CSE472 (Machine Learning Sessional)
Assignment 1: Logistic Regression and AdaBoost for Classification

Asif Imtial
1605017

Instructions for Running Script:

Script Name: 1605017.py

To preprocess 1st Dataset: Uncomment line 275:

```
X_train, X_test, y_train, y_test = preprocess_1()
```

To preprocess 2nd Dataset: Uncomment line 276:

```
X_train, X_test, y_train, y_test = preprocess_2()
```

To preprocess 3rd Dataset: Uncomment line 277:

```
X_train, X_test, y_train, y_test = preprocess_3()
```

To run logistic regression model: Uncomment line 279:

```
run_logistic(X_train, y_train, X_test, y_test)
```

To run adaboost model: Uncomment line 281-282:

```
run_adaboost(X_train, y_train, X_test, y_test, K)
```

Set value of K before calling the function

The Performance Evaluation (Logistic Regression):

Dataset-1: `WA_Fn-UseC_-Telco-Customer-Churn.csv`

Performance measure	Training	Test
Accuracy	0.78576	0.79418
True positive rate (sensitivity, recall, hit rate)	0.66337	0.63505
True negative rate (specificity)	0.83102	0.84637
Positive predictive value (precision)	0.59213	0.57552
False discovery rate	0.40786	0.42447
F1 score	0.62573	0.60382

Dataset-2: adult.data

Performance measure	Training	Test
Accuracy	0.84229	0.84336
True positive rate (sensitivity, recall, hit rate)	0.54839	0.54576
True negative rate (specificity)	0.93551	0.93541
Positive predictive value (precision)	0.72955	0.72329
False discovery rate	0.27044	0.27670
F1 score	0.62613	0.62211

Dataset-3: `creditcard.csv`

Performance measure	Training	Test
Accuracy	0.98129	0.97379
True positive rate (sensitivity, recall, hit rate)	0.58244	0.54310
True negative rate (specificity)	1.0	0.99899
Positive predictive value (precision)	1.0	0.96923
False discovery rate	0.0	0.03076
F1 score	0.73613	0.69613

The Performance Evaluation (Adaboost):

Dataset-1: `WA_Fn-UseC_-Telco-Customer-Churn.csv`

Number of boosting rounds	Training	Test
5	0.80120	0.79772
10	0.80014	0.80056
15	0.79197	0.79772
20	0.79215	0.79630

Dataset-2: adult.data

Number of boosting rounds	Training	Test
5	0.83971	0.83697
10	0.84106	0.83875
15	0.83897	0.83789
20	0.83015	0.82770

Dataset-3: `creditcard.csv`

Number of boosting rounds	Training	Test
5	0.98045	0.97189
10	0.98034	0.97189
15	0.98045	0.97189
20	0.98010	0.97093

Observations:

- Logistic regression is itself a strong classifier. So, Adaboost on logistic regression does not provide any significant accuracy over the accuracy of logistic regression.
- Early stopping causes overall bad performance.
- Third dataset is very biased. That's why the higher accuracy rate is very deceptive as it's performance on predicting positive label is very poor.