Day 52

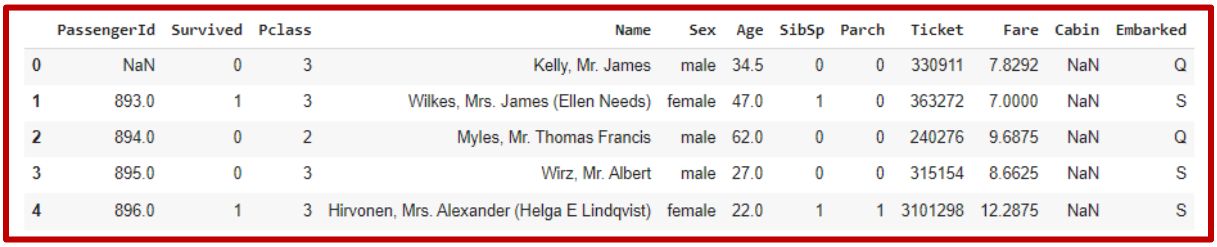
DIY

Q1. Problem Statement: K-Fold Cross-Validation

Load the *'content/titanic.csv'* dataset into a DataFrame and perform the following tasks:

1. Identify the null values and remove the null rows and columns by using the `dropna()` function
2. Considering the *'Survived'* column as the target, separate the target variable from the independent variables
3. Select only the numeric columns from the input variables
4. Split the data into five folds using `KFold()` function
5. Build a decision tree classifier model and print model accuracies for all the data folds
6. Find the accuracies of the model for all the folds using a cross validator and compare the accuracies with the model accuracies

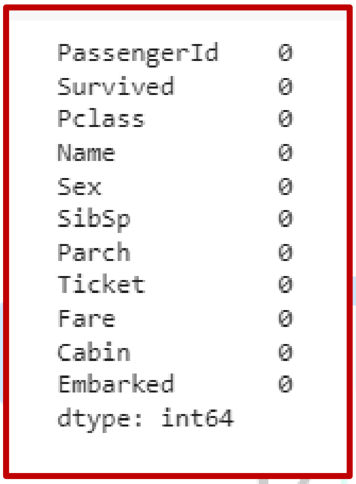
Dataset:



	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	NaN	0	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	Q
1	893.0	1	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	S
2	894.0	0	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	Q
3	895.0	0	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	S
4	896.0	1	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	S

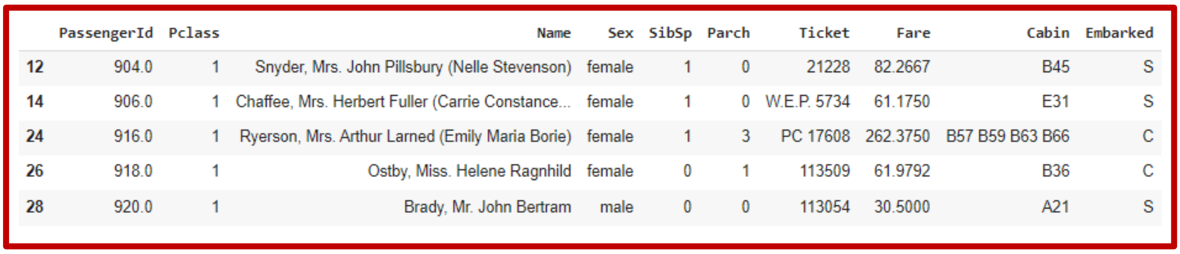
Sample Output:

1. Identify the null values and remove the null rows and columns by using the `dropna()` function

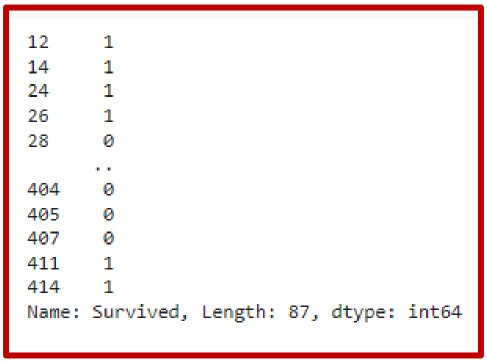


```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Cabin          0
Embarked       0
dtype: int64
```

