深圳大学实验报告

课程编号：**2801000049**

课程名称：机器学习

实验项目名称：**Machine Learning Task 2**

学院：电子与信息工程学院

专业：电子信息工程

指导教师：麦晓春

报告人：廖祖颐 学号：**2022110131** 班级：文华班

实验时间：**2024.5.17**

实验报告提交时间：**2024.5.17**

教务部制

1. Experimental Purposes and Requirements

1. Learn Tutorial 6-11.
2. Experiments

✓ Dataset：https://archive.ics.uci.edu/dataset/236/seeds

✓ Experiment 1: Use logistic regression to implement the classification on the Seeds dataset.

✓ Experiment 2: Use LDA to implement the classification on the Seeds dataset.

✓ Experiment 3: Use KNN to implement the classification on the Seeds dataset.

✓ Experiment 4: Use Naive Bayes classification to implement the classification on the Seeds dataset.

✓ Experiment 5: Use SVM to implement the classification on the Seeds dataset.

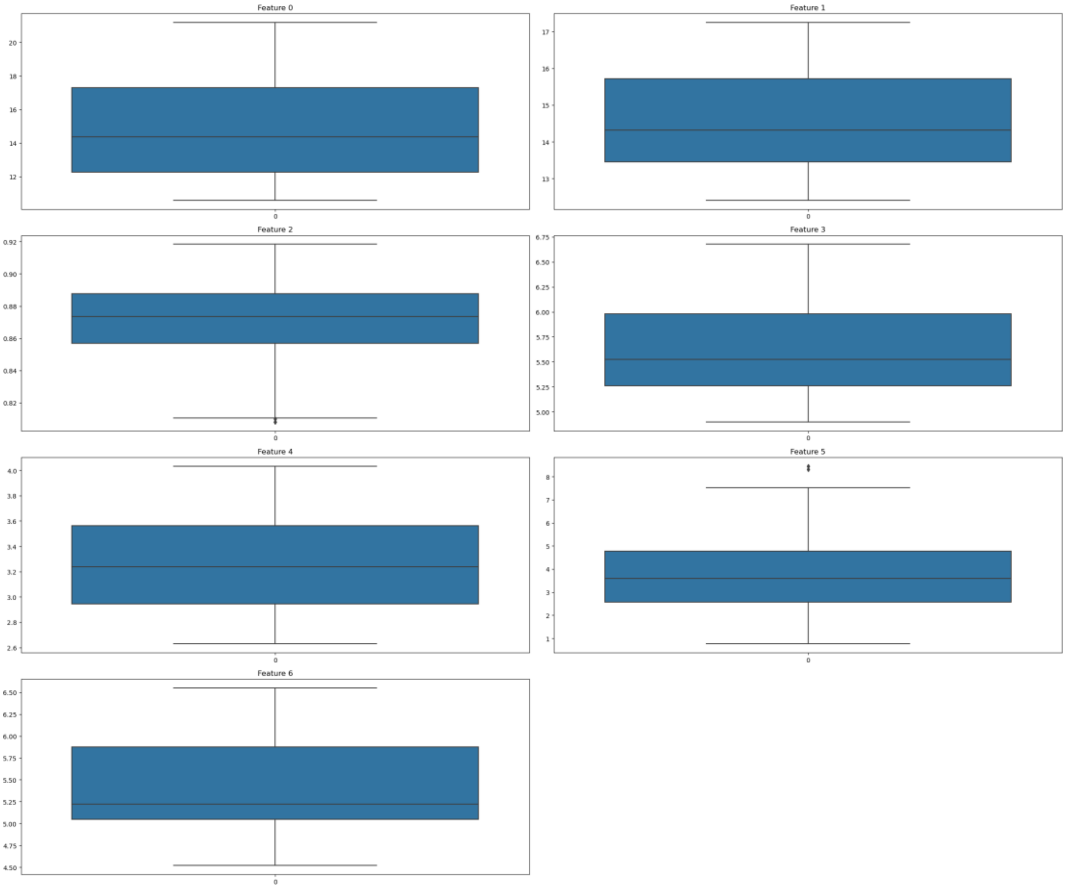
✓ Experiment 6: Use decision tree to implement the classification on the Seeds dataset.

3. In each experiment, report the classification results including quantitative results and visualization results.

2. Experiment Contents and Process

2.1 Data Preprocessing

1. Boxplot



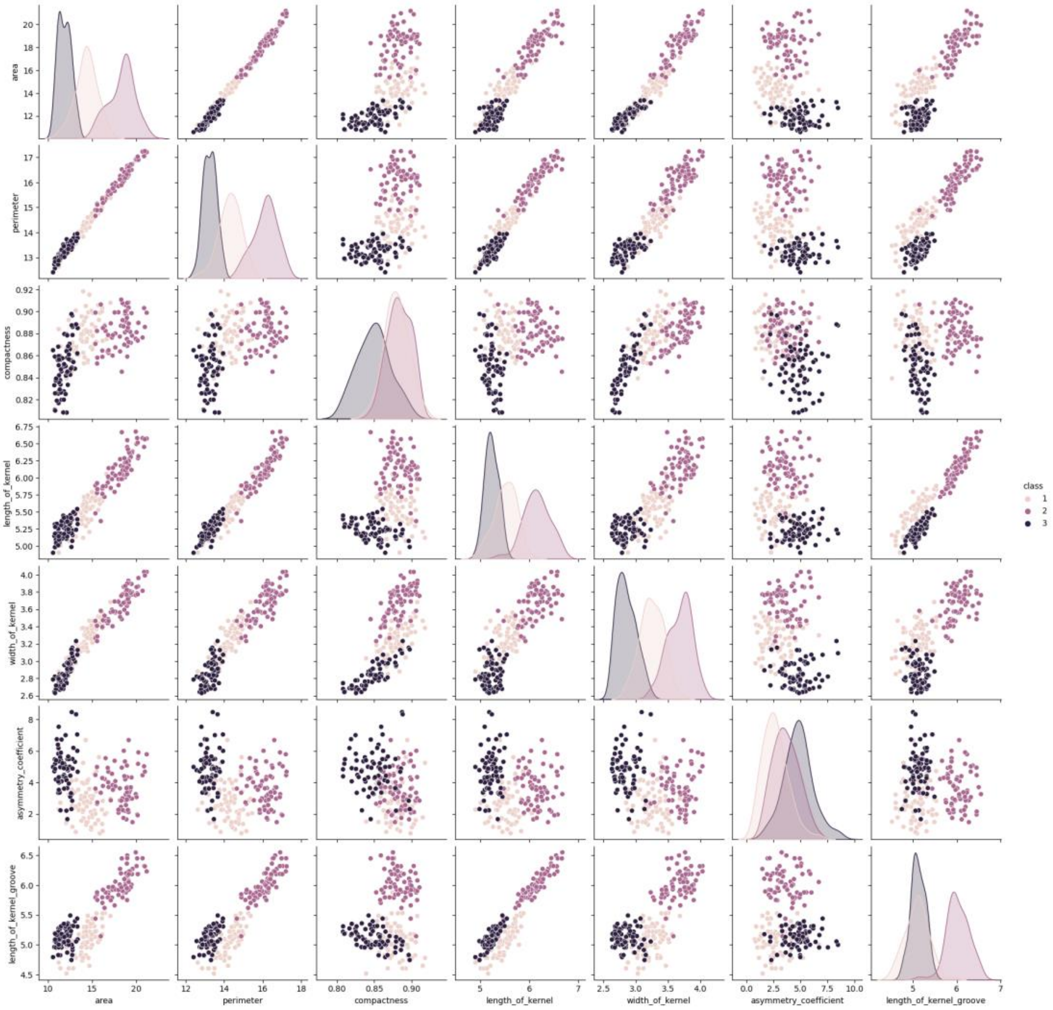
Purpose:

To identify the distribution of each feature, including the median, interquartile range (IQR), and potential outliers.

Analysis:

The boxplots generally show good symmetry and data concentration for most features, with only a few exhibiting slight skewness. Most features do not have outliers, except for Feature 5, which displays at least one low outlier. This requires further analysis to determine whether these outliers need to be addressed (not handled in this experiment). Overall, the dataset appears clean and reliable across these features, making it suitable for further data analysis and modeling.

2. Pairplot



Purpose:

To display the relationships among the features in the dataset and their associations with categories.

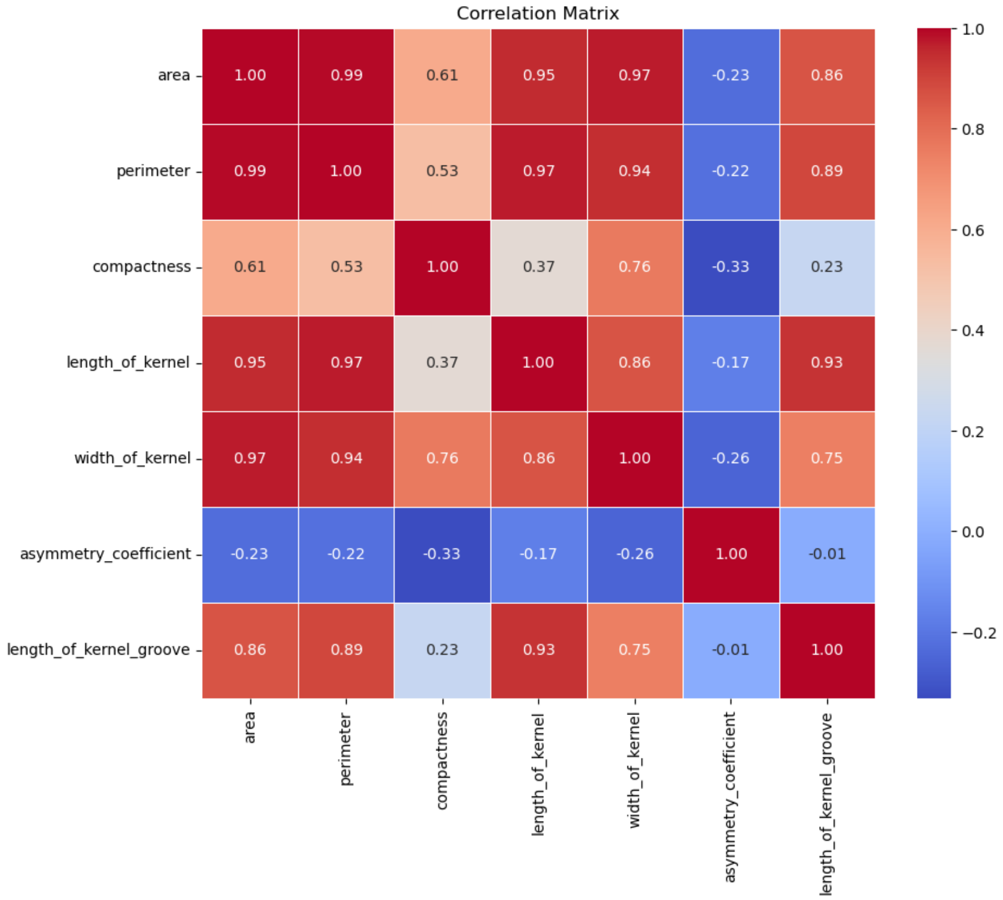
Analysis:

- The plots on the diagonal show the distribution of each feature.
- The off-diagonal plots reveal the bivariate relationships between features, which is very useful for feature selection in classification tasks.
- The area and perimeter exhibit a strong positive correlation, indicated by an almost

linear clustering of points.

- The compactness and asymmetry_coefficient show a scattered distribution of points, suggesting a weak relationship between these two features.

3. Correlation Matrix Heatmap



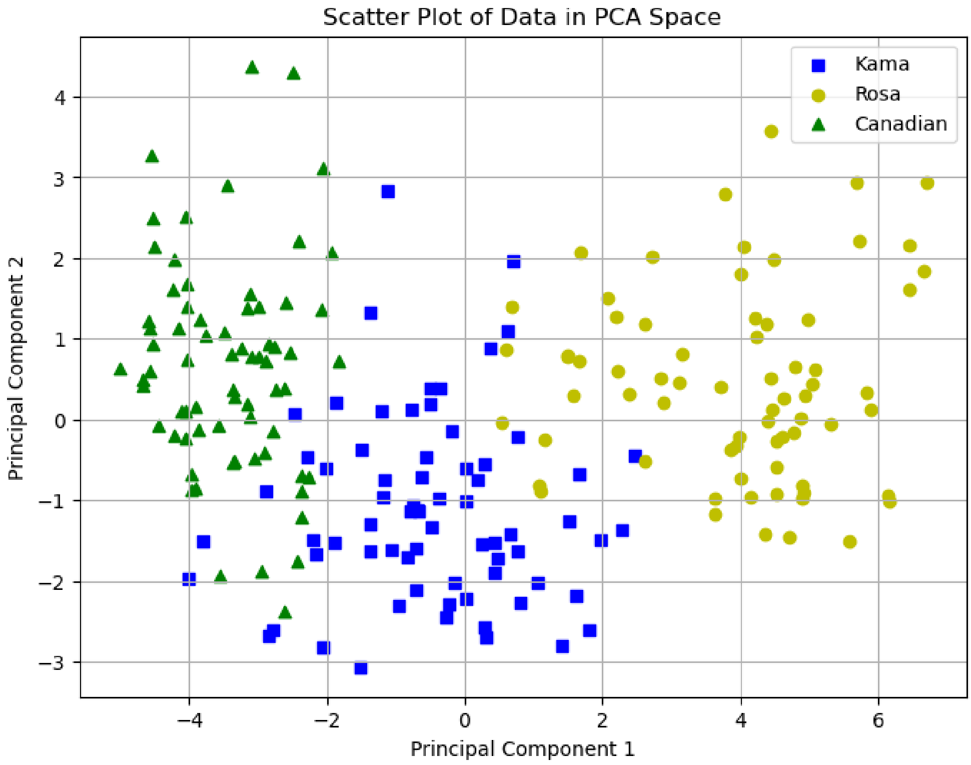
Purpose:

To analyze the strength of linear relationships between features.

