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

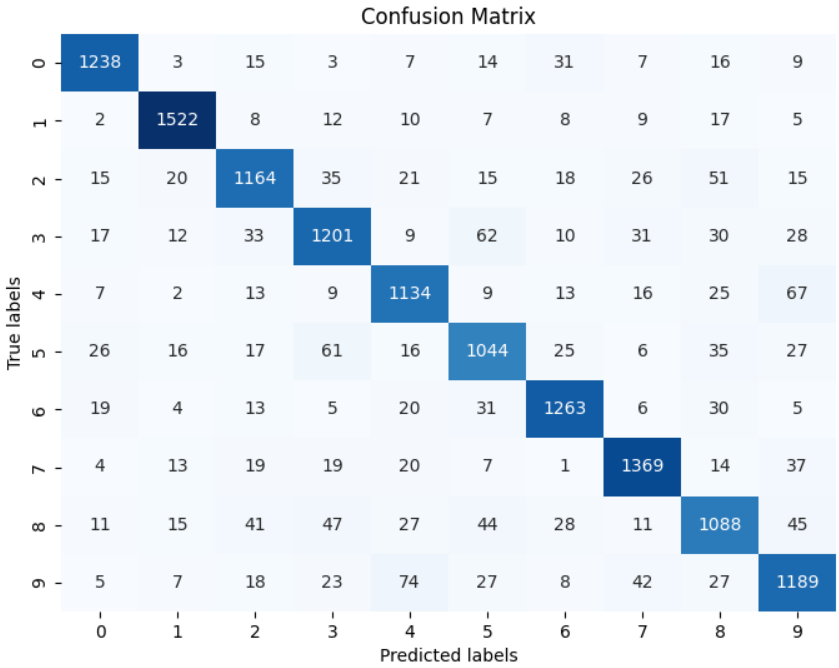

Figure 1: Confusion Matrix : Decision Tree

Best Performance by GridSearch:
Best parameters: {'max_depth': 20,'min_samples_leaf': 4, 'min_samples_split': 5}
Accuracy: 0.876

