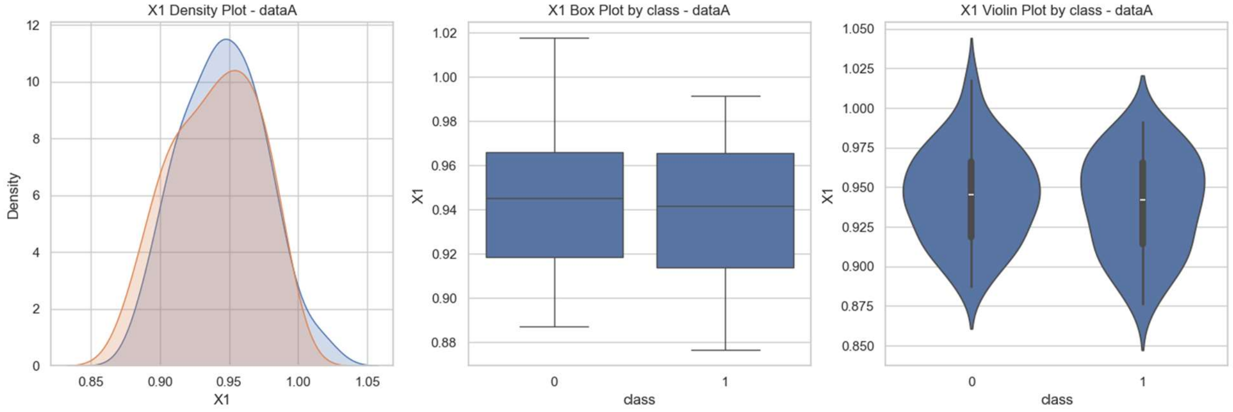
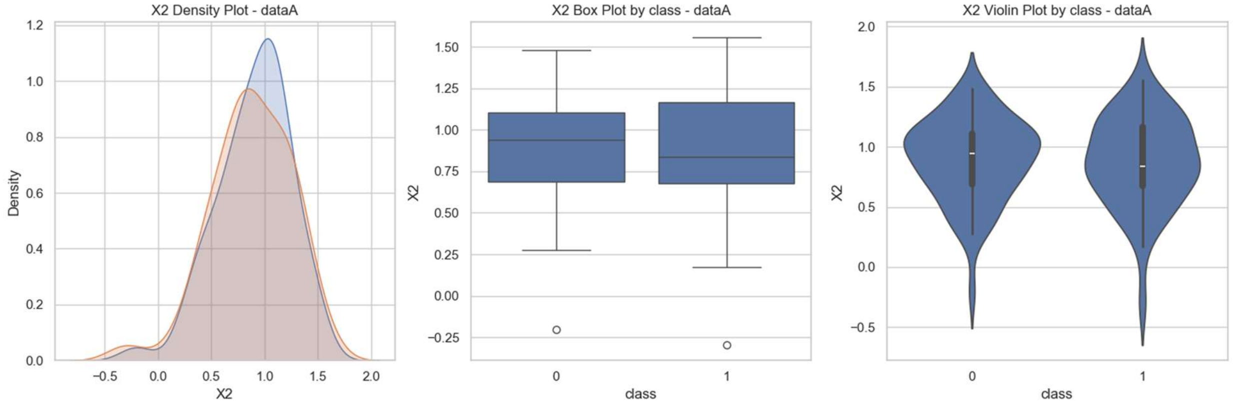
1: Data visualizations generated on step 1 with an explanation of what the visualizations show

