In [ ]:

```
Import library
```

In [1]:

```
import pandas as pd
```

Upload data and test file. clean data and also drop data which are not required to predict data ticket,passengerid, name, and cabin

In [2]:

```
data = pd.read_csv("train.csv")
test = pd.read_csv("test.csv")
test_ids = test["PassengerId"]


def clean (data):
    data = data.drop(["Ticket", "PassengerId", "Name", "Cabin"], axis=1)

    cols = ["SibSp", "Parch", "Fare", "Age"]
    for col in cols:
        data[col].fillna(data[col].median(), inplace=True)

    data.Embarked.fillna("U", inplace=True)
    return data

data = clean(data)
test = clean(test)
```

In [3]:

```
data.head()
```

Out[3]:

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	male	22.0	1	0	7.2500	S
1	1	1	female	38.0	1	0	71.2833	C
2	1	3	female	26.0	0	0	7.9250	S
3	1	1	female	35.0	1	0	53.1000	S
4	0	3	male	35.0	0	0	8.0500	S

From library sklearn import preprocessing

In [4]:

```
from sklearn import preprocessing
le = preprocessing.LabelEncoder()

columns = ["Sex", "Embarked"]

for col in columns:
    data[col] = le.fit_transform(data[col])
    test[col] = le.transform(test[col])
    print(le.classes_)

data.head(5)
```

```
['female' 'male']
['C' 'Q' 'S' 'U']
```

Out[4]:

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	1	22.0	1	0	7.2500	2
1	1	1	0	38.0	1	0	71.2833	0
2	1	3	0	26.0	0	0	7.9250	2
3	1	1	0	35.0	1	0	53.1000	2
4	0	3	1	35.0	0	0	8.0500	2

import logistic regression model

In [5]:

```
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split

y = data["Survived"]
X = data.drop("Survived", axis=1)

X_train, X_val, y_train, y_val = train_test_split(X, y, test_size= 0.2, random_stat
```

In [6]:

```
clf = LogisticRegression(random_state=0, max_iter=1000).fit(X_train, y_train)
```

In [7]:

```
predictions = clf.predict(X_val)
from sklearn.metrics import accuracy_score
accuracy_score(y_val, predictions)
```

Out[7]:

```
0.8100558659217877
```

In [8]:

```
submission_preds = clf.predict(test)
```

In [9]:

```
df = pd.DataFrame({"passengerId": test_ids.values,
                   "Survived": submission_preds,
                  })
```

In [10]:

```
df.to_csv("submission.csv", index=False)
```

In [ ]: