INDIAN INSTITUTE OF TECHNOLOGY, DELHI

ASSIGNMENT REPORT

COL774 : Assignment 3

APAR AHUJA | ENTRY NO. 2019CS10465

Course - COL774 | Prof. Parag Singla

Compiled on October 27, 2021

Contents

Chapter 1

Decision Trees and Random Forests

1.1 Decision Tree

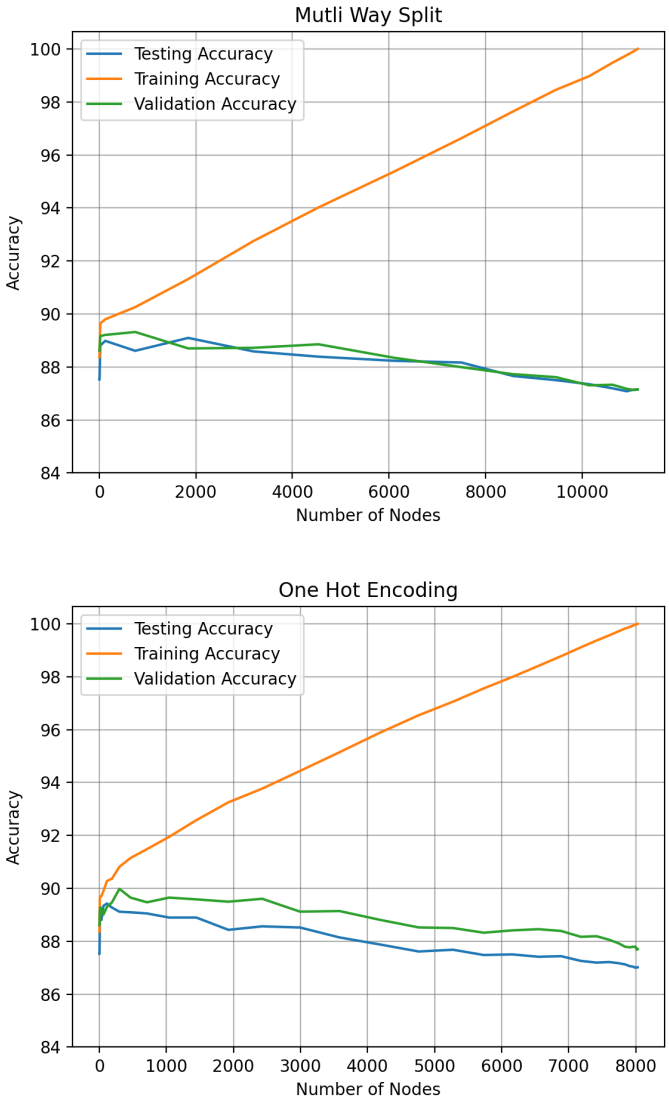


Figure 1.1: Multi way split v/s One Hot Encoding

Analysis:
1. Training accuracy increases to nearly 100 percent with increase in. number of node. The more the nodes the better the model can fit the training data. However this is leads to over-fitting to the training data.
2. The initial training, testing and validation accuracy is low showing the model hasn't learnt a lot and is under-fitting. The training, testing and validation accuracy then increase to a peak and then drop back down with increase in number of nodes. This shows that the model starts over-fitting after a certain point.
3. Using one-hot-encoding gives higher peak accuracy. Infact one hot encoding gave close to 90% validation accuracy at it's peak. Thus, using one hot encoding gives a better model in terms of accuracy.
4. The number of nodes used by one-hot-encoding is less than those needed by multi way split. Thus, this model generalizes better for the same number of nodes.