Analysis:

The heatmap clearly displays the positive and negative correlations among the features, aiding in the selection of classification methods for the data. The heatmap helps in reducing the dimensions from multiple to two main components:

- Principal Component 1: All three categories are broadly covered on this axis, indicating that Principal Component 1 captures a significant dimension of variation in the data, likely related to the length_of_kernel.
- Principal Component 2: Although the distribution of all categories on this axis is not as extensive as Principal Component 1, Principal Component 2 plays a critical role in distinguishing the Canadian category from the other two. This separation suggests that Principal Component 2 is a useful dimension for differentiating the Canadian category.



2.2 Logistic Regression

1. **Data Preprocessing**

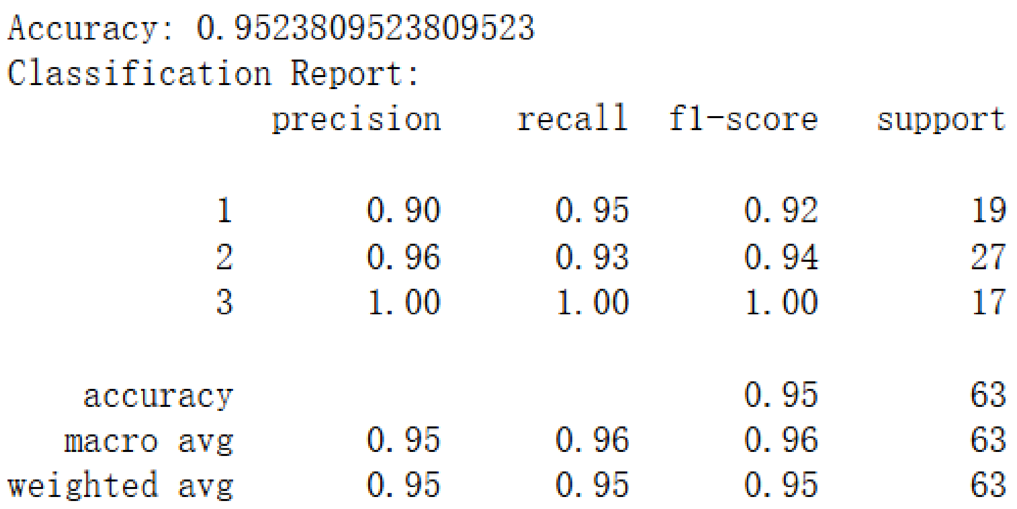
- **Feature Extraction:** Separate the feature variables and target variable from the data. Feature variables include seed area, perimeter, compactness, kernel length, kernel width, asymmetry coefficient, and kernel groove length.
- **Data Standardization:** Use StandardScaler to standardize the features to eliminate the impact of different scales of features, making the model training more effective.

2. **Model Building and Training**

In this experiment, the logistic regression model is configured to handle multi-class tasks using the 'multinomial' option and 'lbfgs' as the solver. The model training includes the following steps:

- **Model Configuration:** Set max_iter=200 to ensure the solver has sufficient iterations to converge.
- **Model Training:** Fit the model using the standardized training data.

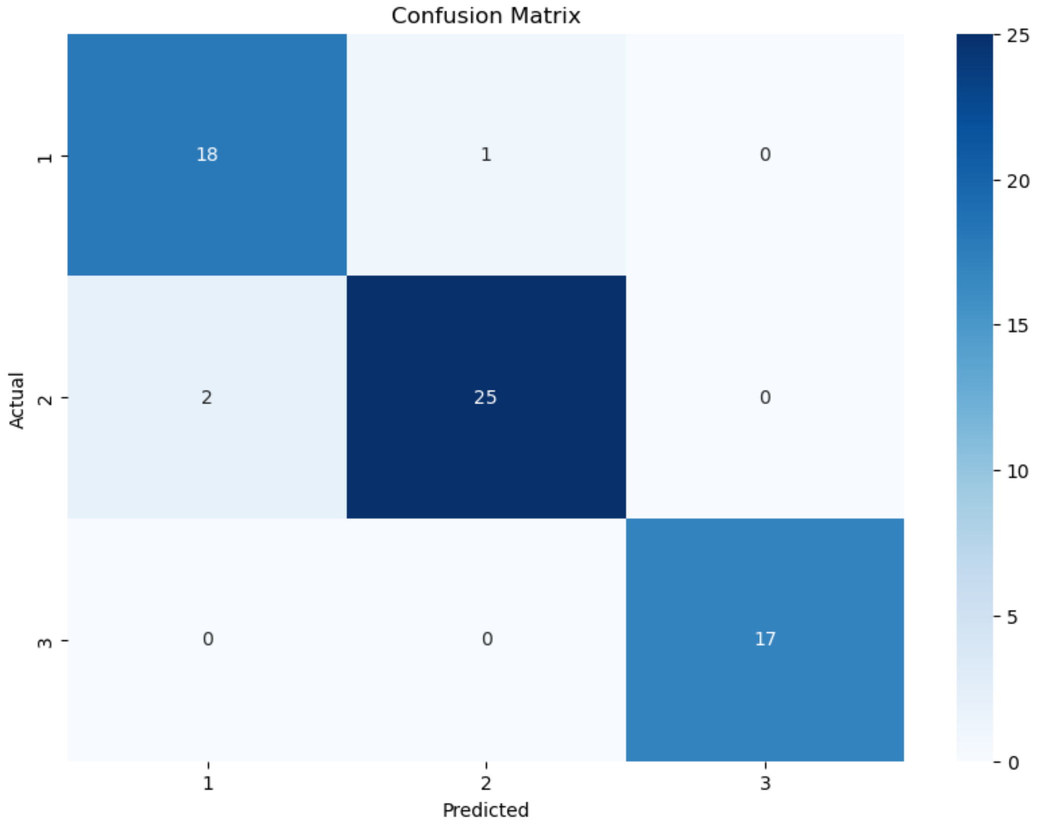
3. **Model Evaluation**



```
Accuracy: 0.9523809523809523
Classification Report:
              precision    recall  f1-score   support

           1       0.90      0.95      0.92        19
           2       0.96      0.93      0.94        27
           3       1.00      1.00      1.00        17

    accuracy                           0.95        63
   macro avg       0.95      0.96      0.96        63
weighted avg       0.95      0.95      0.95        63
```



