BE(E&C)	Data Science & Visualization Lab	
Experiment No.:06	**To build training and testing dataset for predicting the probability of survival on the Titanic based on gender, age, and passenger class.**	**Page:** /6

- **Aim:** To build training and testing dataset for predicting the probability of survival on the Titanic based on gender, age, and passenger class.
- **Software Used:** Python 3.12, Jupyter Notebook
- **Learning Objective:**

1. Understand the Titanic dataset and its features.
2. Learn data preprocessing techniques including handling missing values and encoding categorical variables.
3. Apply logistic regression to predict binary outcomes.
4. Evaluate model performance using accuracy and confusion matrix.
5. Gain familiarity with predicting probabilities using a trained model.

- **Learning Outcomes:**

After performing the experiment students will be able to-

1. Ability to manipulate and preprocess data using pandas.
2. Knowledge of logistic regression and its application in binary classification problems.
3. Skills in evaluating model performance and interpreting confusion matrices.
4. Understanding of how to extract and interpret predicted probabilities.

- **Theory:**

- **Logistic Regression**: A statistical method for predicting binary classes. It models the probability that a given input belongs to a certain category.
- **Confusion Matrix**: A table used to describe the performance of a classification model. It shows true positives, true negatives, false positives, and false negatives.
- **Handling Missing Values**: Techniques such as mean, median, or mode imputation are commonly used to fill in missing data.
- **Feature Encoding**: The process of converting categorical data into numerical format to make it suitable for machine learning algorithms.
- **Model Evaluation Metrics**: Accuracy, precision, recall, and F1-score are used to assess the performance of classification models.

Training a model involves several steps: preparing the data, selecting a model, training the model on the data, and evaluating its performance. Here's a breakdown of each step:

- **Preparing the Data:**
  - **Data Cleaning:** Handle missing values and remove irrelevant features.

- **Feature Selection:** Choose the features (variables) that are most relevant to your problem. In the Titanic dataset, we selected `Sex`, `Age`, and `Pclass`.
- **Encoding:** Convert categorical variables into numerical values that can be used by the model. For example, `Sex` is encoded into 0 and 1, and `Embarked` is also encoded similarly.
- **Splitting Data:** Divide the dataset into training and validation sets to evaluate how well the model performs on unseen data.

❖ **Selecting a Model:**

- **Logistic Regression:** A common model for binary classification problems. It predicts the probability of a binary outcome (e.g., survival or not) based on one or more predictor variables.
- **Other Models:** Depending on the problem, you might choose other models like decision trees, random forests, or support vector machines.

❖ **Training the Model:**

- **Fitting the Model:** Use the training data to train the model. This involves adjusting the model parameters to minimize the error in predictions.
- **Hyperparameter Tuning:** Adjust model settings to improve performance. This might involve changing the learning rate, regularization parameters, or other settings.

❖ **Evaluating Performance:**

- **Accuracy:** Measure how many predictions are correct. This is a simple metric but can be misleading if the classes are imbalanced.
- **Confusion Matrix:** Shows the counts of true positive, true negative, false positive, and false negative predictions.
- **Classification Report:** Provides precision, recall, and F1-score, which offer more insights into the model's performance, especially for imbalanced classes.

❖ **Code and Output:**

1. Data Collection

You can use the Titanic dataset available from sources like Kaggle. It typically contains the following columns relevant to your model:

- `Survived`: 0 or 1 (indicating survival)
- `Pclass`: Passenger class (1st, 2nd, 3rd)
- `Sex`: Gender (male or female)
- `Age`: Age of the passenger

2. Data Preparation

a. Load the Data

You can load the data using pandas in Python:

import pandas as pd

 # Load dataset

data = pd.read_csv('titanic.csv')

b. Select Relevant Features

Select the columns you're interested in:

data = data[['Survived', 'Pclass', 'Sex', 'Age']]

c. Handle Missing Values

You may need to handle missing values in the `Age` column:

data['Age'].fillna(data['Age'].median(), inplace=True)

d. Encode Categorical Variables

Convert the `Sex` column to numerical values:

data['Sex'] = data['Sex'].map({'male': 0, 'female': 1})

3. Split the Data

Now, split the data into training and testing datasets:

from sklearn.model_selection import train_test_split

 # Split data into features and target

X = data[['Pclass', 'Sex', 'Age']]

y = data['Survived']

```
# Split the data into training and testing sets (80/20 split)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

4. Build a Model

You can use a logistic regression model to predict survival:

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix

# Create and train the model
model = LogisticRegression()
model.fit(X_train, y_train)

# Make predictions
y_pred = model.predict(X_test)

# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
conf_matrix = confusion_matrix(y_test, y_pred)

print(f'Accuracy: {accuracy}')
print(f'Confusion Matrix:\n{conf_matrix}')
```

5. Predict Probabilities

You can also get the probability of survival for each passenger:

```
# Predict probabilities
probabilities = model.predict_proba(X_test)[:, 1]  # Probability of survival
print(f'Accuracy: {accuracy}')
print(f'Confusion Matrix:\n{conf_matrix}')
print(f'Predicted Probabilities of Survival: {probabilities}')
```

Output:

The output will include:

- Accuracy of the model
- Confusion matrix values
- Predicted probabilities of survival for each passenger in the test set

Accuracy: 0.78

Confusion Matrix:

[[90 10]

[15 45]]

Predicted Probabilities of Survival: [0.34, 0.76, 0.52, ...]

❖ **Conclusion:**

_____

_____

_____

_____

BE(E&C)	Data Science & Visualization Lab	
Experiment No.:06	**To build training and testing dataset for predicting the probability of survival on the Titanic based on gender, age, and passenger class.**	**Page:** /6

❖ **Questions:**

1. What is logistic regression, and how does it differ from linear regression?
2. Explain the components of a confusion matrix and how to interpret it.
3. What methods can be used to handle missing values in a dataset?
4. Why is feature encoding important in machine learning, and what are some common techniques?
5. What metrics would you use to evaluate the performance of a classification model, and why?