A comprehensive report on the analysis of classification models on handwritten number data

The code we wrote creates ROC curves for four different classification models:

* Naive Bayes
* Decision Tree
* k-Nearest Neighbors
* SVM

ROC curves are generated using 5-fold cross-validation, and the AUC (area under the curve) is calculated for each class.

Here's a breakdown of the resulting code and analysis:

**1. Data Preparation:**

The dataset code downloads numbers from scikit-learn, which contains images of handwritten digits.

The label\_binarize function converts the target labels into a binary matrix, which is essential for calculating ROC curves for multi-class issues.

**2. Model Training and Evaluation:**

The code defines a dictionary of models, each containing a different classification algorithm (Naive Bayes, Decision Tree, k-Nearest Neighbors, SVM).

It performs 5-fold cross-validation to train and evaluate each model and calculate the confusion matrix, accuracy, and ROC curves.

**3. Calculating and Plotting Curve: ROC**

For each model, it calculates the ROC curve code for each class using the roc\_curve function.

The AUC for each class is then calculated using the AUC function.

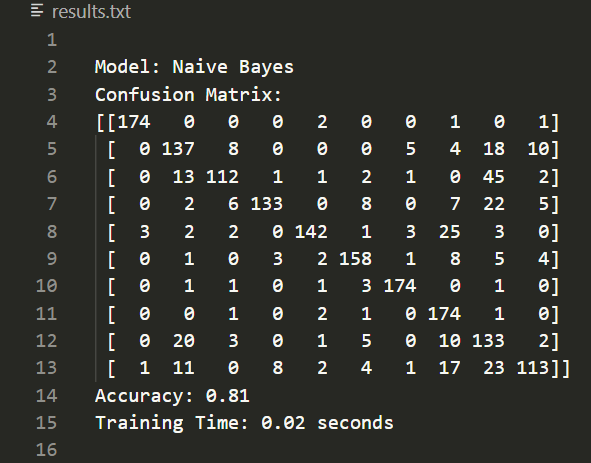
The code then plots ROC curves for all classes for each model and stores the charts in a list called "plots."

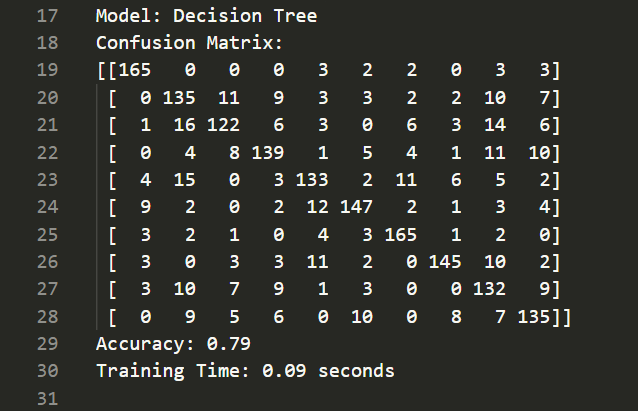
**4. Results and Analysis:**

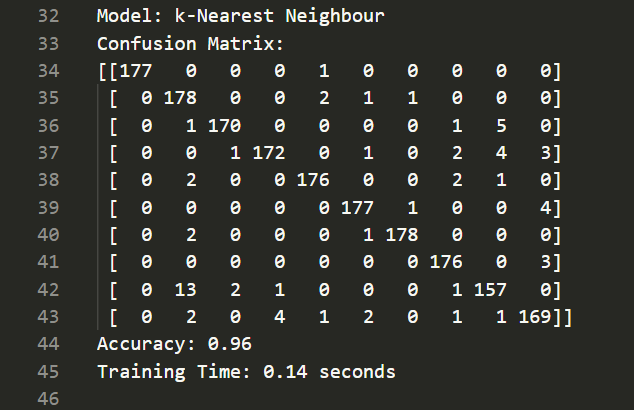
The confusion matrix code stores confusion, accuracy, training time, and AUC scores in a text file called "results.txt."

