BIM 314 - INTRODUCTİON TO MACHINE LEARNING PROJECT

1.Öğrenci(Adı Soyadı – Numarası)	Anisah KASO	030118130
2.Öğrenci(Adı Soyadı – Numarası)	Beyen HAVVA	030119112
3.Öğrenci(Adı Soyadı – Numarası)	Rumeysa EĞİLMEZ	030118056
Konu	Optical Recognition of Handwritten Digits	

1- Select a dataset from the following website:

https://archive.ics.uci.edu/ml/datasets.php

⇨ http://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits

2- Describe the dataset by explaining (at least) the following:

⇨ **The source of the data :** E. Alpaydin, C. Kaynak
Department of Computer Engineering
Bogazici University, 80815 Istanbul Turkey
alpaydin '@' boun.edu.tr

⇨ **The type of the data :** Dataset contains images of hand-written digits: 10 classes where each class refers to a digit.

⇨ **Data Set Characteristics:** Multivariate
Attribute Characteristics: Integer

⇨ **The size of the data** : There are totally 5620 instances of data set, and each instance is a 32x32 bit map image of digit from 0 to 9. Among them, there are 3823 training data samples and 1797 test data. The image data is represented by an 8x8 pixel matrix, with a total of 64 pixel dimensions.

⇨ **Number of Attributes:** 64
Number of Instances: 5620

⇨ **Attribute Information:**

- All input attributes are integers in the range 0...16
- The last attribute is the class code 0...9
- 8x8 image of integer pixels in the range 0…16

⇨ **The format of the data :** 5620 inputs and output represente to 0...9 classes where each class refers to a digit.

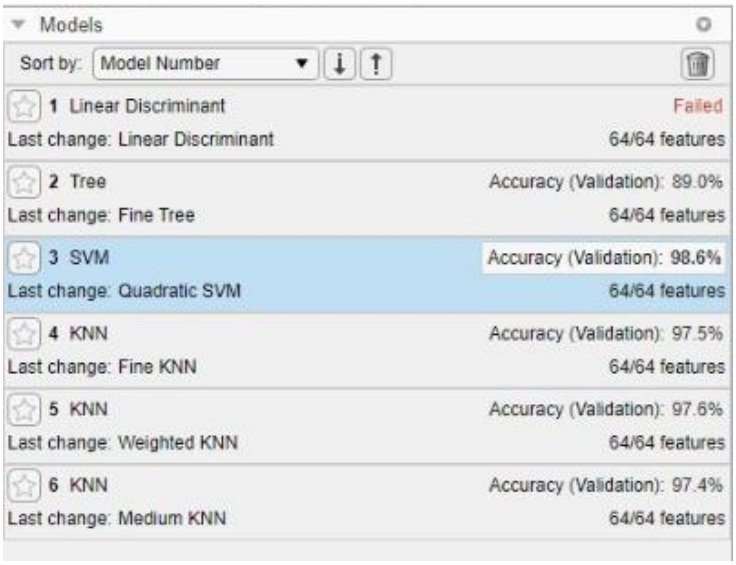
⇨ **The type of the problem :** Classification

⇨ **Data Set Information:** We used preprocessing programs made available by NIST to extract normalized bitmaps of handwritten digits from a preprinted form. From a total of 43 people, 30 contributed to the training set and different 13 to the test set. 32x32 bitmaps are divided into no overlapping blocks of 4x4 and the numbers of on pixels are counted in each block. This generates an input matrix of 8x8 where each element is an integer in the range 0...16. This reduces dimensionality and gives invariance to small distortions.

3- Select the suitable algorithm based on the dataset, explain the reasons for selecting the algorithm.

We select Quadratic SVM model for this data set because after training which many type of training models such as Linear Discriminant, Decision Tree, KNN etc. when compare the accuracy result number of each model the highest accuracy result is accuracy result of Quadratic SVM model.
The given result is 98.6%

The accuracy results for each model are given below:

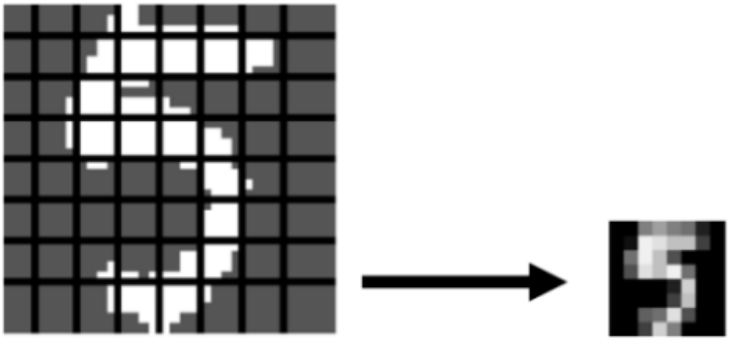


So Quadratic SVM model algorithm is selected to the Optical Recognition of Handwritten Digits Data Set.

4- Explain any pre-processing operation applied on the dataset (why and how).

Preprocessing programs made available by NIST were used to extract normalized bitmaps of handwritten digits from a preprinted form. From a total of 43 people, 30 contributed to the training set and different 13 to the test set. 32x32 bitmaps are divided into nonoverlapping blocks of 4x4 and the number of on pixels are counted in each block. This generates an input matrix of 8x8 where each element is an integer in the range 0..16. This reduces dimensionality and gives invariance to small distortions.

