

MTL782: Data Mining Assignment 2

Saket Kandoi 2021MT60265
Navneet Raj 2021MT10240
Aditya Thomas 2021MT60944

Question 1: MNIST Handwritten Digits

Packages Used: sklearn, keras, pandas, numpy, matplotlib, seaborn

Decision Tree

Accuracy: 0.8730714285714286
Precision: 0.8728356387299575
Recall: 0.8730714285714286
F1 Score: 0.8728940837963073

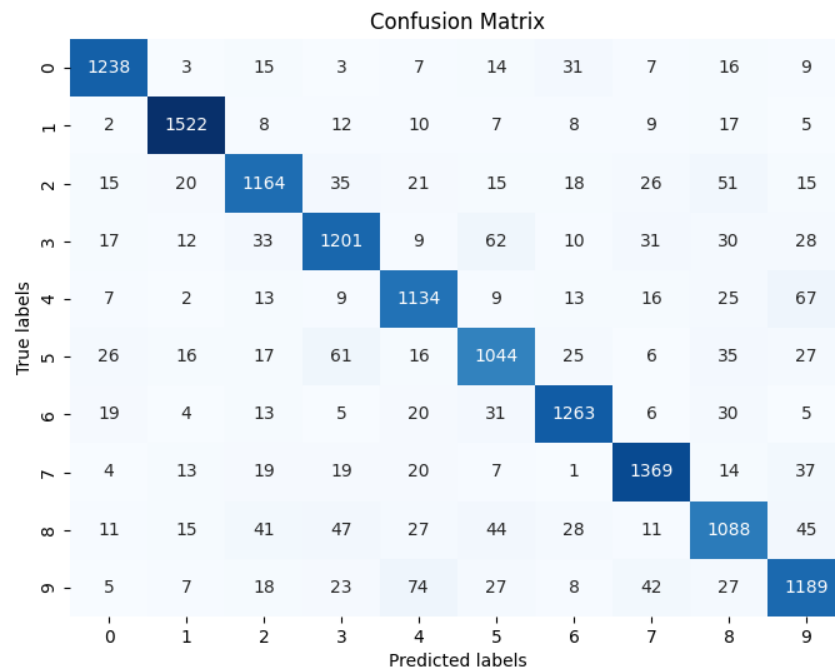


Figure 1: Confusion Matrix : Decision Tree

Best Performance by GridSearch:

Best parameters: {'max_depth': 20, 'min_samples_leaf': 4, 'min_samples_split': 5}

Accuracy: 0.876

MTL782: Assignment 2

Random Forest

Accuracy: 0.9676428571428571
Precision: 0.9676483805024302
Recall: 0.9676428571428571
F1 Score: 0.9676255671575715

Confusion Matrix										
True labels	0	1	2	3	4	5	6	7	8	9
	1325	0	4	0	2	1	2	1	7	1
	0	1576	4	8	2	0	0	6	4	0
	2	4	1339	2	5	0	8	9	8	3
	1	0	24	1357	1	14	3	14	11	8
	3	0	2	0	1257	0	5	3	2	23
	4	3	3	22	3	1217	6	1	13	1
	5	4	2	1	0	5	8	1374	0	2
	6	4	2	1	0	5	8	1374	0	2
	7	2	5	16	1	4	0	0	1458	1
	8	0	4	12	13	4	15	4	7	1288
	9	5	7	2	15	14	4	1	12	11
Predicted labels										

Figure 2: Confusion Matrix : Random Forest

Best Performance by GridSearch:
Best parameters: {'max_depth': None, 'n_estimators': 300}
Accuracy: 0.9676428571428571

Naive Bayes

We tested three cases, and selected Bernoulli for cross-validation since it had the best performance.

GaussianNB

0.5566 accuracy with a standard deviation of 0.0063

MultinomialNB

0.8256 accuracy with a standard deviation of 0.0104

MTL782: Assignment 2

BernoulliNB

Accuracy: 0.8348571428571429
Precision: 0.8367482492867026
Recall: 0.8348571428571429
F1 Score: 0.8346397011541737

Confusion Matrix										
True labels	0	1	2	3	4	5	6	7	8	9
	1198	1	6	13	3	63	31	1	25	2
	0	1540	8	8	1	18	3	1	17	4
	17	25	1133	39	29	7	57	12	55	6
	8	36	67	1148	3	31	13	22	60	45
	4	7	8	0	1042	8	22	6	26	172
	31	22	9	190	38	906	25	7	17	28
	15	42	32	2	19	33	1250	0	3	0
	8	28	11	5	34	3	0	1274	30	110
	17	55	23	103	14	45	6	6	1033	55
	11	26	8	11	101	7	0	55	37	1164
Predicted labels										

Figure 3: Confusion Matrix : Naive Bayes - Bernoulli

Best Performance by GridSearch:
Best parameters: {'alpha': 0.1, 'binarize': 0.0}
Accuracy: 0.8352857142857143

MTL782: Assignment 2

KNN

Accuracy: 0.9700714285714286
Precision: 0.9702368001894589
Recall: 0.9700714285714286
F1 Score: 0.9700163750952855

Confusion Matrix										
True labels	0	1	2	3	4	5	6	7	8	9
	1336	0	3	0	0	0	2	1	1	0
	0	1592	2	0	1	1	0	3	0	1
	7	17	1323	1	4	1	5	17	3	2
	0	2	12	1384	1	8	2	7	7	10
	3	8	1	0	1251	0	2	3	1	26
	2	5	0	16	2	1232	13	0	1	2
	5	1	0	0	5	6	1379	0	0	0
	1	21	4	0	4	0	0	1458	1	14
	4	13	6	22	2	19	4	10	1267	10
9	6	5	2	11	19	0	0	17	1	1359
Predicted labels										

Figure 4: Confusion Matrix : KNN

Best Performance by GridSearch:
Best parameters: {'n_neighbors': 3, 'p': 2, 'weights': 'distance'}
Accuracy: 0.9728571428571429

MTL782: Assignment 2

Neural Networks

Accuracy: 0.9625714285714285
Precision: 0.9628620900925863
Recall: 0.9625714285714285
F1 Score: 0.9625253215986614

Confusion Matrix										
True labels	0	1	2	3	4	5	6	7	8	9
	1314	0	2	2	1	0	9	1	11	3
	0	1565	7	7	6	0	2	6	6	1
	4	5	1304	19	13	1	9	5	17	3
	2	0	13	1383	2	15	0	2	9	7
	2	0	2	0	1255	1	5	1	8	21
	3	2	2	31	1	1202	13	1	14	4
	2	0	3	1	7	5	1369	0	8	1
	3	4	13	15	9	1	1	1439	4	14
	4	6	10	24	9	4	9	6	1276	9
9	4	2	1	14	26	8	0	13	14	1338
Predicted labels										

Figure 5: Confusion Matrix : Neural Networks

Best Performance by GridSearch:

Best parameters: {'activation': 'relu', 'alpha': 0.001, 'hidden_layer_sizes': (500,)}

Accuracy: 0.9741428571428571

Grid Search Cross Validation

Model	Accuracy
Decision Tree	0.876
Random Forest	0.9676
Bernoulli NB	0.8353
KNN	0.9728
Neural Network	0.9741

Table 1: Model Comparison

Glossary

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1 Score} = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$