Comprehensive Report on Breast Cancer Prediction Machine Learning Analysis

1. Introduction

Breast cancer is one of the most common cancers among women worldwide. Early detection is crucial for effective treatment and improving survival rates. This report provides a comprehensive analysis of various machine learning models used for breast cancer detection.

- Problem Statement: Predicting breast cancer status (malignant or benign) based on tumor features.
    - Objective: Develop a machine learning model to accurately classify tumors.

2. Data Overview

- Data Source: Breast Cancer Wisconsin (Diagnostic) Dataset (UCI Machine Learning
- Repository).
- Features: Radius, texture, perimeter, area, smoothness, compactness, concavity, etc.
- Target Variable: Diagnosis (Malignant or Benign).

3. Data Preprocessing

The dataset used for this analysis includes various features related to breast cancer tumors, such as radius, perimeter, area, compactness, concavity, and concave points, among others. The target variable is the diagnosis of the tumor (malignant or benign).

Feature Selection: Features highly correlated with the diagnosis were selected for model training. The correlation values are as follows:

- Radius Mean: 0.730029
- Perimeter Mean: 0.742636
- Area Mean: 0.708984
- Compactness Mean: 0.596534
- Concavity Mean: 0.696360
- Concave Points Mean: 0.776614
- Radius SE: 0.567134
- Perimeter SE: 0.556141
- Area SE: 0.548236
- Radius Worst: 0.776454
- Perimeter Worst: 0.782914

- Area Worst: 0.733825
- Compactness Worst: 0.590998
- Concavity Worst: 0.659610
- Concave Points Worst: 0.793566 .

4. **Exploratory Data Analysis:**

- Summary Statistics: Mean, median, min, max, etc., for each feature.
- Visualizations: Histograms and correlation matrix.

5. **Model Selection & Training:**

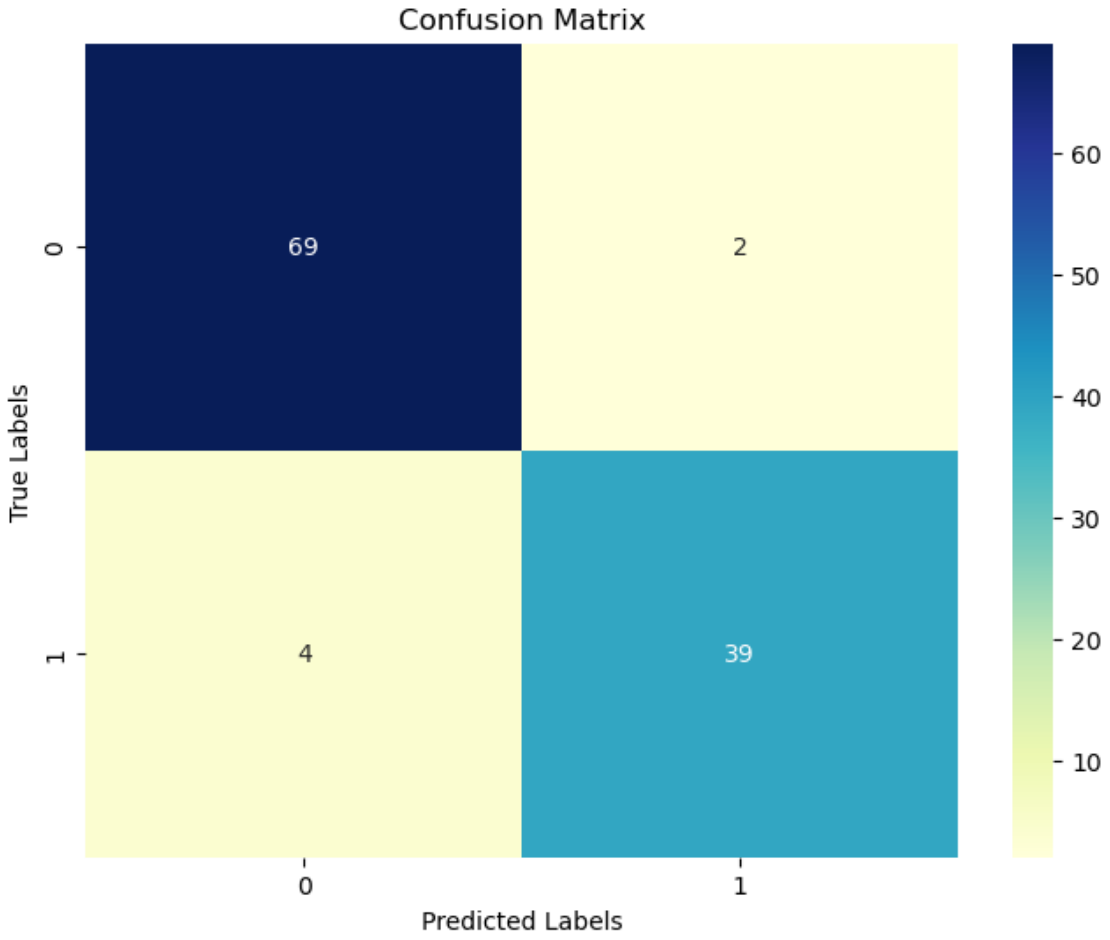
- Models Considered: Logistic Regression, Decision Trees, Random Forests, Naïve Bayes, MLP Classifier, Support Vector Classifier.
- Best Model: Logistic Regression (accuracy: 0.99).
- Data Splitting: The dataset was split into training and testing sets with an 80-20 ratio using the train_test_split method.

