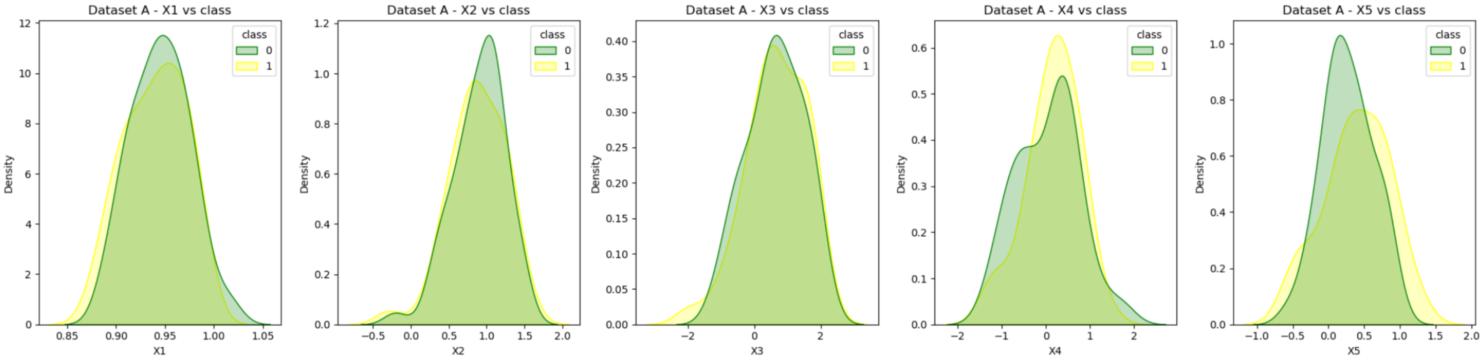
Assignment 1

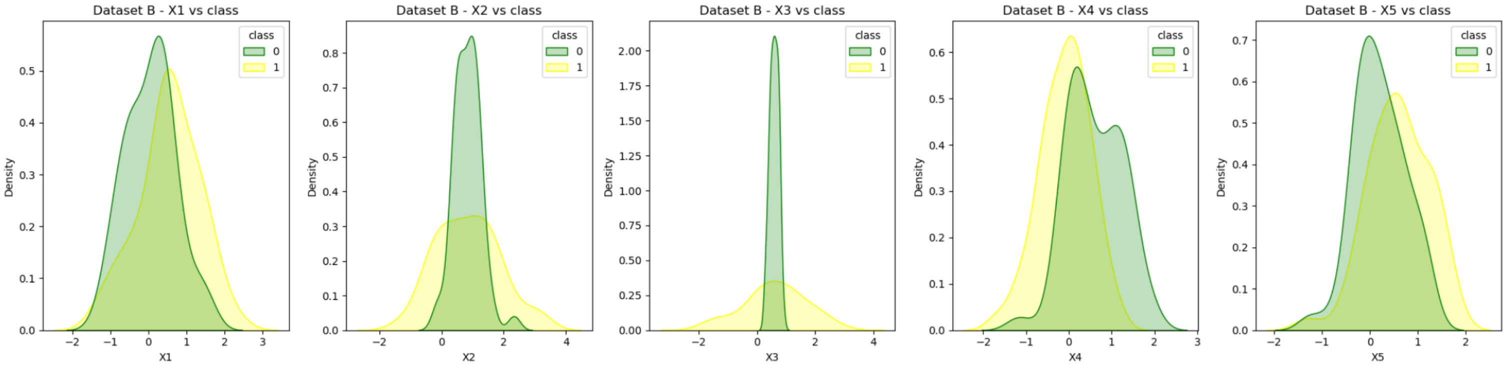
Muhammad Bilal Naseer
202046892

Density Plots

Dataset A

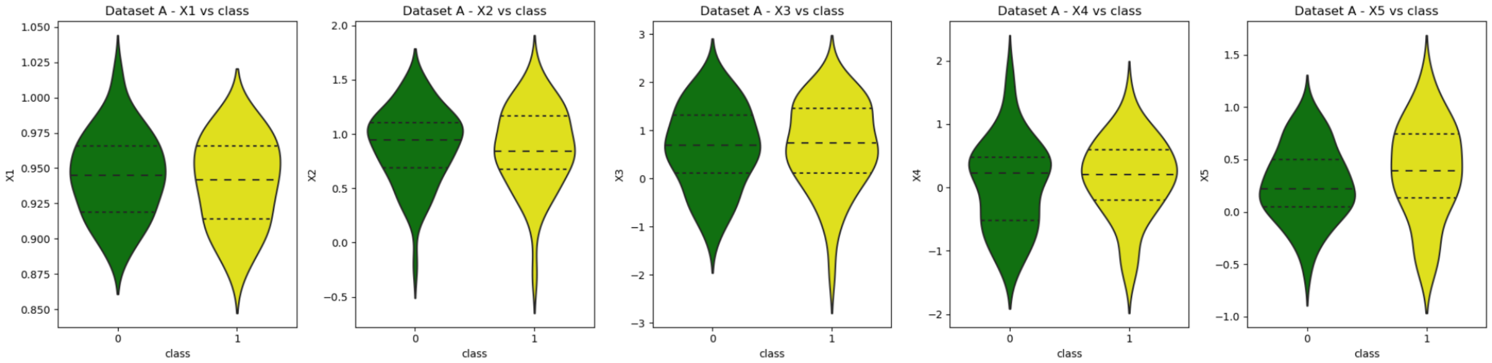


Dataset B

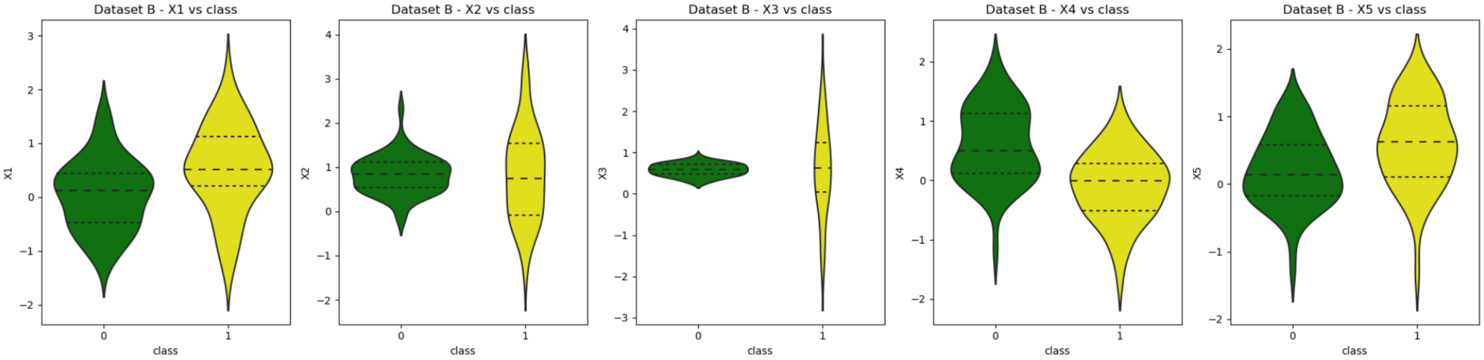


Violin Plots

Dataset A

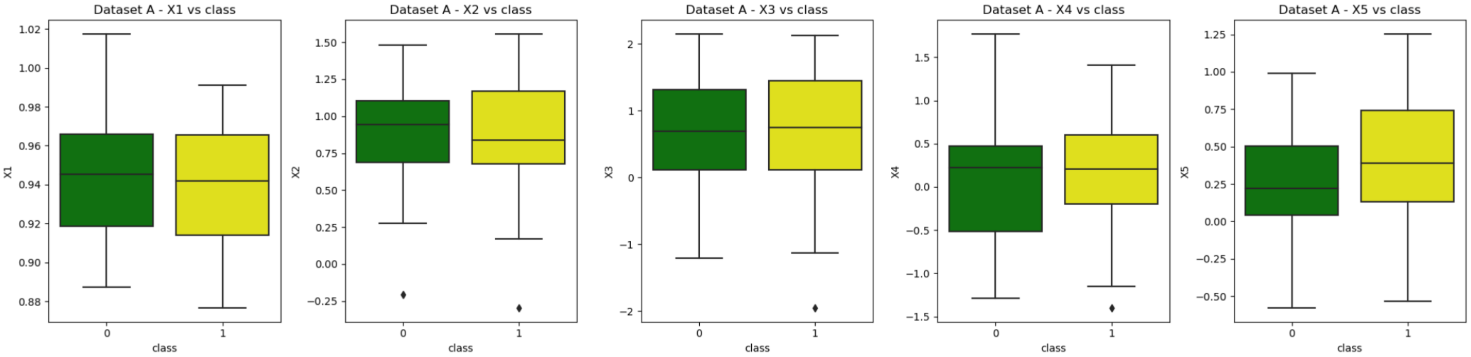


Dataset B

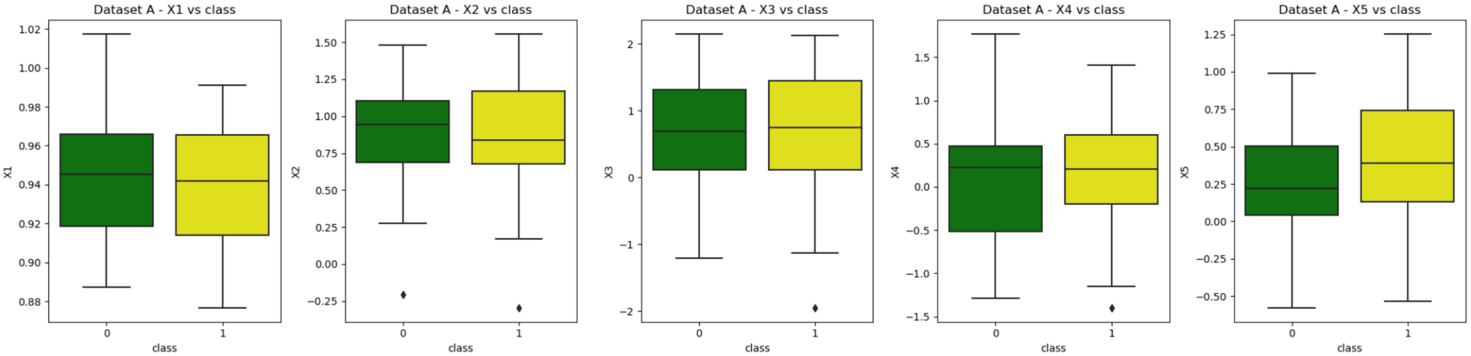


Boxplot

Dataset A

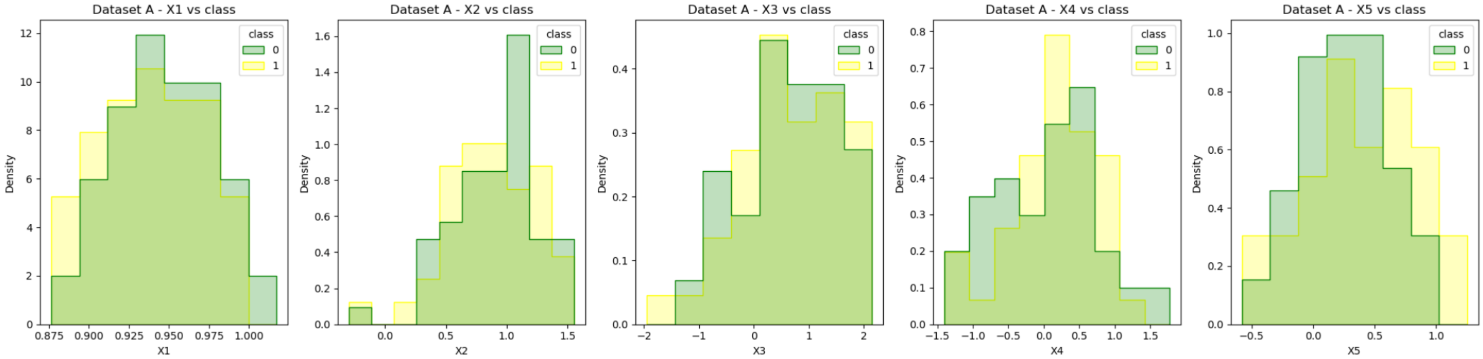


Dataset B

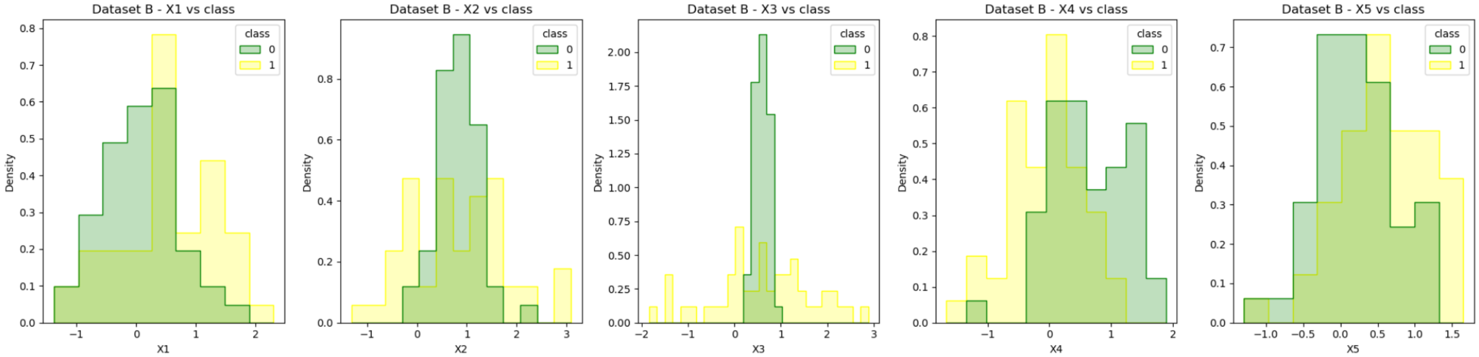


Distribution plot

Dataset A

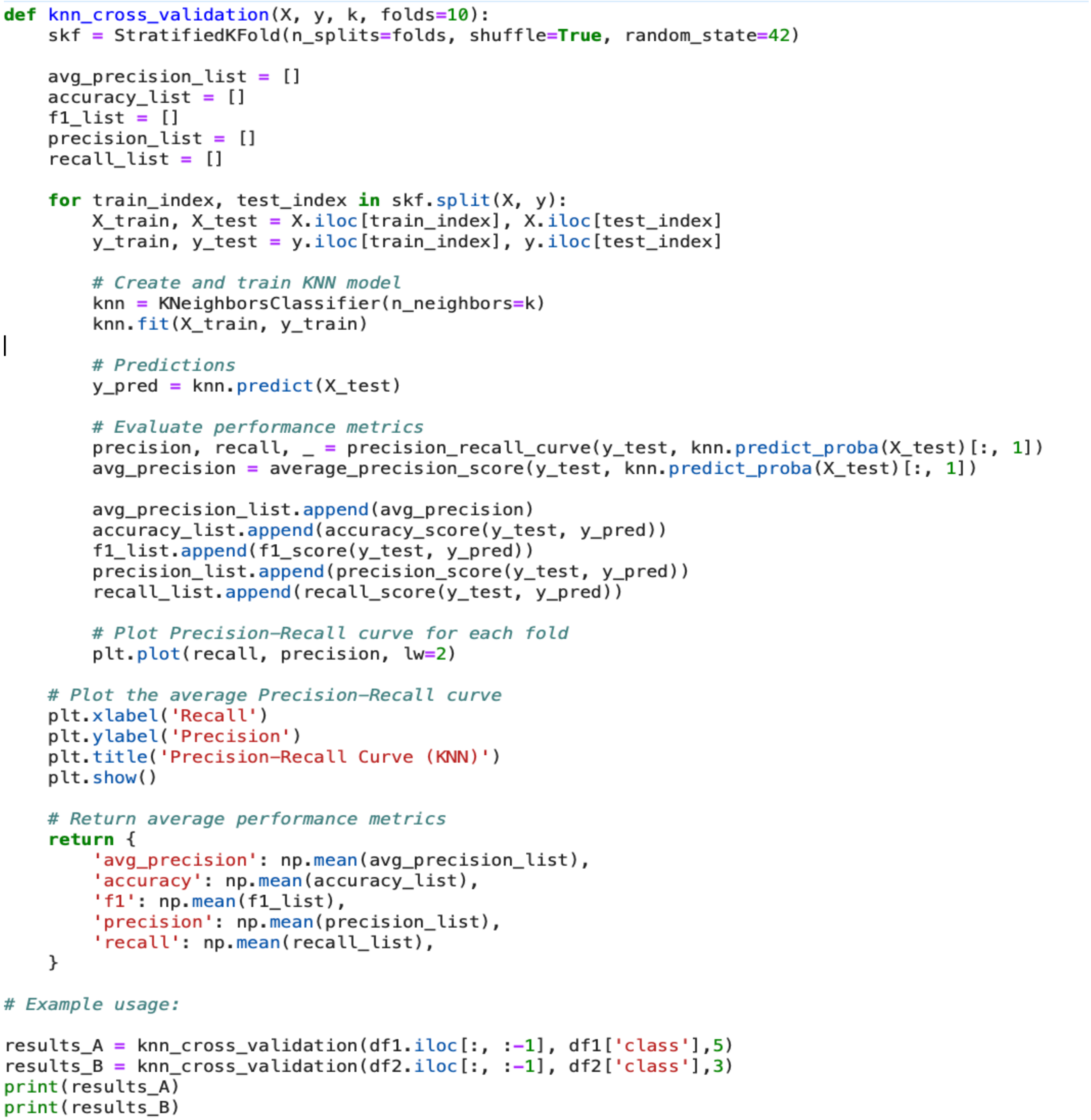


Dataset B

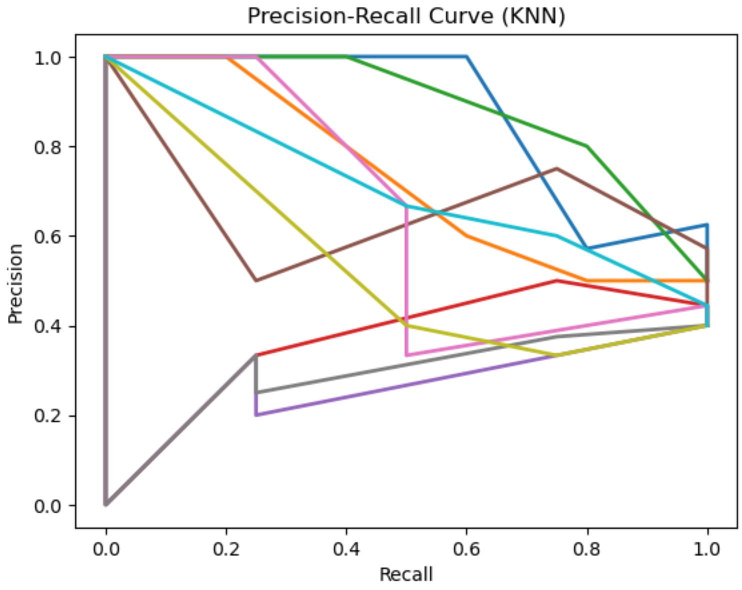


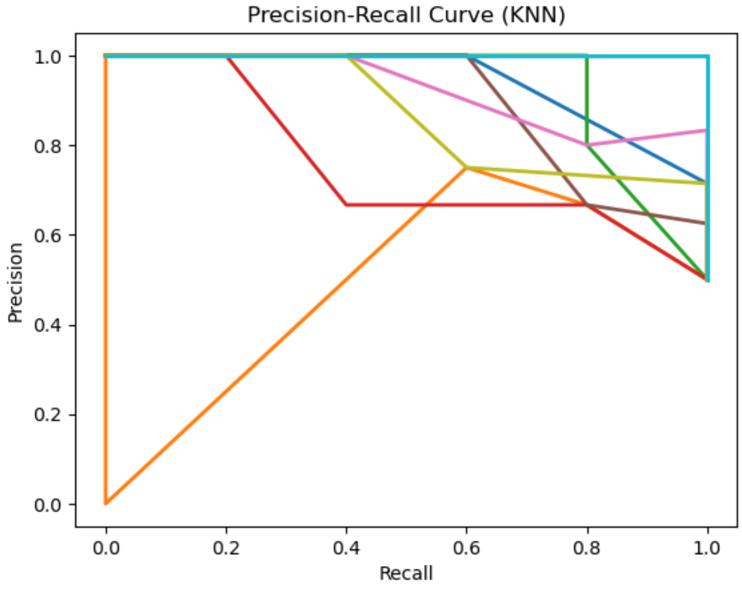
Hypothesis

When considering the above plots for both dataset A and dataset B we can assert that Machine Learning (ML) model is expected to perform better with dataset B than dataset A through the visual examination of the plots depicting