Lung Cancer Prediction Using Machine Learning

Objective:

The objective of this project is to compare various classification algorithms on a lung cancer dataset and identify which one performs best in predicting lung cancer.

Dataset:

The dataset used for this project is the **Lung Cancer Dataset**, which has been sourced from **data.world**. You can access it via the following link:

Survey Lung Cancer Dataset

Methodology:

For this project, we have employed **10 different classification algorithms**. Below is the list of classifiers used:

1. **Logistic Regression**
2. **K-Nearest Neighbors (KNN)**
3. **Decision Tree**
4. **Support Vector Machine (SVM)**
5. **Naive Bayes**
6. **Random Forest**
7. **Gradient Boosting**
8. **Neural Networks**
9. **AdaBoost**
10. **XGBoost**

For each classifier, we trained the model and evaluated its performance using the following **evaluation metrics**:

1. **Accuracy**
2. **Precision**
3. **F1 Score**
4. **Recall Score**
5. **Confusion Matrix**

Algorithm Performance:

Below are the accuracy scores achieved by each classification algorithm:

- **Logistic Regression**: 90.29%
- **K-Nearest Neighbors (KNN)**: 87.37%
- **Decision Tree**: 87.37%
- **Support Vector Machine (SVM)**: 84.46%
- **Naive Bayes**: 86.4%
- **Random Forest**: 89.32%
- **Gradient Boosting**: 89.32%
- **Neural Networks**: 84.46%
- **AdaBoost**: 84.46%
- **XGBoost**: 84.46%

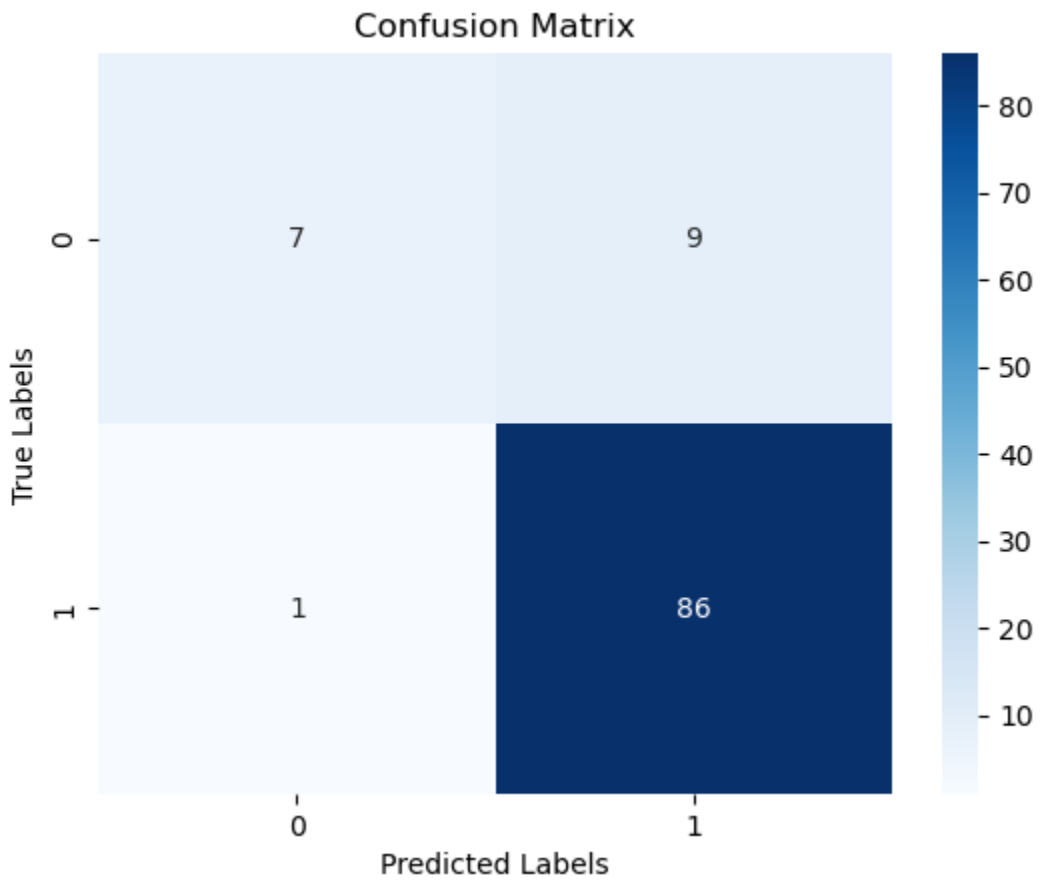
Conclusion:

Among all the algorithms tested, **Logistic Regression** yielded the highest performance with the following evaluation metrics:

- **Accuracy**: 0.9029
- **Precision**: 0.9053
- **Recall**: 0.9885
- **F1 Score**: 0.9451

Confusion Matrix:

The confusion matrix for Logistic Regression is as follows:



In summary, **Logistic Regression** proved to be the most effective classifier for this dataset, achieving the highest accuracy and performing well across other metrics as well.