جامعة أم القرى
UMM AL-QURA UNIVERSITY

Data Analysis 2

Naive Bayes Classifier

Dataset:

iris

Student Name	ID
Alaa Khalil Bondagji	443000351
Jana Maher Jamal	443010542

Naive Bayes

Report: Data Processing, Model Choice, Performance Evaluation, and Insights.

1. Data Processing:

The Iris dataset was used for this classification task. The dataset contains 150 samples of iris flowers, divided into three species: Setosa, Versicolor, and Virginica. Four features were used (sepal length, sepal width, petal length, and petal width) to predict the species of each flower.

Key data processing steps:

- The dataset was split into training and testing sets using an 70/30 split. This ensures that the model is trained on 70% of the data and evaluated on the remaining 30%.
- No feature scaling or normalization was required since Gaussian Naive Bayes handles continuous variables based on the Gaussian distribution.

2. Model Choice:

The Gaussian Naive Bayes classifier was chosen for this classification problem. This model is well-suited for datasets like Iris due to the assumption that features follow a normal distribution within each class. Additionally, Naive Bayes models are efficient and work well with relatively small datasets.

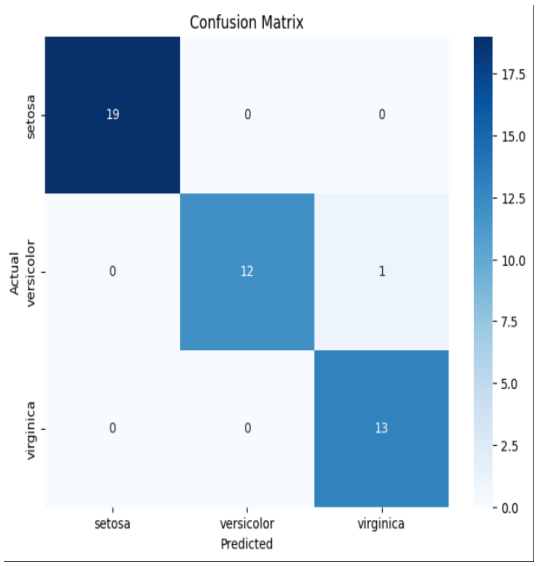
3. Performance Evaluation:

To evaluate the model's performance, a confusion matrix was used to visualize how well the model predicted each class. The confusion matrix displayed correct classifications and misclassifications for each flower species.

The model achieved high accuracy, correctly classifying most of the test set. However, a few misclassifications were observed between the Versicolor and Virginica species.

Key performance metrics include:

- **Accuracy**: The model accurately predicted most of the samples.
- **Confusion Matrix**: Most errors occurred when distinguishing between Versicolor and Virginica, which is expected due to the similarity between these two species in the feature space.



4. Insights Gained:

- **Model Strength**: The Gaussian Naive Bayes model performed well, achieving a high level of accuracy in classifying Iris flowers, particularly for the Setosa species, which is linearly separable from the other two species.
- **Misclassifications**: The few misclassifications primarily occurred between Versicolor and Virginica, suggesting some overlap in their feature distributions. This could indicate that a more complex model (such as a decision tree or SVM) might better capture the differences between these two species.
- **Feature Importance**: While feature selection wasn't explicitly performed, the confusion between Versicolor and Virginica indicates that petal-related features (length and width) are likely more informative in distinguishing between these two species compared to sepal measurements.

Overall, the Gaussian Naive Bayes model is effective for the Iris dataset, particularly in classifying the Setosa species, but struggles slightly with Versicolor and Virginica. Further tuning or the use of additional models could improve performance.