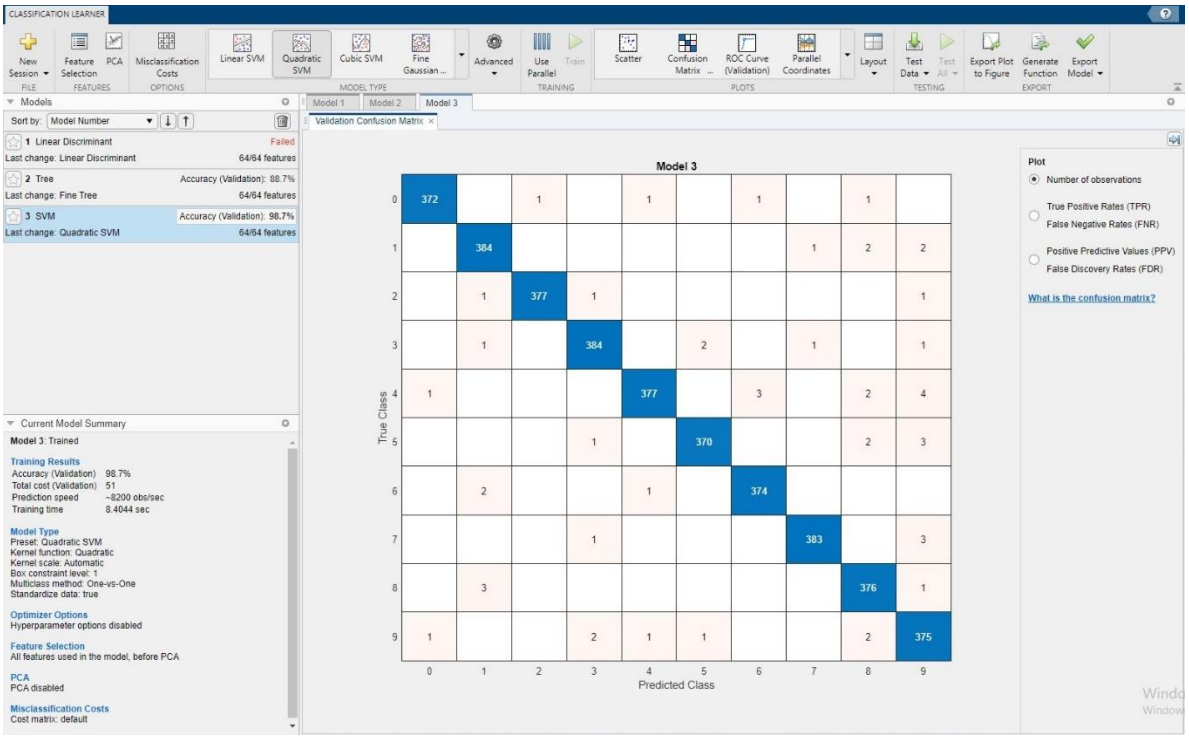
Figure shows the two steps of preprocess. The first step is shown in Figure(a), where the original image is divided into 8*8 blocks, with 4*4 pixels in each block. Figure(b) shows the second step, where the number of written pixels is counted in each block of 4*4 pixels, and we got a 8*8 matrix with value ranges from 0 to 16.



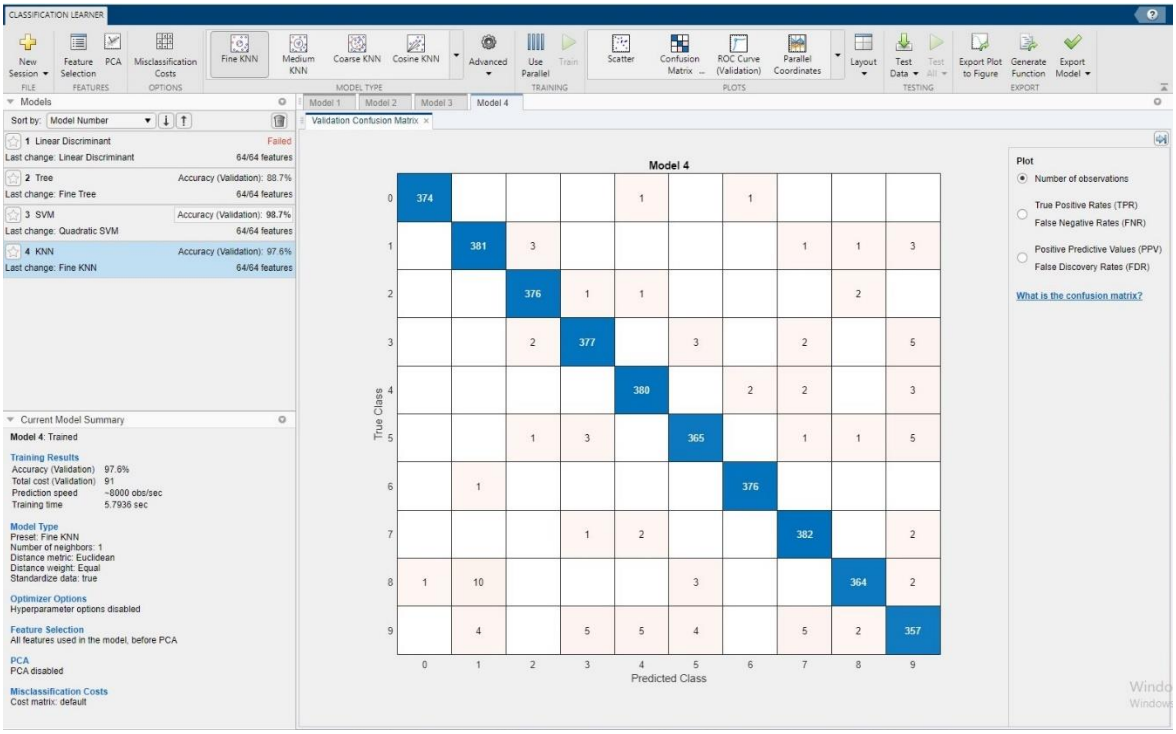
(a) (b)

