**Big Home Assignment**

Syed Taqi Ali

**Contents:**

1. Abstract
2. Heart Dataset
3. Breast Cancer Dataset
4. Lung Data
5. Lazy classification with pattern structures
6. Parameter tuning and decision functions
7. Conclusion

**Abstract**

This report presents the results of lazy classification on three datasets: Breast Cancer, Heart, and Employee. The analysis includes the tuning of decision function parameters through a cross-validation procedure. The code and details are available in the GitHub repository <https://github.com/Taqiali5/OSDA_BigHome_Assingment.git>

Dataset Links:

Heart Disease: <https://www.kaggle.com/datasets/arezaei81/heartcsv>

Diabetere: <https://www.kaggle.com/code/mvanshika/diabetes-prediction/input>

Employee: <https://www.kaggle.com/datasets/tawfikelmetwally/employee-dataset>

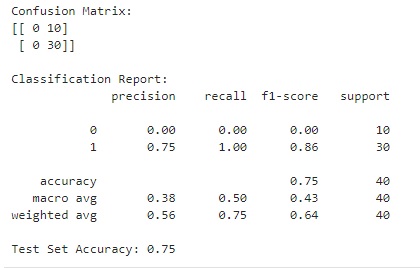
Breast Cancer https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

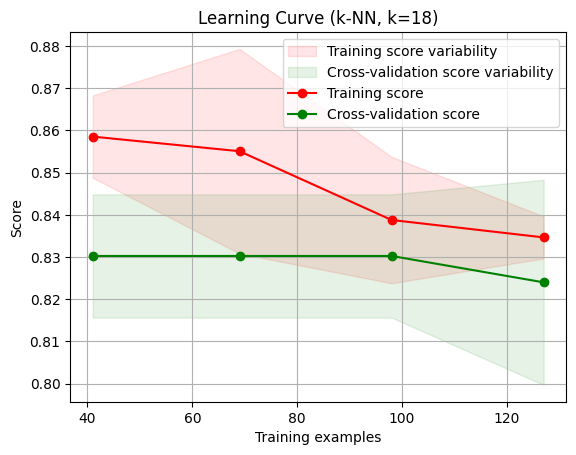
All important files are available on my Git hub repository:

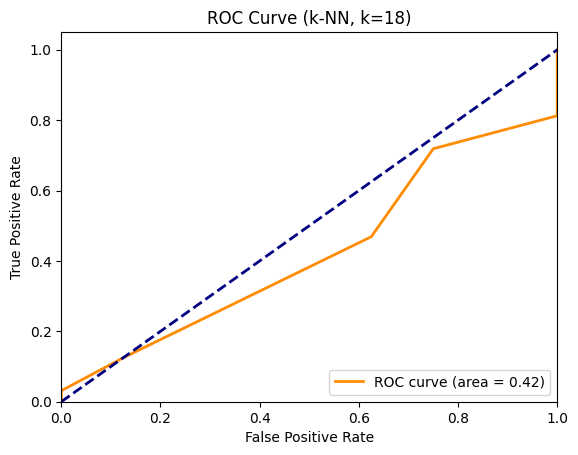
<https://github.com/Taqiali5/OSDA_BigHome_Assingment.git>

