## Breast Cancer Dataset Classification Report Using Multiple Machine Learning Approaches

**Abstract**

Breast cancer is a serious health concern, and accurate prediction of cancer recurrence is crucial for effective treatment and patient management. This study aims to analyze the breast cancer dataset and evaluate the performance of three different classification algorithms: Decision Tree, Logistic Regression, and MLP Classifier. By leveraging various data preprocessing techniques, including handling missing values, encoding categorical variables, feature scaling, and balancing the class distribution through oversampling, the study seeks to uncover insights and develop robust predictive models for breast cancer recurrence.

### 1. Dataset Overview

The breast cancer dataset used in this study consists of 286 instances with 9 input features and a binary target variable indicating whether the cancer recurred or did not recur. The dataset exhibits a class imbalance, with 201 instances (70.0%) belonging to the "no-recurrence-events" class and 85 instances (30.0%) belonging to the "recurrence-events" class.

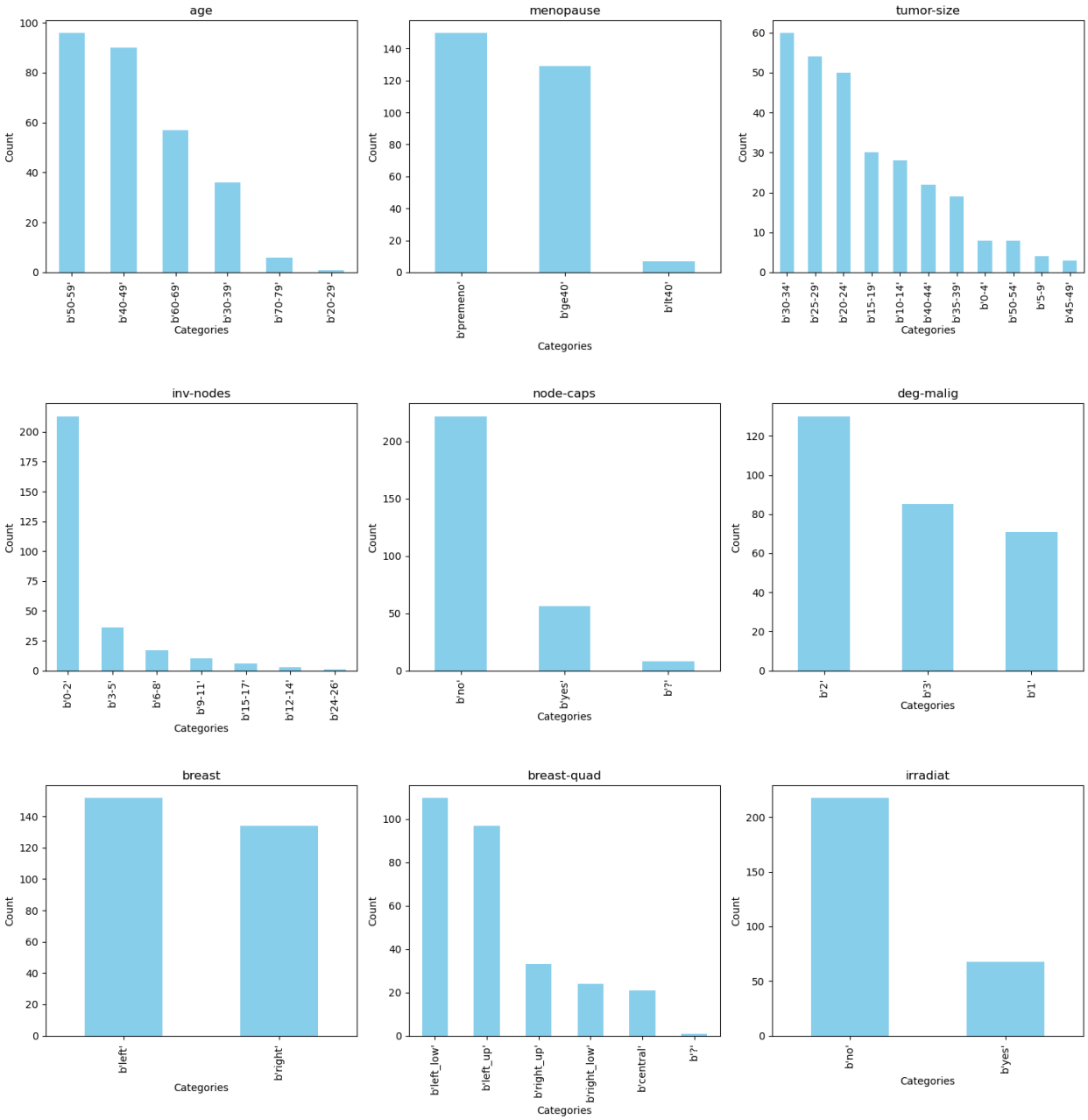
First few rows of the dataset look like below:

```
First few rows of the DataFrame:
     age menopause tumor-size inv-nodes node-caps deg-malig breast  \
0  40-49   premeno      15-19       0-2       yes         3  right
1  50-59      ge40      15-19       0-2        no         1  right
2  50-59      ge40      35-39       0-2        no         2   left
3  40-49   premeno      35-39       0-2       yes         3  right
4  40-49   premeno      30-34       3-5       yes         2   left

  breast-quad irradiat                Class
0     left_up       no    recurrence-events
1     central       no  no-recurrence-events
2    left_low       no    recurrence-events
3    left_low      yes  no-recurrence-events
4    right_up       no    recurrence-events
```

The input features include characteristics such as the patient's age, menopausal status, tumor size, number of involved lymph nodes, node capsule status, degree of malignancy, breast location, and radiation therapy.

The distribution of features in the dataset looks like below:




The dataset contains both categorical and continuous variables, which require appropriate preprocessing techniques to be applied.

## 2. Methodology

### 2.1 Experimental Design

#### 2.1.1 Experimental Setup

In this study, the experimental setup involved training and evaluating machine learning models to predict breast cancer recurrence using a balanced dataset. The class imbalance in the target variable was addressed through oversampling of the minority class (recurrence events). Furthermore, a 10-fold cross-validation (CV) strategy was implemented to evaluate model performance, ensuring the results' robustness by training and validating on multiple subsets of the data.

#### 2.1.2 Evaluation Metrics

Models were assessed using various metrics, including accuracy, precision, recall, F1-score, and confusion matrices, which provide a comprehensive understanding of the models' predictive performance.