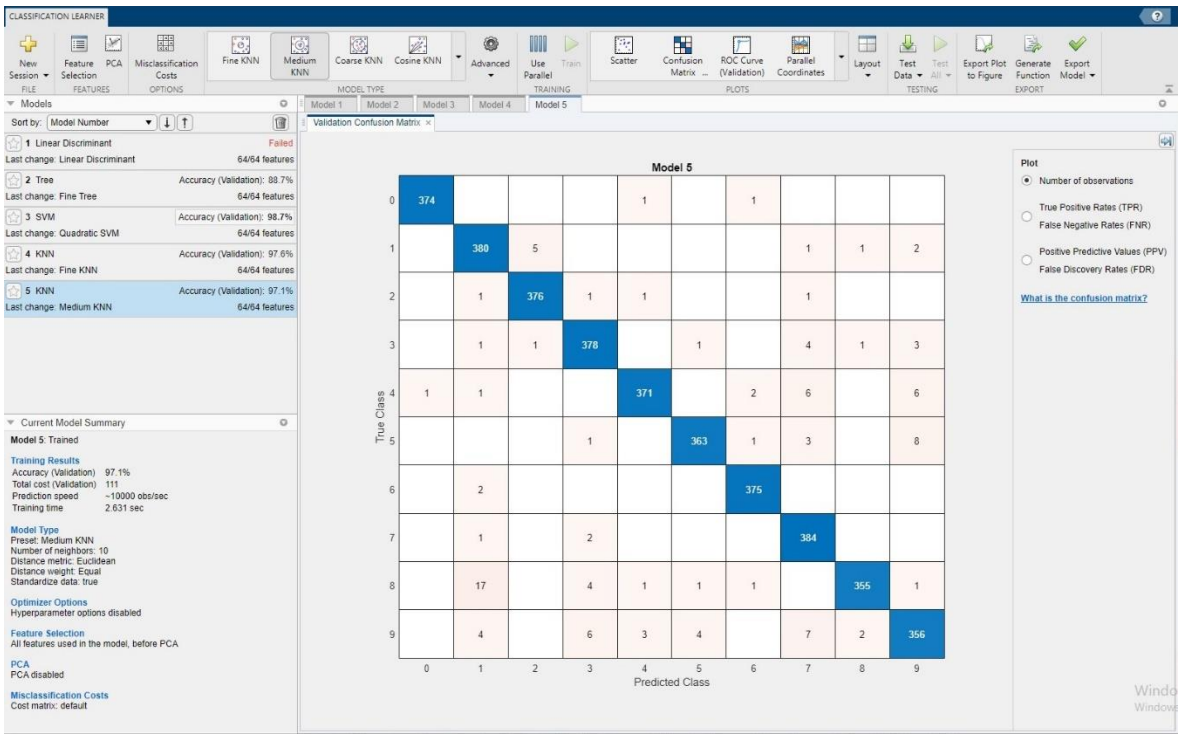
5- Apply the algorithm and discuss the steps and the parameters that you use (why and how)

