

Machine Learning Assignment – 2 Report

Uday Bhaskar Valapadasu

11696364

The study used logistic regression to predict diabetes on a dataset with 768 samples and 9 features. Three data splits were tested: 80-20, 70-30, and 60-40 (train-test).

Key Findings:

1. The 80-20 and 60-40 splits both achieved the highest accuracy of 75.32%.
2. The 80-20 split was selected for further analysis.
3. For the 80-20 split:
 - a. Accuracy: 75.32%
 - b. AUC: 0.8165
 - c. Confusion Matrix: $\begin{bmatrix} 80 & 19 \\ 19 & 36 \end{bmatrix}$

Bootstrap Analysis (80-20 split):

- 1000 iterations performed
- P-value: 0.51
- 95% Confidence Interval: [0.708, 0.792]

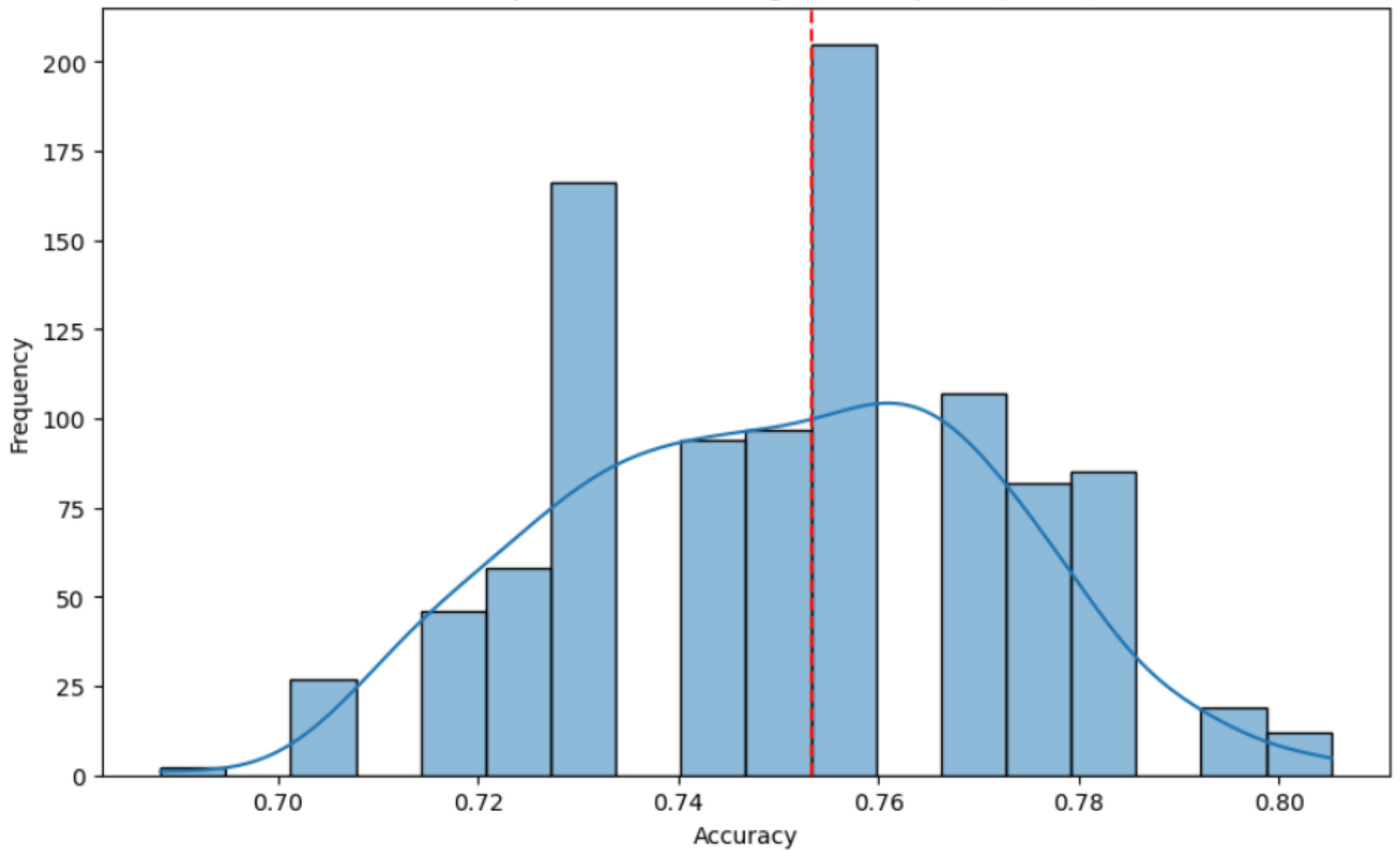
The model shows good predictive performance with an AUC of 0.8165, indicating strong discrimination ability. The bootstrap analysis suggests the accuracy estimate is robust, with a narrow confidence interval.

```
Confusion Matrix and Accuracy for Split-1 Training:Test(80:20)
Confusion Matrix:
[[80 19]
 [19 36]]
Accuracy: 0.7532467532467533
AUC: 0.8165289256198347
```

```
Confusion Matrix, Accuracy, and AUC for Split-2 Training:Test(70:30)
Confusion Matrix:
[[121 30]
 [ 30 50]]
Accuracy: 0.7402597402597403
AUC: 0.7964403973509934
```

```
Confusion Matrix, Accuracy, and AUC for Split-3 Training:Test(60:40) Ratio
Confusion Matrix:
[[166 40]
 [ 36 66]]
Accuracy: 0.7532467532467533
AUC: 0.8213401865600609
```

Bootstrap Accuracies Histogram for Split-1 (80:20)



Selected Data Split Ratio (80:20) Metrics:

Confusion Matrix:

[[80 19]

[19 36]]

Accuracy: 0.7532467532467533

AUC: 0.8165289256198347

ROC AUC: 0.8165289256198347

ROC Thresholds: [1.97062516 0.97062516 0.96040546 0.87284274 0.87075569 0.78031972

0.77299541 0.77209963 0.76866991 0.72395837 0.70889906 0.70883013

0.69285014 0.66792411 0.66001388 0.65390286 0.65205796 0.63552015

0.62281485 0.61803155 0.57749256 0.56826283 0.55545133 0.54548516

0.5319286 0.5140412 0.49408098 0.49150511 0.4436292 0.42606807

0.40481265 0.37964985 0.36247904 0.327527 0.28621408 0.27489661

0.26885186 0.26091845 0.26086909 0.25787497 0.23546844 0.23035508

0.21334583 0.20356076 0.15930276 0.15109309 0.1416267 0.14056454

0.13153163 0.12834498 0.12069508 0.11672998 0.04015114 0.0384741

0.00415931]

Bootstrap Analysis for Split-1:

P-Value: 0.51

Confidence Interval: [0.70779221 0.79220779]