6. **Model Performance and Evaluation:**

Several machine learning models were trained and evaluated for their performance on the breast cancer detection task. The models include Random Forest Classifier, Decision Tree Classifier, KNeighbors Classifier, Logistic Regression, Support Vector Classifier, and Naive Bayes.

1. Random Forest Classifier:

- Accuracy: 94.74%
- Precision, Recall, F1-Score:
  - Benign (0): 0.95, 0.97, 0.96
  - Malignant (1): 0.95, 0.91, 0.93
- Confusion Matrix:



2. Decision Tree Classifier:

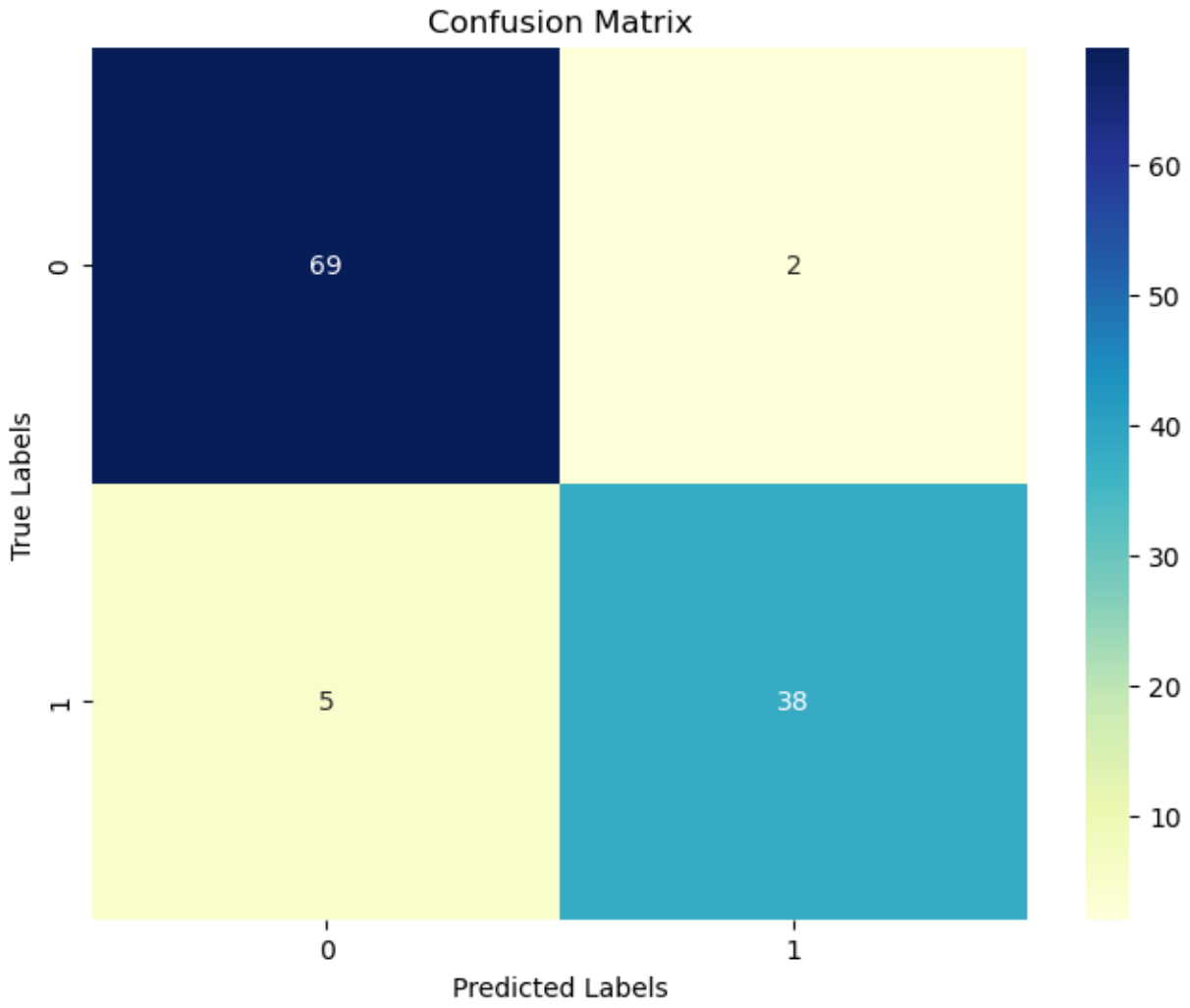
o Accuracy: 93.86%

o Precision, Recall, F1-Score:

 Benign (0): 0.93, 0.97, 0.95

 Malignant (1): 0.95, 0.88, 0.92

o Confusion Matrix:



3. KNeighbors Classifier:

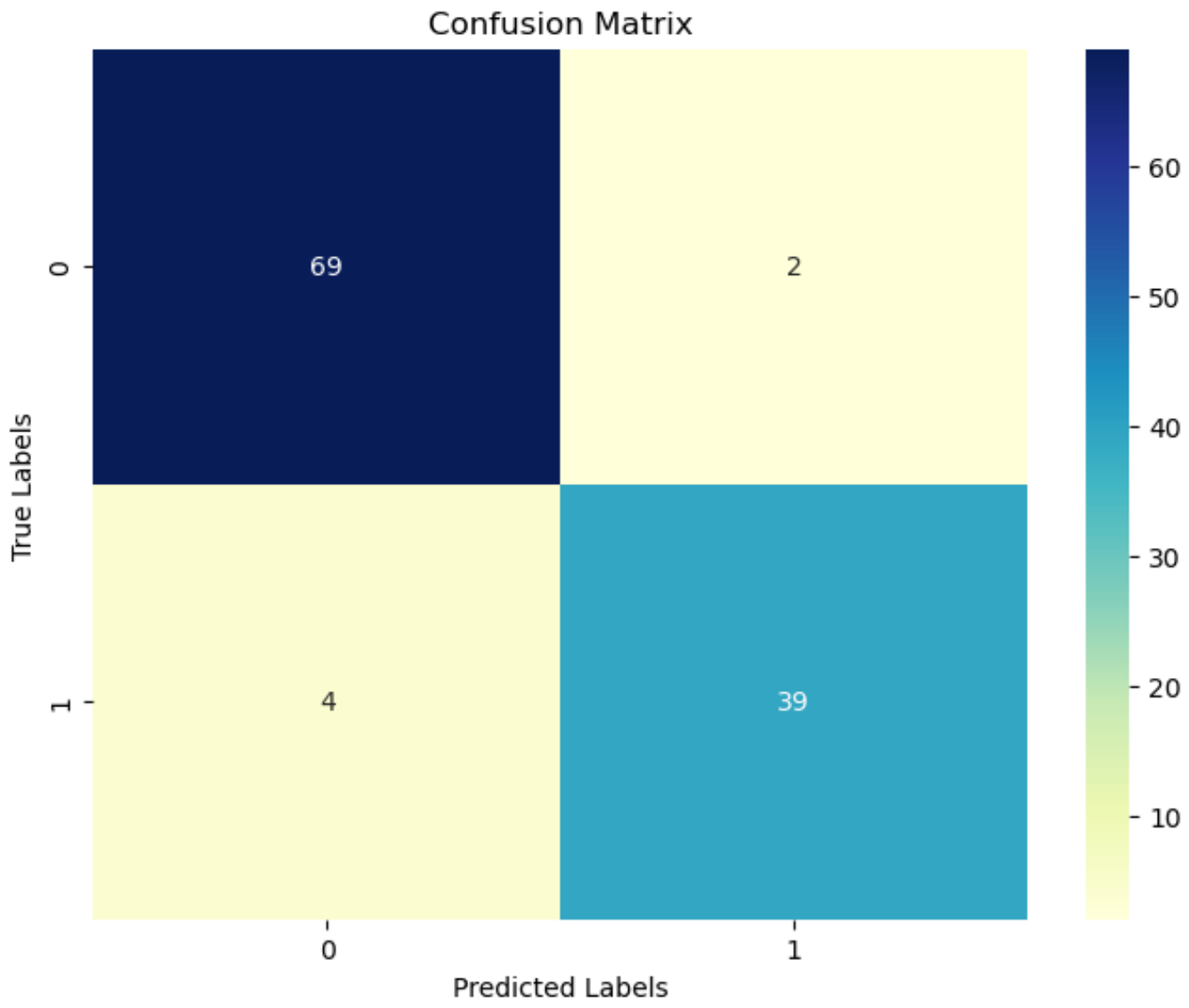
o Accuracy: 94.74%

o Precision, Recall, F1-Score:

 Benign (0): 0.93, 0.99, 0.96

 Malignant (1): 0.97, 0.88, 0.93

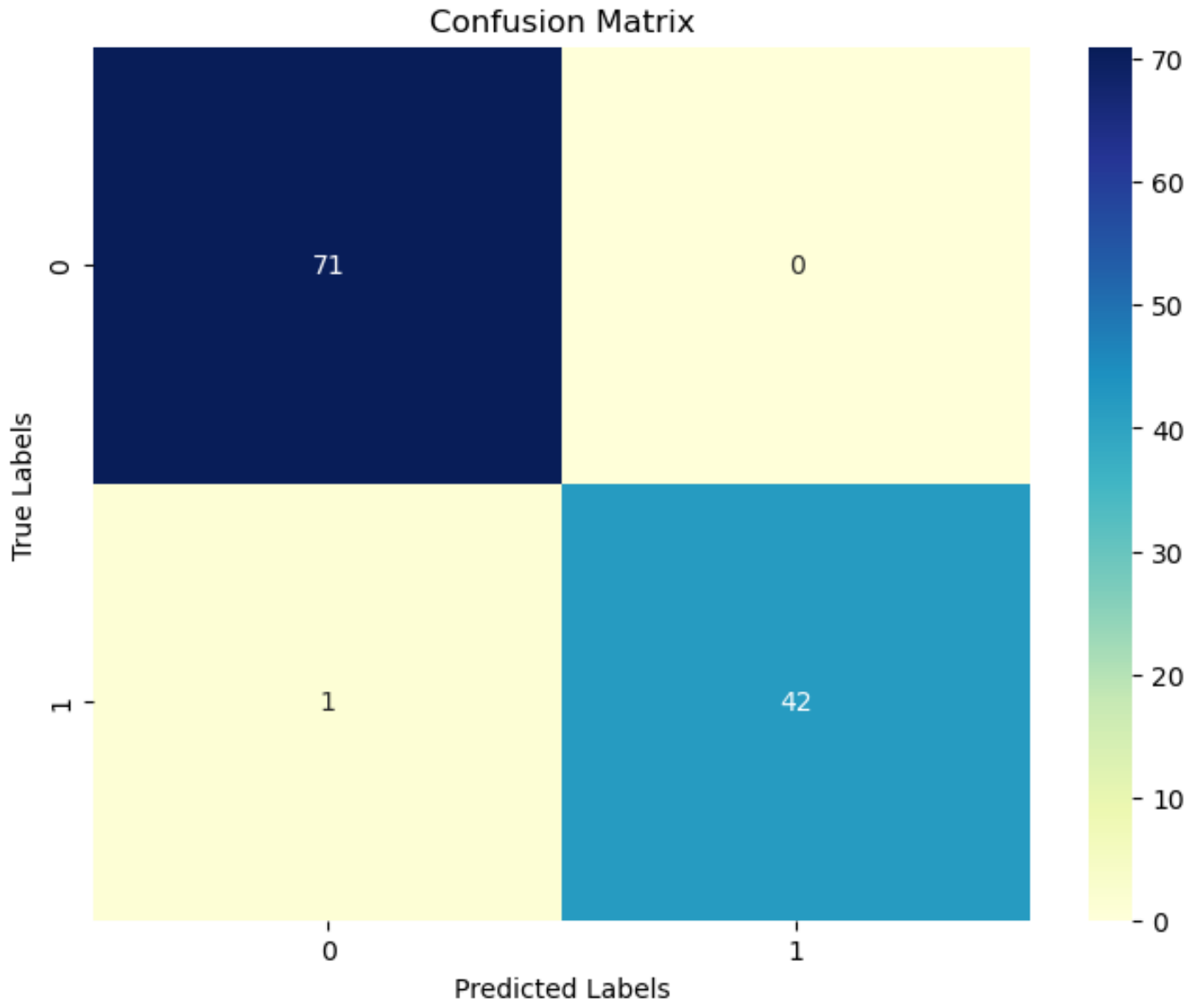
o Confusion Matrix:



4. Logistic Regression:

- Accuracy: 99.12%
- Precision, Recall, F1-Score:
- Benign (0): 0.99, 1.00, 0.99
- Malignant (1): 1.00, 0.98, 0.99