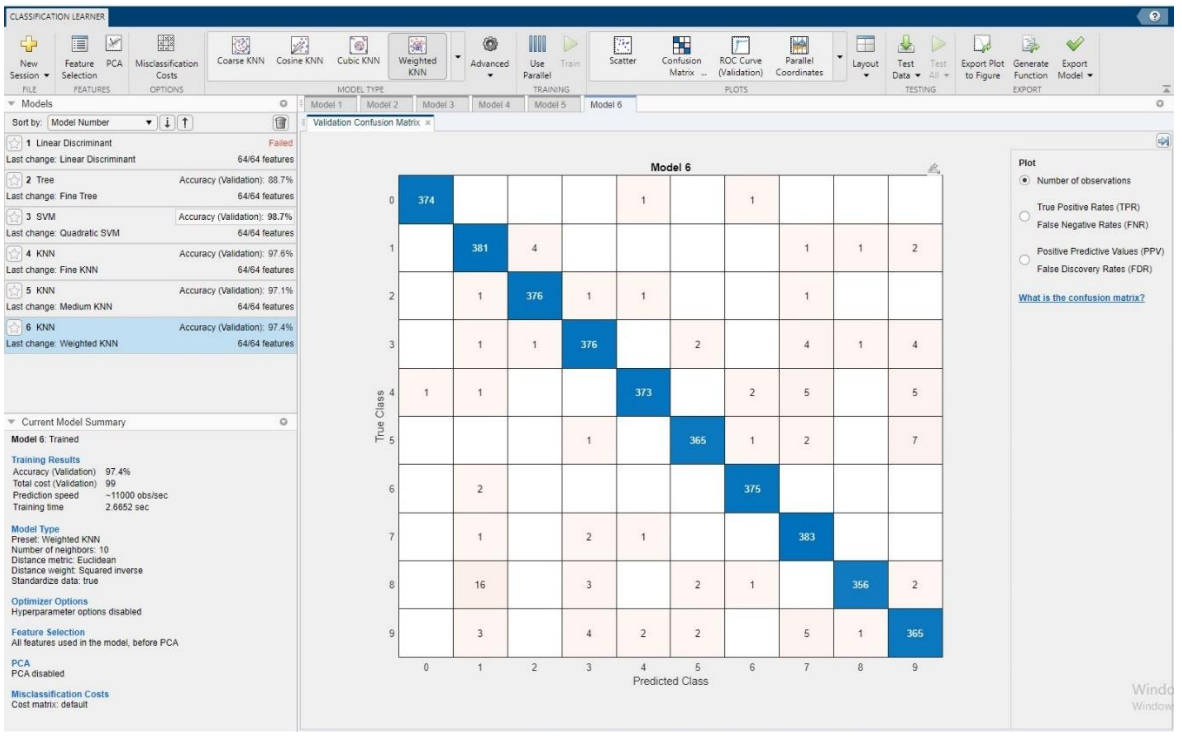
⇨ **SVM :**



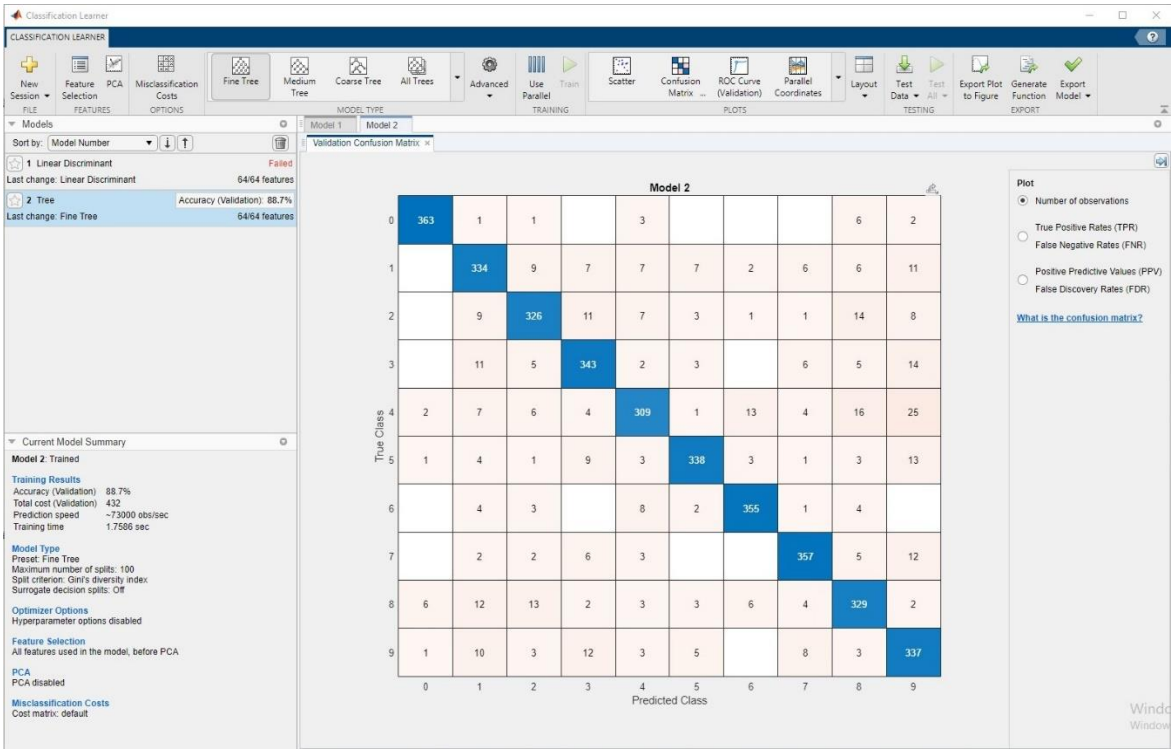
⇨ **KNN:**




CLASSIFICATION LEARNER
Models
Sort by: Model Number
1 Linear Discriminant Failed
Last change: Linear Discriminant 64/64 features
2 Tree Accuracy (Validation): 88.7%
Last change: Fine Tree 64/64 features
3 SVM Accuracy (Validation): 98.7%
Last change: Quadratic SVM 64/64 features
4 KNN Accuracy (Validation): 97.6%
Last change: Fine KNN 64/64 features
5 KNN Accuracy (Validation): 97.1%
Last change: Medium KNN 64/64 features
Current Model Summary
Model 5: Trained
Training Results
Accuracy (Validation) 97.1%
Total cost (Validation) 111
Prediction speed ~10000 obs/sec
Training time 2.631 sec
Model Type
Preset: Medium KNN
Number of neighbors: 10
Distance metric: Euclidean
Distance weight: Equal
Standardize data: true
Optimizer Options
Hyperparameter options disabled
Feature Selection
All features used in the model, before PCA
PCA
PCA disabled
Misclassification Costs
Cost matrix: default
Model 5
True Class
Predicted Class
Plot
Number of observations
True Positive Rates (TPR)
False Negative Rates (FNR)
Positive Predictive Values (PPV)
False Discovery Rates (FDR)
What is the confusion matrix?