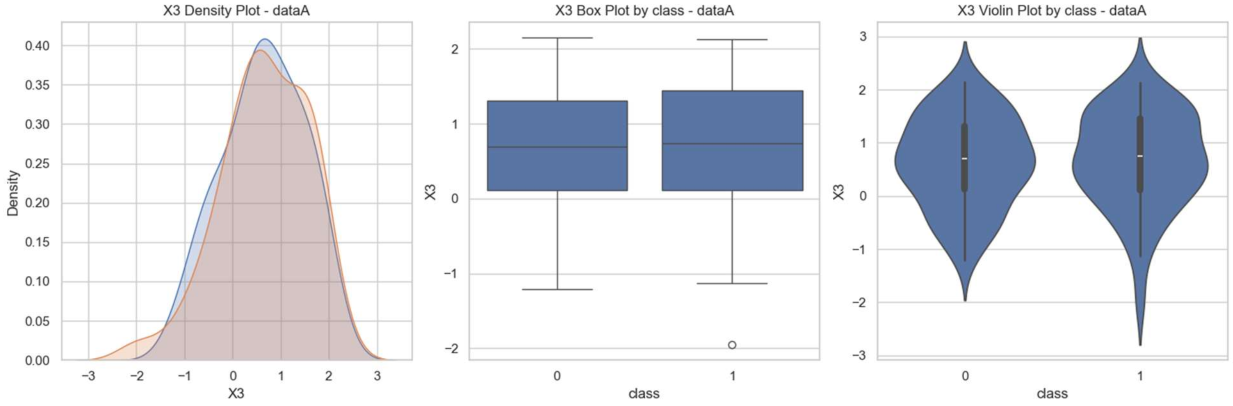
X1 – Dataset A



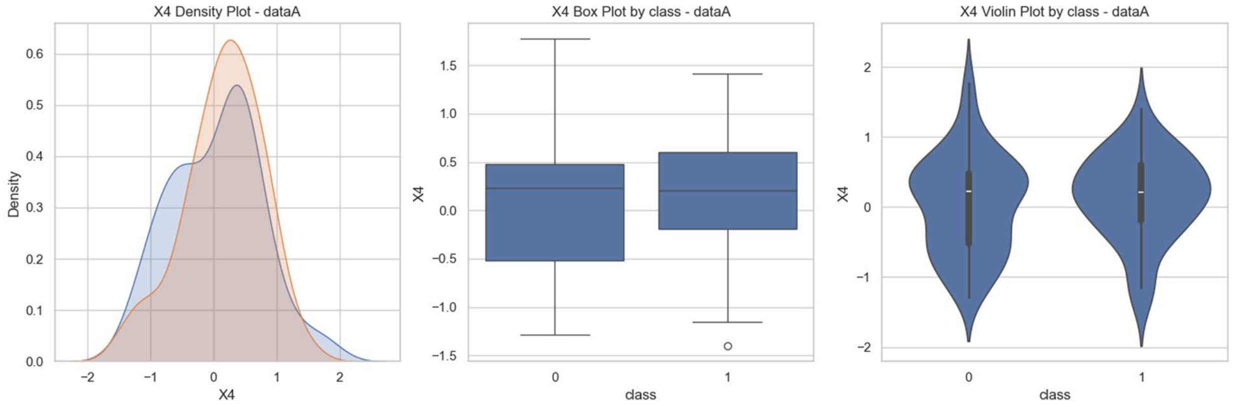
X2 – Dataset A



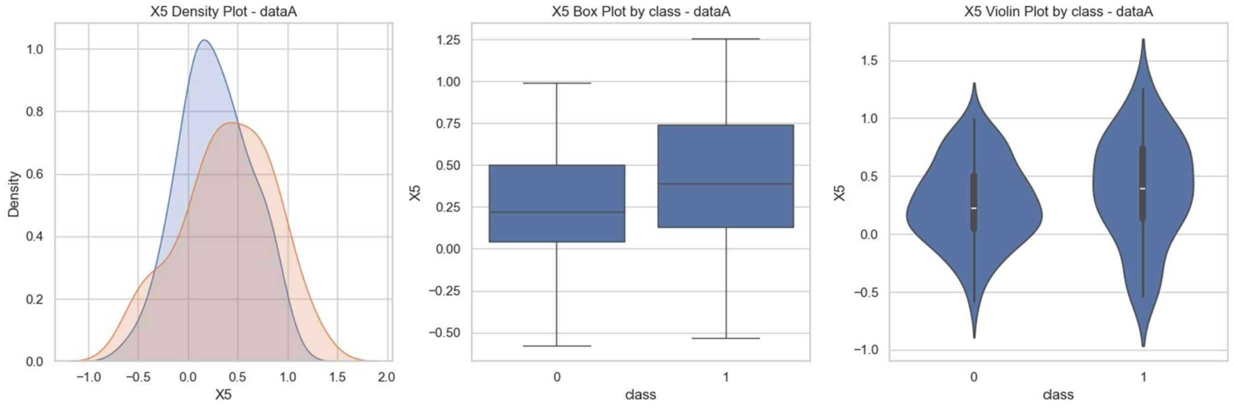
X3 – Dataset



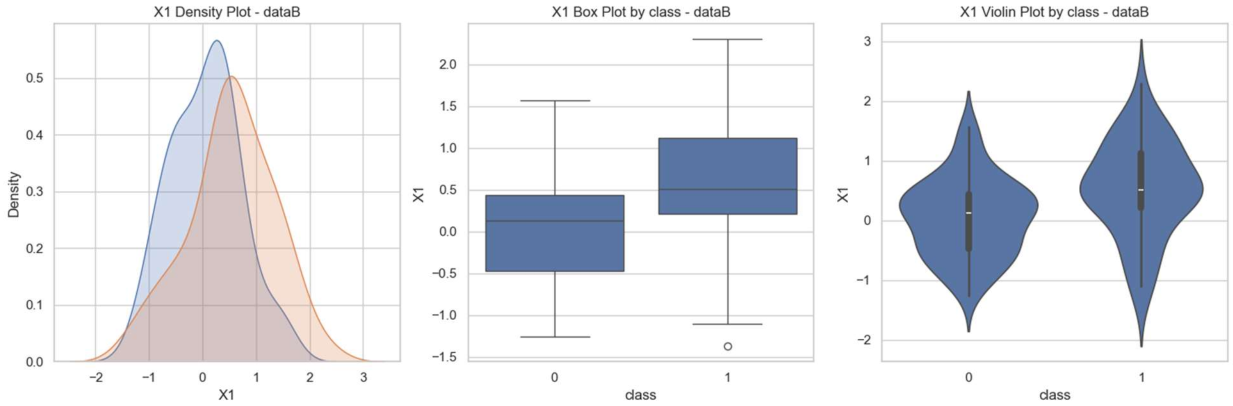
X4 - Dataset A



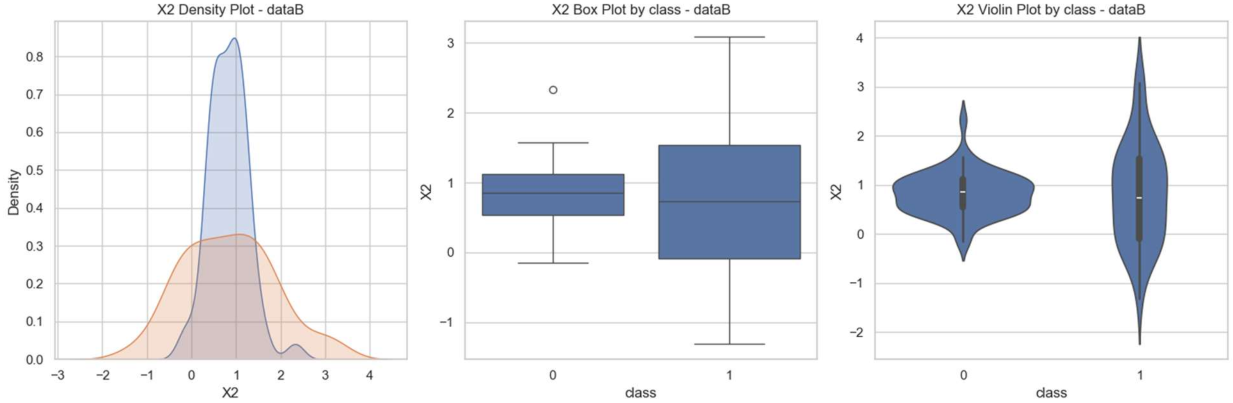
X5 – Dataset A



X1 - Dataset B



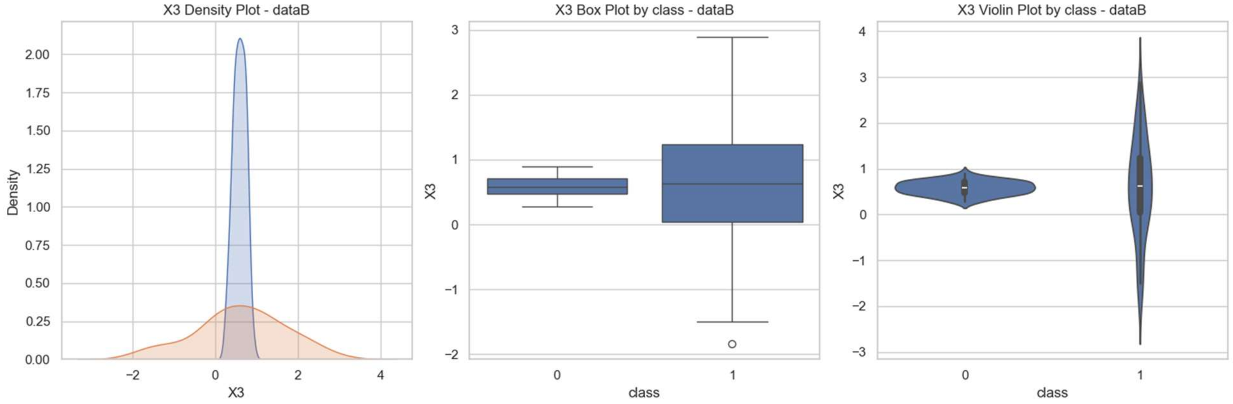
X2 – Dataset B


X2 Density Plot - dataB

X2 Box Plot by class - dataB

X2 Violin Plot by class - dataB

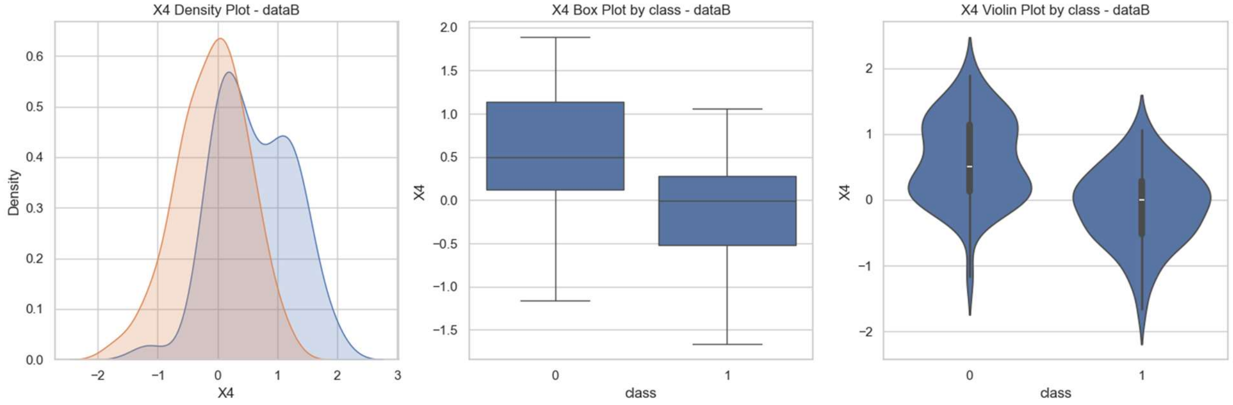
X3 – Dataset B


X3 Density Plot - dataB

X3 Box Plot by class - dataB

X3 Violin Plot by class - dataB

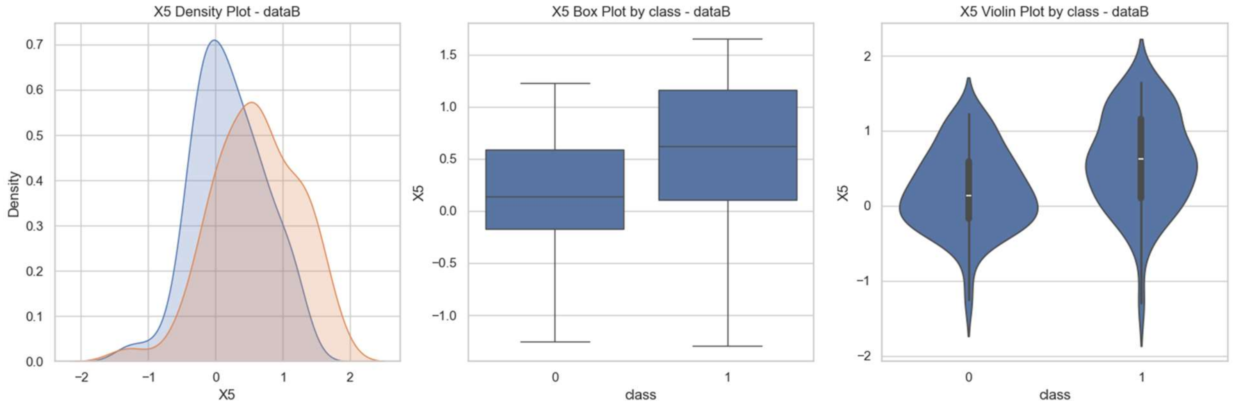
X4 – Dataset B


X4 Density Plot - dataB

X4 Box Plot by class - dataB

X4 Violin Plot by class - dataB

X5 – Dataset B



- Density Plots:
  - The density plots provide insights into the distribution of a continuous variable (in this case, X1 to X5).
  - The plots show how frequently different values of X variable occur within each class.
  - The blue curve represents class 0, and the orange curve represent class 1.
  - The peak of each curve indicates the most common value of X variable for that class.
  - The y-axis represents the density, which is a measure of how densely the data points are distributed along the x-axis.
  - In summary, the density plot helps us understand where the data tends to cluster for each class.
- Box Plots:
  - The box plots summarize the distribution of a variable.
  - For each class, the box plots display:
    - The median (the line inside the box).
    - The interquartile range (IQR), which spans from the 25th percentile (lower quartile) to the 75th percentile (upper quartile).
    - Any outliers (individual data points beyond the whiskers).
  - The x-axis represents the class (with values 0 or 1), and the y-axis represents the X1 values.
  - The boxes provide information about the central tendency and spread of the data.
- Violin Plot:
  - A violin plot combines features of a density plot and a box plot.
  - It shows the distribution of data points for each class.
  - The width of the violin at any given point represents the density of data at that value.
  - The white dot inside each violin marks the median.
  - The violin plot provides a visual representation of the data’s shape and spread.
  - In summary, it complements the box plot by showing the entire distribution.