#### 2.1.3 Tools and Libraries

The analysis was conducted using Python and the *scikit-learn* library for data preprocessing, model training, and evaluation. Visualization tools like *matplotlib* and *seaborn* were employed to interpret the data and model outcomes.

### 2.2 Data Preprocessing

To address the challenges posed by the dataset, the following data preprocessing steps were undertaken:

1. **Handling Missing Values:** The dataset had 9 instances with missing values, which were replaced with the mode of the respective features.
2. **Categorical Variable Encoding:** The categorical features were encoded using *one-hot encoding*, ensuring that the classifier can effectively utilize the information contained in these variables.
3. **Feature Scaling:** The features were normalized using the *MinMaxScaler* and standardized using the *StandardScaler* to ensure that all variables are on a similar scale, which is essential for many machine learning algorithms.
4. **Distribution Balancing with Oversampling:** To address the class imbalance, the minority class ("recurrence-events") was oversampled by duplicating samples from this class until the class distribution was balanced. This was achieved using the *resample()* function from the *sklearn.utils* module.

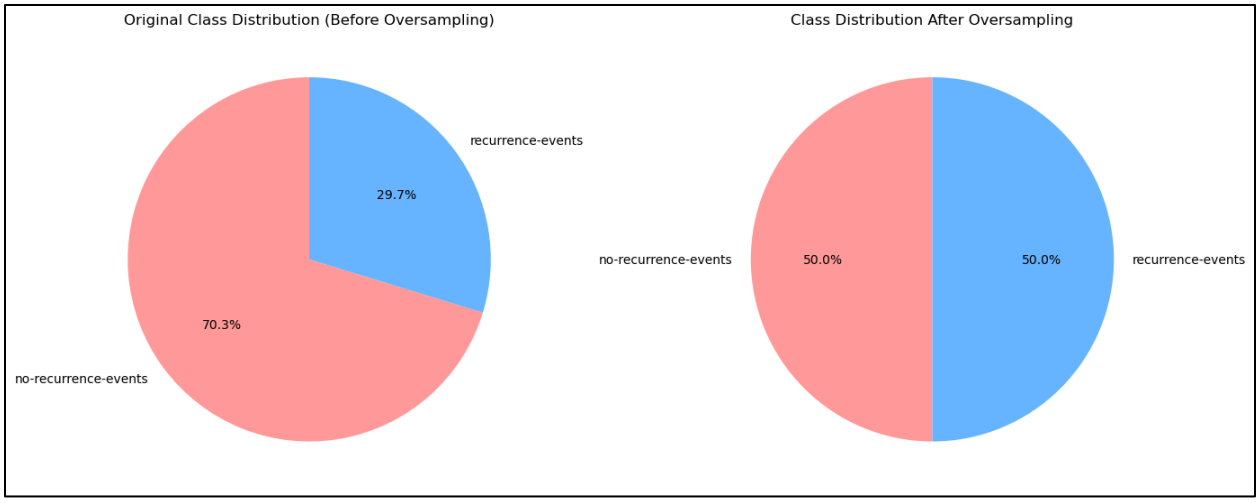



**Fig.1: Class Distribution Before and After Resampling**

The pie charts provide a clear visual representation of the class distribution in the breast cancer dataset before and after oversampling the minority class.

By implementing these data preprocessing steps, the study aimed to create a robust and balanced dataset that can be effectively utilized for the subsequent classification tasks.

### 2.3 Classification Methods

#### 2.3.1 Decision Tree Classifier

The Decision Tree Classifier was selected for this analysis due to several key factors:

**Interpretability**: Decision trees are inherently easy to understand and visualize, making them an ideal choice for explaining model decisions. The hierarchical, tree-like structure provides a clear and intuitive way to understand how the model arrives at its predictions.

**Non-linear Relationships**: Decision trees have the ability to capture complex, non-linear decision boundaries within the data, allowing them to handle datasets with intricate patterns and relationships among features.

**Versatility**: Decision trees can work with both categorical and numerical data without the need for feature scaling, providing a flexible approach to the breast cancer dataset.

**Robustness to Missing Data**: The decision tree model can effectively deal with missing values by learning from the available data, making it a suitable choice for the dataset where some values were replaced with the mode.

**Architecture**: The Decision Tree Classifier in *scikit-learn* consists of a root node, internal nodes, and leaf nodes. The root node uses the most informative feature to split the data, while the internal nodes represent decision points that further split the data based on specific feature conditions. The leaf nodes assign the final class labels. The splitting criteria, such as Gini impurity or entropy, are used to determine the best feature and threshold for splitting the data.

### 2.3.2 Logistic Regression

Logistic Regression was selected for the following reasons:

**Simplicity and Efficiency**: Logistic Regression is a computationally efficient and widely-used classification algorithm, making it a suitable baseline model to compare against more complex techniques.

**Probabilistic Output**: Logistic Regression provides probability estimates for each class, which can be helpful in setting decision thresholds and understanding the model's confidence in its predictions.

**Interpretability**: The linear nature of Logistic Regression makes it easier to interpret how the input features contribute to the final classification, providing insights into the underlying relationships.

**Generalization**: Logistic Regression is less prone to overfitting compared to more complex models, especially when working with simpler datasets, as it can generalize well with proper regularization.

**Architecture**: Logistic Regression is a linear classification model that uses the *logistic (sigmoid)* function to transform a linear combination of features into probability scores. The key components include the linear predictor (weighted sum of input features plus a bias term), the *sigmoid* function that transforms the linear predictor into a probability between 0 and 1, and the decision boundary (a hyperplane that separates different classes). The model parameters are optimized using gradient descent to minimize the prediction error.

### 2.3.3 MLP Classifier

The Neural Network classifier, implemented using MLP Classifier, was chosen for the following reasons:

**Powerful for Complex Patterns**: MLP Classifier excels at learning intricate, non-linear relationships within the data, making them well-suited for complex classification tasks like the breast cancer dataset.

**High Flexibility**: MLP Classifier can adapt to various data types and automatically learn feature interactions, allowing it to handle the preprocessed dataset with one-hot encoded features.

**Suitability for Large Datasets**: MLP Classifier is known to perform well on large datasets, where their ability to model complex patterns can be fully leveraged, particularly when other models might underperform.

**Non-linear Activation Functions**: The use of non-linear activation functions, such as *ReLU* or *sigmoid*, in the MLP classifier architecture allows for the creation of highly flexible decision boundaries, improving the model's prediction accuracy.