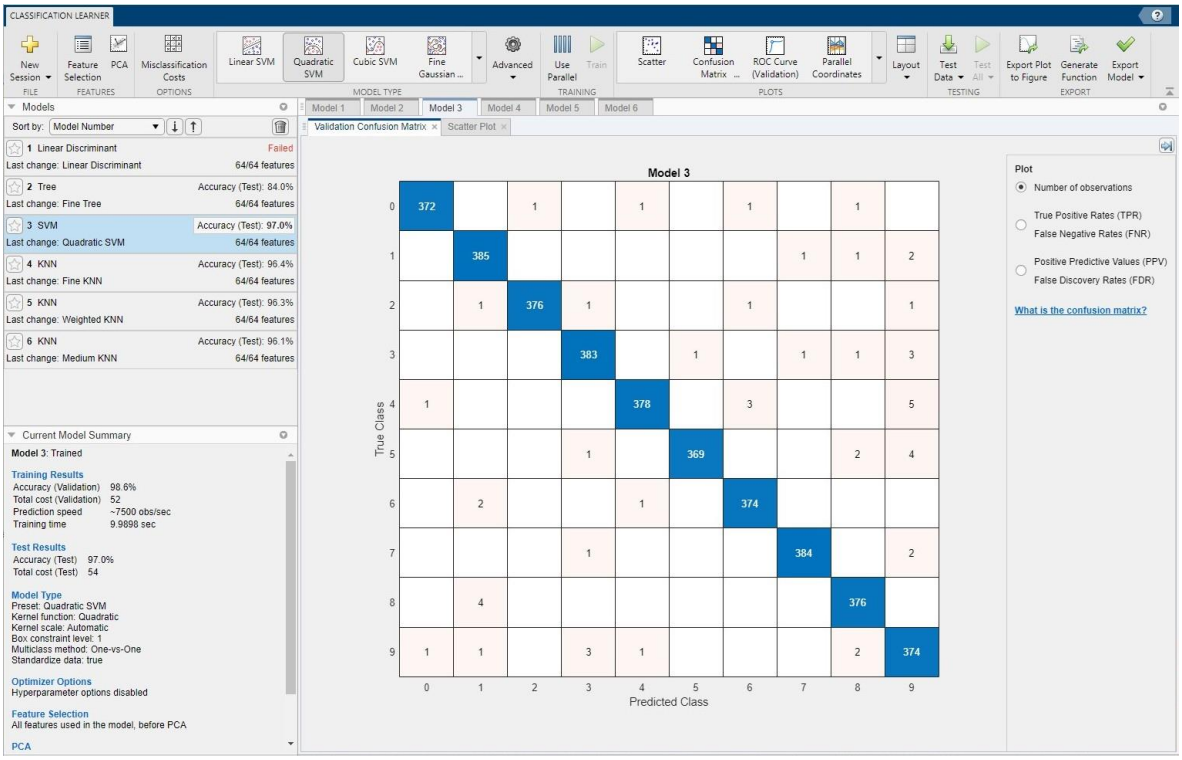

CLASSIFICATION LEARNER
Models
Sort by: Model Number
1 Linear Discriminant Failed
Last change: Linear Discriminant 64/64 features
2 Tree Accuracy (Validation): 88.7%
Last change: Fine Tree 64/64 features
3 SVM Accuracy (Validation): 98.7%
Last change: Quadratic SVM 64/64 features
4 KNN Accuracy (Validation): 97.6%
Last change: Fine KNN 64/64 features
5 KNN Accuracy (Validation): 97.1%
Last change: Medium KNN 64/64 features
6 KNN Accuracy (Validation): 97.4%
Last change: Weighted KNN 64/64 features
Current Model Summary
Model 6: Trained
Training Results
Accuracy (Validation) 97.4%
Total cost (Validation) 99
Prediction speed ~11000 obs/sec
Training time 2.6652 sec
Model Type
Preset: Weighted KNN
Number of neighbors: 10
Distance metric: Euclidean
Distance weight: Squared inverse
Standardize data: true
Optimizer Options
Hyperparameter options disabled
Feature Selection
All features used in the model, before PCA
PCA
PCA disabled
Misclassification Costs
Cost matrix: default
Model 6
True Class
Predicted Class
Plot
Number of observations
True Positive Rates (TPR)
False Negative Rates (FNR)
Positive Predictive Values (PPV)
False Discovery Rates (FDR)
What is the confusion matrix?