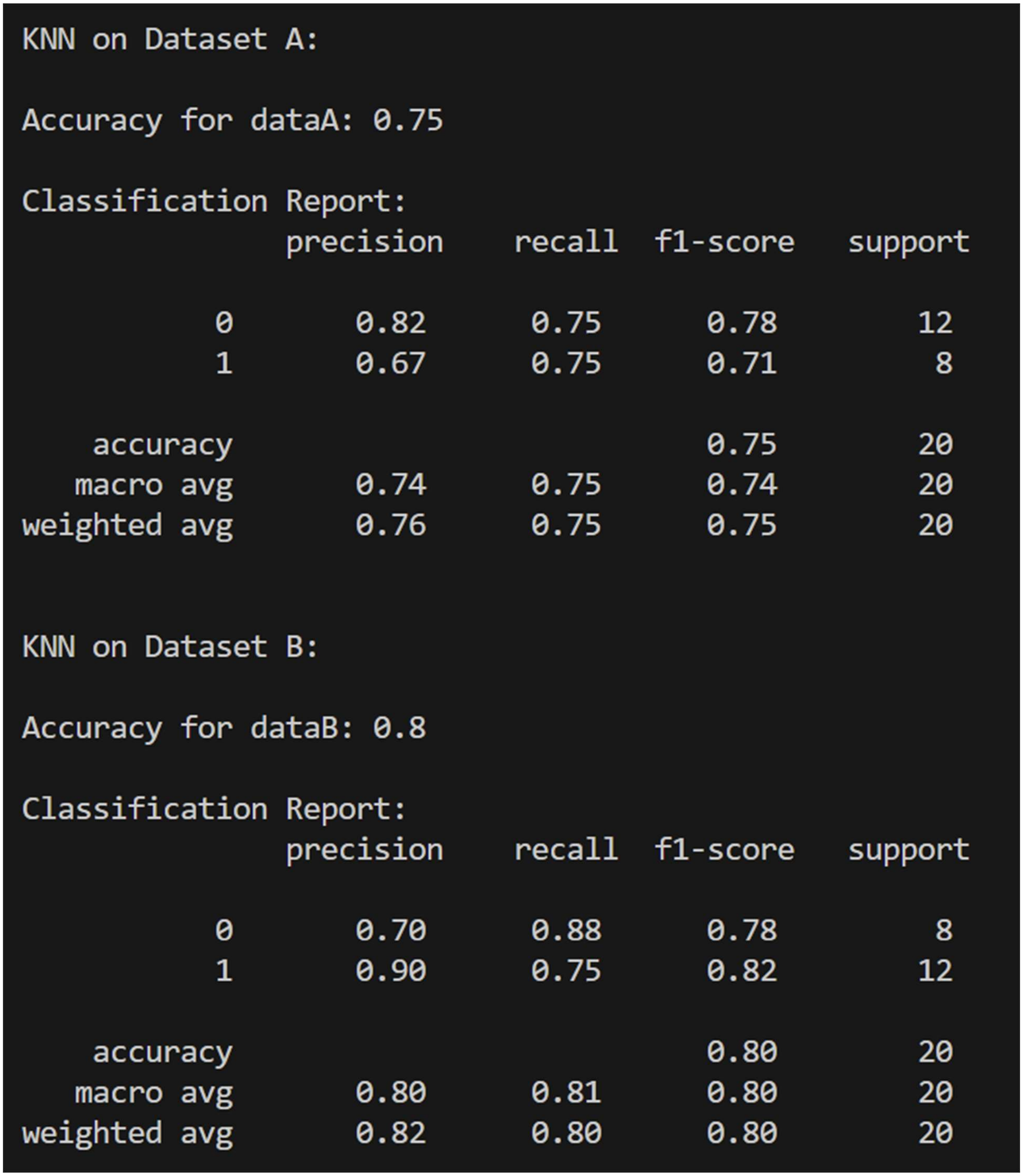
2: Hypotheses formulated on step 2 explaining how you came up with each hypothesis.

Dataset A Hypothesis: Before obtaining the Precision-Recall curve for Dataset A, one could hypothesize that the ML model is likely to exhibit high precision but at the expense of lower recall. This hypothesis is based on the observation that the model starts with high precision at lower recall values, indicating a conservative approach in making positive predictions. The expectation would be that the model tends to be cautious and selective in labeling instances as positive, leading to a lower number of false positives but potentially missing some true positive cases. Therefore, the anticipation is for a Precision-Recall curve that demonstrates a trade-off between precision and recall, with an initial emphasis on high precision.