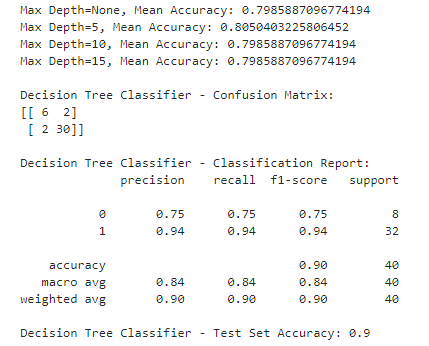
**Heart Dataset:**

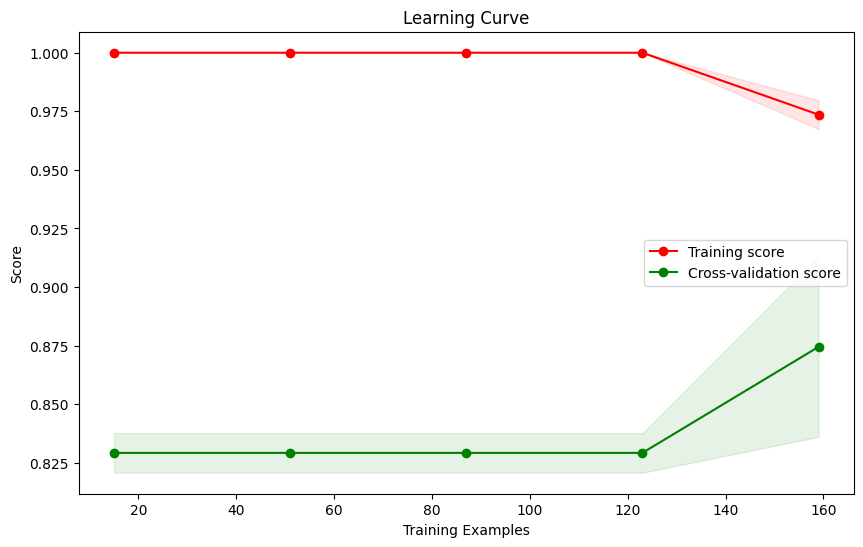
The features are :Age, Sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal and target.