1.2 Decision Tree Post Pruning

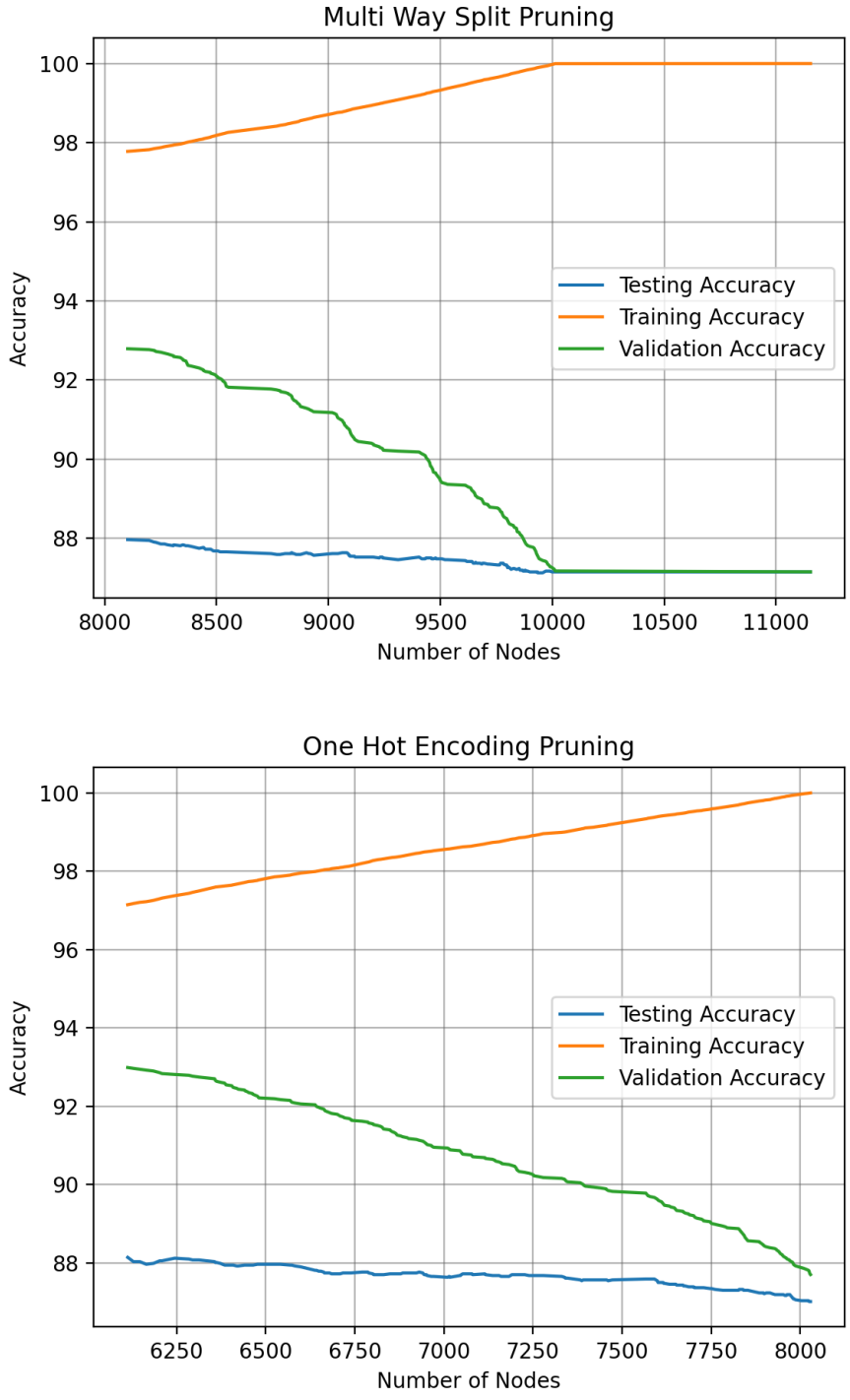


Figure 1.2: Multi way split v/s One Hot Encoding Pruning

Analysis:
1. Pruning leads to a reduce in training accuracy for both the models. The models unlearn the over-fitting and remove nodes that aren't as relevant to begin with. The final training accuracy is still around 97% - 98% which is really nice and shows the model isn't over-fitting to the validation set. The increase in the testing accuracy also suggests the model didn't over-fit the validation data but rather it generalized better.
2. Pruning increases the validation accuracy to well over 92 %. This was expected by the very nature of the implementation where we try to pruned nodes that increase the validation accuracy.
3. The testing accuracy increases consistently with decrease in number of nodes. The models ended up with a near 88% accuracy. Simple pruning led to a jump of nearly a percent from it's previous values.

1.3 Random Forests

Note: One-hot-encoding was used to encode categorical data to integers.
Best Parameters: {max_features: 0.3, min_samples_split: 10, n_estimators: 350}
Out-of-bag Accuracy = 90.38%
Testing Accuracy = 89.91 %
Training Accuracy = 97.92 %
Validation Accuracy = 90.69 %
I used 5 fold validation during the grid search and multi-core processing to increase the search speed. Using random forests led to a significant increase in testing, training and validation accuracy. Growing multiple decision trees in parallel with bootstrapping leads to better trees/models.