**Architecture**: The Multilayer Perceptron classifier in *scikit-learn* consists of an input layer, one or more hidden layers, and an output layer. The input layer receives the preprocessed feature data, the hidden layers transform the input features using weighted connections and non-linear activation functions, and the output layer produces the class probability predictions. The model is trained using backpropagation, which adjusts the weights to minimize the prediction error.

## 3. Results and Discussion

The implementation of the three classification models - Decision Tree, Logistic Regression, and Multilayer Perceptron Classifier - on the breast cancer dataset provided valuable insights into their performance and trade-offs.

### 3.1 Model Performances

#### 3.1.1 Decision Tree Classifier Model Performance

The decision tree classifier performed the best among the three models tested. The output for decision tree looks like below:

```
Decision Tree Classifier
Cross-Validation Accuracy Scores: [0.87804878 0.87804878 0.8        0.65       0.95       0.775
 0.75       0.875      0.875      0.8        ]
Mean Accuracy: 0.8231
Standard Deviation: 0.0814

Overall Performance (on the entire dataset):
Accuracy: 0.9850746268656716
Classification Report:
                     precision    recall  f1-score   support

no-recurrence-events       0.99      0.98      0.98       201
   recurrence-events       0.98      1.00      0.99       201

            accuracy                           0.99       402
           macro avg       0.99      0.99      0.99       402
        weighted avg       0.99      0.99      0.99       402

Confusion Matrix:
[[196   5]
 [  1 200]]
```

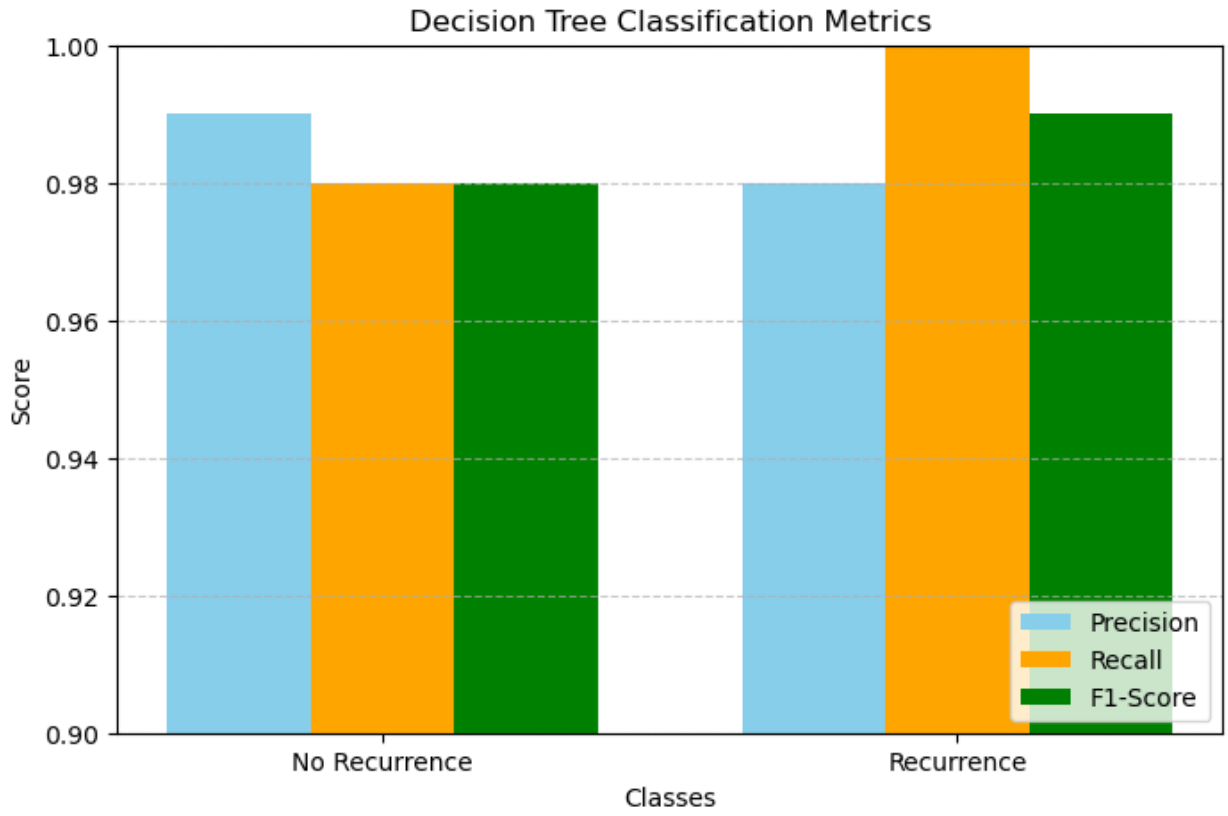



**Fig.2: Decision Tree Classification Metrices**

Looking at the classification report, the decision tree had excellent precision, recall, and F1-score for both the "no-recurrence-events" and "recurrence-events" classes. The precision for the "no-recurrence-events"

class was 0.99, and the recall was 0.98, indicating that the model was able to correctly identify the majority of the "no-recurrence-events" cases. Similarly, for the "recurrence-events" class, the precision was 0.98, and the recall was 1.00, meaning the model was able to identify all the "recurrence-events" cases correctly. The high F1 scores for both categories (0.98 and 0.99) indicate that the model is effectively balancing precision and recall.