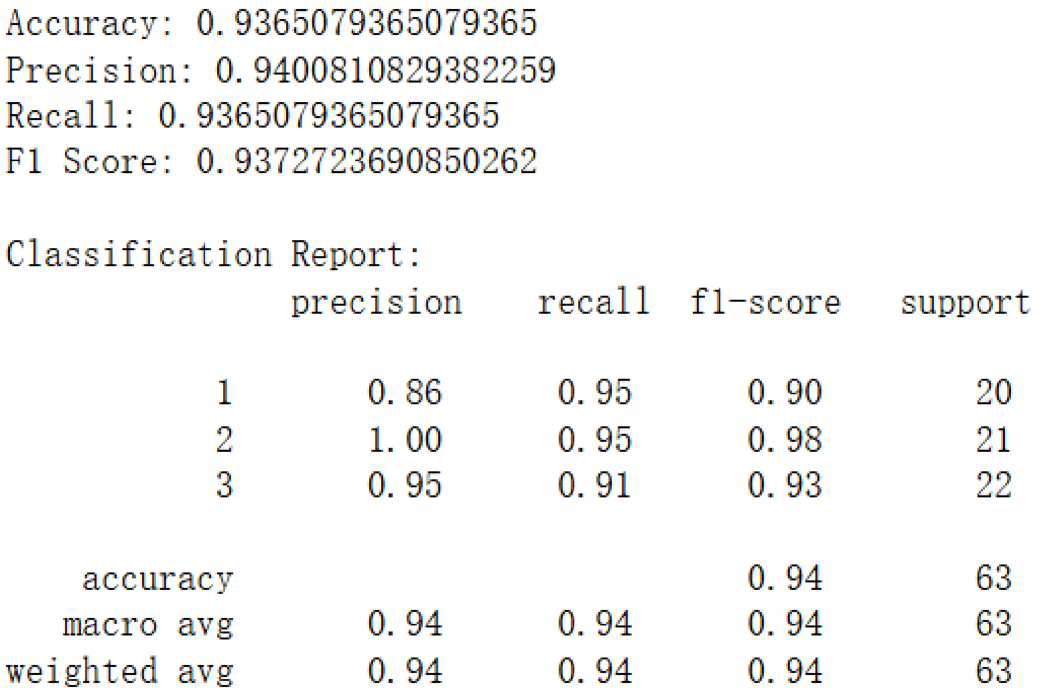
1. **Accuracy Analysis**
   The model achieved an accuracy of 95.24% on the test set, demonstrating good classification performance.
2. **Confusion Matrix**
   The confusion matrix shows that most predictions accurately fall on the diagonal, indicating correct classifications:
   - Category 1 has 18 correct classifications and 1 misclassified as Category 2.
   - Category 2 has 25 correct classifications and 2 misclassified as Category 1.
   - Category 3 perfectly achieved all correct classifications (17).
3. **Classification Report**

The classification report further details the performance for each category:

- Category 1 has a precision of 0.90, a recall of 0.95, and an F1 score of 0.92.
- Category 2 has a precision of 0.96, a recall of 0.93, and an F1 score of 0.94.
- Category 3 has a precision and a recall of 1.00, resulting in an F1 score of 1.00.

2.3 LDA

1. **Data Preprocessing**
   - No special processing was done.
2. **Model Building and Training**
   - No special processing was done.
3. **Model Evaluation**

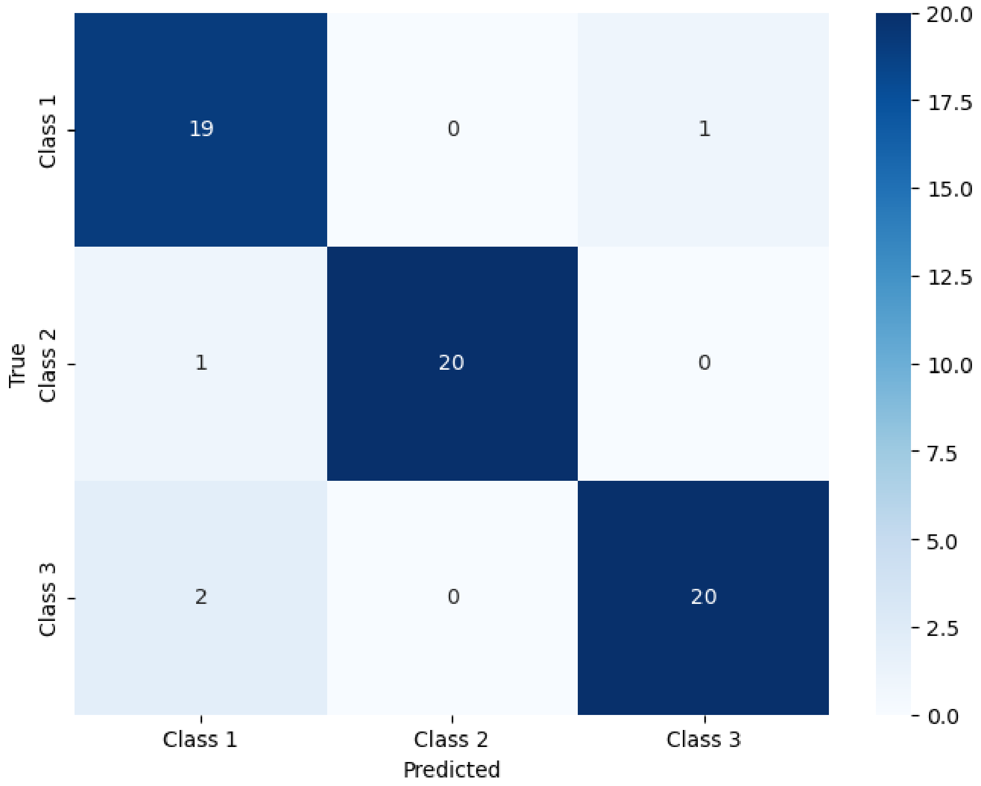


```
Accuracy: 0.9365079365079365
Precision: 0.9400810829382259
Recall: 0.9365079365079365
F1 Score: 0.9372723690850262

Classification Report:
              precision    recall  f1-score   support

           1       0.86      0.95      0.90        20
           2       1.00      0.95      0.98        21
           3       0.95      0.91      0.93        22

    accuracy                           0.94        63
   macro avg       0.94      0.94      0.94        63
weighted avg       0.94      0.94      0.94        63
```



1)Confusion Matrix

- Category 1 was correctly predicted 19 times, misclassified as Category 3 once.
- Category 2 was perfectly predicted 20 times.
- Category 3 was correctly predicted 20 times, misclassified as Category 1 twice.