1.4 Parameter Sensitivity

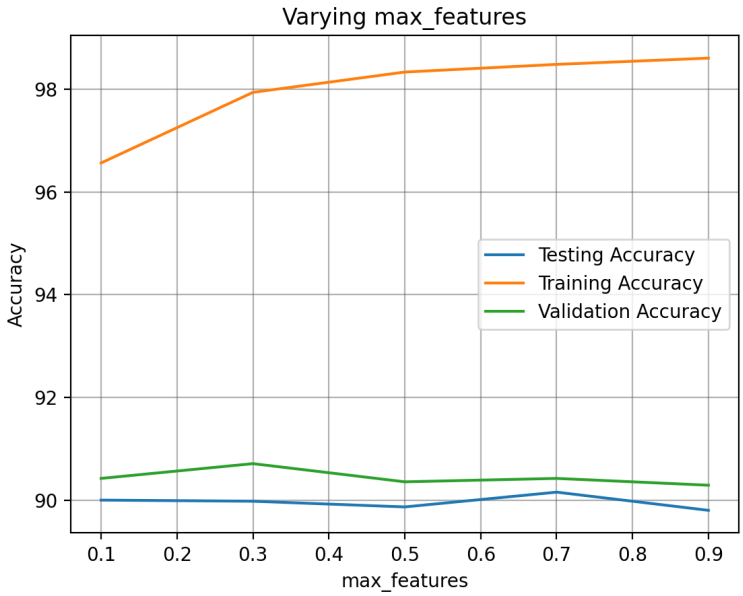


Figure 1.3: Parameter Sensitivity Analysis

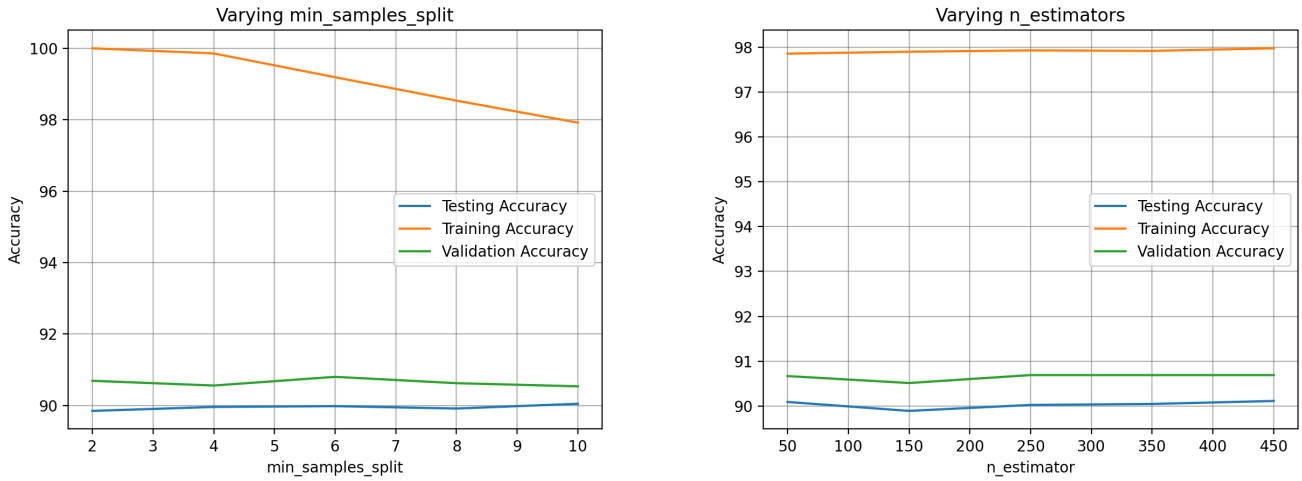


Figure 1.4: Parameter Sensitivity Analysis

Analysis:

1. There is a possible increase in training accuracy with variation in optimal parameters even though the model was fit on training data. This is because of the fact that we used out-of-bag accuracy as a scoring mechanism in the grid search.
2. The validation and testing accuracy don't change a lot. A less than 1% change is observed in them showing the quality of the model stays put even with minor variations in a certain parameter.
3. The n_estimators parameter is particularly stable and doesn't lead to any major change in testing, training and validation accuracy with local changes in it's value.

Chapter 2

Neural Networks

2.1 C

I have attached at the following pages the data with learning rate = 0.1 as specified in the question. However, I noticed that the model converged even for a rate = 1 in far lesser iterations without any compromise in accuracy and an even stricter stopping criteria, mentioned below. So, I've attached my observation for 5 and 10 hidden layer nodes with a learning rate = 1.

Stricter Stopping Criteria with learning rate = 1:
a. The model must traverse the complete dataset atleast once.
b. The difference in average batch MSE for two consecutive epochs[1] must fall below 5x10^{-7} atleast 10 times and the maximum number of epochs is 10000.

Variation with number of hidden layer nodes for learning rate = 1:
1. Number of nodes = 5:
Number of epochs before stopping : 6838
Training Time : 308.79 secs
Training MSE : 0.2214
Training Accuracy : 67.9%
Testing MSE : 0.2286
Testing Accuracy : 66.31%
Testing Confusion Matrix

$$\begin{bmatrix} 432565 & 191925 & 6876 & 6559 & 1952 & 1708 & 28 & 42 & 4 & 3 \\ 68644 & 230573 & 40746 & 14562 & 1933 & 288 & 1396 & 188 & 8 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