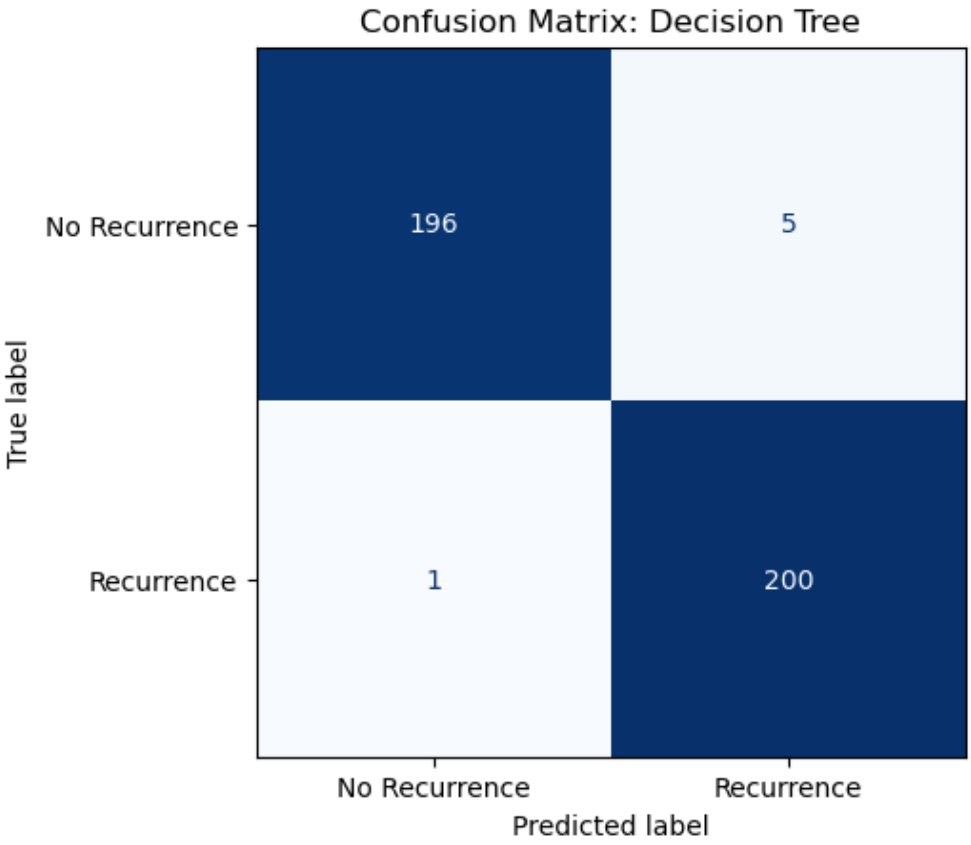



**Fig.3: Decision Tree Confusion Matrix**

The confusion matrix further reinforces the strong performance of the decision tree. It shows that out of 201 "no-recurrence-events" cases, the model correctly identified 196 of them, and out of 201 "recurrence-events" cases, the model correctly identified 200 of them.

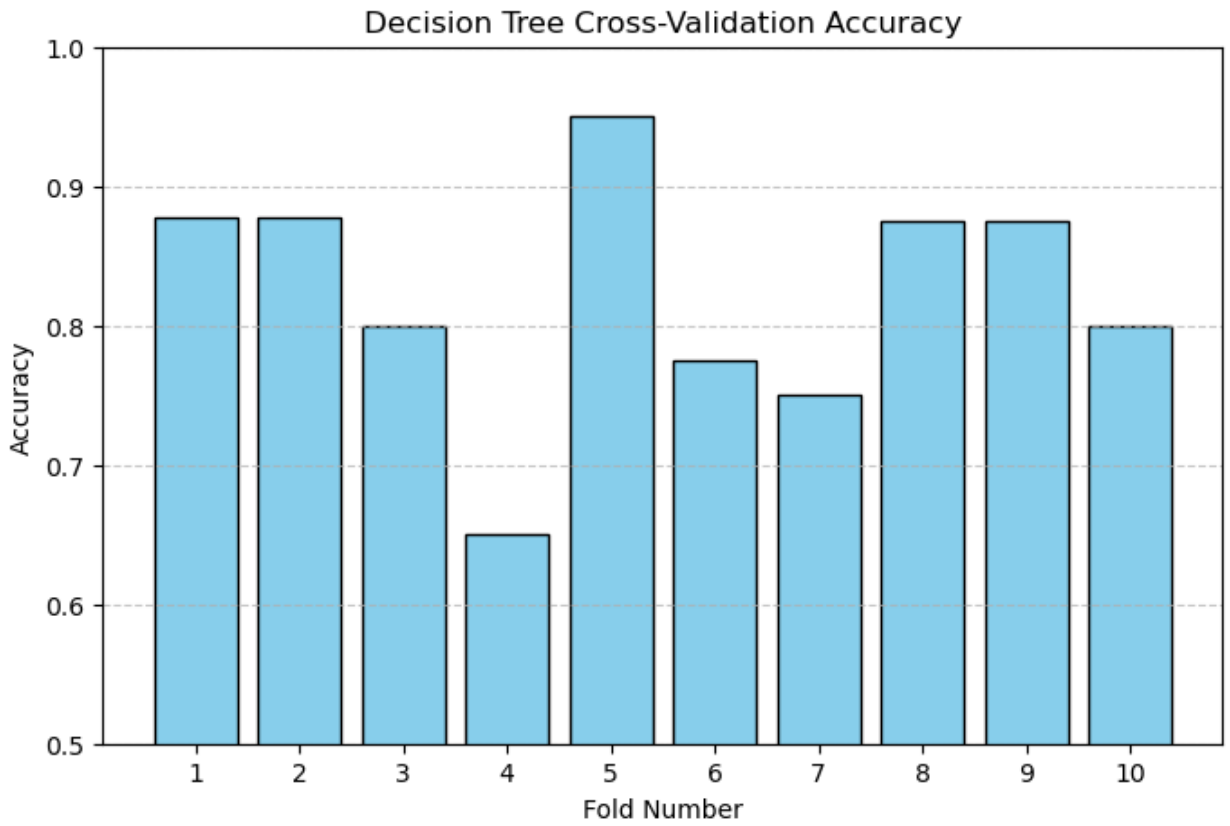



**Fig.4: Decision Tree Cross-Validation Result**

The Decision Tree Cross-Validation Accuracy remains high across all 10 folds, with a mean cross-validation accuracy of 0.8231 and a standard deviation of 0.0814. On the full dataset, the decision tree achieved an impressive accuracy of 0.985. The results indicates that the decision tree model is robust and consistent in its performance.

### 3.1.2 Logistic Regression Classifier Model Performance

The logistic regression model performed relatively well, but not as strongly as the decision tree classifier. The output for Logistic Regression looks like below:

```
Logistic Regression Classifier
Cross-Validation Accuracy Scores: [0.7804878  0.56097561 0.625      0.6        0.725      0.625
 0.65       0.675      0.65       0.375     ]
Mean Accuracy: 0.6266
Standard Deviation: 0.1025

Overall Performance (on the entire dataset):
Accuracy: 0.6965174129353234
Classification Report:
                     precision    recall  f1-score   support

no-recurrence-events       0.71      0.66      0.69       201
   recurrence-events       0.68      0.73      0.71       201

            accuracy                           0.70       402
           macro avg       0.70      0.70      0.70       402
        weighted avg       0.70      0.70      0.70       402

Confusion Matrix:
[[133  68]
 [ 54 147]]
```