2)Other Classification Metrics

- **Accuracy:** The model achieved an accuracy of 93.65% on the test set, indicating high overall predictive accuracy.
- **Precision:** The weighted average precision is 94.01%, meaning that 94.01% of the samples predicted as positive are true positives.
- **Recall:** The weighted average recall is 93.65%, indicating that 93.65% of all true positives were correctly predicted by the model.
- **F1 Score:** The weighted average F1 score is 93.73%, which is the harmonic mean of precision and recall, used to evaluate the robustness of the model.

3)Classification Report

- The classification report provides precision, recall, and F1 scores for each category, showing that Categories 2 and 3 have better predictive performance, while Category 1 has a slightly lower recall, possibly due to a few misclassifications of Category 3.

2.4 KNN

1. Data Preprocessing

- No special processing was performed.

2. Model Building and Training

- KNN (k-Nearest Neighbors) is an instance-based learning method, assuming that similar instances are located close to each other in space. It calculates the distance between the test sample and all samples in the training set, then selects the nearest K neighbors to vote and decide the final category.

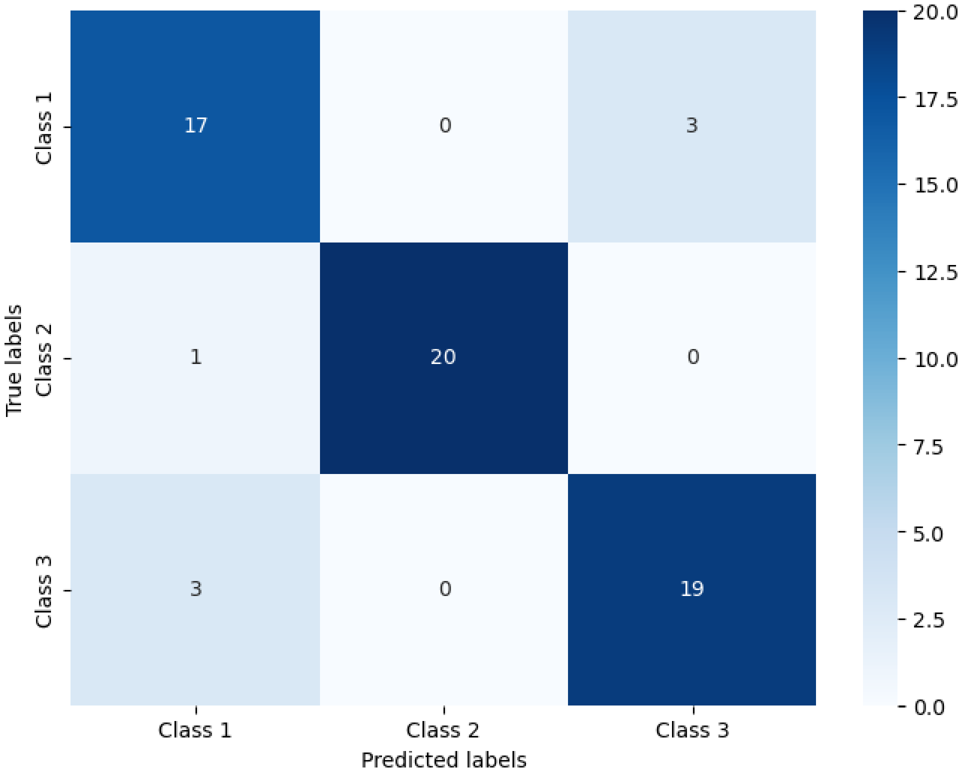
3. Model Evaluation

```
Accuracy: 0.89
Precision: 0.89
Recall: 0.89
F1 Score: 0.89

Classification Report:
              precision    recall  f1-score   support

           1       0.81      0.85      0.83        20
           2       1.00      0.95      0.98        21
           3       0.86      0.86      0.86        22

    accuracy                           0.89        63
   macro avg       0.89      0.89      0.89        63
weighted avg       0.89      0.89      0.89        63
```



