Implementing and Evaluating Logistic Regression Algorithm

1. Introduction

This project focused on building and evaluating a machine-learning classification model using the Iris dataset. The primary goal was to implement a classification algorithm and understand the relationship between different training set sizes and test accuracy, visualized through a learning curve. Additionally, a confusion matrix was generated to assess the model's performance in classifying different Iris species.

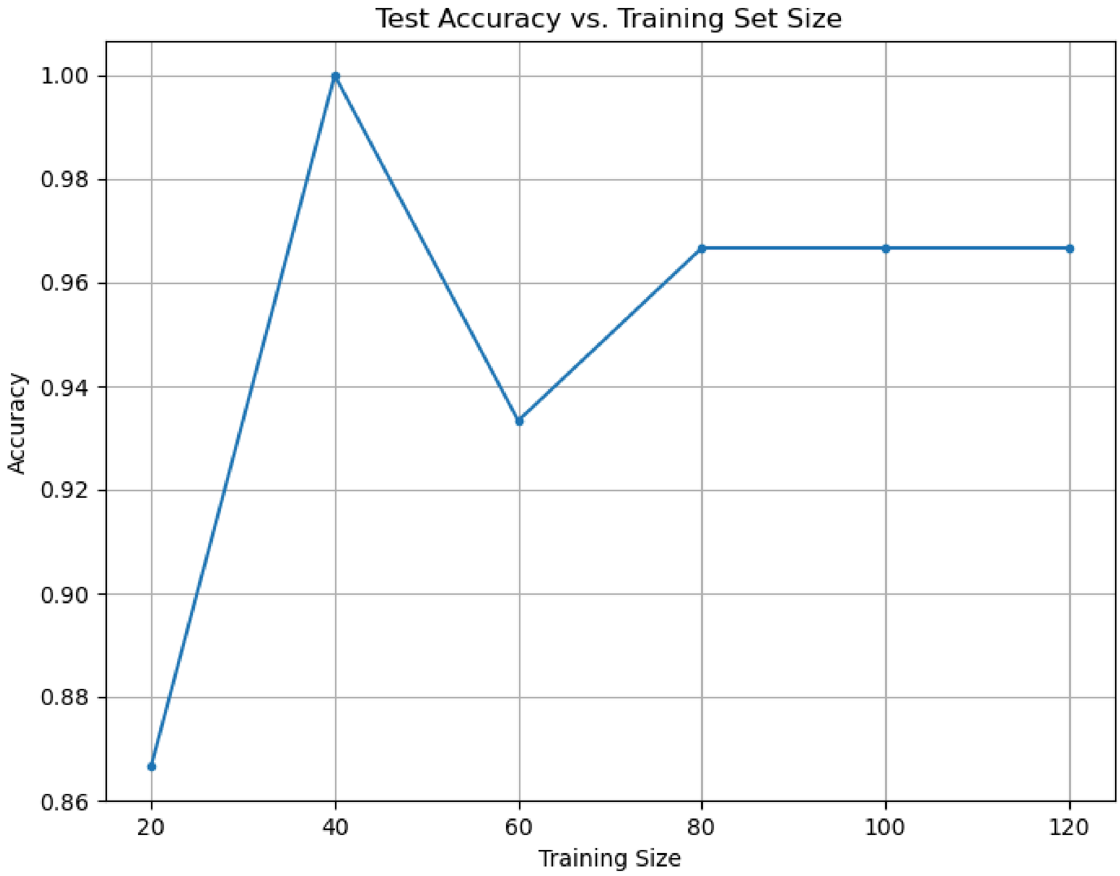
2. Methodology

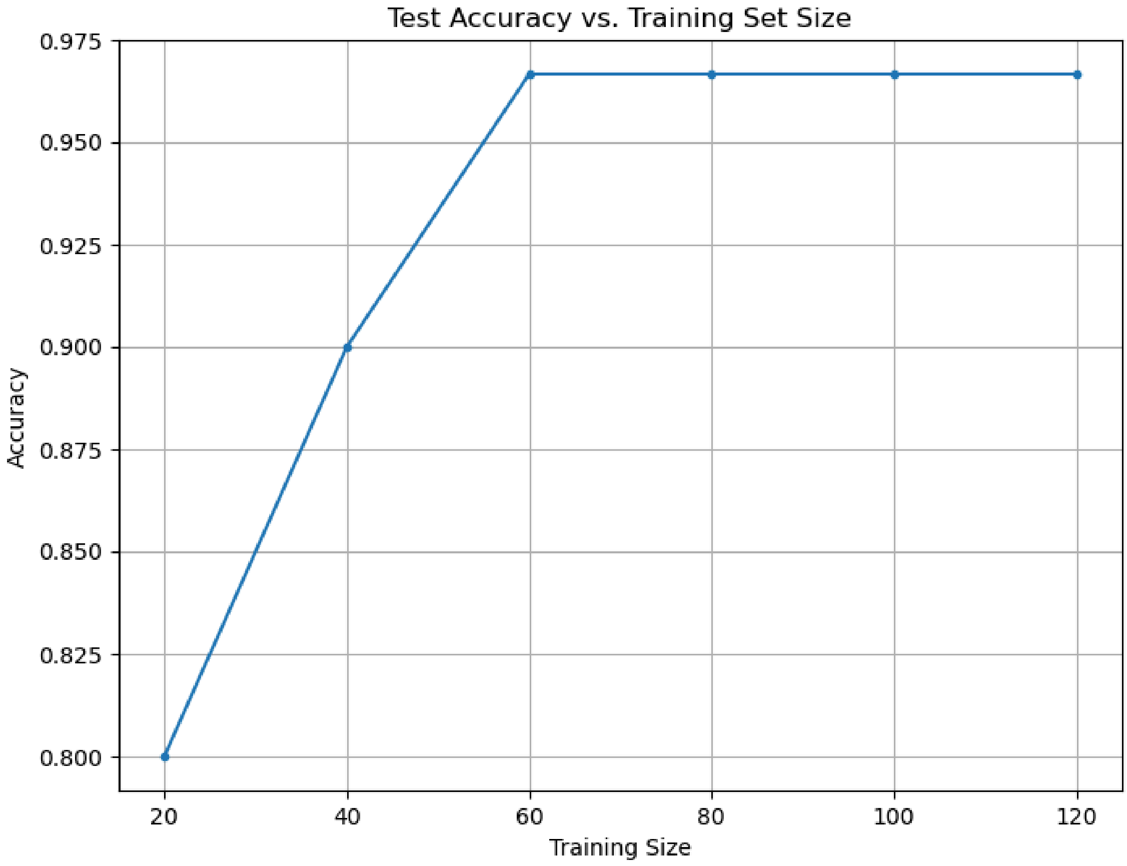
- Dataset: The Iris dataset was used, containing measurements of sepal and petal lengths and widths for three Iris species (setosa, versicolor, and virginica).
- Model: A logistic regression model was trained and evaluated.
- Training and Evaluation: The dataset was split into 80% training and 20% testing sets, ensuring a stratified split to maintain class proportions. The model was trained on varying sizes of the training set (20, 40, 60, 80, 100, and 120 samples) and evaluated on the fixed 20% test set.
- Confusion Matrix: A confusion matrix was generated to visualize the model's classification performance on the full training set.