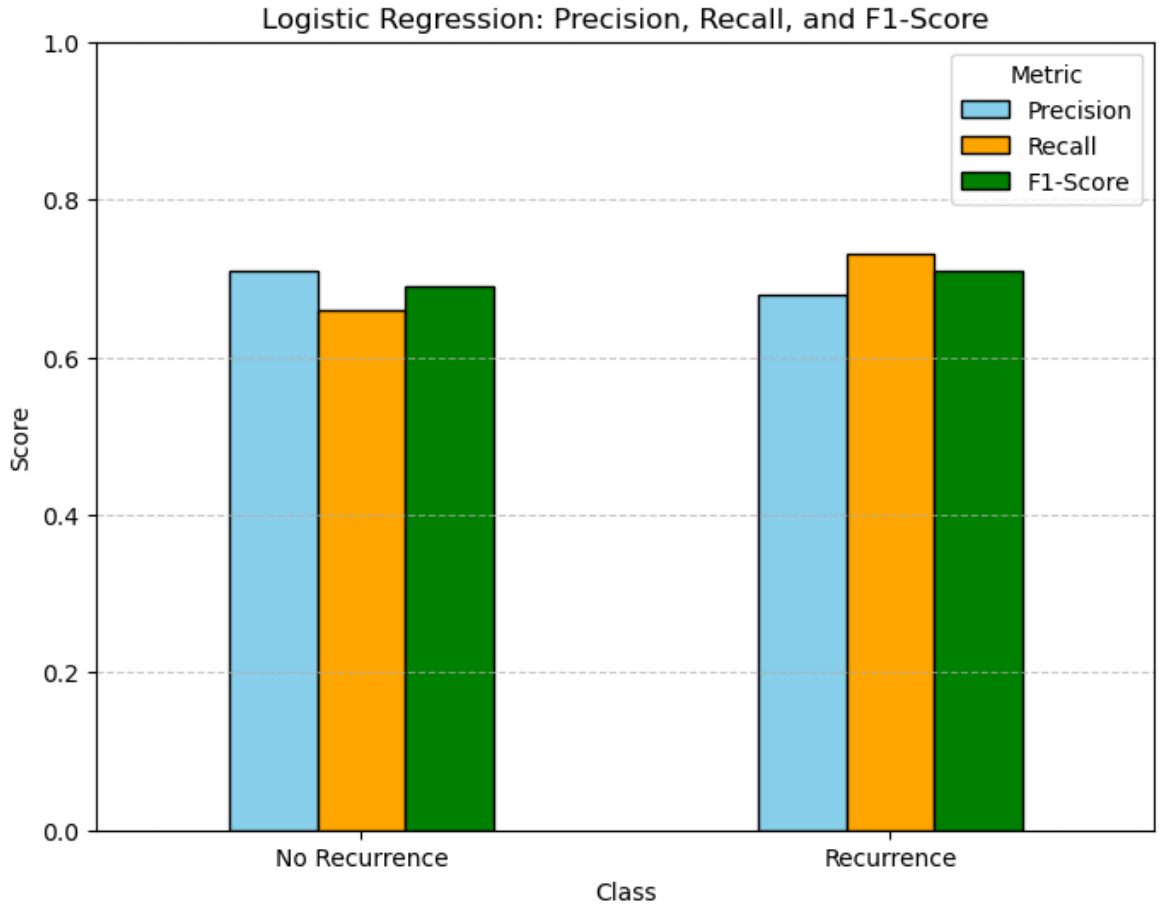



**Fig.5: Logistic Regression Classification Report**

The classification report provides more detailed insights. The logistic regression had a precision of 0.71 and a recall of 0.66 for the "no-recurrence-events" class, and a precision of 0.68 and a recall of 0.73 for the "recurrence-events" class. The F1-scores for both classes were around 0.70, indicating that the model was not as effective in correctly identifying both classes compared to the decision tree.

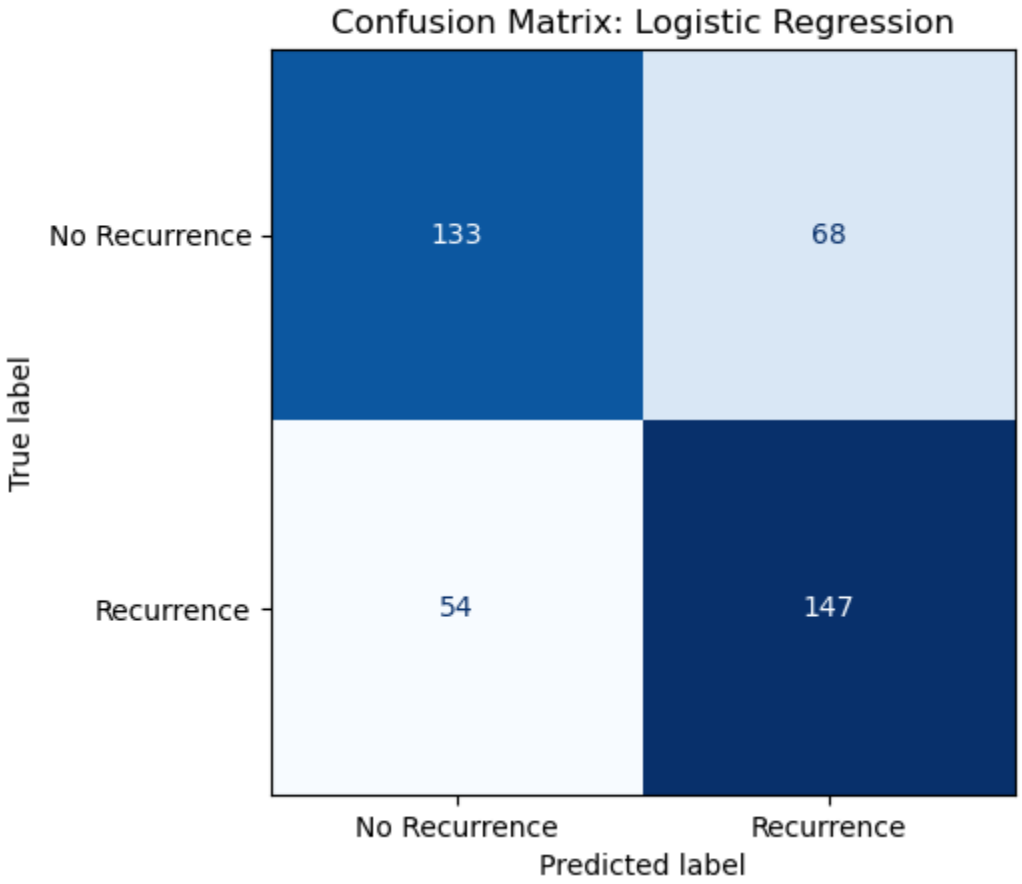



**Fig.6: Logistic Regression Confusion Matrix**

The confusion matrix for the logistic regression shows that it had a higher number of false positives and false negatives compared to the decision tree. The model misclassified 68 "no-recurrence-events" cases as "recurrence-events" and 54 "recurrence-events" cases as "no-recurrence-events".