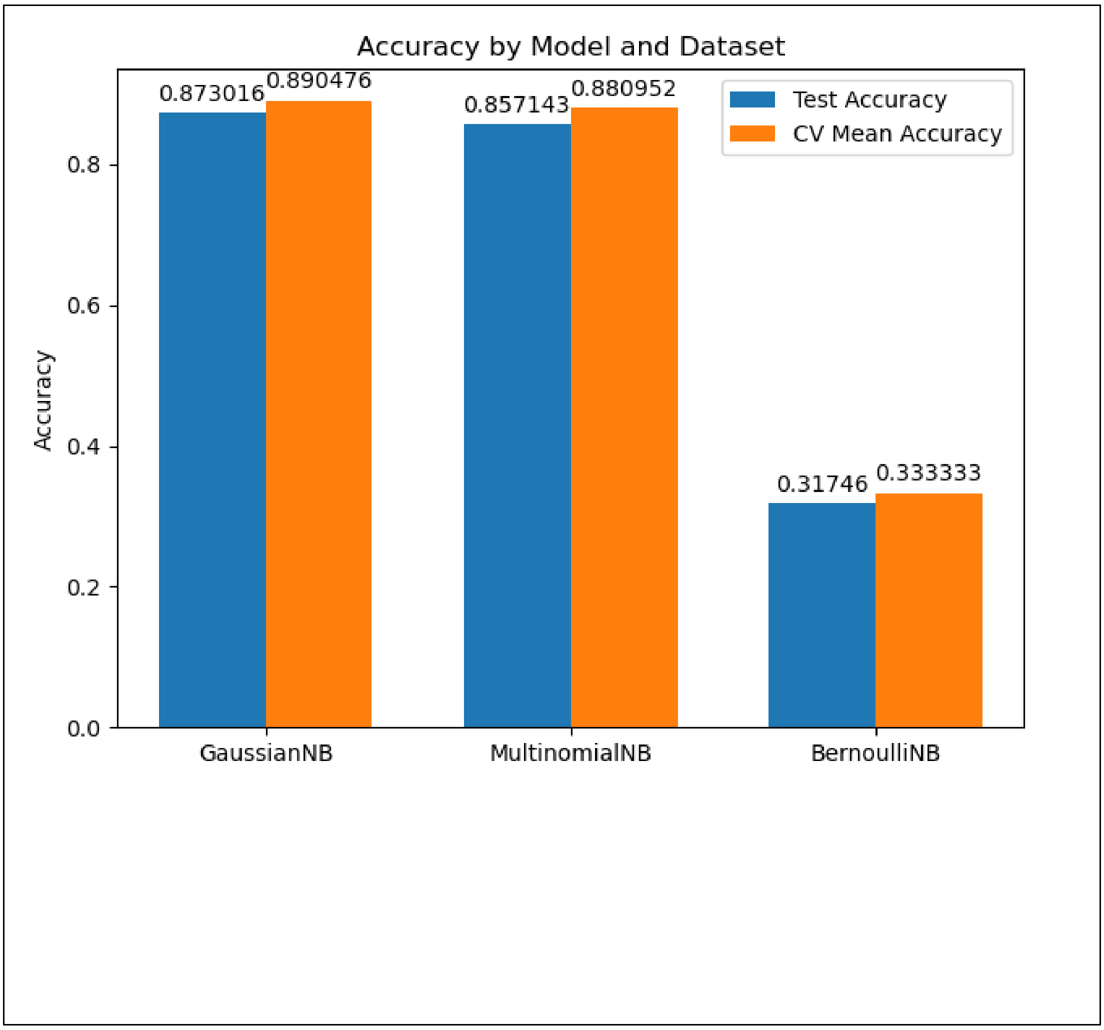
- **Confusion Matrix**

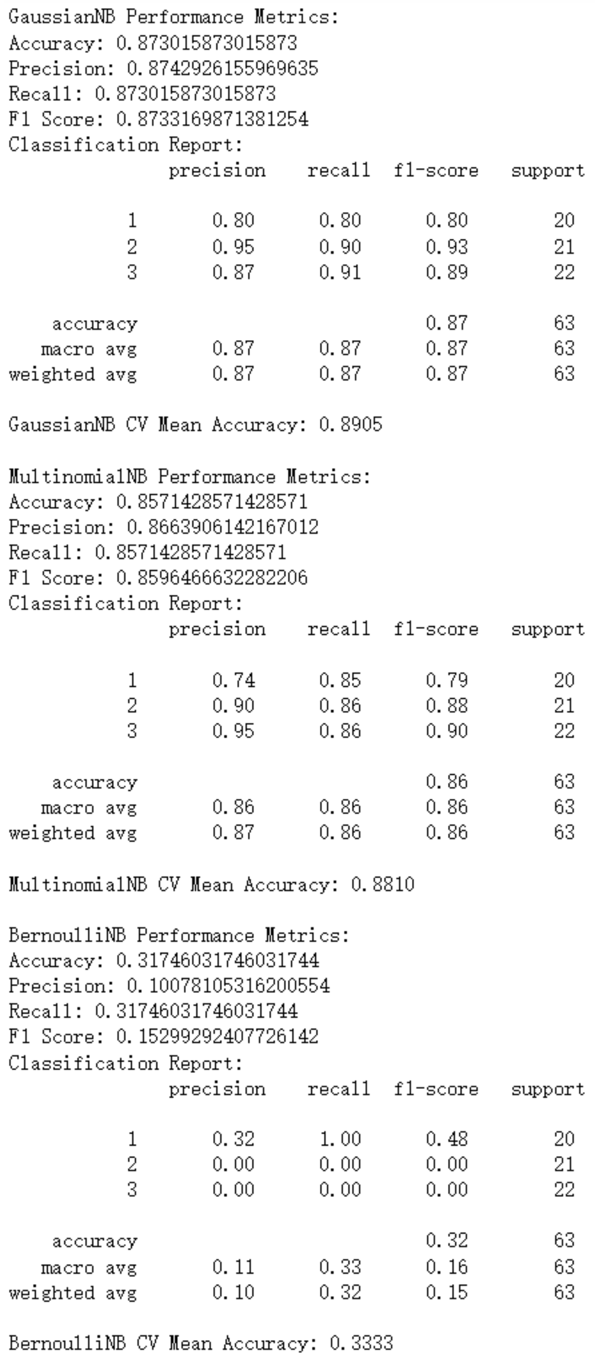
  The visualization of the confusion matrix shows the performance of the model on the test data:
  - Category 1 has 17 correct classifications, with 3 misclassifications as Category 3.
  - Category 2 has 20 correct classifications, with 1 misclassification as Category 1.
  - Category 3 has 19 correct classifications, with 3 misclassifications as Category 1.
- **Other Classification Metrics**
  - **Accuracy:** The model achieved an accuracy of 89% on the test set, indicating good predictive accuracy.
  - **Precision:** The weighted average precision is 89%, indicating that 89% of the samples identified as positive are true positives.
  - **Recall:** The weighted average recall is also 89%, showing that 89% of all true positives were correctly predicted by the model.
  - **F1 Score:** The weighted average F1 score is 89%, the harmonic mean of precision and recall, used to evaluate the robustness of the model.
- **Classification Report**

  The classification report provides precision, recall, and F1 scores for each category, showing very good predictive performance for Category 2, while the performance of Categories 1 and 3 is slightly lower, possibly due to uneven sample distribution or feature overlap.

2.5 Naive Bayes

1. Data Preprocessing
   - For the multinomial Naive Bayes classifier, data normalization was performed to suit the input requirements of the model.
2. Model Building and Training
   - In the experiment, three types of Naive Bayes models were constructed: Gaussian Naive Bayes, Multinomial Naive Bayes, and Bernoulli Naive Bayes. The Gaussian model assumes that features follow a normal distribution, the multinomial model is used for discrete features, and the Bernoulli model is suitable for binary features.
3. Model Evaluation


Accuracy by Model and Dataset
0.873016
0.890476
0.857143
0.880952
0.31746
0.333333
Test Accuracy
CV Mean Accuracy
Accuracy
0.0
0.2
0.4
0.6
0.8
GaussianNB
MultinomialNB
BernoulliNB



```
GaussianNB Performance Metrics:
Accuracy: 0.873015873015873
Precision: 0.8742926155969635
Recall: 0.873015873015873
F1 Score: 0.8733169871381254
Classification Report:
              precision    recall  f1-score   support

           1       0.80      0.80      0.80        20
           2       0.95      0.90      0.93        21
           3       0.87      0.91      0.89        22

    accuracy                           0.87        63
   macro avg       0.87      0.87      0.87        63
weighted avg       0.87      0.87      0.87        63

GaussianNB CV Mean Accuracy: 0.8905

MultinomialNB Performance Metrics:
Accuracy: 0.8571428571428571
Precision: 0.8663906142167012
Recall: 0.8571428571428571
F1 Score: 0.8596466632282206
Classification Report:
              precision    recall  f1-score   support

           1       0.74      0.85      0.79        20
           2       0.90      0.86      0.88        21
           3       0.95      0.86      0.90        22

    accuracy                           0.86        63
   macro avg       0.86      0.86      0.86        63
weighted avg       0.87      0.86      0.86        63

MultinomialNB CV Mean Accuracy: 0.8810

BernoulliNB Performance Metrics:
Accuracy: 0.31746031746031744
Precision: 0.10078105316200554
Recall: 0.31746031746031744
F1 Score: 0.15299292407726142
Classification Report:
              precision    recall  f1-score   support

           1       0.32      1.00      0.48        20
           2       0.00      0.00      0.00        21
           3       0.00      0.00      0.00        22

    accuracy                           0.32        63
   macro avg       0.11      0.33      0.16        63
weighted avg       0.10      0.32      0.15        63

BernoulliNB CV Mean Accuracy: 0.3333
```

- **Model Performance Comparison** The performance of each model on the test set was evaluated using confusion matrices, accuracy, precision, recall, and F1 scores. Additionally, the average accuracy of the models was calculated using 10-fold cross-validation to verify their stability and reliability.
  - **Gaussian Naive Bayes** showed higher performance, with an accuracy of 87.30% and an average accuracy of 89.05% in cross-validation.
  - **Multinomial Naive Bayes** also performed well, with an accuracy of 85.71% and an average accuracy of 88.10% in cross-validation.
  - **Bernoulli Naive Bayes** performed significantly worse, with an accuracy of only 31.75% and an average accuracy of 33.33% in cross-validation, which may be due to the model's unsuitability for continuous or multi-class feature

data.

