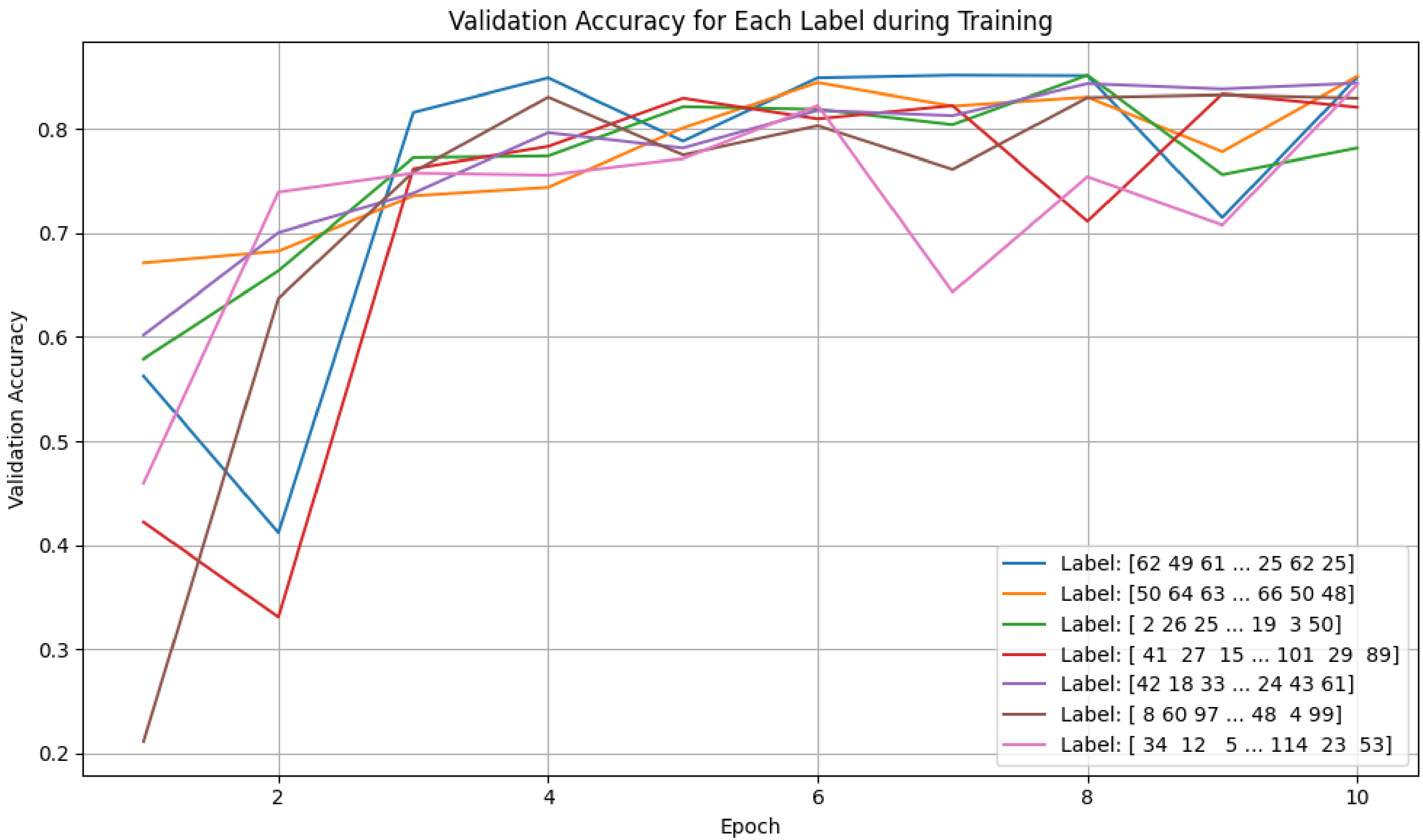
Adding more machine learning model





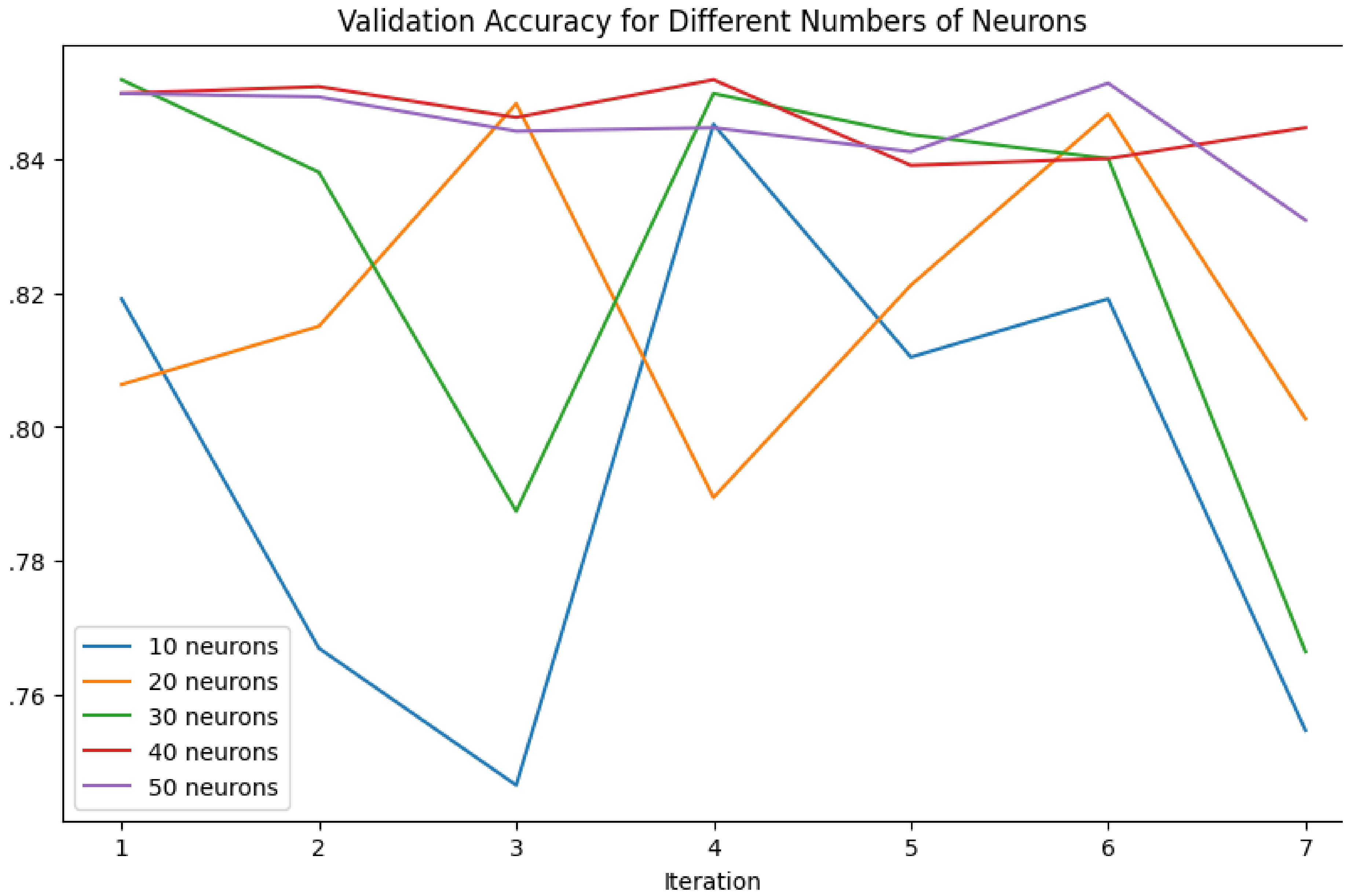
- Flow Chart for Question 4,5

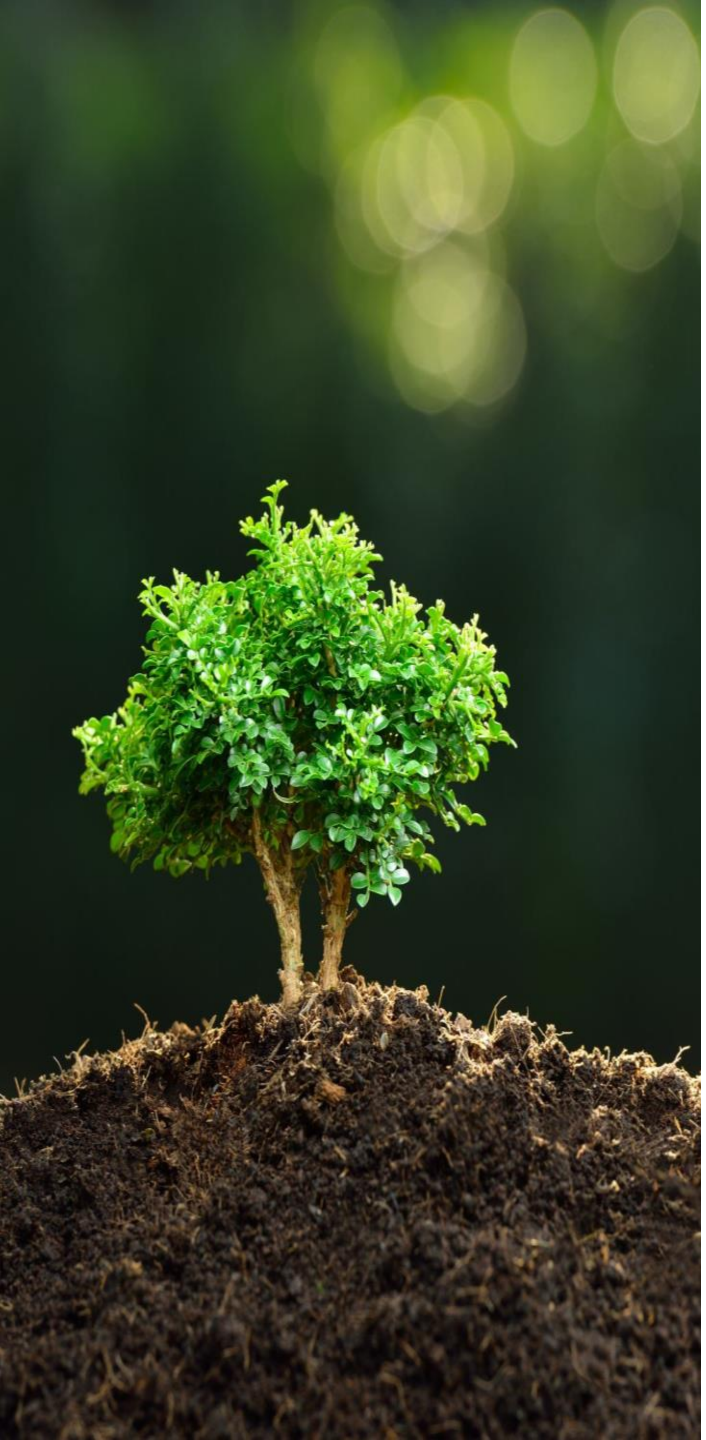
PKI



Validation Accuracy for Each Label during Training
Validation Accuracy
Epoch
0.2
0.3
0.4
0.5
0.6
0.7
0.8
2
4
6
8
10
Label: [62 49 61 ... 25 62 25]
Label: [50 64 63 ... 66 50 48]
Label: [2 26 25 ... 19 3 50]
Label: [41 27 15 ... 101 29 89]
Label: [42 18 33 ... 24 43 61]