**Fig.7: Logistic Regression Cross-Validation Results**

The Logistic Regression Cross-Validation Accuracy data reveals some inconsistency in the model's performance across the 10 folds, with accuracy ranging from 0.375 to 0.78. The mean cross-validation accuracy for the logistic regression was 0.6266, with a standard deviation of 0.1025, which is lower than the decision tree's performance. This suggests that the logistic regression model may not be as stable and consistent as the decision tree.

### 3.1.3 MLP Classifier Model Performance

The MLP Classifier performed very well, the output of this model is as follows:

```
Neural Network Classifier
Cross-Validation Accuracy Scores: [0.95121951 0.85365854 0.875      0.725      0.925      0.95
 0.8        0.9        0.85       0.75       ]
Mean Accuracy: 0.8580
Standard Deviation: 0.0752

Overall Performance (on the entire dataset):
Accuracy: 0.9104477611940298
Classification Report:
                     precision    recall  f1-score   support

no-recurrence-events       0.96      0.86      0.91       201
   recurrence-events       0.87      0.96      0.91       201

            accuracy                           0.91       402
           macro avg       0.91      0.91      0.91       402
        weighted avg       0.91      0.91      0.91       402

Confusion Matrix:
[[173  28]
 [  8 193]]
```

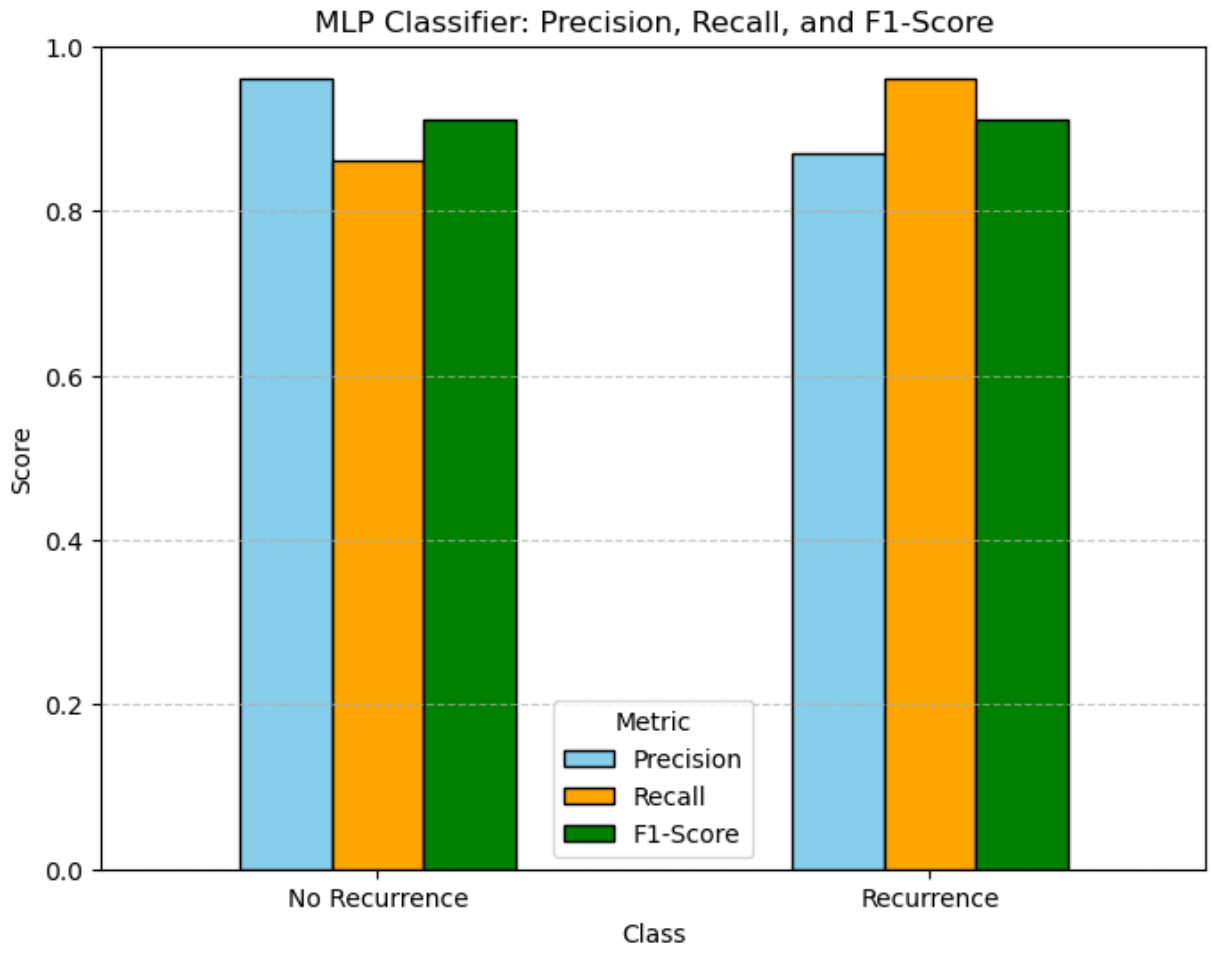



**Fig.8: MLP Classifier Classification Report**

Examining the classification report, the MLP Classifier had excellent precision, recall, and F1-score for both the "no-recurrence-events" and "recurrence-events" classes. The precision for the "no-recurrence-events" class was 0.96, and the recall was 0.86, indicating that the model was able to correctly identify the majority of the "no-recurrence-events" cases. For the "recurrence-events" class, the precision was 0.87, and the recall was 0.96, meaning the model was able to identify almost all of the "recurrence-events" cases correctly.