2. Considering the *'Survived'* column as the target, separate the target variable from the independent variables

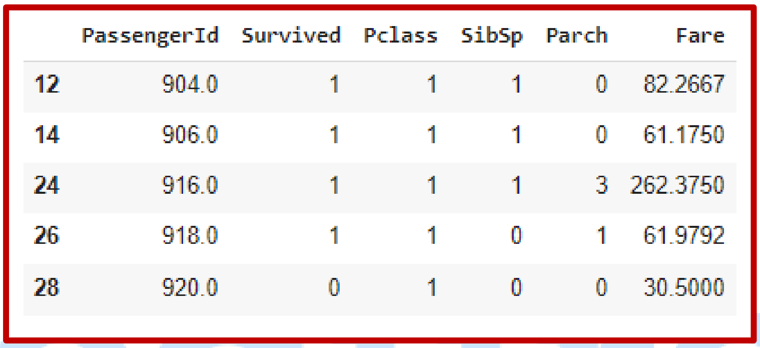


	PassengerId	Pclass	Name	Sex	SibSp	Parch	Ticket	Fare	Cabin	Embarked
12	904.0	1	Snyder, Mrs. John Pillsbury (Nelle Stevenson)	female	1	0	21228	82.2667	B45	S
14	906.0	1	Chaffee, Mrs. Herbert Fuller (Carrie Constance...	female	1	0	W.E.P. 5734	61.1750	E31	S
24	916.0	1	Ryerson, Mrs. Arthur Larned (Emily Maria Borie)	female	1	3	PC 17608	262.3750	B57 B59 B63 B66	C
26	918.0	1	Ostby, Miss. Helene Ragnhild	female	0	1	113509	61.9792	B36	C
28	920.0	1	Brady, Mr. John Bertram	male	0	0	113054	30.5000	A21	S



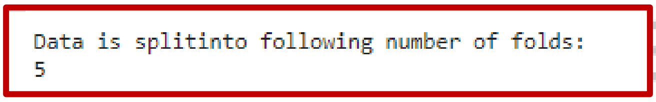
```
12     1
14     1
24     1
26     1
28     0
      ..
404    0
405    0
407    0
411    1
414    1
Name: Survived, Length: 87, dtype: int64
```

3. Select only the numeric columns from the input variables



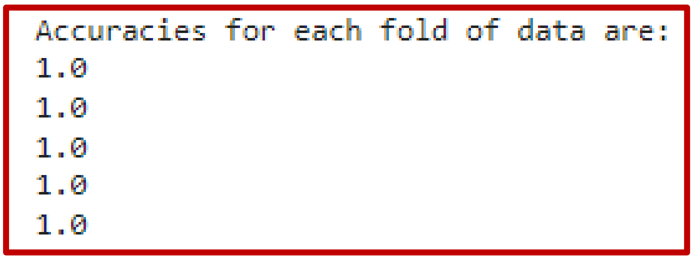
	PassengerId	Survived	Pclass	SibSp	Parch	Fare
12	904.0	1	1	1	0	82.2667
14	906.0	1	1	1	0	61.1750
24	916.0	1	1	1	3	262.3750
26	918.0	1	1	0	1	61.9792
28	920.0	0	1	0	0	30.5000

4. Split the data into five folds using `KFold()` function



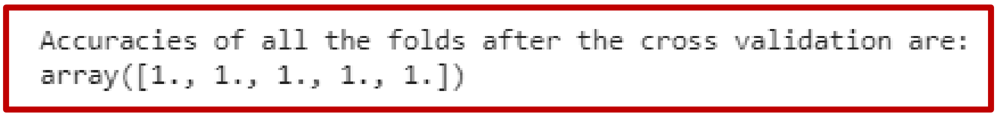
```
Data is splitinto following number of folds:
5
```

5. Build a decision tree classifier model and print model accuracies for all the data folds



```
Accuracies for each fold of data are:
1.0
1.0
1.0
1.0
1.0
```

6. Find the accuracies of the model for all the folds using a cross validator and compare the accuracies with the model accuracies



```
Accuracies of all the folds after the cross validation are:
array([1., 1., 1., 1., 1.])
```