[1] one epoch refers to traversing the whole dataset once

2. Number of nodes = 10:
Number of epochs before stopping : 10000
Training Time : 565.1 secs
Training MSE : 0.1414
Training Accuracy : 84.17%
Testing MSE : 0.1529
Testing Accuracy : 82.4%
Testing Confusion Matrix

$$\begin{bmatrix} 472944 & 71478 & 1512 & 1813 & 3331 & 1887 & 10 & 2 & 12 & 0 \\ 28265 & 351020 & 46110 & 19308 & 554 & 109 & 1414 & 228 & 0 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Stopping Criteria for learning rate = 0.1:
a. The model must traverse the complete dataset atleast once.
b. The difference in average batch MSE for two consecutive epochs[2] must fall below 10^{-6} atleast 5 times and the maximum number of epochs is 10000.
c. The intuition behind crossing the threshold difference 5 times is that the difference sometime fall just below the threshold for a single value and increases back again. Hence, a single drop below the threshold is not a consistent convergence criteria.

Data for learning rate = 0.1 is in the following pages.

[2] one epoch refers to traversing the whole dataset once

Variation with number of hidden layer nodes for learning rate = 0.1:

1. Number of nodes = 5:
Number of epochs before stopping : 6056
Training Time : 364.76 secs
Training MSE : 0.2267
Training Accuracy : 66.4%
Testing MSE : 0.2327
Testing Accuracy : 65.23%
Testing Confusion Matrix

$$\begin{bmatrix} 436801 & 207033 & 10283 & 7815 & 2763 & 1734 & 146 & 40 & 8 & 2 \\ 64408 & 215465 & 37339 & 13306 & 112 & 262 & 1278 & 190 & 4 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

2. Number of nodes = 10:
Number of epochs before stopping : 5353
Training Time : 329.81 secs
Training MSE : 0.1406
Training Accuracy : 82.4%
Testing MSE : 0.1555
Testing Accuracy : 80.09%
Testing Confusion Matrix

$$\begin{bmatrix} 462964 & 84532 & 828 & 1300 & 3613 & 1866 & 2 & 3 & 11 & 3 \\ 38245 & 337966 & 46794 & 19821 & 272 & 130 & 1422 & 227 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

3. Number of nodes = 15:
Number of epochs before stopping : 5842
Training Time : 434.06 secs
Training MSE : 0.0619
Training Accuracy : 91.9%
Testing MSE : 0.066
Testing Accuracy : 91.62%
Testing Confusion Matrix

$$\begin{bmatrix}
500268 & 6538 & 125 & 320 & 3840 & 1994 & 4 & 29 & 12 & 3 \\
941 & 415960 & 47497 & 20801 & 45 & 2 & 1420 & 201 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}$$

4. Number of nodes = 20:
Number of epochs before stopping : 6119
Training Time : 474.28 secs
Training MSE : 0.0602
Training Accuracy : 92.33%
Testing MSE : 0.0632
Testing Accuracy : 92.27%
Testing Confusion Matrix

$$\begin{bmatrix}
501167 & 1013 & 2 & 4 & 3866 & 1996 & 0 & 0 & 12 & 3 \\
42 & 421485 & 47620 & 21117 & 19 & 0 & 1424 & 230 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}$$

5. Number of nodes = 25:
Number of epochs before stopping : 7599
Training Time : 629.22 secs
Training MSE : 0.0564
Training Accuracy : 92.33%
Testing MSE : 0.0638
Testing Accuracy : 92.28%
Testing Confusion Matrix

$$\begin{bmatrix} 501089 & 741 & 0 & 142 & 3785 & 1995 & 0 & 0 & 12 & 3 \\ 120 & 421757 & 47622 & 20979 & 100 & 1 & 1424 & 230 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Plots