- **Classification Report**
    - **Gaussian Naive Bayes** has balanced performance across all three categories, especially performing well in categories 2 and 3.
    - **Multinomial Naive Bayes** also has good predictive effects for categories 2 and 3, but has a slightly lower recall for category 1.
    - **Bernoulli Naive Bayes** is very inaccurate in predicting all categories, emphasizing the importance of selecting a model that matches the data characteristics.

2.6 SVM

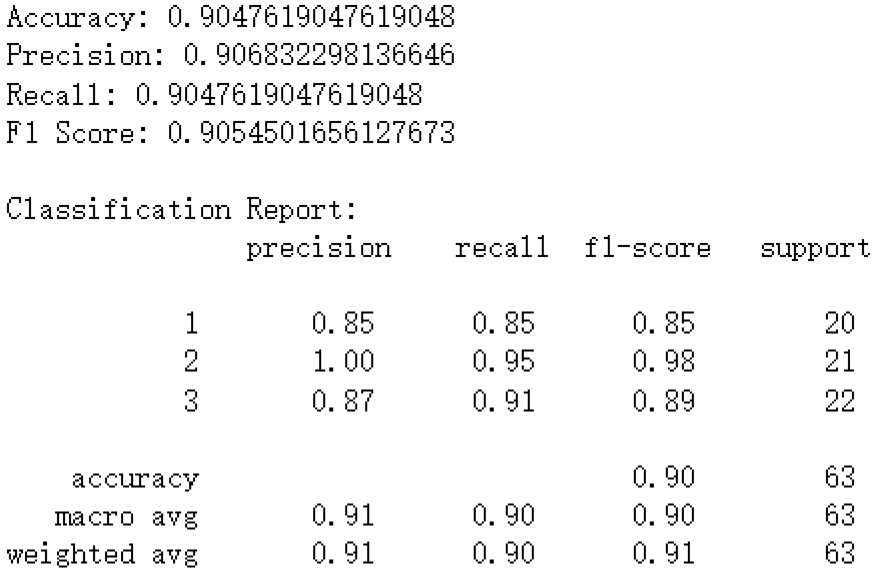
1. **Data Preprocessing**

   No special processing was performed.

2. **Model Building and Training**

   SVM (Support Vector Machine) is a powerful classifier that classifies by finding the maximum margin between different categories. In this experiment, the SVC class was used with the parameter kernel='linear' specifying the use of a linear kernel.

3. **Model Evaluation**

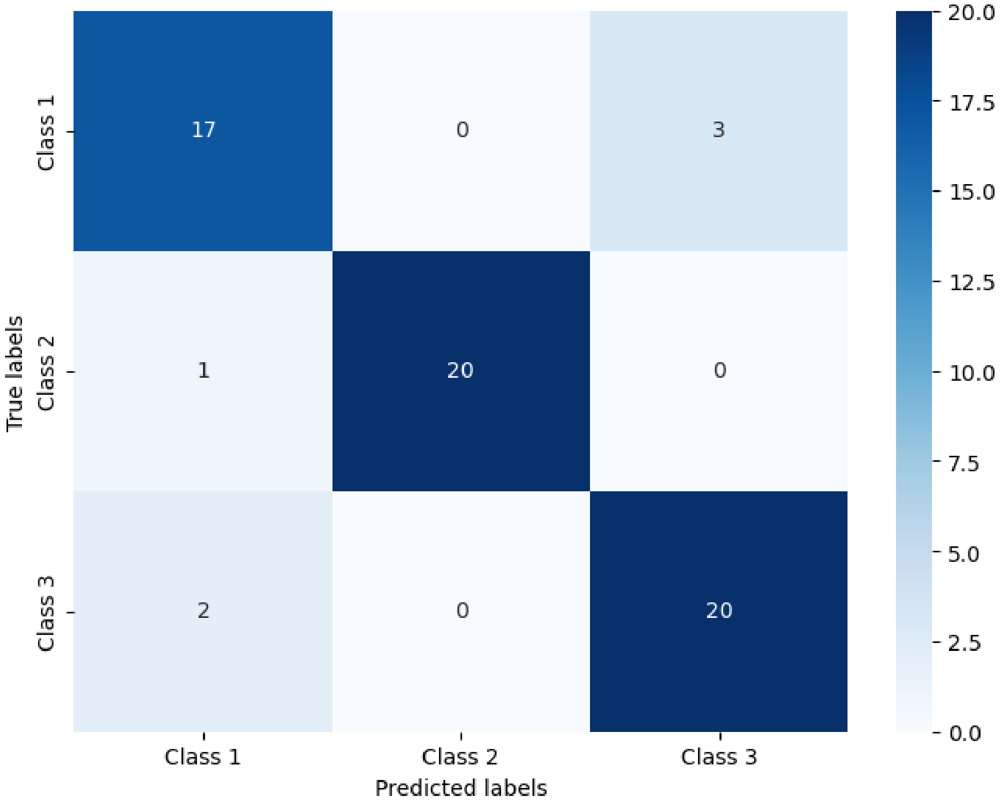


```
Accuracy: 0.9047619047619048
Precision: 0.906832298136646
Recall: 0.9047619047619048
F1 Score: 0.9054501656127673

Classification Report:
              precision    recall  f1-score   support

           1       0.85      0.85      0.85        20
           2       1.00      0.95      0.98        21
           3       0.87      0.91      0.89        22

    accuracy                           0.90        63
   macro avg       0.91      0.90      0.90        63
weighted avg       0.91      0.90      0.91        63
```



- **Confusion Matrix**

  The visualization of the confusion matrix shows the performance of the model on the test data:

  - Category 1 had 17 correct classifications and 3 misclassifications as Category 3.
  - Category 2 had 20 correct classifications and 1 misclassification as Category 1.
  - Category 3 had 20 correct classifications and 2 misclassifications as Category 1.
- **Classification Metrics**
  - **Accuracy:** The model achieved an accuracy of 90.48% on the test set, indicating high predictive accuracy.
  - **Precision:** The weighted average precision is 90.68%, meaning that 90.68% of the samples identified as positive are true positives.
  - **Recall:** The weighted average recall is 90.48%, indicating that 90.48% of all true positives were correctly predicted by the model.
  - **F1 Score:** The weighted average F1 score is 90.55%, the harmonic mean of precision and recall, used to evaluate the robustness of the model**.**
- **Classification Report**

  The classification report provides precision, recall, and F1 scores for each category, showing very good predictive performance for Category 2, while the performance of Categories 1 and 3 is slightly lower, but still maintains a high level.