both the datasets. The argument centers around the basis that dataset B exhibits more distinctive and vivid features with very few overlaps therefore conveying that classes in dataset B are better separated. This separation in dataset B implies that a ML model should perform better and generalize patterns associated with different classes.



```python
def knn_cross_validation(X, y, k, folds=10):
    skf = StratifiedKFold(n_splits=folds, shuffle=True, random_state=42)

    avg_precision_list = []
    accuracy_list = []
    f1_list = []
    precision_list = []
    recall_list = []

    for train_index, test_index in skf.split(X, y):
        X_train, X_test = X.iloc[train_index], X.iloc[test_index]
        y_train, y_test = y.iloc[train_index], y.iloc[test_index]

        # Create and train KNN model
        knn = KNeighborsClassifier(n_neighbors=k)
        knn.fit(X_train, y_train)

        # Predictions
        y_pred = knn.predict(X_test)

        # Evaluate performance metrics
        precision, recall, _ = precision_recall_curve(y_test, knn.predict_proba(X_test)[:, 1])
        avg_precision = average_precision_score(y_test, knn.predict_proba(X_test)[:, 1])

        avg_precision_list.append(avg_precision)
        accuracy_list.append(accuracy_score(y_test, y_pred))
        f1_list.append(f1_score(y_test, y_pred))
        precision_list.append(precision_score(y_test, y_pred))
        recall_list.append(recall_score(y_test, y_pred))

        # Plot Precision-Recall curve for each fold
        plt.plot(recall, precision, lw=2)

    # Plot the average Precision-Recall curve
    plt.xlabel('Recall')
    plt.ylabel('Precision')
    plt.title('Precision-Recall Curve (KNN)')
    plt.show()

    # Return average performance metrics
    return {