Dataset B Hypothesis: Like Dataset A, a hypothesis for Dataset B would suggest that the ML model is expected to achieve high precision but at the cost of lower recall. The anticipation is based on the observation that, for individual folds, the model starts with high precision at lower recall values, indicating an initial ability to identify positive cases accurately. However, as recall increases, precision decreases, suggesting a conservative approach like Dataset A. Therefore, the hypothesis would be that the Precision-Recall curve for Dataset B will illustrate a trade-off between precision and recall, emphasizing high precision initially with a decline as recall increases.

Overall, both hypotheses suggest that the ML models in both datasets are likely to be conservative in making positive predictions, prioritizing precision over recall. The expectation is that the Precision-Recall curves will exhibit the typical trade-off pattern observed in models with such characteristics.

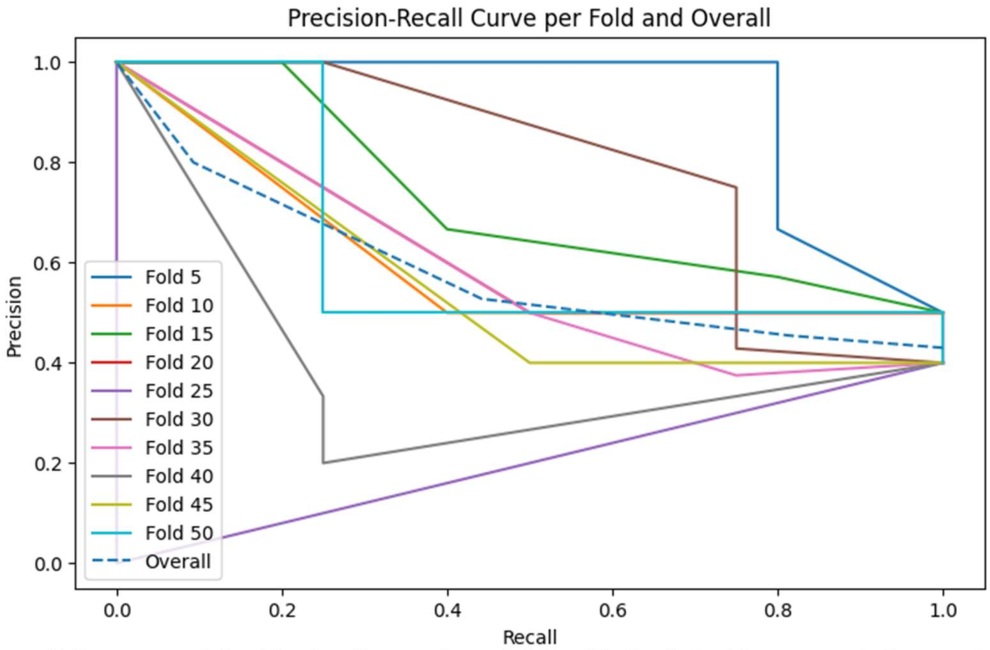
3: A screenshot of a run of your program showing its output.



```
KNN on Dataset A:

Accuracy for dataA: 0.75

Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.75      0.78        12
           1       0.67      0.75      0.71         8

    accuracy                           0.75        20
   macro avg       0.74      0.75      0.74        20
weighted avg       0.76      0.75      0.75        20


KNN on Dataset B:

Accuracy for dataB: 0.8

Classification Report:
              precision    recall  f1-score   support

           0       0.70      0.88      0.78         8
           1       0.90      0.75      0.82        12

    accuracy                           0.80        20
   macro avg       0.80      0.81      0.80        20
weighted avg       0.82      0.80      0.80        20
```

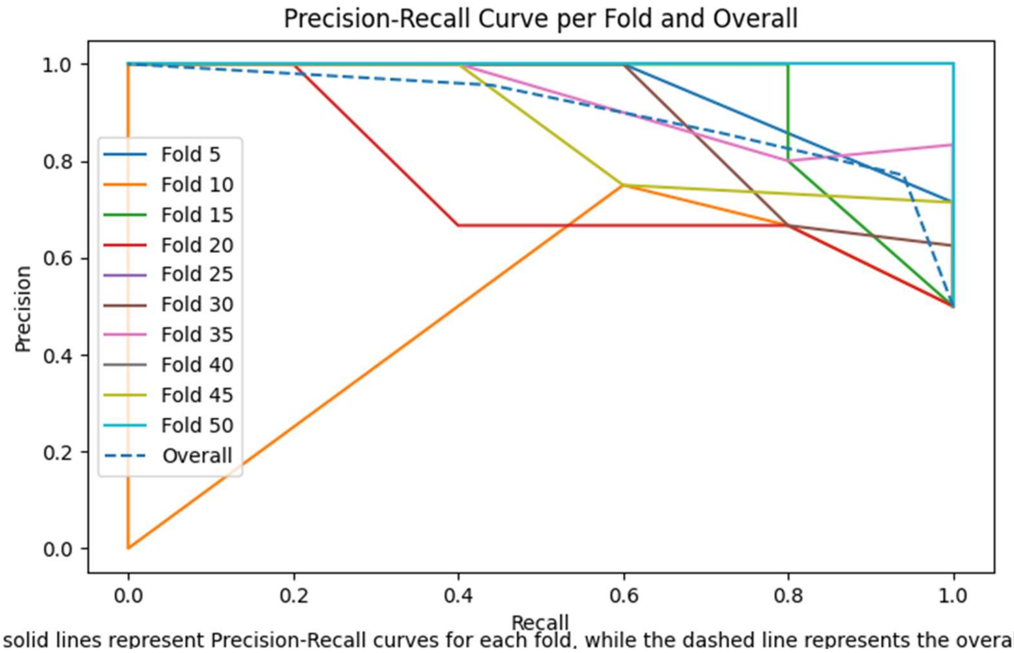
4: The precision-recall curves per dataset. Figures should contain axis labels and a caption explaining the curves.