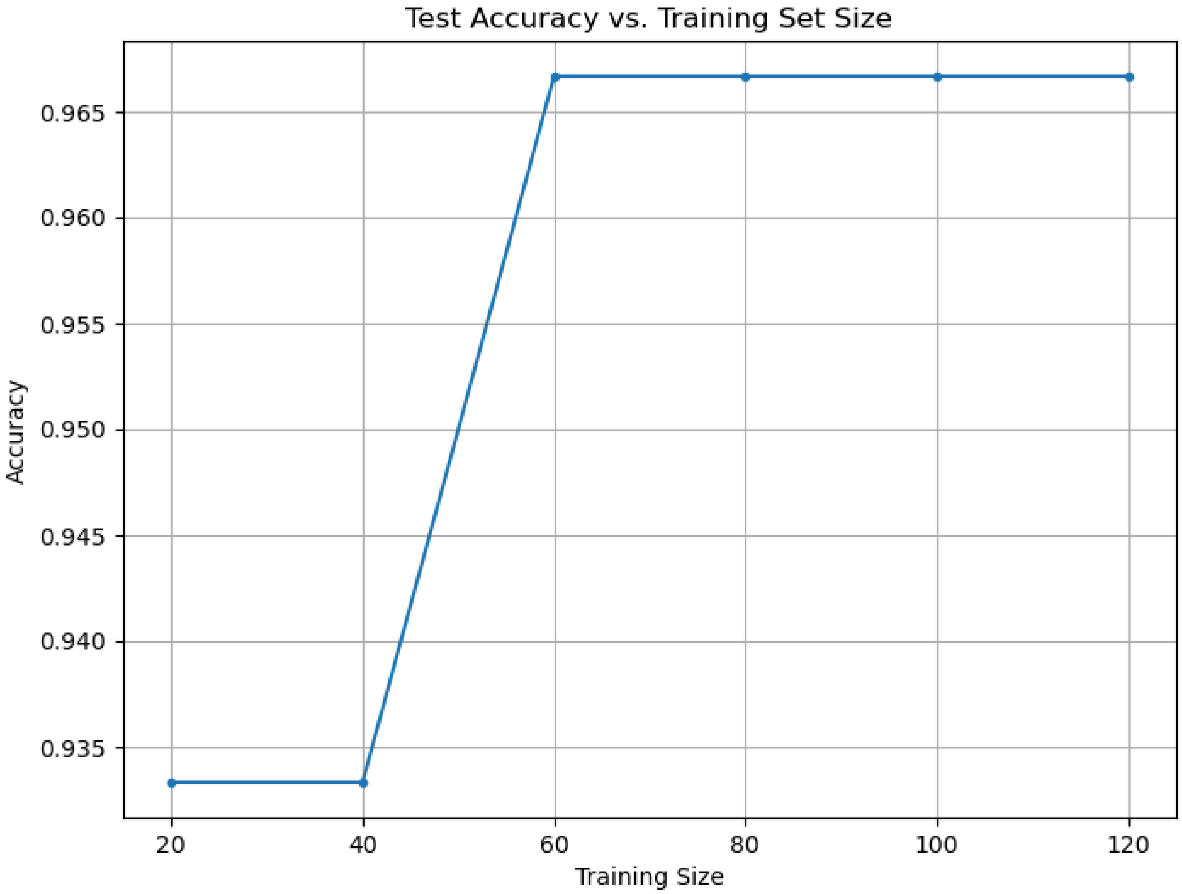
3. Results

Learning curves:

Different learning curves were obtained when the program was run each time. Some instances had very similar curves while some had different patterns. Nevertheless, they all reach high accuracy when the highest data size 120 is used for training.

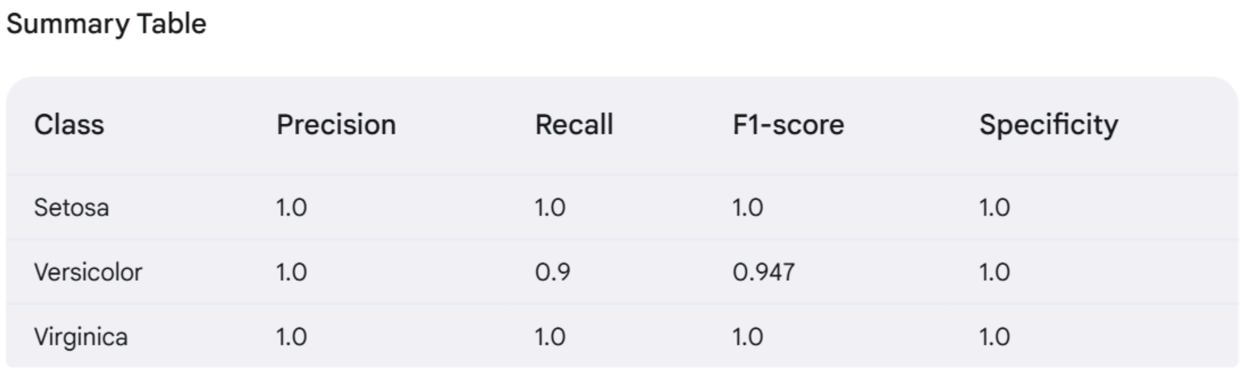

Test Accuracy vs. Training Set Size
Accuracy
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
20
40
60
80
100
120
Training Size