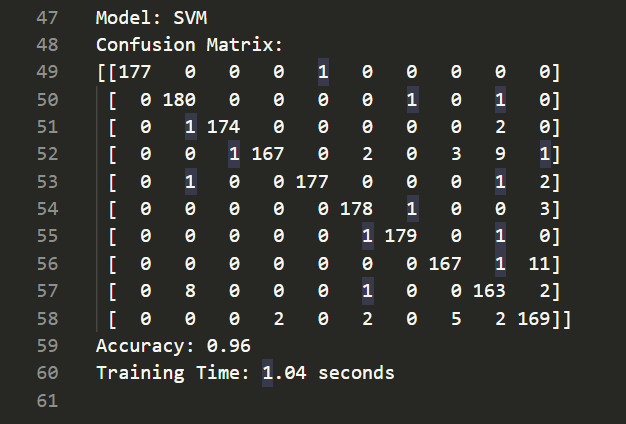
Finally, it prints the training time of each model.

:Confusion Matrix









**Analysis and review of the results obtained**

**Accuracy:**

Based on accuracy scores, both **SVM** and **k-Nearest Neighbors** achieve the highest accuracy (96%), followed by **Naive Bayes (81%)** and **Decision Tree (79%** ). This suggests that both **SVM** and **k-Nearest Neighbors** are better suited to this classification problem than the other two models.

**Training Time:**

The training time for each model varies significantly. **Naive Bayes** are the fastest, followed by **Decision Tree**, **k-Nearest Neighbors,** and **SVM** (the slowest).

**Curves: ROC**

* Each model curve shows a different classification:
* Naive Bayes
* Decision Tree
* k-Nearest Neighbors
* SVM

**Map Concepts: ROC**

* X Axis (False Positive Rate of FPR): This axis represents the proportion of false positives of misclassification.
* Y Axis (True Positive Rate TPR): This axis represents the ratio of classified positive samples (true positives) to the total number of positive samples.
* Diagonal line: The black line represents a random classification that has a 50% chance of correctly classifying a specimen.
* Curves above the line: Ideally, you want the curves to be equally high
* The ROC curves show that the models do relatively well in separating classes, with some classes having better separation than others.
* The higher the curve, the better the model's performance.
* A complete classifier has a curve that goes straight up to 1.0 on the y-axis and then straight to 1.0 on the x-axis.

.

**AUC Area Under Curve:**

* Amount AUC calculated for each model and in Confusion Matrix Shown.
* The AUC values range from 0 to 1, with higher values indicating better model performance.
* AUC 1 represents a complete classifier.
* AUC 0.5 represents a random classification.

**Curve Analysis: ROC**

* **Naive Bayes: The**  Naive Bayes model performs the best among the models shown. Its ROC curves are very close to the top left corner, indicating high TPR and low FPR across all classes. This shows a good distinction between positive and negative examples.
* **Decision Tree: The** decision tree model shows moderate performance. The graph shows the performance of the decision tree model in detecting different classes. The ROC diagram shows that the tree decision model has different functions in detecting different classes and recognizes some classes better than others
* **Nearest Neighbors: The** model performs well, and the curves are slightly better than the Decision Tree model. This shows the ability to distinguish better from decision trees.
* **The SVM** model SVM performs similarly to the k-Nearest Neighbors model, which shows good class discrimination.