Label: [8 60 97 ... 48 4 99]
Label: [34 12   5 ... 114 23 53]

Fine tuning


Validation Accuracy for Different Numbers of Neurons
.84
.82
.80
.78
.76
10 neurons
20 neurons
30 neurons
40 neurons
50 neurons
1
2
3
4
5
6
7
Iteration

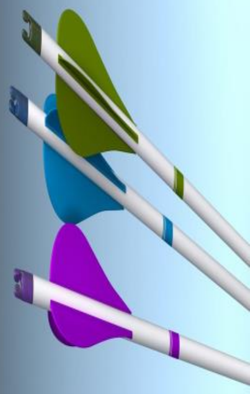


Conclusion (Question1)

- Decision Tree emerged as the best model with an accuracy of 76.2%.
- KNN and Naive Bayes also performed reasonably well, achieving accuracies ranging from 0.6 to 0.76.
- Logistic Regression and SVM showed relatively poor results, with low accuracies due to misclassifications in specific classes.
- The most common misclassifications occurred in classes 0, 1, and 5.
- Further experimentation using Dimensionality Reduction and Feature selection could improve the models' performance.
- The obtained results serve as a starting point for enhancing the classifiers' accuracy on the dataset.
- Overall, the current accuracies indicate room for improvement in future iterations.