Test Accuracy vs. Training Set Size
Accuracy
0.975
0.950
0.925
0.900
0.875
0.850
0.825
0.800
20
40
60
80
100
120
Training Size



4. Analysis

- **Learning Curve:** The learning curve demonstrates a general trend of increasing test accuracy with larger training set sizes. When the program was run multiple times, the learning curve was slightly different for each run but most of the time appeared to converge around 96-97% accuracy with 120 training samples.
- **Confusion Matrix:** The confusion matrix indicates that the model performs well in classifying the different Iris species, with high accuracy across all classes. There is a slight confusion between versicolor and virginica but overall good performance.



Summary Table

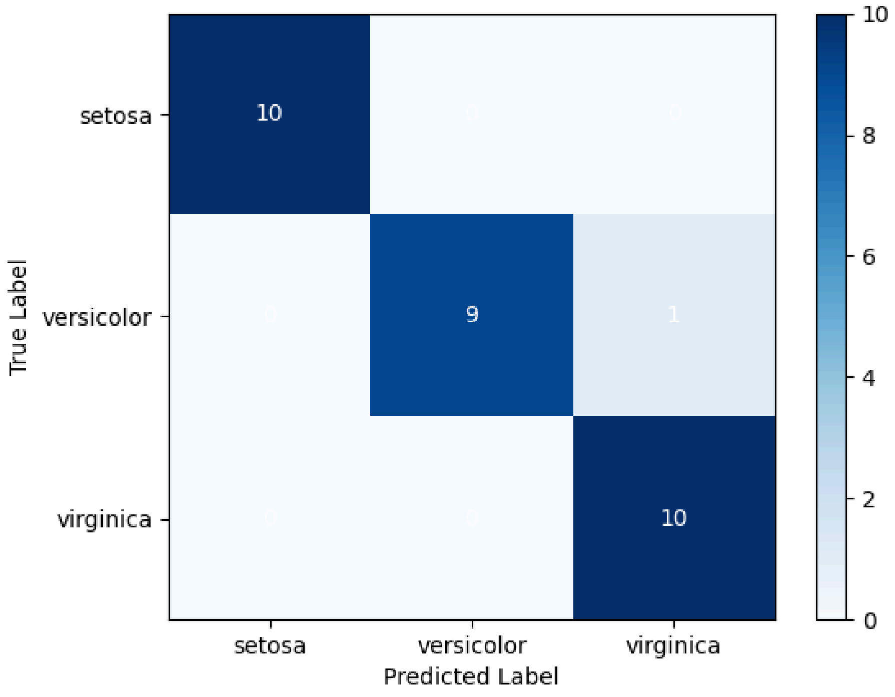
Class	Precision	Recall	F1-score	Specificity
Setosa	1.0	1.0	1.0	1.0
Versicolor	1.0	0.9	0.947	1.0
Virginica	1.0	1.0	1.0	1.0

High Scores: The model performs exceptionally well, with perfect precision and recall for setosa and virginica. Versicolor has a slightly lower recall, indicating that one actual versicolor flower was misclassified as virginica.

Balanced Performance: The F1-scores are very high overall, showing a good balance between precision and recall for all classes.

Specificity: The specificity is perfect for all classes, meaning that the model is excellent at correctly identifying flowers that do *not* belong to a particular class.

Confusion Matrix:



5. Conclusion

This project provided valuable insights into the impact of training set size on model performance. The learning curve effectively visualized the relationship between training data and accuracy, while the confusion matrix offered a detailed view of the model's classification strengths and weaknesses. This hands-on learning experience solidified my understanding of key classification concepts and evaluation matrices. Working with the Iris dataset and visualizing the learning curve and confusion matrix was particularly helpful. If I were to further improve on this work, I would explore techniques to address the slight confusion between versicolor and virginica, potentially by adding more features or a higher volume of observations where classes are fairly proportional or by using a more complex model.