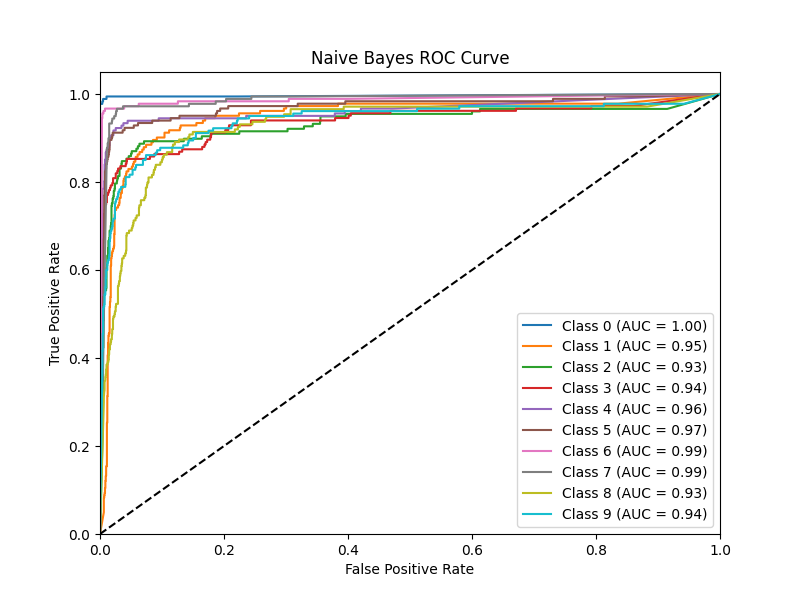
**In general:**

Based on ROC curves and AUC values, the Naive Bayes model performs better than other models in this particular dataset. However, it is important to note that model performance can vary depending on the dataset and the specific problem being addressed.

The provided text file supports these findings with additional information such as confusion matrices, accuracy scores, and further training time. This indicates that Naive Bayes has the highest accuracy and fastest practice time, making it a desirable option in this scenario.

**Results and Analysis of Charts:**

* **Naive Bayes**



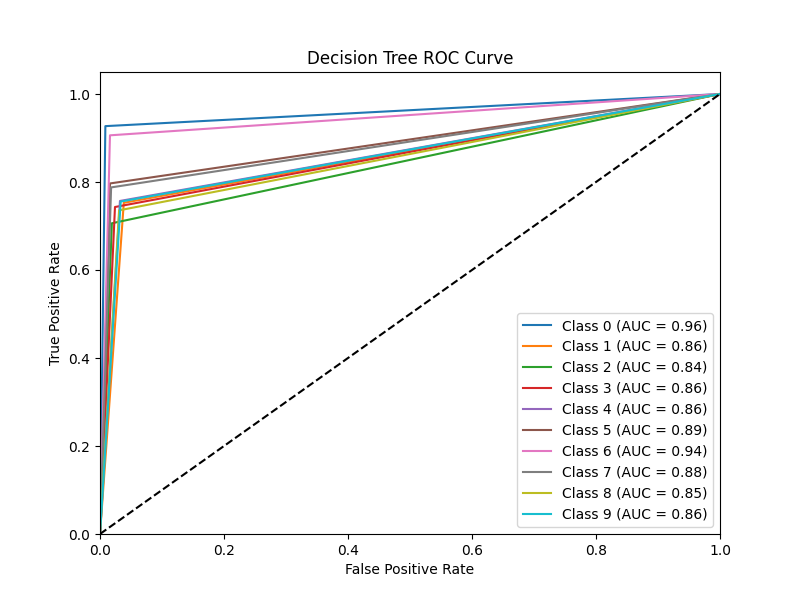
The ROC plan is presented that is drawn for different classes. In this assessment, a false positive evaluation (FPR) versus a true class evaluation (TPR) is drawn for each class.

Based on the design, it can be seen that Class 6 and Class 7 have the best performance, as their curves are closer to the top-left corner of the image. This indicates that these classes have a high real value and a false positive low.

Classes 0, 1, 2, 3, 4, 5, 8, and 9 have moderate performance because their curves are in the middle of their placement. This indicates that these classes have a moderate true value and a false positive.

Finally, it can be seen that the Naive Bayes model performs best, as the curves of all classes in this model are closer to the top-left corner of the images.