2.6 Decision Tree

1. Data Preprocessing

   No special processing was performed.

2. Model Building and Training

   Decision trees predict the value of a target variable by creating a tree-like structure of decision rules.

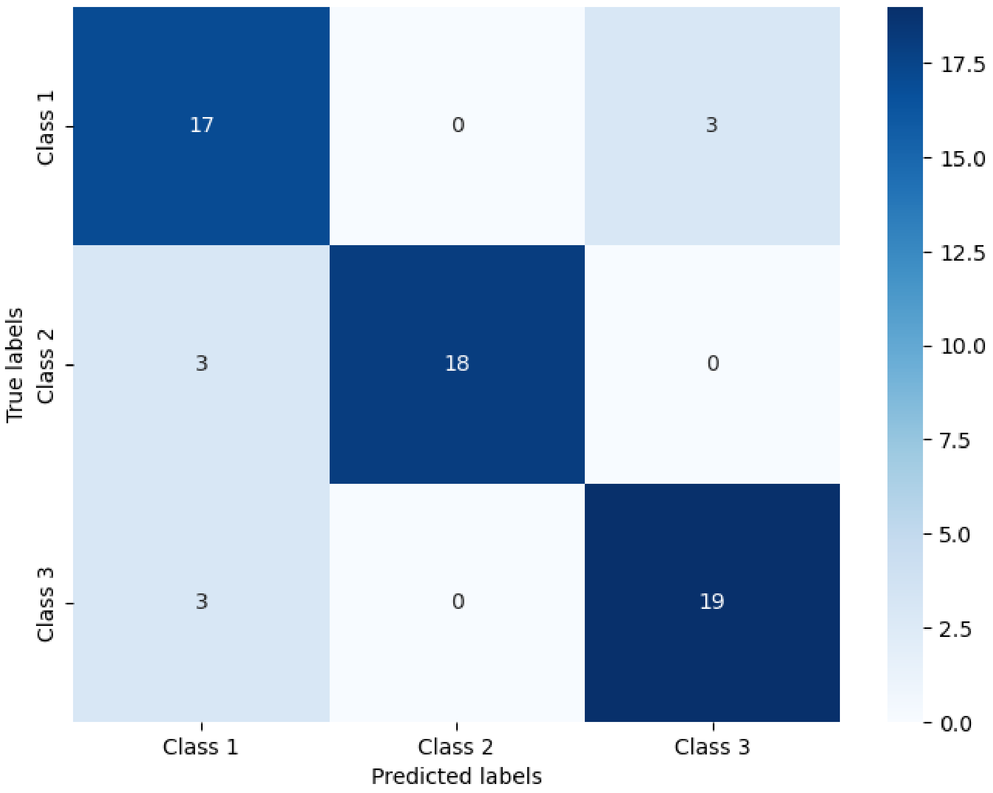
3. Model Evaluation

```
Accuracy: 0.8571428571428571
Precision: 0.8695652173913043
Recall: 0.8571428571428571
F1 Score: 0.8602947440156743

Classification Report:
              precision    recall  f1-score   support

           1       0.74      0.85      0.79        20
           2       1.00      0.86      0.92        21
           3       0.86      0.86      0.86        22

    accuracy                           0.86        63
   macro avg       0.87      0.86      0.86        63
weighted avg       0.87      0.86      0.86        63
```



- **Confusion Matrix**

The visualization of the confusion matrix shows the performance of the model on the test data:

  - Category 1 had 17 correct classifications, with 3 misclassifications as Category 3.
  - Category 2 had 18 correct classifications, with 3 misclassifications as Category 1.
  - Category 3 had 19 correct classifications, with 3 misclassifications as Category 1.

- **Classification Metrics**
  - **Accuracy:** The model achieved an accuracy of 85.71% on the test set, indicating good predictive accuracy.
  - **Precision:** The weighted average precision is 86.96%, meaning that 86.96% of the samples identified as positive are true positives.
  - **Recall:** The weighted average recall is 85.71%, indicating that 85.71% of all true positives were correctly predicted by the model.
  - **F1 Score:** The weighted average F1 score is 86.03%, the harmonic mean of precision and recall, used to evaluate the robustness of the model.
- **Classification Report** The classification report provides precision, recall, and F1 scores for each category, showing very good predictive performance for Category 2, while the performance of Categories 1 and 3 is slightly lower, which may be due to uneven sample distribution or feature overlap.

3. Discussion and Conclusions

1. **Logistic Regression :**Overall, the logistic regression model performed excellently in the seed classification task, especially for Category 3, achieving perfect accuracy and recall rates. Although there were a few misclassifications in Categories 1 and 2, the overall precision, recall, and F1 scores were still very high. These misclassifications could be caused by feature overlap or sample imbalance in the data.
2. **LDA :**The LDA model showed good performance in the seed classification task, particularly in predicting Categories 2 and 3 with high precision and recall. The slightly lower performance in Category 1 may require further investigation and improvement, such as adjusting the balance of the training dataset or experimenting with other feature selection techniques.
3. **KNN :**The K-nearest neighbors model performed well in the seed classification task, particularly in predicting Category 2 with high precision and recall. For misclassifications in Categories 1 and 3, adjusting the number of neighbors or considering the introduction of weights to improve classification accuracy might be necessary.
4. **Naive Bayes :**The experimental results indicate that choosing the appropriate Naive Bayes classifier model is crucial for specific dataset characteristics. Gaussian Naive Bayes and Multinomial Naive Bayes demonstrated good classification capabilities for continuous data, while Bernoulli Naive Bayes performed poorly on this dataset due to the limitations of the model's assumptions.
5. **SVM :**The Support Vector Machine model excelled in the seed classification task, particularly in predicting Category 2 with high precision and recall. For the few misclassifications in Categories 1 and 3, further adjustments to the model parameters or the use of different kernel functions might be needed to optimize performance.
6. **Decision Tree**: The decision tree model showed good performance in the seed classification task, particularly in predicting Category 2 with high precision and recall. For misclassifications in Categories 1 and 3, adjusting the depth of the decision tree or considering the introduction of pruning techniques to prevent overfitting might be necessary.

指导教师批阅意见： 成绩评定： 指导教师签字： 年　　月　　日
备注：

注：1、报告内的项目或内容设置，可根据实际情况加以调整和补充。

2、教师批改学生实验报告时间应在学生提交实验报告时间后 10 日内。