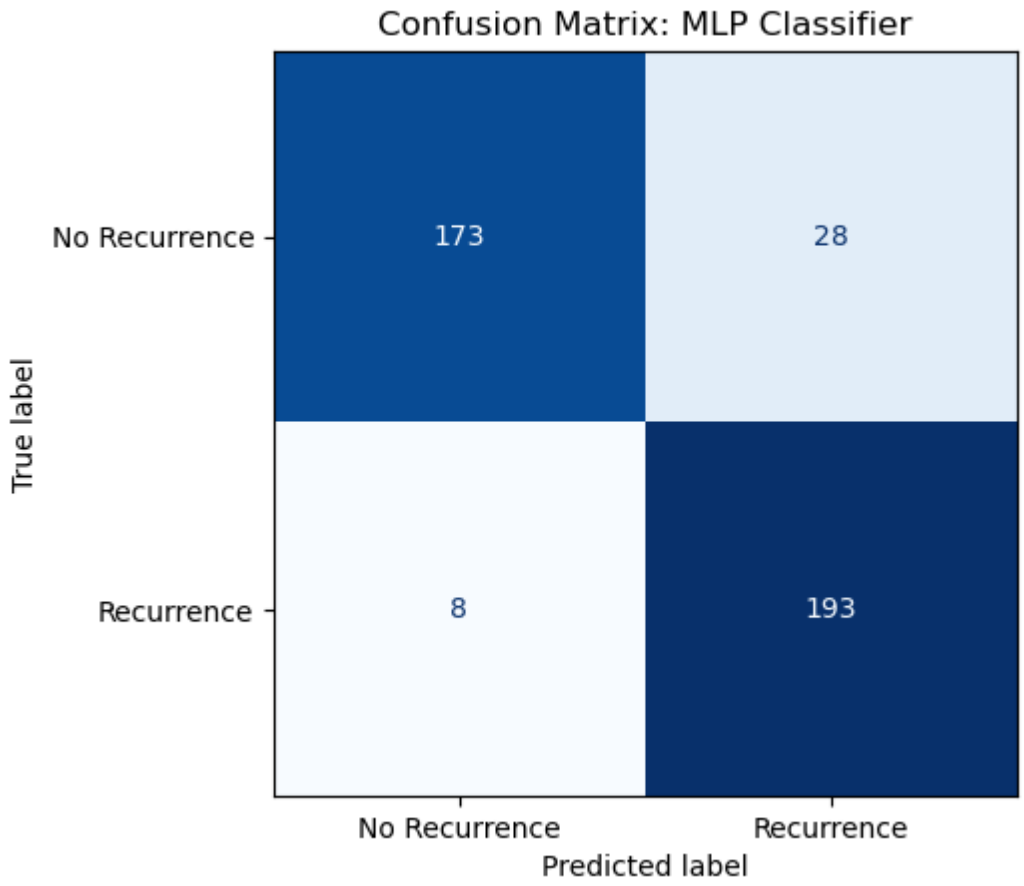



**Fig.9: MLP Classifier Confusion Matrix**

The confusion matrix for the MLP Classifier shows a very low number of false positives and false negatives, with 173 true positives for "no-recurrence-events" and 193 true positives for "recurrence-events". This aligns with the high precision and recall scores observed in the classification report.

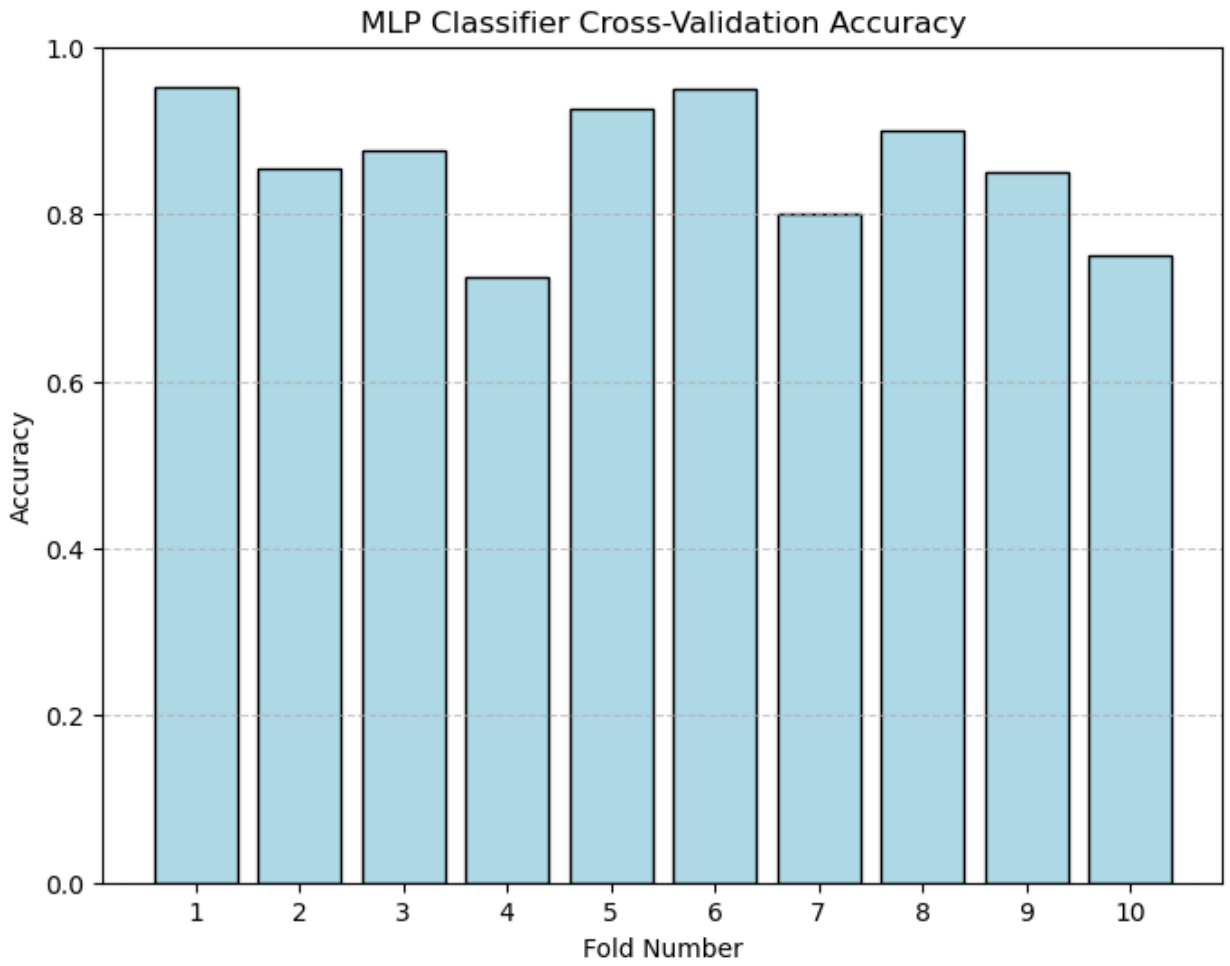



**Fig.10: MLP Classifier Cross-Validation Results**

The " MLP Classifier Cross-Validation Accuracy" plot demonstrates the model's consistent performance across the 10 folds, with a minimum accuracy of 0.725 and a maximum of 0.95 and with a mean cross-validation accuracy of 0.8580 and a standard deviation of 0.0752. This suggests that the MLP Classifier model is robust and stable in its predictions.