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

Random Forest

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

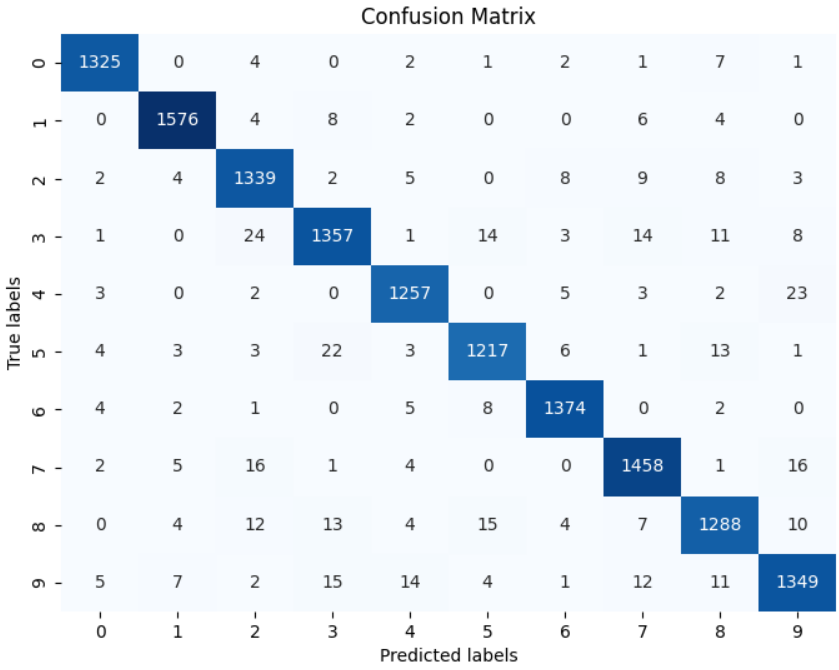

Figure 2: Confusion Matrix : Random Forest

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

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

BernoulliNB

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

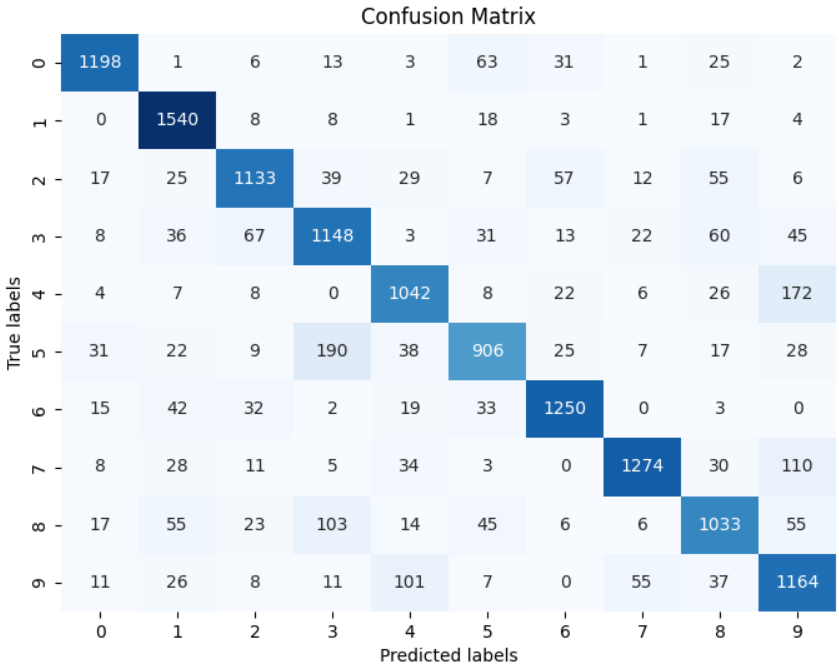

Figure 3: Confusion Matrix : Naive Bayes - Bernoulli

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

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

KNN

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

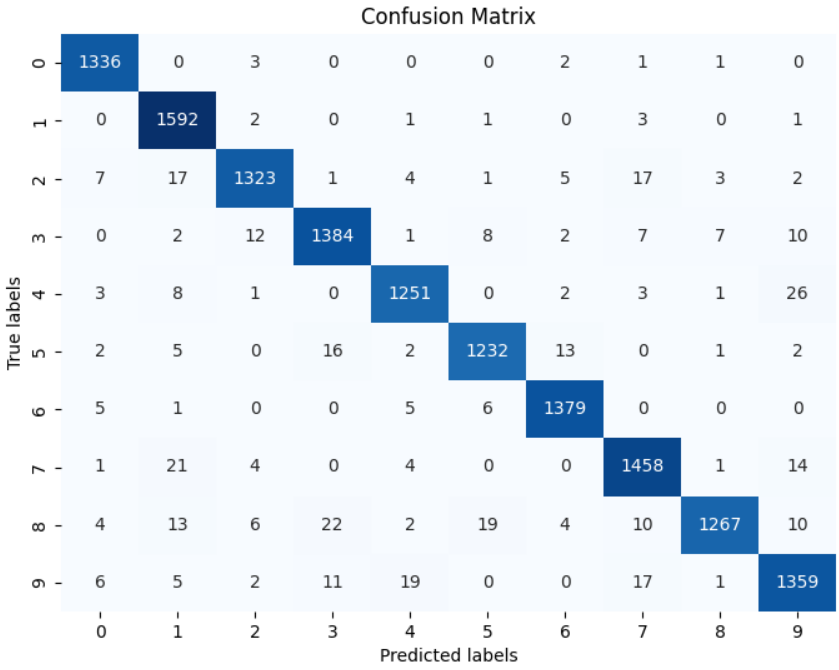

Figure 4: Confusion Matrix : KNN

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

Neural Networks

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

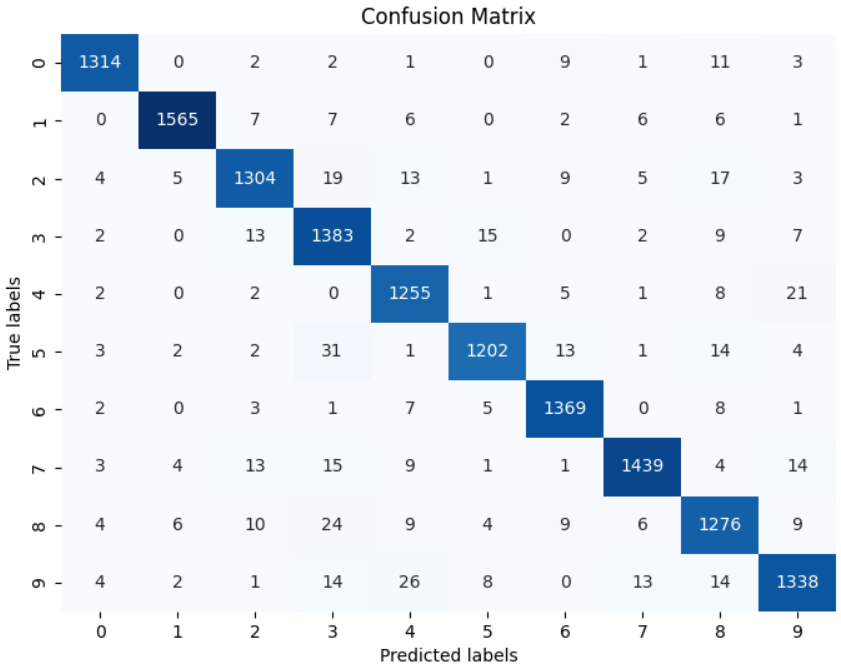

Figure 5: Confusion Matrix : Neural Networks

Best Performance by GridSearch:
Best parameters: {'activation': 'relu', 'alpha': 0.01, 'hidden_layer_sizes': (500,)}
Accuracy: 0.9741428571428571

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

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

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

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

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

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