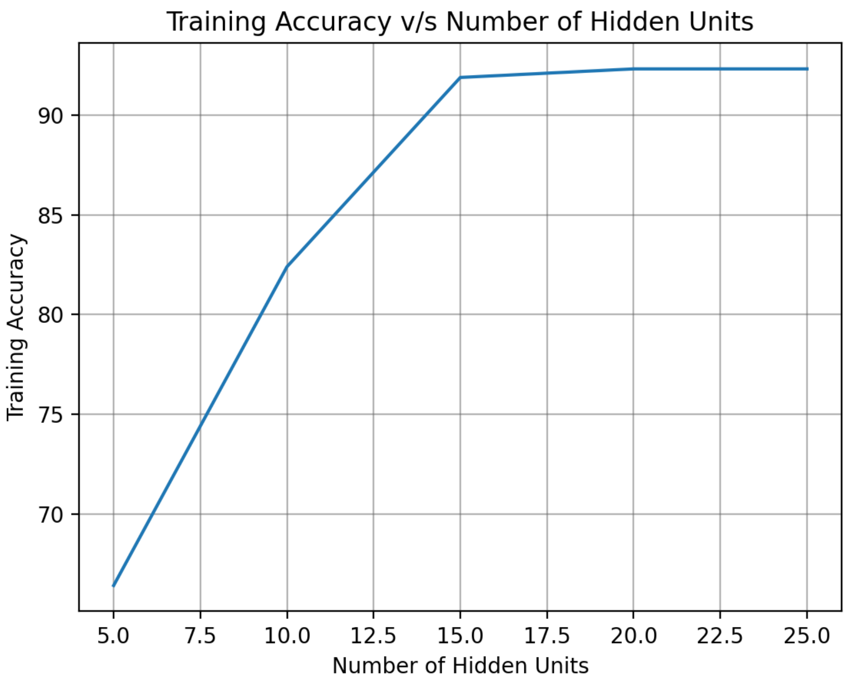
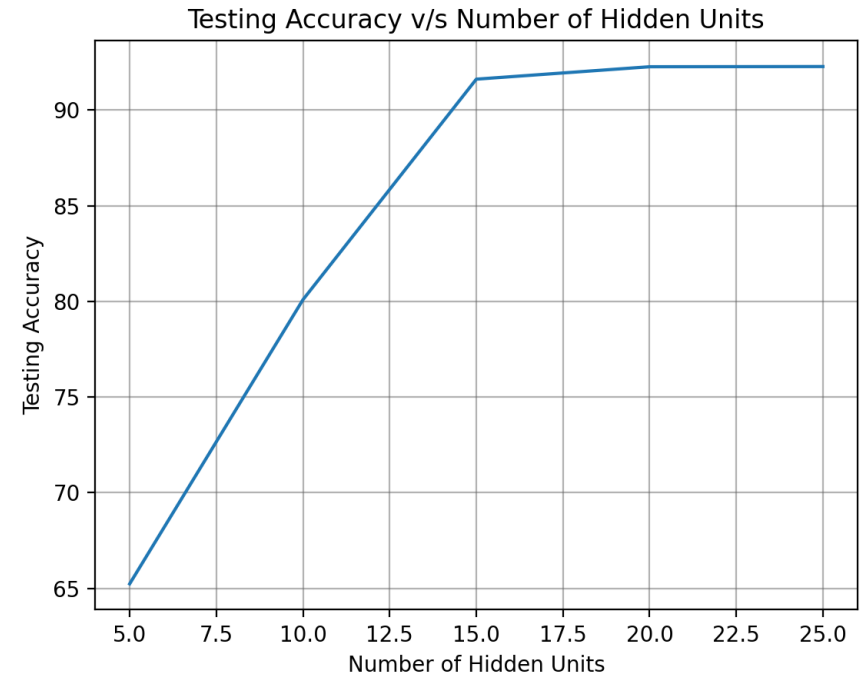


Figure 2.1: Training Accuracy



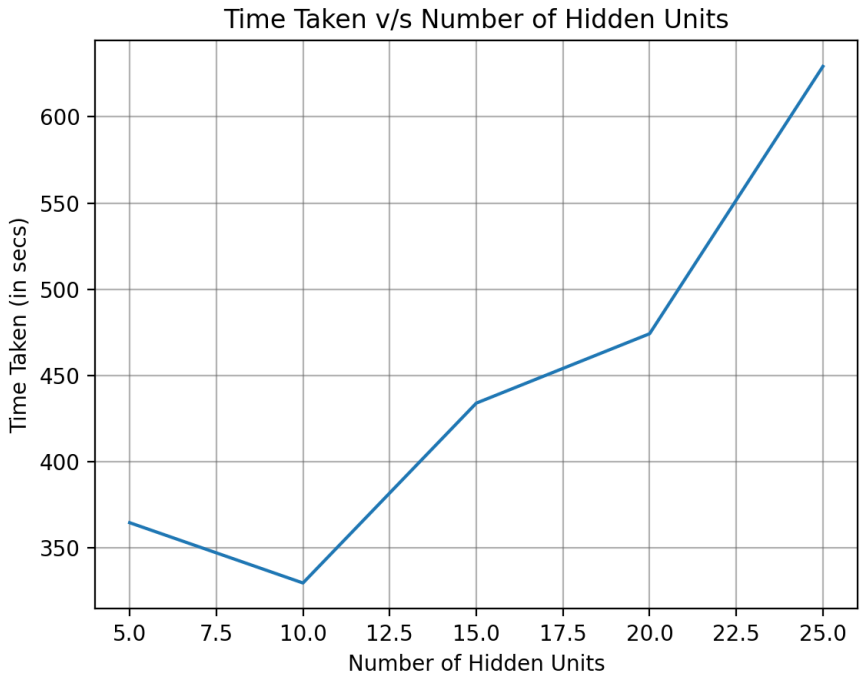


Figure 2.2: Testing Accuracy and Time Taken Plots

Analysis:
1. The model predicts the first 2 classes all the time. The training data for other classes is not sufficient for the model to learn correctly.
2. Number of epochs before convergence and total time taken roughly increase with increase in number of nodes in the layer.
3. The data is heavily skewed and accuracy might not be the right measure for quality of a model here. We might want to use macro F1 score instead to compare models and their quality.