        'avg_precision': np.mean(avg_precision_list),
        'accuracy': np.mean(accuracy_list),
        'f1': np.mean(f1_list),
        'precision': np.mean(precision_list),
        'recall': np.mean(recall_list),
    }

# Example usage:

results_A = knn_cross_validation(df1.iloc[:, :-1], df1['class'],5)
results_B = knn_cross_validation(df2.iloc[:, :-1], df2['class'],3)
print(results_A)
print(results_B)
```





The Precision-Recall Curves for K-Nearest Neighbors (KNN) classification models on datasets A and B reveal distinct performance characteristics. In the first graph representing dataset A, there is greater variability in performance across different classes or thresholds, with some curves exhibiting notably low precision values for specific recall levels. Conversely, the second graph depicting dataset B displays a consistently higher overall performance, with most curves consistently surpassing the 0.5 precision level across various recall levels. The concentration of more lines in the top-right corner of the plot for dataset B, which is considered the ideal scenario in precision-recall analysis, further supports the notion of superior performance compared to dataset A. This observation suggests that the KNN model performs more reliably on dataset B, as evidenced by its ability to achieve higher precision at various recall levels, highlighting its efficacy in classification tasks on that dataset.

Dataset A	
Average Precision	0.58
Accuracy	0.63
F1-Score	0.48
Precision	0.65
Recall	0.42

Dataset B	
Average Precision	0.87
Accuracy	0.8
F1-Score	0.77
Precision	0.86
Recall	0.72

Conclusion

In conclusion, the visual analysis of Precision-Recall Curves for K-Nearest Neighbors (KNN) classification models on datasets A and B supports the hypothesis that the machine learning model is likely to perform better with dataset B. The distinct performance characteristics observed in the precision-recall plots indicate that dataset B exhibits clearer and more separated class features compared to dataset A. The consistently higher overall performance, as reflected in more curves surpassing across various recall levels, further reinforces the superiority of dataset B.

Attributions

For this assignment I used the following resources:
https://scikit-learn.org/stable/modules/cross_validation.html
https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
https://stackoverflow.com/questions/29656550/how-to-plot-pr-curve-over-10-folds-of-cross-validation-in-scikit-learn
https://youtu.be/-8s9KuNo5SA?si=faFV88g6oFiR8b1E

I also discussed a few concepts with my classmates Sara Hamid, Basim Ali and Muneeb-gpt ur-Rehman as well as used chatgpt to optimize my code.