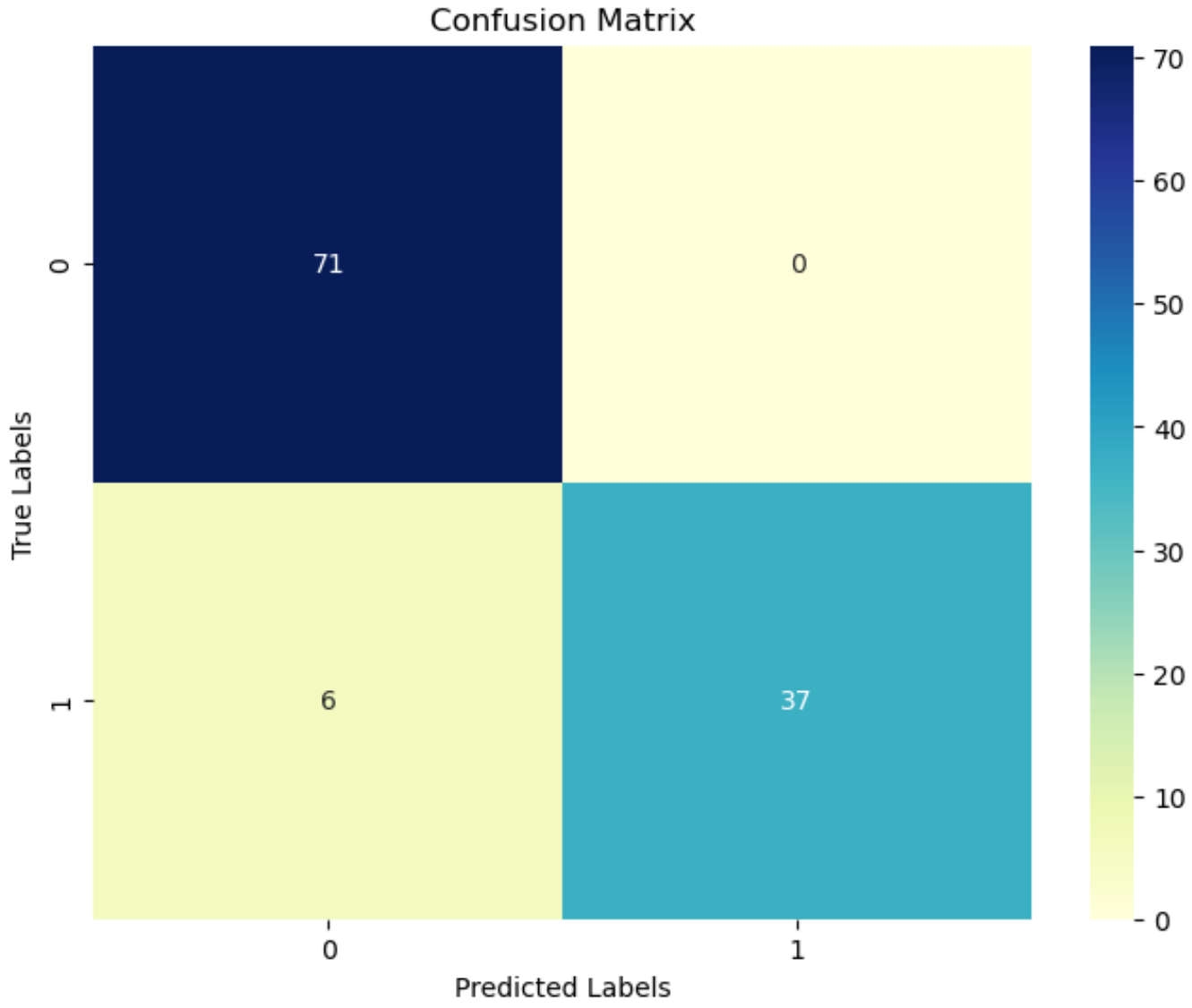
o Confusion Matrix:



5. Support Vector Classifier:

- Accuracy: 94.74%
- Precision, Recall, F1-Score:
- Benign (0): 0.92, 1.00, 0.96
- Malignant (1): 1.00, 0.86, 0.92