⇨ Decision Tree:

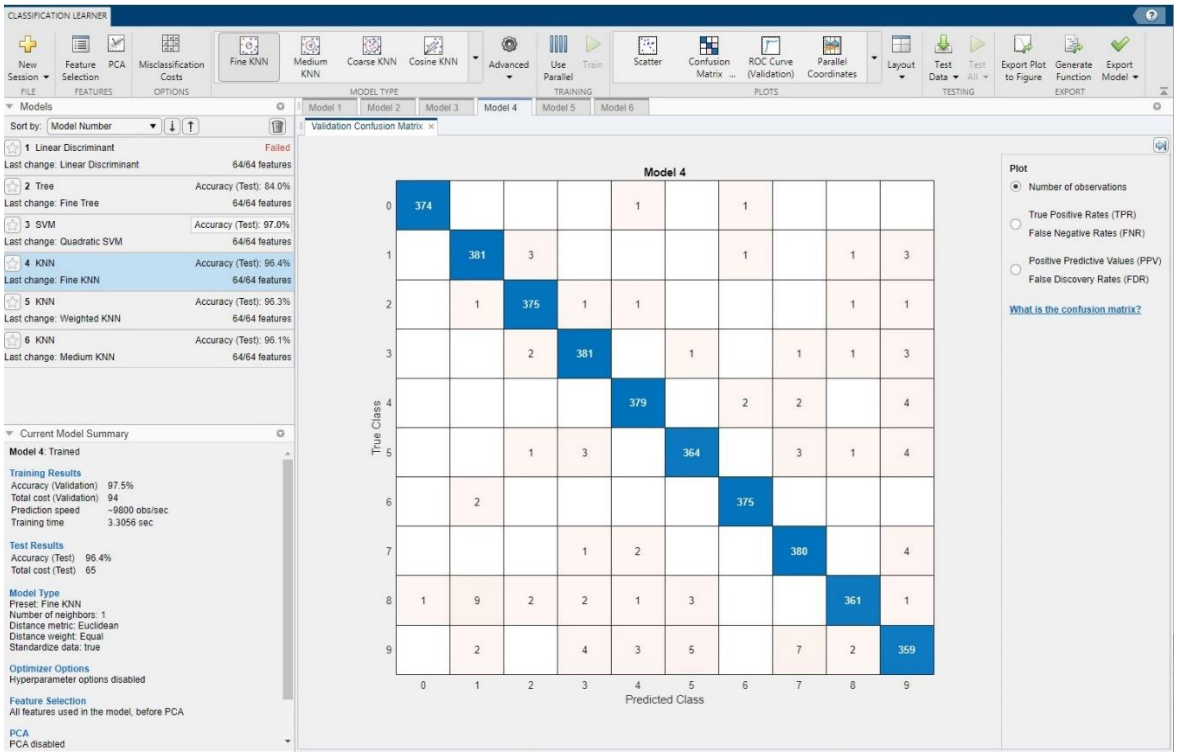


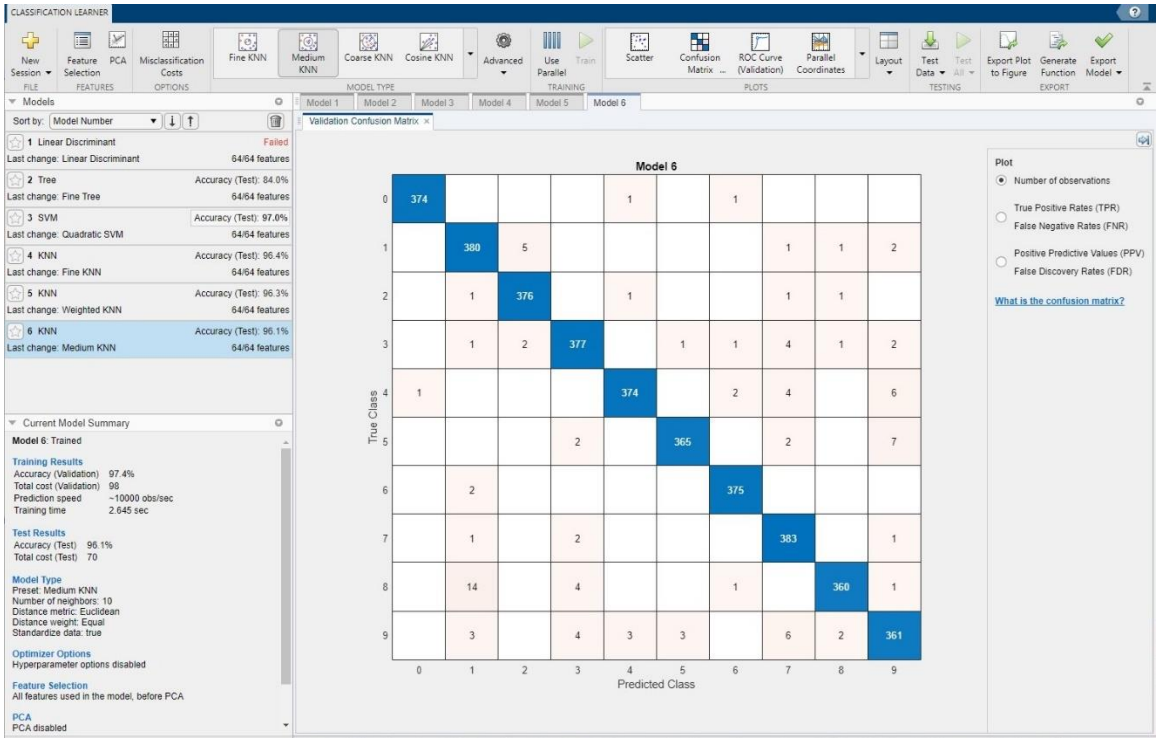
6- Discuss the evaluation and testing step as well as the results of your Project.

