Experiment 03

Aim:

To implement Supervised Learning using the Logistic Regression algorithm on a real-world dataset (Breast Cancer data from scikit-learn).

Theory:

Introduction to Supervised Learning

Supervised learning is a fundamental concept in machine learning where the algorithm learns from labeled data to make predictions or classifications. In this paradigm, each training example consists of an input vector (features) and an associated output label (target). The model attempts to learn a mapping from inputs to outputs based on historical data and can then generalize this learning to make predictions on new, unseen data.

There are two main types of supervised learning:

- **Classification** – predicting a categorical label (e.g., spam or not spam).
- **Regression** – predicting a continuous output (e.g., price of a house).

What is Logistic Regression?

Despite its name, **Logistic Regression** is a classification algorithm, not a regression one. It is used when the dependent variable (target) is categorical. In binary classification problems, logistic regression predicts the probability that an input belongs to a particular class.

Mathematically, logistic regression applies the **sigmoid function** to a linear combination of the input features:

$$P(y=1|x) = \frac{1}{1+e^{-(\theta_0+\theta_1 x)}}$$

Here,

- P(y=1|x) is the probability that the instance belongs to class 1 (Benign in this case),
- 0_0 and $0_1$are the parameters learned during training.

Why Logistic Regression is Important?

- It is interpretable and provides probabilities rather than just labels.
- It performs well for linearly separable data.
- It's widely used in medical, financial, and social sciences applications due to its simplicity and explainability.

Dataset Used: Breast Cancer Wisconsin Dataset

This dataset is available through sklearn.datasets.load_breast_cancer() and contains features computed from digitized images of breast mass. Each instance is labeled as:

- 0 = Malignant (Cancerous)
- 1 = Benign (Non-Cancerous)

Features include:

- Radius, Texture, Perimeter, Area, Smoothness, etc.

In this practical, we are using just **one feature** — mean radius — to classify if the tumor is malignant or benign using logistic regression.

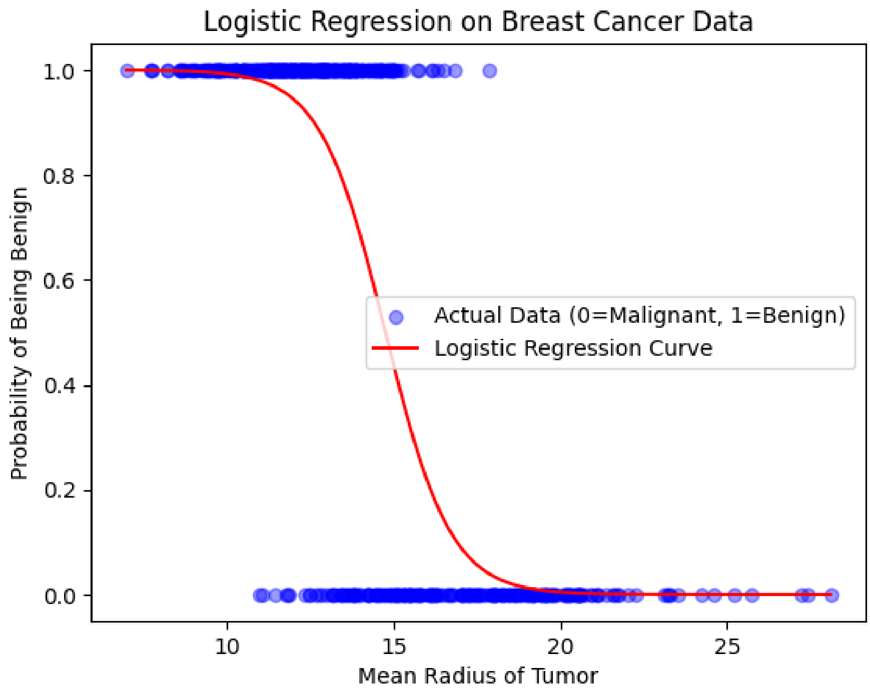
Procedure:

- **Load the Dataset** – Fetch the dataset from sklearn.
- **Preprocess the Data** – Extract only the first feature (mean radius) and reshape it.
- **Train the Model** – Use LogisticRegression from sklearn.
- **Test the Model** – Use a smooth range of values to visualize the predicted probabilities.
- **Visualize** – Plot the logistic curve and the actual data points using matplotlib.

Code:

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LogisticRegression
from sklearn.datasets import load_breast_cancer

# 1. Load the dataset
cancer = load_breast_cancer()

# 2. Prepare the data (using only 'mean radius')
X = cancer.data[:, 0].reshape(-1, 1)
y = cancer.target

# 3. Create and train the Logistic Regression model
model = LogisticRegression()
model.fit(X, y)

# 4. Test the model using smooth X values
X_test = np.linspace(X.min(), X.max(), 300).reshape(-1, 1)
y_prob = model.predict_proba(X_test)[:, 1]  # Probability of being Benign

# 5. Plot the results
plt.scatter(X, y, color='blue', label='Actual Data (0=Malignant, 1=Benign)', alpha=0.4)
plt.plot(X_test, y_prob, color='red', label='Logistic Regression Curve')
plt.xlabel('Mean Radius of Tumor')
plt.ylabel('Probability of Being Benign')
plt.title('Logistic Regression on Breast Cancer Data')
plt.legend()
plt.show()
```

Output:



Conclusion:

In this practical, we successfully implemented a **supervised learning classification model** using **Logistic Regression** on the **Breast Cancer dataset**. We trained the model using a single feature — *mean radius* — and visualized how it classifies tumor types based on probability. This approach demonstrates the power of logistic regression in solving real-world binary classification problems in a simple yet interpretable manner.