2.2 D

I used an initial learning rate = 5 as the rate decays way to quickly to cause any change of values in later iterations. And the updated stopping criteria is -
a. The model must traverse the complete dataset atleast once.
b. The difference in average batch MSE for two consecutive epochs[3] must fall below 10^{-6} atleast 5 times and the maximum number of epochs is 10000.

Variation with number of hidden layer nodes for adaptive learning rate:

1. Number of nodes = 5:
Number of epochs before stopping : 2661
Training Time : 324.93 secs
Training MSE : 0.2333
Training Accuracy : 65.45%
Testing MSE : 0.2389
Testing Accuracy : 63.9%
Testing Confusion Matrix

$$\begin{bmatrix} 423479 & 207018 & 10507 & 7525 & 2311 & 1718 & 153 & 39 & 5 & 3 \\ 77730 & 215480 & 37115 & 13596 & 1574 & 278 & 1271 & 191 & 7 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

2. Number of nodes = 10:
Number of epochs before stopping : 4763
Training Time : 340.08 secs
Training MSE : 0.1479
Training Accuracy : 82.59%
Testing MSE : 0.1614
Testing Accuracy : 80.4%
Testing Confusion Matrix

[3] one epoch refers to traversing the whole dataset once

$$\begin{bmatrix} 470098 & 88627 & 2132 & 1687 & 3072 & 1873 & 17 & 5 & 9 & 3 \\ 31111 & 333871 & 45490 & 19434 & 813 & 123 & 1407 & 225 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

3. Number of nodes = 15:
Number of epochs before stopping : 4976
Training Time : 378.05 secs
Training MSE : 0.0661
Training Accuracy : 92.1%
Testing MSE : 0.0702
Testing Accuracy : 91.67%
Testing Confusion Matrix

$$\begin{bmatrix} 499342 & 5132 & 44 & 88 & 3762 & 1993 & 1 & 1 & 9 & 3 \\ 1867 & 417366 & 47578 & 21033 & 123 & 3 & 1423 & 229 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

4. Number of nodes = 20:
Number of epochs before stopping : 6704
Training Time : 450.39 secs
Training MSE : 0.0584
Training Accuracy : 92.33%
Testing MSE : 0.0636
Testing Accuracy : 92.14%
Testing Confusion Matrix

$$\begin{bmatrix} 501078 & 2197 & 5 & 214 & 3862 & 1996 & 0 & 14 & 12 & 3 \\ 131 & 420301 & 47617 & 20907 & 23 & 0 & 1424 & 216 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

5. Number of nodes = 25:
Number of epochs before stopping : 3823
Training Time : 419.45 secs
Training MSE : 0.0562
Training Accuracy : 92.33%
Testing MSE : 0.0616
Testing Accuracy : 92.22%
Testing Confusion Matrix

$$\begin{bmatrix} 500615 & 927 & 0 & 23 & 3831 & 1994 & 0 & 0 & 12 & 3 \\ 594 & 421571 & 47622 & 21098 & 54 & 2 & 1424 & 230 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Plots