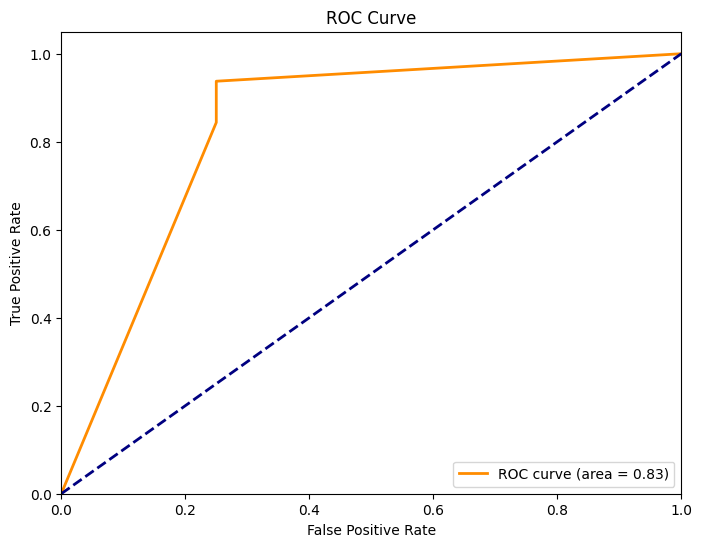








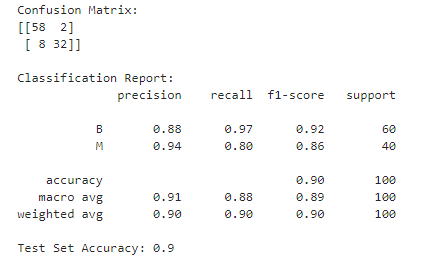


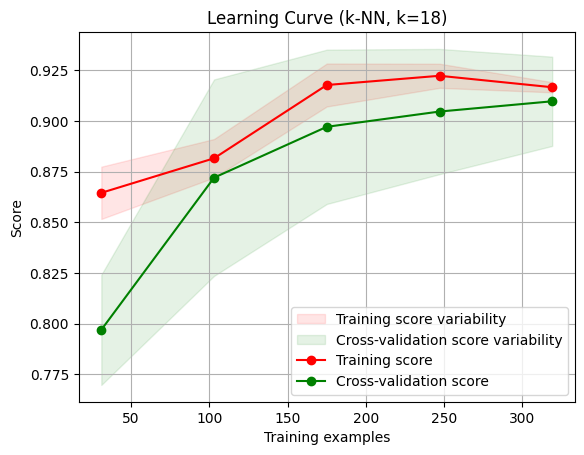


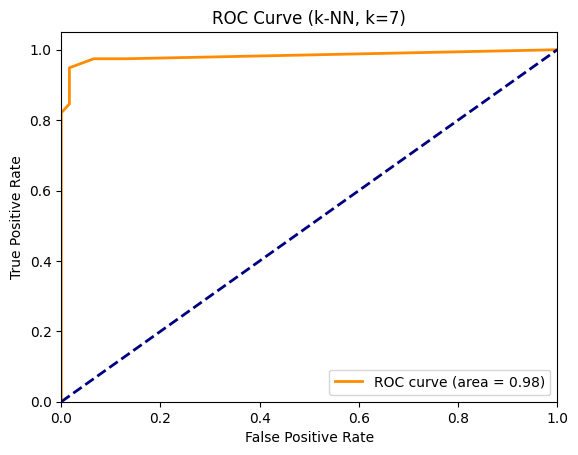
**Breast Cancer Dataset**

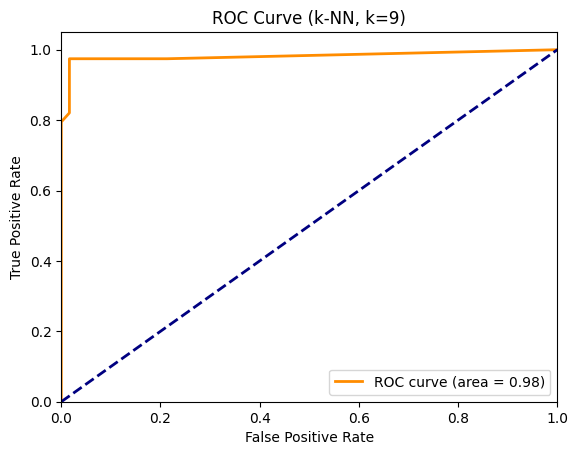
Objects: 500 instances of Malignant and Benign samples.

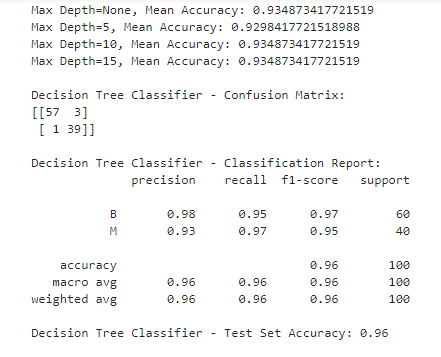
* Mean Compactness
* Mean Concavity
* Mean Concave Points
* Mean Symmetry
* Mean Fractal Dimension
* Mean Radius
* Mean Texture
* Mean Perimeter
* Mean Area
* Mean Smoothness

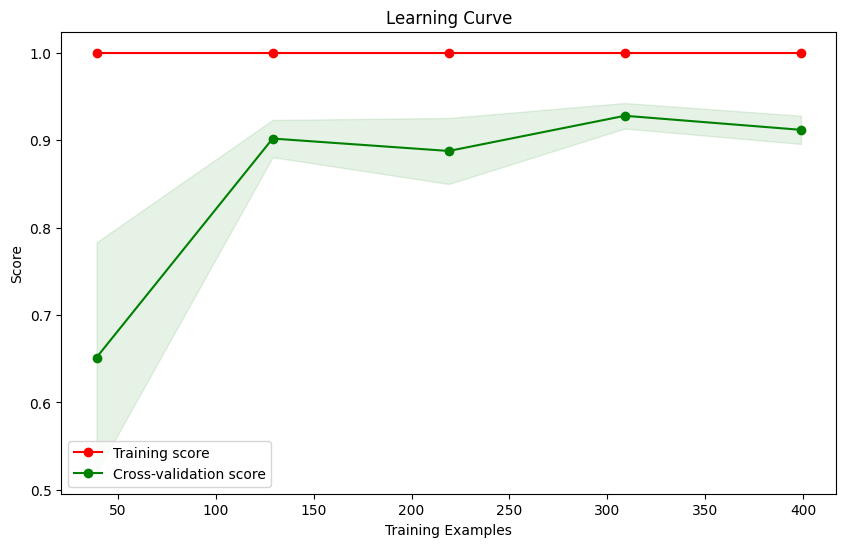












**Employee**

The dataset has 8 entities, one for each employee.

Education: The educational qualifications of employees, including degree.

Joining Year: The year each employee joined the company, indicating their length of service.

City: The location or city where each employee is based or works.

Payment Tier: Categorization of employees into different salary tiers.