### 3.2 Comparative Model Analysis

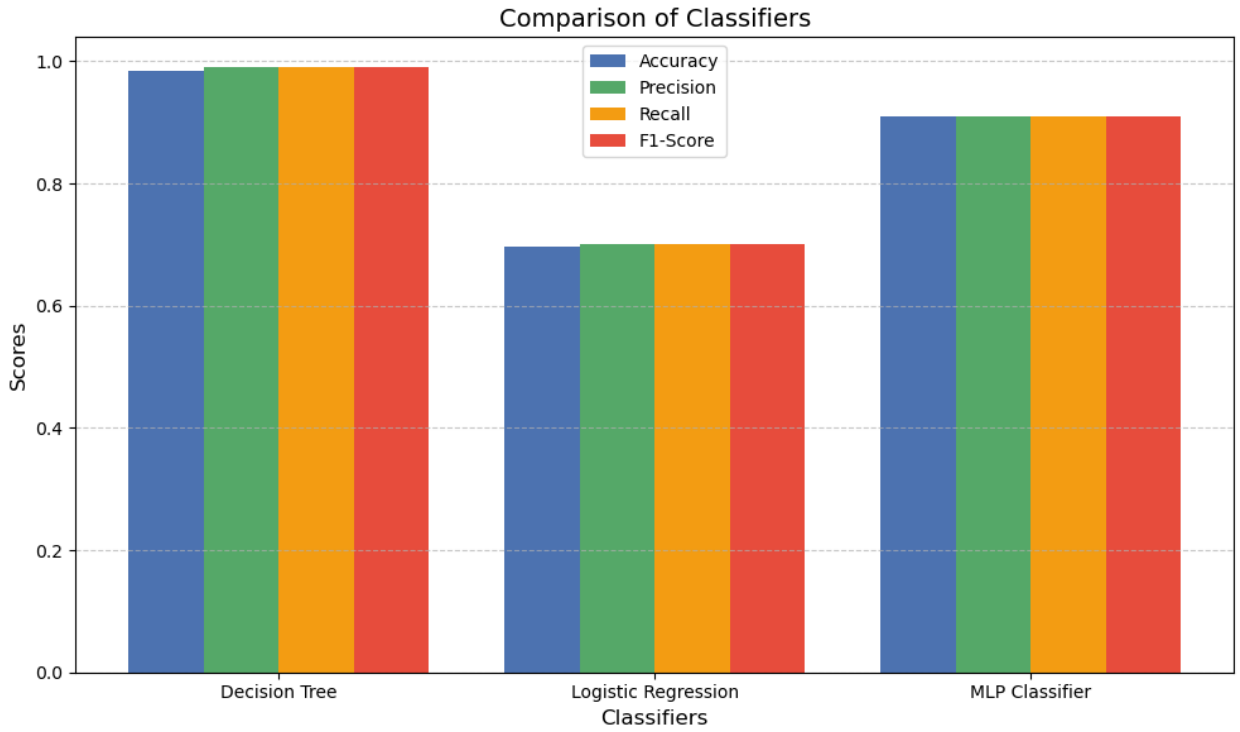



**Fig.11: Model Performance Comparison**

Fig. 11 provides a comparative analysis of the performance metrics for the three classification models implemented - Decision Tree, Logistic Regression, and MLP Classifier. The metrics include accuracy, precision, recall, and F1-score, which are visualized in a bar chart format.

The Decision Tree Classifier is shown to have the highest accuracy among the three models, reaching approximately 0.99. Its precision, recall, and F1-score are also the highest, indicating a well-balanced performance across both the majority and minority classes.

The Logistic Regression Classifier, on the other hand, demonstrates the lowest overall performance, with accuracy, precision, recall, and F1-score all below 0.75. This suggests the Logistic Regression model struggled to capture the complexities within the dataset effectively.

The MLP Classifier falls in the middle, with accuracy, precision, recall, and F1-score ranging from 0.85 to 0.91. While not as high as the Decision Tree, the MLP Classifier still exhibits strong performance, particularly in its ability to learn intricate patterns and relationships within the data.

The comparative analysis provides a clear visual representation of the trade-offs between the three classification methods. It highlights the Decision Tree Classifier as the top performer, the Logistic Regression Classifier as the weakest, and the MLP Classifier as a capable middle ground between the two.

This information can inform the selection of the most appropriate classification model for the given problem, taking into account factors such as the dataset's complexity, the need for interpretability, and the desired balance between performance metrics.

## 4. Key takeaways and Future Work

### 4.1. Key Takeaways

The analysis of the three classification models has revealed several important insights. The Decision Tree classifier emerges as the most effective approach for predicting recurrence events, delivering high accuracy with minimal errors. This makes it a reliable choice for practical application in medical settings, where the ability to accurately identify high-risk patients is critical.

In contrast, while the MLP Classifier model also demonstrated strong overall performance, it presented a higher risk of false positives, which could lead to unnecessary treatments and associated impacts on patients. This trade-off between accuracy and precision should be carefully considered when selecting the most appropriate model for the specific requirements of the breast cancer diagnosis and treatment process.

The Logistic Regression model, on the other hand, performed the least effectively, with lower accuracy and precision compared to the other two classifiers. This suggests its limitations in capturing the complexities inherent in the dataset.

### 4.2. Future Work

Building on these findings, future improvements could focus on fine-tuning the MLP Classifier model to enhance its performance and address the false positive concerns. Exploring alternative architectures, hyperparameter optimization, or additional feature engineering techniques may help the MLP Classifier model achieve a better balance between accuracy and precision.

Additionally, testing other classification algorithms, such as ensemble methods or more advanced techniques, could provide further insights and potentially yield even stronger predictive capabilities. Incorporating domain-specific medical knowledge into the model development and feature selection processes may also lead to improved performance and practical relevance.

By continuing to refine the modeling approaches and exploring innovative solutions, the breast cancer recurrence prediction task can be tackled with even greater effectiveness, ultimately leading to more reliable and impactful results for medical professionals and patients.