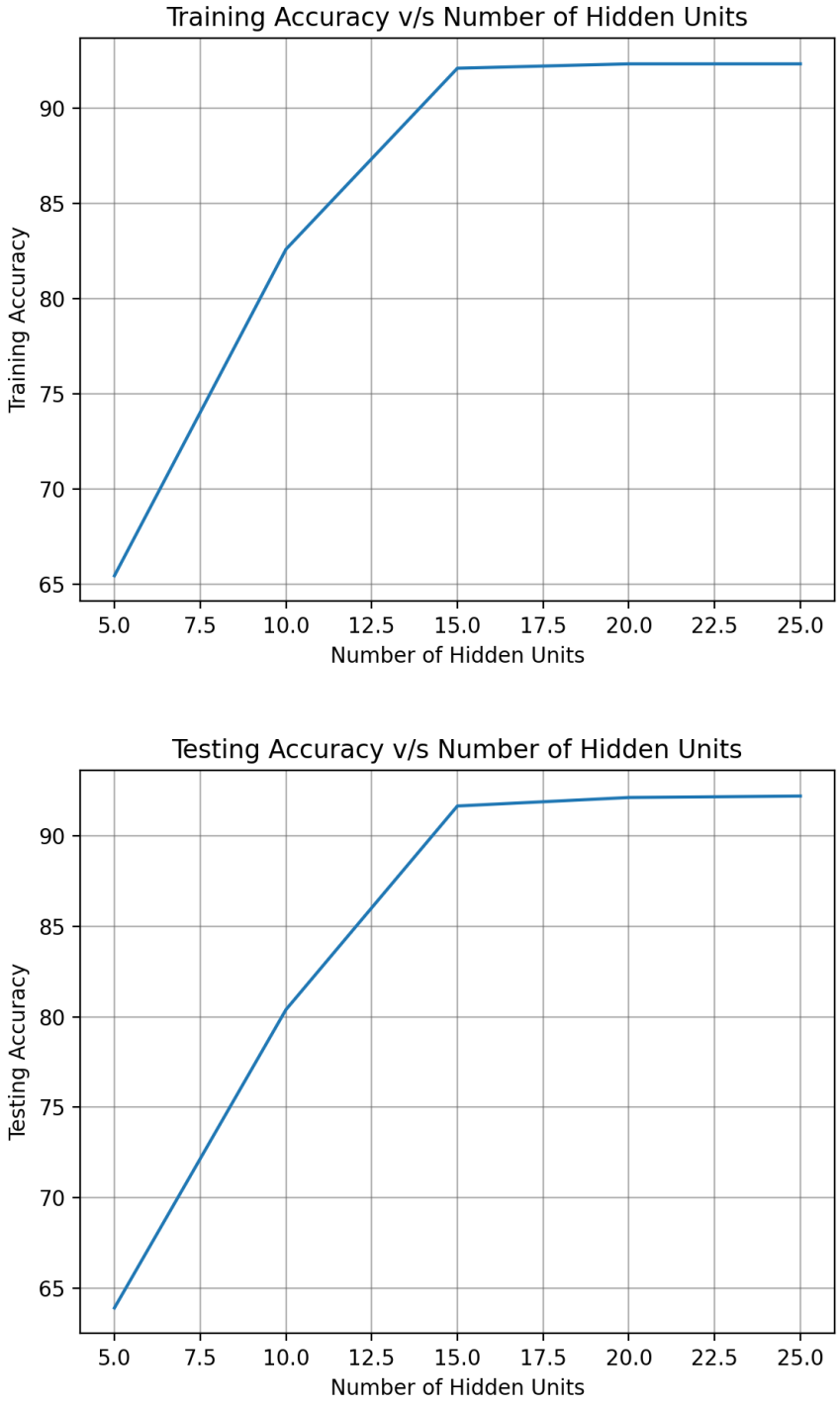


Figure 2.3: Training and Testing Accuracy Plots

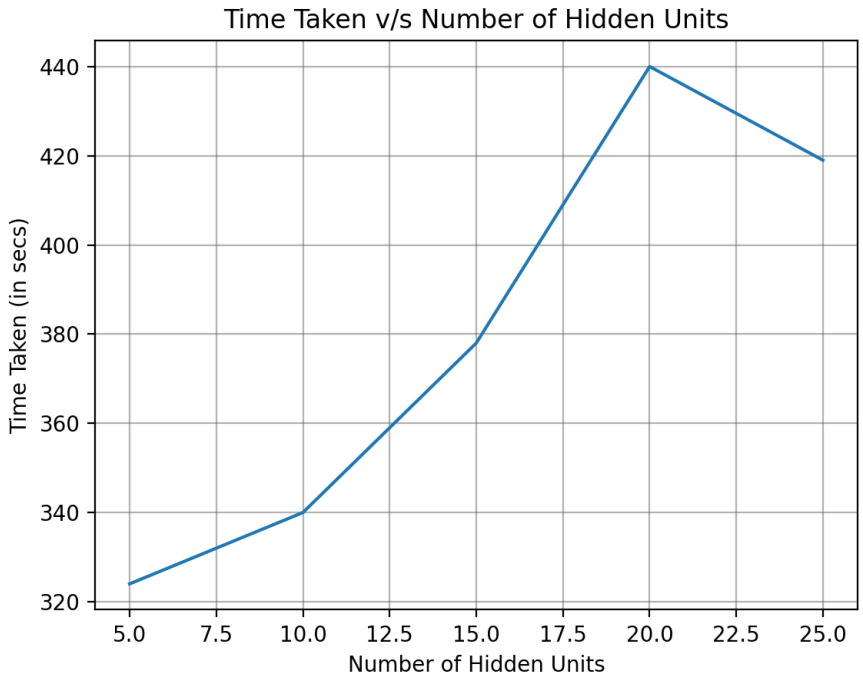


Figure 2.4: Time Taken Plot

Analysis:
1. No change in stopping criteria was necessary, the model converged correctly with the same criteria.
2. The results are very similar to the part C. The accuracies are with 1% error and the confusion matrices look very close to each other.
3. Adaptive learning rate provides faster convergence as compared to the previous part without much compromise in performance quality.
4. The initial learning rate was set to 5 instead of 0.1 due to the fast decreasing nature of inverse adaptive learning.

