

Introduction: Among women worldwide, breast cancer is the most common and one of the leading causes of death due to cancer. The mortality and prognosis of the disease can be enhanced only with its diagnosis at an early stage. Various machine learning approaches have been discussed that analyze clinical data in order to predict the nature of a breast tumor, whether benign or malignant. Logistic regression is a type of supervised learning that would be applied in this project to classify instances of breast cancer based on characteristics of the cells, such as clump thickness, uniformity of cell size, and presence of mitoses. The dataset contains several features describing these properties of cells, and the aim is to develop a predictive model which will efficiently classify the tumors as either benign or malignant. The contribution of logistic regression towards the accuracy and reliability of the diagnosis of breast cancer is aimed to be highlighted in the study by using some certain data preprocessing, splitting the dataset into both training and testing sets, and the usage of metrics evaluation.


```
In [6]: import pandas as pd

# Load the dataset
file_path = 'C:\\Users\\smita\\Downloads\\breast_cancer_bd.csv'
data = pd.read_csv(file_path)

# Display the first few rows of the dataset
data.head()
```

```
Out[6]:
```

	Sample code number	Clump Thickness	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	I N
0	1000025	5	1	1	1	2	1	3	
1	1002945	5	4	4	5	7	10	3	
2	1015425	3	1	1	1	2	2	3	
3	1016277	6	8	8	1	3	4	3	
4	1017023	4	1	1	3	2	1	3	



```
In [7]: # Convert 'Bare Nuclei' to numeric and handle errors
data['Bare Nuclei'] = pd.to_numeric(data['Bare Nuclei'], errors='coerce')

# Drop rows with missing values
data = data.dropna()

# Verify if there are any missing values left
data.isnull().sum()
```

```
Out[7]: Sample code number      0
        Clump Thickness        0
        Uniformity of Cell Size  0
        Uniformity of Cell Shape 0
        Marginal Adhesion       0
        Single Epithelial Cell Size 0
        Bare Nuclei             0
        Bland Chromatin         0
        Normal Nucleoli         0
        Mitoses                 0
        Class                   0
        dtype: int64
```

```
In [8]: # Features and target variable
X = data.drop(columns=['Class', 'Sample code number'])
y = data['Class']

# Splitting the dataset into training and testing sets
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_sta
```

```
In [9]: from sklearn.linear_model import LogisticRegression

# Create the model
model = LogisticRegression(max_iter=200)

# Train the model
model.fit(X_train, y_train)
```

```
Out[9]: LogisticRegression
LogisticRegression(max_iter=200)
```

```
In [10]: from sklearn.metrics import mean_squared_error

# Make predictions
y_pred = model.predict(X_test)

# Calculate the Mean Squared Error (MSE)
mse = mean_squared_error(y_test, y_pred)
print(f'Mean Squared Error: {mse}')
```

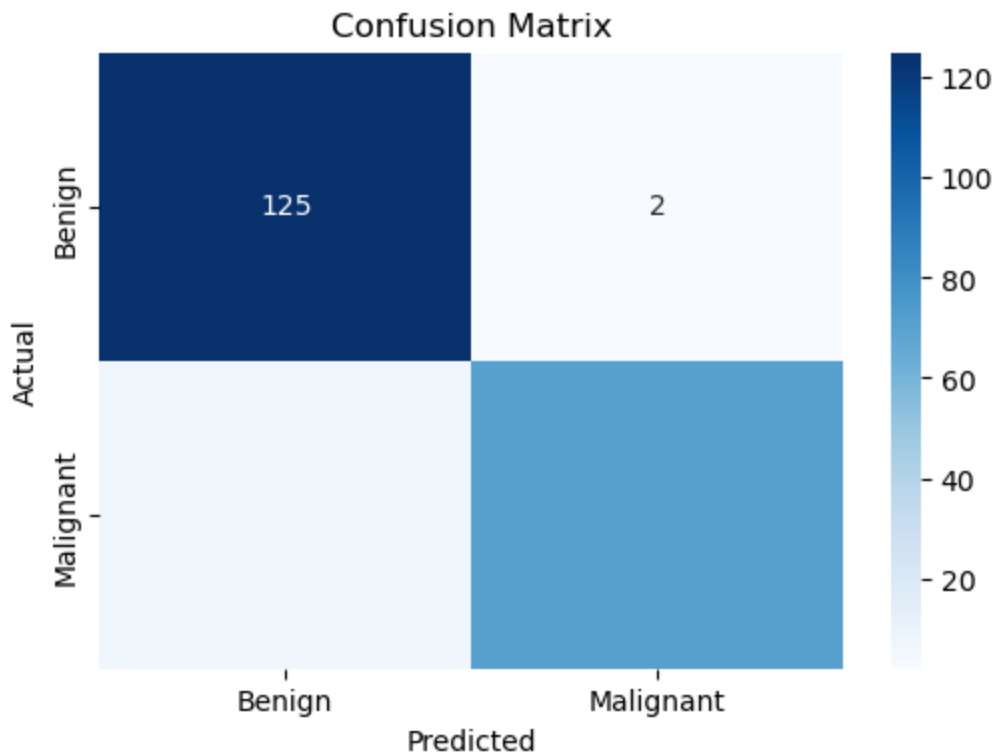
Mean Squared Error: 0.17560975609756097

```
In [11]: from sklearn.metrics import confusion_matrix
import seaborn as sns
import matplotlib.pyplot as plt

# Generate confusion matrix
cm = confusion_matrix(y_test, y_pred)

# Visualize confusion matrix
plt.figure(figsize=(6, 4))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', xticklabels=['Benign', 'Malignan
plt.xlabel('Predicted')
```

```
plt.ylabel('Actual')
plt.title('Confusion Matrix')
plt.show()
```



Reflection: Problem: Diagnosis of the disease was usually based on several medical parameters, which, most of the time, proved difficult to analyze by hand. There are attributes in this dataset regarding cells in the feature space, like clump thickness, uniformity of cell size and shape, and mitoses. The work here is to classify these features into tumor types, namely benign or malignant, which will again be a binary classification problem.

Solution: The solution utilizes Logistic Regression, a supervised machine learning model very suitable for binary classification problems. The code does preprocessing on the dataset: cleaning by removing missing values and trains the logistic regression model on the given cell characteristics to evaluate and predict the tumor class.

Reflection for both Problem and Solution : Diagnosis of breast cancer requires analyzing complex cell features to classify tumors as either benign or malignant, a process that is often extremely time-consuming and very susceptible to human error if performed manually. With patient data volumes increasing day by day, the requirement is all the greater for an automated, efficient, and reliable diagnostic tool. Logistic regression provides a structured solution to such problems by automatically classifying the tumors as malignant or benign, based on key attributes about the tumor, such as clump thickness and uniformity of cell size. Once appropriate pre-processing of the data is done, the model is then trained using logistic regression to determine relationships between such different attributes and tumor classification; hence, it provides predictions that are interpretable and trustworthy by medical professionals. This simple, interpretable model therefore presents a useful diagnosis

tool, with the assessment of its accuracy supported by such evaluation metrics as the Mean Squared Error and confusion matrix. Given the binary classification nature of the problem, logistic regression was a good baseline to work from, though performance may be improved with more complex models. Be that as it may, this approach serves to show how machine learning can be used to orchestrate and improve the accuracy of diagnoses of breast cancer.