

Breast Cancer Classification Using Logistic Regression

Avril Luo

Problem & Data

Problem

1. Predict whether a breast tumor is malignant or benign
2. Binary classification task

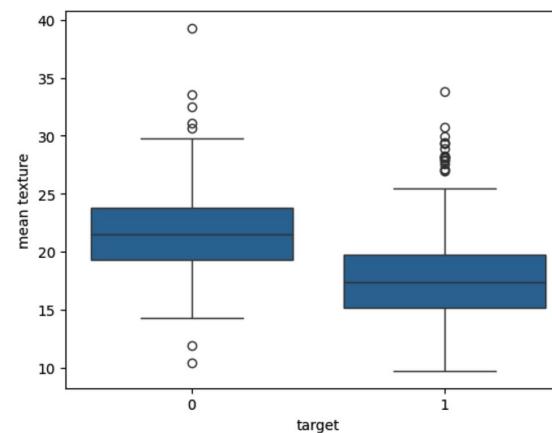
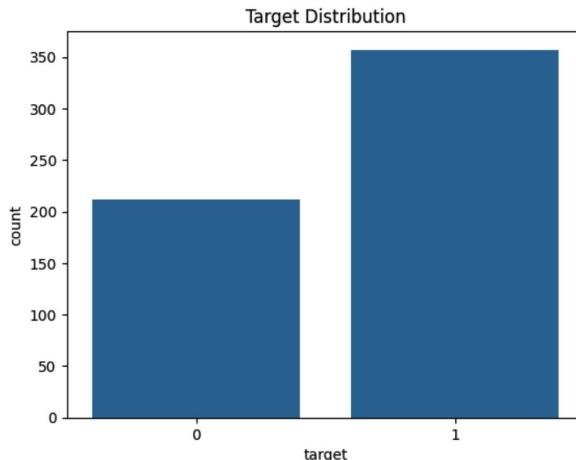
Data

1. Source: `sklearn.datasets.load_breast_cancer`
2. Samples: 569
3. Features: 30 numerical features
4. Target: Malignant vs Benign

Exploratory Data Analysis (EDA)

Key EDA Findings

1. Target variable is relatively balanced
2. Features such as mean radius and mean texture show clear differences between classes
3. No missing values in the dataset



Model & Method

Model: Logistic Regression (baseline and final model)

Method

1. Train / test split (80% / 20%)
2. Evaluation metrics: Accuracy & Confusion Matrix

Why Logistic Regression?

1. Simple and interpretable
2. Well-suited for binary classification

Results

Model Performance

1. **Test Accuracy: ~95.6%**
2. **Most tumors were correctly classified**
3. **Low number of false negatives**
4. **Strong overall performance**

```
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	0.97	0.91	0.94	43
1	0.95	0.99	0.97	71
accuracy			0.96	114
macro avg	0.96	0.95	0.95	114
weighted avg	0.96	0.96	0.96	114

```
confusion_matrix(y_test, y_pred)
```

```
array([[39,  4],  
       [ 1, 70]])
```

Conclusion & Next Steps

Conclusion

1. Logistic Regression performs well on this dataset
2. Simple models can achieve strong performance on structured data

Next Steps

1. Feature selection
2. Hyperparameter tuning
3. Try ensemble models (e.g. Random Forest)