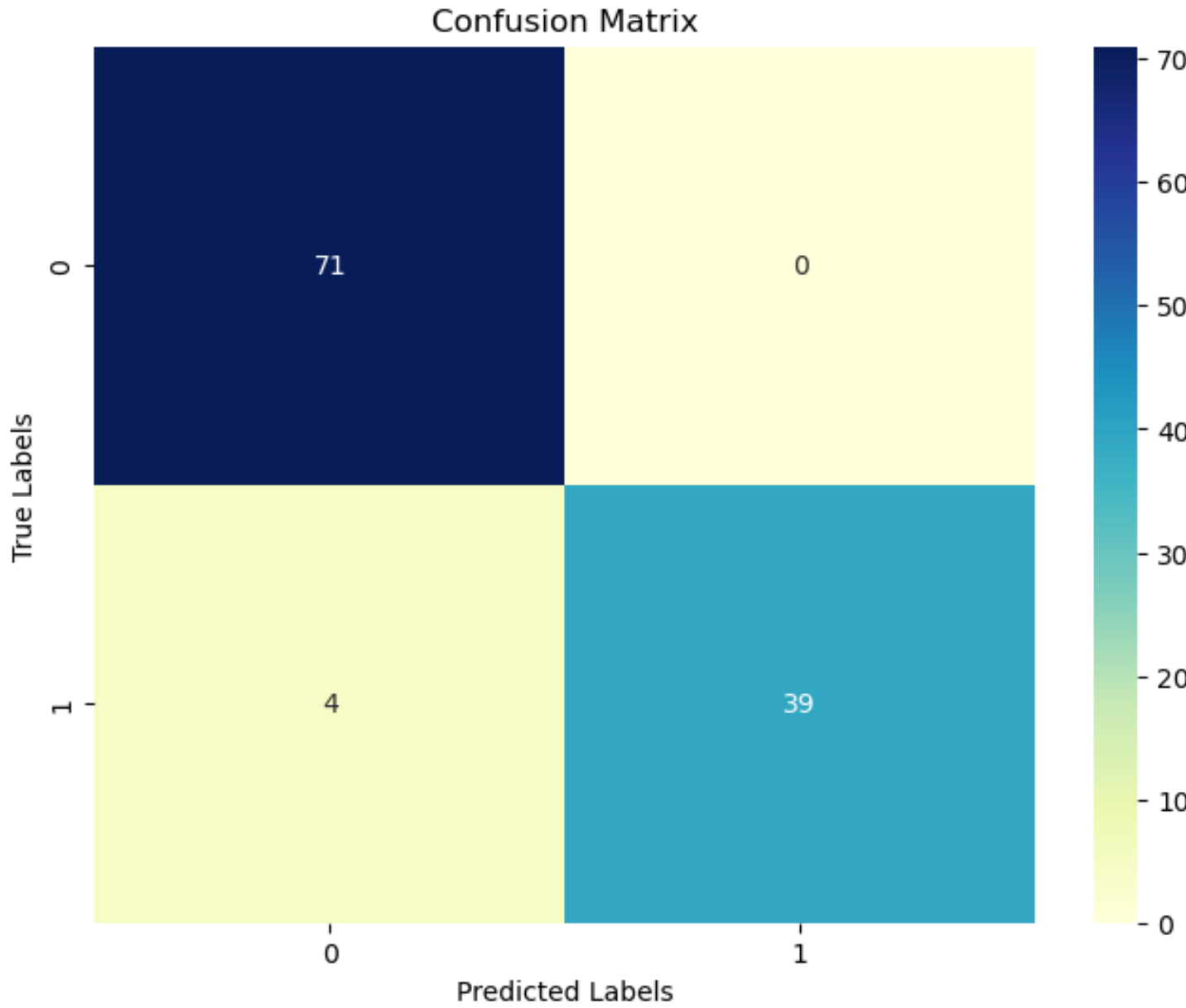
o Confusion Matrix:



6. Naive Bayes:

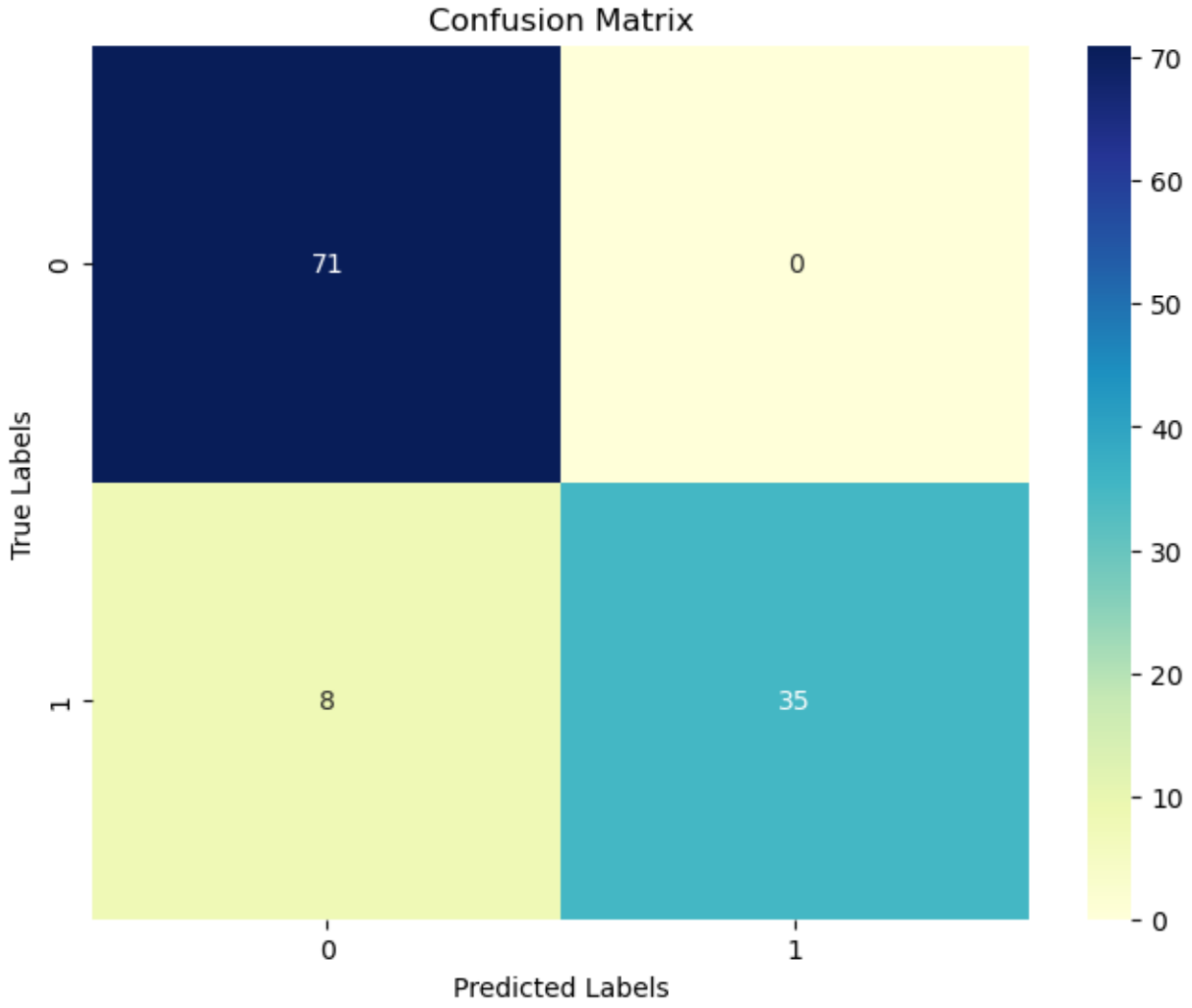
- Accuracy: 96.49%
- Precision, Recall, F1-Score:
- Benign (0): 0.95, 1.00, 0.97
- Malignant (1): 1.00, 0.91, 0.95

o Confusion Matrix:



7. MLP Classifier

- Accuracy: 92.98%
- Precision, Recall, F1-Score:
- Benign (0): 0.90, 1.00, 0.95
- Malignant (1): 1.00, 0.81, 0.90



7. **Model Interpretation:**
    - Feature Importance: Radius, concavity, and perimeter were most important.
    - Insights: Larger tumor size and irregular shape correlate with malignant tumors.

8. **Deployment Consideration:**
    - Scalability: Model performs efficiently on new data.
    - Interpretability: Provide probability scores for predictions.
    - Website Deployment: Streamlit was used to develop a website.

9. **Conclusion & Future Work**

The analysis shows that all the evaluated models performed well, with accuracy scores above 90%. However, Logistic Regression achieved the highest accuracy of 99.12%, making it the best-performing model for this dataset. The confusion matrices indicate that most models have high precision and recall for both benign and malignant tumors, with Logistic Regression showing the best balance between precision and recall.

The final model choice should consider not only the accuracy but also the interpretability and computational efficiency. Logistic Regression, being both interpretable and computationally efficient, is recommended for deployment in breast cancer detection systems.

9b. Future Work:

Future enhancements could include:

- **Feature Engineering**: Exploration of additional features or transformation techniques to improve model performance.
- **Hyperparameter Tuning**: Optimization of model parameters using techniques such as Grid Search or Random Search.
- **Cross-Validation**: Implementation of k-fold cross-validation to ensure robustness of results.
- **Deep Learning Models**: Exploration of advanced neural networks for potentially higher accuracy.

This comprehensive approach underscores the importance of machine learning in the early detection and treatment of breast cancer, ultimately contributing to better patient outcomes

10. **References:**

Dataset: Breast Cancer Wisconsin (Diagnostic) Dataset, UCI Machine Learning Repository.
• Libraries: Scikit-learn, Pandas, Matplotlib.