PR-Curve for Dataset A:



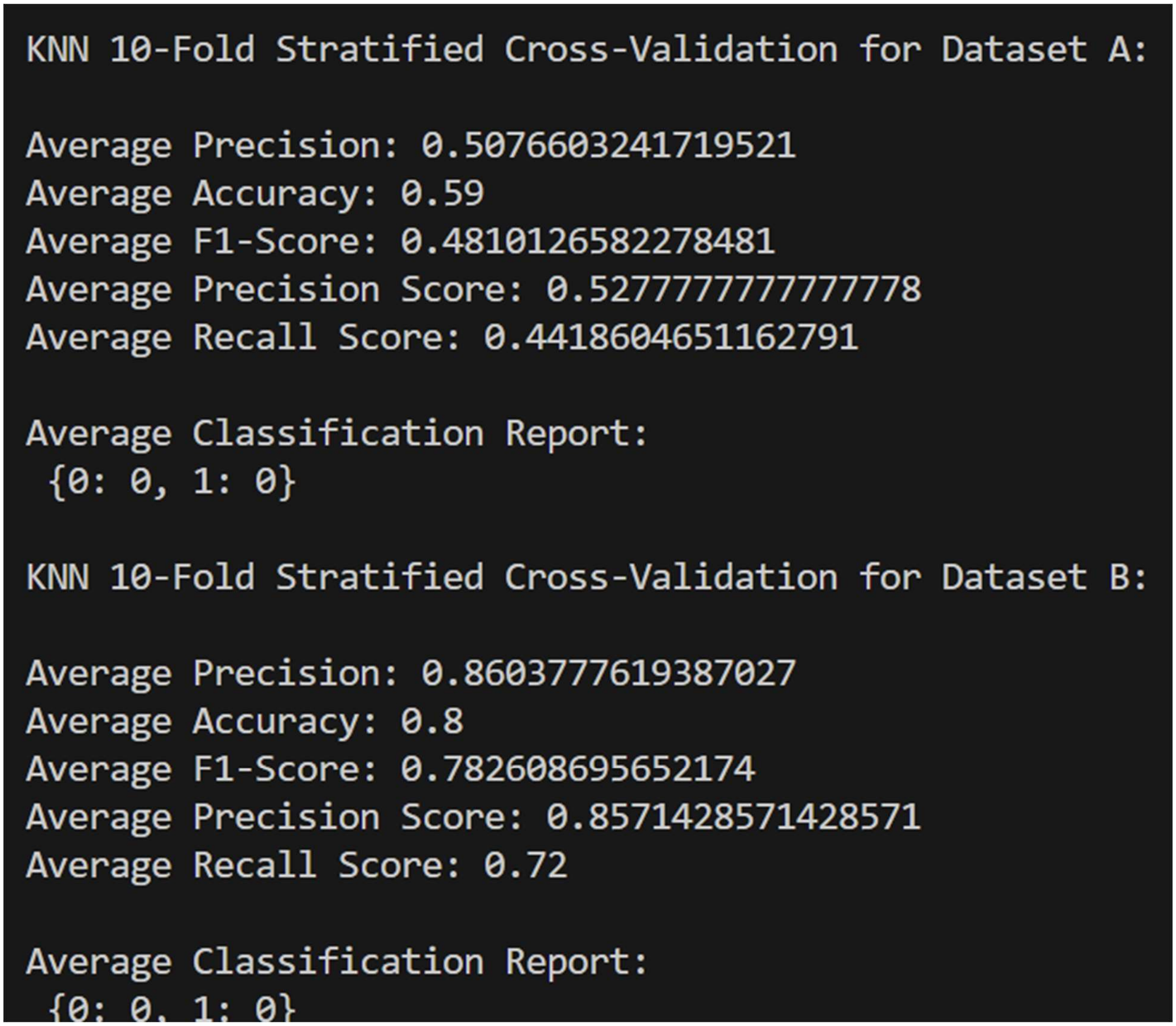
The solid lines represent Precision-Recall curves for each fold, while the dashed line represents the overall curve.