2.3 E

Layer Architecture = $[100, 100]$
Learning Rate = $3/\sqrt{epoch}$
Stopping Criteria −
a. The model must traverse the complete dataset atleast once.
b. The difference in average batch MSE for two consecutive epochs[4] must fall below 10^{-6} atleast 5 times and the maximum number of epochs is 3000.

[4]one epoch refers to traversing the whole dataset once

ReLU
Number of epochs = 253
Training Time : 59.78 secs
Training MSE : 0.0397
Training Accuracy : 92.33%
Testing MSE : 0.0423
Testing Accuracy : 92.37%
Testing Confusion Matrix

$$\begin{bmatrix} 501186 & 0 & 0 & 0 & 3806 & 1987 & 0 & 0 & 11 & 1 \\ 23 & 422498 & 47622 & 21121 & 79 & 9 & 1424 & 230 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Sigmoid
Number of epochs = 3000 (= MAX_ITERATIONS)
Training Time : 1019.79 secs
Training MSE : 0.0471
Training Accuracy : 92.34%
Testing MSE : 0.0709
Testing Accuracy : 90.86%
Testing Confusion Matrix

$$\begin{bmatrix} 494612 & 8545 & 77 & 89 & 3731 & 1961 & 0 & 3 & 11 & 3 \\ 6595 & 413945 & 47545 & 21030 & 153 & 35 & 1424 & 227 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Analysis:
1. The models still classifies mostly into the first 2 classes.
2. Nonetheless, the training and testing accuracies are high (90+).
3. Max iterations were reduced to 3000 due to the fast convergence observed for both the models.
4. Both the activations work equally well although the ReLU activation converged very fast when compared to sigmoid which took 3000 epochs.
5. The results are good when compared to 2c with a single layer but the difference is statistically insignificant. Thus, a single layer with 25 neurons is learning equally well as a 2 - 100 node - layers.

2.4 F

MLP with ReLU (constant learning rate)
Training Accuracy is : 100.0%
Testing Accuracy is : 99.28%
Learning Rate is constant : 0.1
Layer Architecture is : $[100, 100]$
Confusion Matrix :

$$\begin{bmatrix} 501160 & 7 & 5 & 201 & 3849 & 1782 & 455 & 0 & 9 & 3 \\ 0 & 422491 & 29 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 47588 & 0 & 0 & 0 & 560 & 0 & 0 & 0 \\ 0 & 0 & 0 & 20918 & 0 & 0 & 8 & 230 & 0 & 0 \\ 44 & 0 & 0 & 0 & 36 & 0 & 0 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 & 214 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

MLP with ReLU (adaptive[5] learning rate)
Training Accuracy is : 100.0%
Testing Accuracy is : 99.22%
Adaptive Learning Rate with initialization : 0.1
Layer Architecture is : $[100, 100]$
Confusion Matrix :

[5] adpative here does not mean the same as $n_0/\sqrt{epoch}$

$$\begin{bmatrix} 501064 & 33 & 305 & 307 & 3690 & 1925 & 566 & 0 & 11 & 3 \\ 1 & 422458 & 134 & 14 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 7 & 47181 & 21 & 0 & 0 & 357 & 0 & 0 & 0 \\ 0 & 0 & 2 & 20779 & 0 & 0 & 89 & 230 & 0 & 0 \\ 141 & 0 & 0 & 0 & 195 & 0 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 0 & 71 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 412 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

MLP with sigmoid (adaptive learning rate)
Training Accuracy is : 99.94%
Testing Accuracy is : 98.26%
Adaptive Learning Rate with initialization : 0.1
Layer Architecture is : $[100, 100]$
Confusion Matrix :

$$\begin{bmatrix} 500972 & 510 & 1440 & 1998 & 2914 & 1725 & 110 & 0 & 6 & 1 \\ 11 & 421620 & 2615 & 145 & 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 367 & 42915 & 3171 & 0 & 0 & 73 & 0 & 0 & 0 \\ 0 & 1 & 650 & 15756 & 0 & 0 & 1138 & 230 & 0 & 0 \\ 218 & 0 & 0 & 0 & 968 & 0 & 0 & 0 & 4 & 0 \\ 8 & 0 & 0 & 0 & 0 & 271 & 0 & 0 & 2 & 2 \\ 0 & 0 & 2 & 51 & 0 & 0 & 103 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Analysis:
1. The learning time is very low when compared to the parts above. This is possibly due to the adaptive learning rate strategy used by MLP which is much superior than using the inverse decay rate.
2. The confusion matrix has values across all the classes thus, the model is not over-fitting to the first 2 classes as in previous parts.
3. The models have a very high accuracy and are learning very well from the little data available for the last 8 classes.
4. The MLP model has a much better macro F1 score as compared to the previous parts. The change is much more significant than the change in accuracy due to the skewed nature of the datasets.