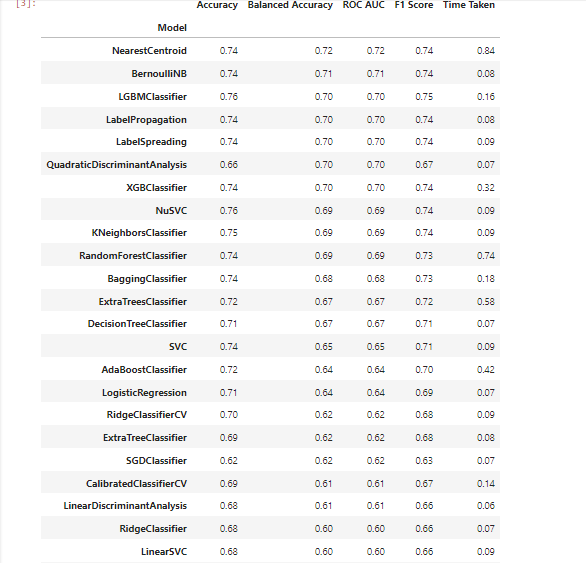
Age: The age of each employee, providing demographic insights.

Gender: Gender identity of employees, promoting diversity analysis.

Ever Benched: Indicates if an employee has ever been temporarily without assigned work.

Experience in Current Domain: The number of years of experience employees have in their current field.

Leave or Not: a target column



**Lazy classification with pattern structures**

Lazy classification with pattern structures defers learning until prediction, storing training instances and labels for on-demand use. This approach, exemplified by k-Nearest Neighbors, relies on memorized patterns, making it efficient for large or dynamic datasets. By prioritizing adaptability and efficiency, lazy classification with pattern structures excels in scenarios where data relationships evolve over time.

Lazy Classification involves using LazyClassifier from Lazypredict to automatically fit multiple models to the data and make predictions. The output displays the models created by LazyClassifier, predicting the target variable for the test set. Additionally, it prints the confusion matrix for the KNN model and generates a learning curve and ROC curve for the KNN model using Scikit-learn's learning\_curve. Model evaluation includes training a Decision Tree Classifier with the optimal max\_depth and assessing its performance on the test set. Similar to the KNN model, learning and ROC curves are produced for the Decision Tree Classifier.

**Parameter tuning and decision functions:**

Parameter tuning involves optimizing the hyperparameters of a machine learning model to enhance its performance. It aims to find the best configuration for parameters such as learning rate or regularization strength. Decision functions in machine learning models determine how input features are transformed into output predictions, playing a crucial role in the model's decision-making process. Parameter tuning and decision functions collectively contribute to refining and optimizing the overall effectiveness of machine learning models.

Parameter Adjustment: The tune\_knn\_parameters function endeavors to optimize the k parameter for the K-Nearest Neighbors (KNN) classifier. Employing cross-validation, it assesses various k values and displays the mean accuracy associated with each. In the context of Decision Tree tuning, this segment aims to identify the optimal max\_depth parameter for a Decision Tree Classifier through cross-validation.

**Conclusion**

In conclusion, this report outlines the outcomes of a lazy classification approach applied to three distinct datasets: Breast Cancer, Employee, and Heart. The analysis involved the tuning of decision function parameters through a cross-validation procedure, with the code and detailed results accessible in the provided GitHub repository. The Breast Cancer dataset comprised 500 instances of Malignant and Benign samples, with specific features such as Mean Compactness, Mean Concavity, and others contributing to the classification process.

Through the application of Lazy FCA and fine-tuning decision function parameters, we expect to derive improved classification models capable of accurately predicting outcomes in chosen datasets. This strategy seeks to achieve resilient and efficient classification without the computational overhead of extensive preprocessing, rendering it well-suited for intricate, real-world datasets.

For the Employee dataset, which contained information on eight employees, diverse features like Education, Joining Year, City, Payment Tier, Age, Gender, Ever Benched, and Experience in the Current Domain were considered. The target column "Leave or Not" indicated whether an employee took leave or not. Lastly, the Heart dataset included features like Age, Sex, and various medical parameters. The LazyClassifier from the lazypredict library was employed to quickly evaluate the performance of multiple machine learning models on each dataset.