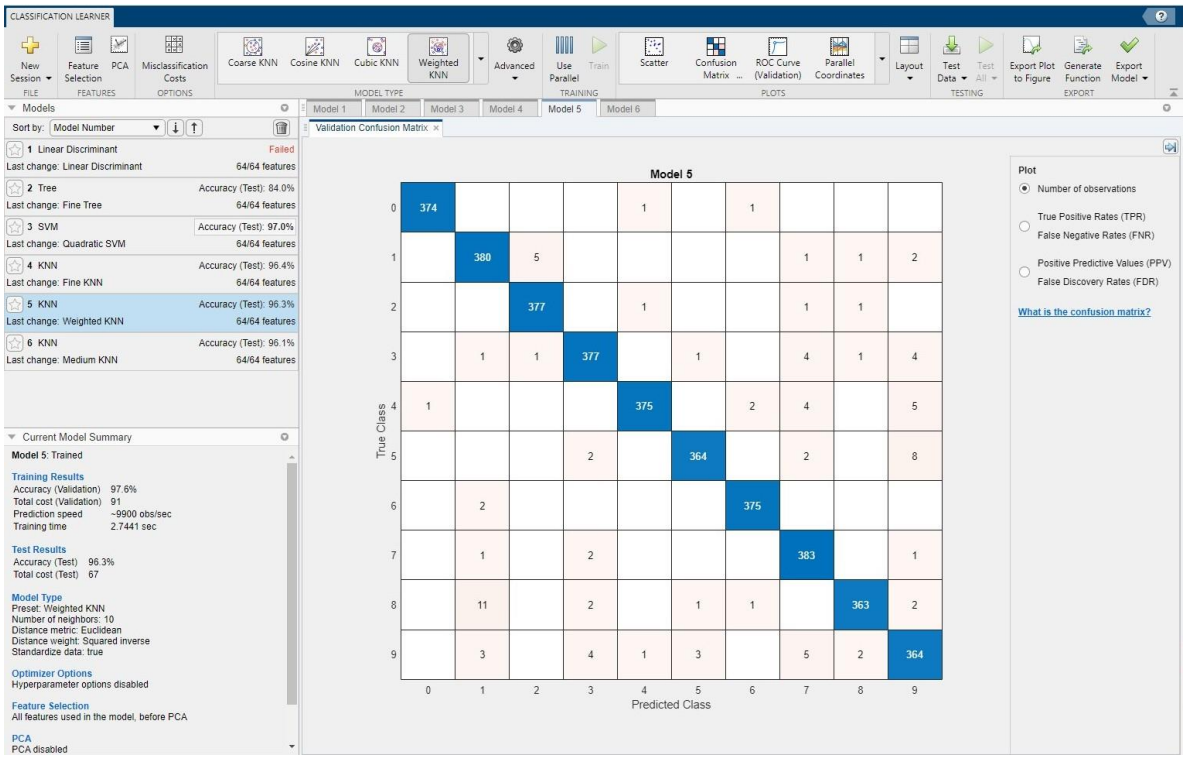
⇨ SVM :