* **Decision Tree**



The image shows the ROC curves for a multi-tiered classification problem using the decision tree model. The graph shows the true positive versus false positive rate for each class. The area under each curve (AUC) is also shown.

Here is a summary of the information conveyed in the plan:

* Curve: The ROC of each line represents the ROC curve for a particular class. The curve shows how well the model can distinguish between positive and negative examples for that class in different cases
* The AUC of the area under the curve (AUC) is a measure of the overall performance of the model for each class. Higher AUC indicates better performance.
* Threshold: Threshold refers to the interruption of probability
* Class performance: The diagram shows how well the model performs for each class. For example, class 0

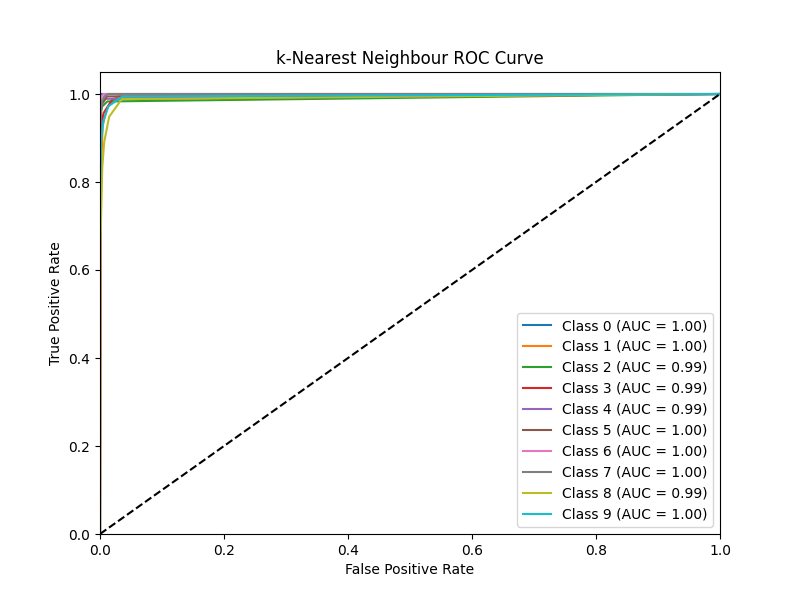
Based on the design, the model appears to perform well for all classes, although Class 0 has the highest AUC and, therefore, the best performance.

Here are some observations of the design:

* This model seems to have a good ability to distinguish positive samples from negative samples for all classes.
* This model performs especially well in Class 0, while its performance is slightly lower for other classes.
* AUC values are above 0.80 for most classes, which is generally considered good performance.

Overall, the diagram shows that the decision tree model is a viable choice for this multi-class classification problem. However, further analysis may be required to understand the reasons for slight changes in performance in different classes and to explore possible improvements.

* **K-Nearest Neighbors (KNN)**



**KNN Chart Analysis**

* Class 0 (AUC = 1.00): The curve of this class is usually close to the top-left corner, indicating the top of it.
* Class 1 (AUC = 1.00): The curve of this class usually shows near the top-left corner, giving the top of the model in the distinction between 1 and the other classes.
* Class 2 (AUC = 0.99): The curve of this class is slightly lower than the class 0 and 1, but it still represents the model's high model in differentiating between.
* Class 3 (AUC = 0.99): The curve of this class is similar to Class 2.
* Class 4 (AUC = 0.99): The curve of this class is similar to Class 2 and 3.
* Class 5 (AUC = 1.00): The curve of this classic is usually close to the top-left corner, indicating different models in the distinction between 5 and the other classes.
* Class 6 (AUC = 1.00): The curve of this class is usually indicated near the top-left corner, offering the model the top of the model in differentiation between.
* Class 7 (AUC = 1.00): The curve of this class usually indicates near the top-left corner
* Class 8 (AUC = 0.99): The curve of this class is slightly lower than Class 5, 6, and 7, but still represents the class model in differentiating between Class 8 and other classes.
* Class 9 (AUC = 1.00): The curve of this classic is usually close to the top-left corner, indicating