Conclusion (Question 2)

- The filter method for feature selection was implemented, selecting features based on individual importance without considering feature interactions.
- Results indicate that the accuracy of both the KNN and Decision Tree models increases as the number of selected features increases, but plateaus after a certain threshold.
- There are a few crucial features significantly contributing to the models' accuracy, while others have less impact.
- Best k values for KNN and Decision Tree were found to be 22 and 15, respectively.
- The accuracy of the KNN model with the best k value is 73.7%, and the Decision Tree model achieves 76.3%.
- The filter method can be a valuable approach to enhance model accuracy, but careful selection of the optimal number of features is essential for best results.



Conclusion (Question 2)

- Implemented the wrapper method for feature selection using Recursive Feature Elimination with Cross-Validation (RFECV) algorithm.
- The wrapper method iteratively adds features to the model and evaluates its performance, improving accuracy for both KNN and Decision Tree models.
- Best k values for KNN and Decision Tree are 10 and 23, respectively, with accuracies of 0.738 and 0.766.
- The accuracy of models plateaus after selecting a certain number of features, indicating important features and less significant ones.
- The wrapper method is a valuable approach to enhance model accuracy, though more complex to implement compared to the filter method.
- In comparison, the wrapper method proves to be more powerful, but the choice between the two depends on the specific problem requirements.

Conclusion (Question 3)

- Implemented ensemble methods to improve model accuracy, including KNN, Decision Tree, Random Forest, Bagging, and AdaBoost.
- Results show that Random Forest achieved the highest accuracy at 80.59%, followed by Bagging at 79.48%.
- KNN and Decision Tree achieved moderate accuracies of 73.84% and 75.80%, respectively.
- AdaBoost had the lowest accuracy at 34.43%.
- Ensemble methods, particularly Random Forest and Bagging, are effective in enhancing model accuracy compared to baseline.
- The choice of the best model depends on specific problem requirements - Random Forest for highest accuracy and Decision Tree for interpretability.

Conclusion (Question 4)

- Implemented a Self-Organizing Map (SOM) for clustering the data into 7 regions.
- Trained a Deep Neural Network (DNN) on the SOM cluster labels, achieving a validation accuracy of 85.18%.
- DNN trained on the original data labels had a lower validation accuracy of 72.82%, indicating the effectiveness of SOM clustering.
- Best validation accuracy achieved with a 6x6 grid of SOM neurons, suggesting larger grids may further improve DNN performance.
- SOM clustering identified meaningful data patterns, enhancing DNN learning and performance.
- Potential improvements include using a more sophisticated SOM algorithm (e.g., GrowingSOM), utilizing convolutional neural networks (CNNs), and exploring various data preprocessing and regularization techniques to optimize DNN performance.

Conclusion (Question 5)

- The improved code incorporates two additional callbacks, ReduceLROnPlateau and Early Stopping, to improve the accuracy of the Deep Neural Network (DNN) model during training.
- The model architecture includes Batch Normalization and Dropout layers to stabilize training, prevent overfitting, and improve convergence.
- The Adam optimizer with a learning rate of 0.001 is used for optimization, enhancing the model's ability to converge efficiently.
- The utilization of Self-Organizing Maps (SOM) for data mapping allows the DNN to handle complex data distributions and capture intricate patterns, leading to better data representation.
- Early Stopping prevents overfitting by monitoring the validation loss and stopping training when there is no improvement for a certain number of epochs.
- ReduceLROnPlateau reduces the learning rate when the validation loss plateaus, fine-tuning the model and aiding convergence.
- The combination of these techniques results in a more accurate DNN model compared to the previous version, as it prevents overfitting, fine-tunes the model, and enhances data representation and feature learning.

Thanks