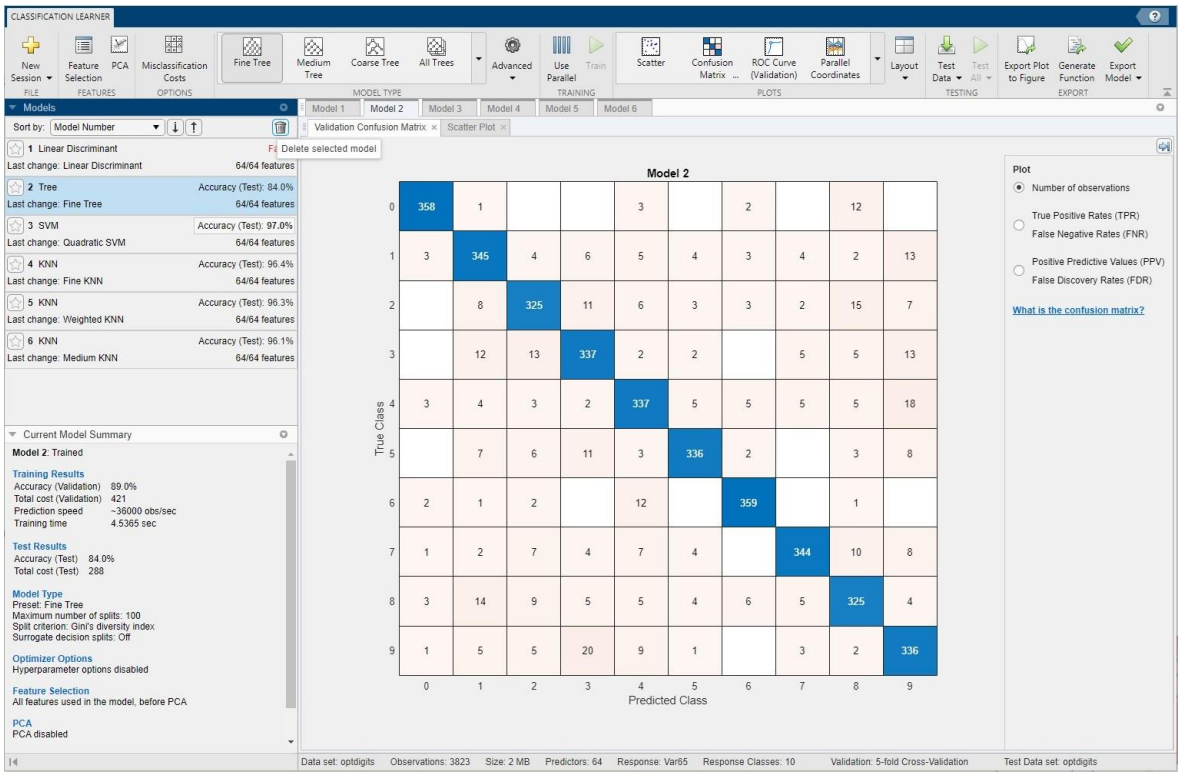
⇨ KNN:


CLASSIFICATION LEARNER
Models
Sort by: Model Number
1 Linear Discriminant Failed
Last change: Linear Discriminant 64/64 features
2 Tree Accuracy (Test): 84.0%
Last change: Fine Tree 64/64 features
3 SVM Accuracy (Test): 97.0%
Last change: Quadratic SVM 64/64 features
4 KNN Accuracy (Test): 96.4%
Last change: Fine KNN 64/64 features
5 KNN Accuracy (Test): 96.3%
Last change: Weighted KNN 64/64 features
6 KNN Accuracy (Test): 96.1%
Last change: Medium KNN 64/64 features
Current Model Summary
Model 4: Trained
Training Results
Accuracy (Validation) 97.5%
Total cost (Validation) 94
Prediction speed ~9800 obs/sec
Training time 3.3056 sec
Test Results
Accuracy (Test) 96.4%
Total cost (Test) 65
Model Type
Preset: Fine KNN
Number of neighbors: 1
Distance metric: Euclidean
Distance weight: Equal
Standardize data: true
Optimizer Options
Hyperparameter options disabled
Feature Selection
All features used in the model, before PCA
PCA
PCA disabled
Model 4
True Class
Predicted Class
Plot
Number of observations
True Positive Rates (TPR)
False Negative Rates (FNR)
Positive Predictive Values (PPV)
False Discovery Rates (FDR)
What is the confusion matrix?


CLASSIFICATION LEARNER
Models
Sort by: Model Number
1 Linear Discriminant Failed
Last change: Linear Discriminant 64/64 features
2 Tree Accuracy (Test): 84.0%
Last change: Fine Tree 64/64 features
3 SVM Accuracy (Test): 97.0%
Last change: Quadratic SVM 64/64 features
4 KNN Accuracy (Test): 96.4%
Last change: Fine KNN 64/64 features
5 KNN Accuracy (Test): 96.3%
Last change: Weighted KNN 64/64 features
6 KNN Accuracy (Test): 96.1%
Last change: Medium KNN 64/64 features
Current Model Summary
Model 6: Trained
Training Results
Accuracy (Validation) 97.4%
Total cost (Validation) 98
Prediction speed ~10000 obs/sec
Training time 2.645 sec
Test Results
Accuracy (Test) 96.1%
Total cost (Test) 70
Model Type
Preset: Medium KNN
Number of neighbors: 10
Distance metric: Euclidean
Distance weight: Equal
Standardize data: true
Optimizer Options
Hyperparameter options disabled
Feature Selection
All features used in the model, before PCA
PCA
PCA disabled
Model 6
True Class
Predicted Class
Plot
Number of observations
True Positive Rates (TPR)
False Negative Rates (FNR)
Positive Predictive Values (PPV)
False Discovery Rates (FDR)
What is the confusion matrix?



⇨ Decision Tree:



7- Comment on the whole Project

We used six algorithms for this project. Such as KNN, Quadratic SVM, and Decision Tree. Some of the results given by these algorithms are good but others are not. The Quadratic SVM model fits with this data set the best. Either from the training validation accuracy result which is 98.6% and the test accuracy result which is 97.0% and when we compare these two numbers, there are only 1.6% differences. This difference means the model is give a good result.

Name of Methods	Accuracy	
	Training	Testing
Quadratic SVM	98.6	97
Fine KNN	97.6	96.4
Weighted KNN	97.6	96.3
Medium KNN	97.4	96.1
Decision Tree	89	84
Linear Discriminant	Failed	Failed