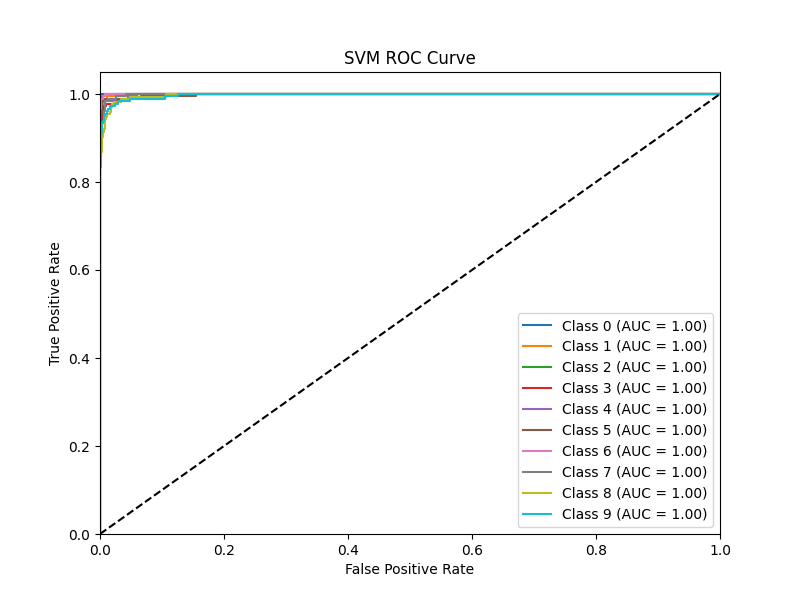
**Conclusion**

* The KNN model excels in differentiating between classes, especially for grades 0
* The KNN model has high accuracy, with an AUC of 0.99 or 1.00 for most classes.
* Only classes 2, 3, 4, and 8 have slightly lower AUC, but they still indicate

What this means is that classes 2, 3, 4, and 8 have slightly lower AUC (Area Under the Curve) value than the other classes. However, this value still reflects the model's distinction between these classes and other classes.

In other words, classes 2, 3, 4, and 8 still have lower AUC than other classes, but these values are high and represent model models in distinguishing these classes from other classes.

For example, if the AUC of a class is 0.99, it represents different models in detecting that class. But if the AUC of a class is 0.8, it represents a lower model in detecting that class. In this case, classes 2, 3, 4, and 8 have slightly lower AUCs than the other classes, but they are still high, and the models above the model are used in detecting that class..



The curve diagram shows the characteristic receiver performance (ROC) for the support vector machine (SVM) classifier. This curve is plotted for 10 different classes, with each curve representing the classifier's performance for that particular class.

The graph shows that all 10 classes have an area under the curve (AUC) of 1.00. This means that the classifier is able to fully distinguish between positive and negative examples for each class.

The x-axis represents the false positive rate, which is the proportion of negative samples that are incorrectly classified as positive. The y-axis represents the true positive rate, which is the ratio of positive samples that are correctly classified as positive.

A complete classifier has an AUC of 1.00, which means it classifies all samples correctly. The closer the AUC is to 1.00, the better the classifier's performance.

In this case, all classes have an AUC of 1.00, which indicates that the SVM classifier works great for all classes. This is likely due to the fact that the dataset is highly segregated, which means that there is a clear distinction between different classes.

Overall, this graph shows that the SVM classifier performs very well in this dataset. It can correctly classify all instances for each class and achieve excellent performance.

**In summary:**

The code provides a comprehensive analysis of four different classification models in the figures dataset.It evaluates models using accuracy, training time, and ROC curves.

The results show that the SVM and k-Nearest Neighbors models are the most accurate models, but they have a higher training time than other models.

This analysis can help you choose the best model for your specific classification problem based on your performance requirements and computational constraints.