PR-Curve for Dataset B:



The solid lines represent Precision-Recall curves for each fold, while the dashed line represents the overall curve.

5: A table with the performance metrics obtained in step 4 per dataset.



```
KNN 10-Fold Stratified Cross-Validation for Dataset A:

Average Precision: 0.5076603241719521
Average Accuracy: 0.59
Average F1-Score: 0.4810126582278481
Average Precision Score: 0.5277777777777778
Average Recall Score: 0.4418604651162791

Average Classification Report:
 {0: 0, 1: 0}

KNN 10-Fold Stratified Cross-Validation for Dataset B:

Average Precision: 0.8603777619387027
Average Accuracy: 0.8
Average F1-Score: 0.782608695652174
Average Precision Score: 0.8571428571428571
Average Recall Score: 0.72

Average Classification Report:
 {0: 0, 1: 0}
```

6: The conclusion obtained in step 5.

PR-Curve for Dataset A:

Precision (Y-axis) represents the proportion of true positive predictions among all positive predictions made by the model. Higher precision indicates fewer false positives.

Recall (X-axis) represents the proportion of true positive predictions among all actual positive instances. Higher recall indicates better ability to capture positive cases.

Observations: For individual folds (represented by colored solid lines), the model starts with high precision at lower recall values. However, as recall increases, precision decreases. This suggests that the model may perform well in identifying positive cases initially but struggles to maintain precision as recall improves. The dashed line, which represents the overall performance across all folds, follows a similar trend. It generally has lower precision compared to individual folds at corresponding recall levels.

Interpretation: The model seems to achieve moderate to high precision at the cost of recall. It may be conservative in making positive predictions. If the application requires high recall (capturing most positive cases), the model might need adjustments to improve recall, even if it leads to more false positives.

PR-Curve for Dataset B:

Observations: For individual folds (represented by colored solid lines), the model starts with high precision at lower recall values. However, as recall increases, precision decreases. This suggests that the model may perform well in identifying positive cases initially but struggles to maintain precision as recall improves. The dashed line, which represents the overall performance across all folds, follows a similar trend. It generally has lower precision compared to individual folds at corresponding recall levels.

Interpretation: The model seems to achieve moderate to high precision at the cost of recall. It may be conservative in making positive predictions. If the application requires high recall (capturing most positive cases), the model might need adjustments to improve recall, even if it leads to more false positives.

Comparing the two curves:

1. **Similarities**:
    - Both curves represent the relationship between **precision** and **recall** for the same machine learning model.
    - They share the same x-axis (recall) and y-axis (precision) scales.

    - The overall shape of both curves follows the typical Precision-Recall trade-off: as recall increases, precision tends to decrease.

2. **Differences**:
    - **Individual Folds vs. Overall Curve**:
        - The **colored solid lines** correspond to individual folds, while the **dashed line** represents the overall performance across all folds.
        - Individual folds exhibit varying levels of precision and recall, whereas the overall curve combines these results.
    - **Variability**:
        - Among individual folds, some curves start with high precision and gradually decline, while others have different patterns.
        - The overall curve may exhibit different trends due to aggregation.
    - **Threshold Setting**:
        - The individual folds might have different threshold settings, leading to varying precision-recall trade-offs.
        - The overall curve represents a more generalized view, considering a broader range of threshold values.
    - **Performance Consistency**:
        - The variability among folds suggests inconsistency in model performance.
        - The overall curve's decline in precision as recall increases indicates potential limitations in reliability.

The provided observations and interpretations of the Precision-Recall (PR) curves for both Dataset A and Dataset B generally align with the hypotheses formulated before obtaining the curves.

Dataset A: Confirmation of Hypothesis: The observed behavior of the model in Dataset A supports the hypothesis that the model achieves high precision but at the cost of lower recall. The initial high precision at lower recall values, followed by a decrease in precision as recall increases, is consistent with the expectation of a conservative model.

Dataset B: Confirmation of Hypothesis: Like Dataset A, the results for Dataset B also confirm the hypothesis that the model achieves high precision at the expense of lower recall. The observed behavior of starting with high precision at lower recall values, and a subsequent decrease in precision as recall increases, aligns with the expectation of a conservative model.

Conclusion: The patterns observed in both datasets, where precision tends to decrease as recall increases, are consistent with the hypothesis of a Precision-Recall trade-off in favor of precision. The mention of variability among individual folds and the overall decline in precision as recall increases supports the hypothesis that the models might be conservative and exhibit inconsistency in performance.