📊 Breast Cancer Prediction using Logistic Regression

📌 Introduction

This project aims to build a machine learning model that predicts whether a tumor is malignant or benign using the Breast Cancer dataset from Kaggle. The model is built using **Logistic Regression**, and the focus is on proper data preprocessing, exploratory data analysis (EDA), and thorough evaluation using multiple metrics.

📁 Dataset Description

The dataset consists of **569 instances** and **32 columns**, including:

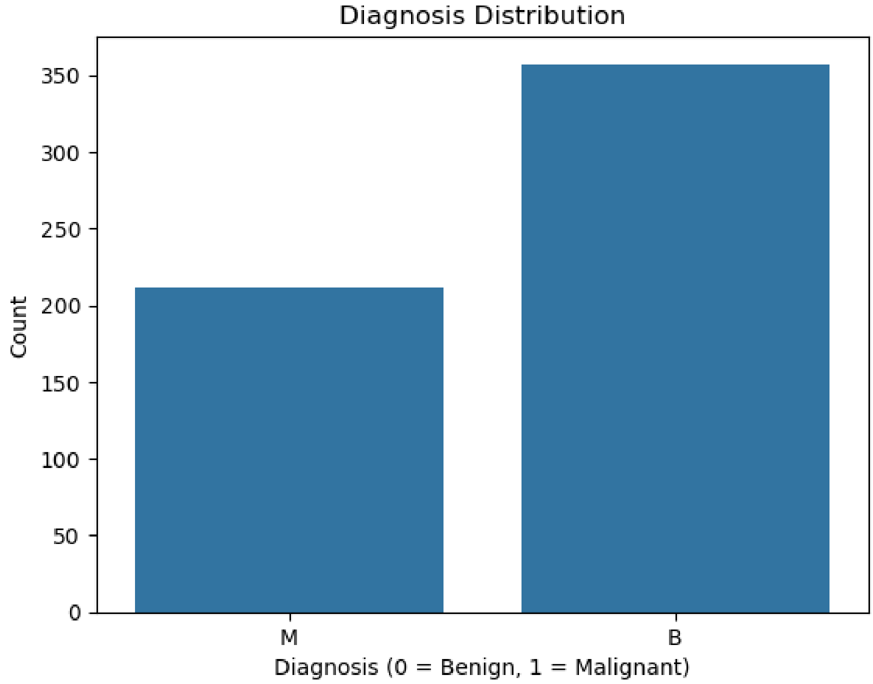
- **ID column** *(unique identifier, already removed during preprocessing)*
- **Diagnosis column** *(target: M = Malignant, B = Benign)*
- **30 numeric features** describing characteristics of the cell nuclei (e.g., radius, texture, perimeter)

🔍 Exploratory Data Analysis (EDA)

Basic insights from the dataset:

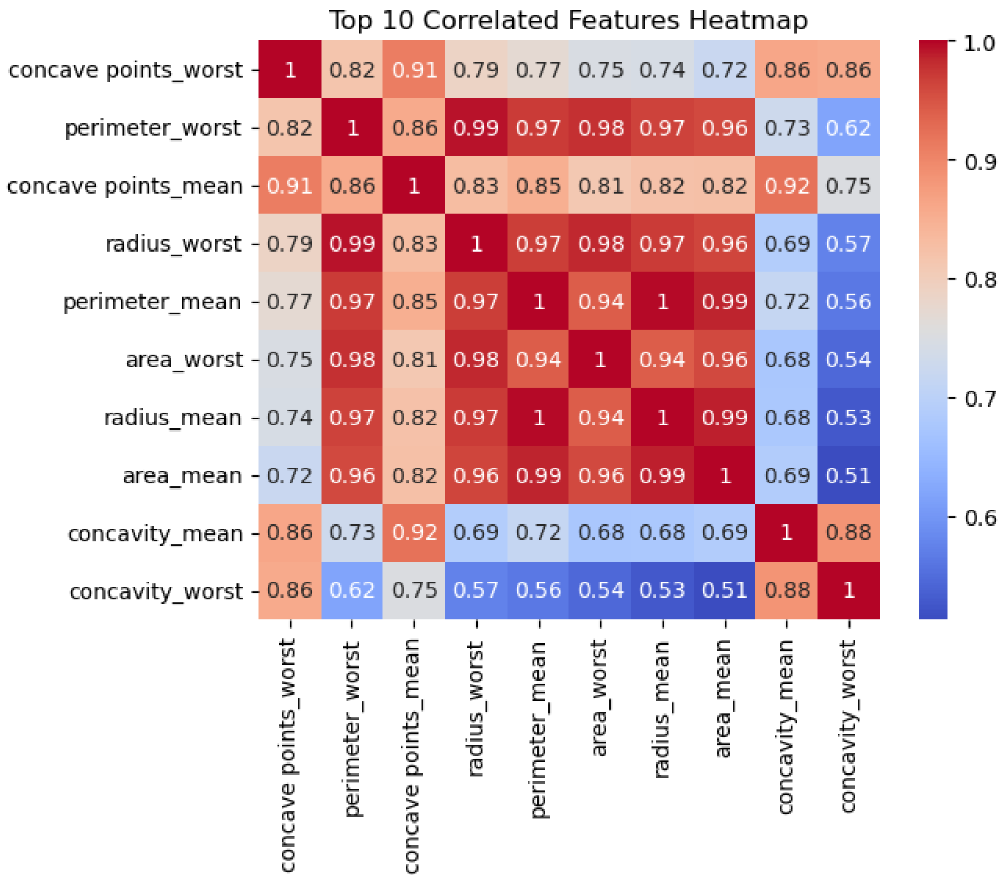
- df.info() confirms that all features are numerical, and there are no missing values.
- df.describe() gives a summary of statistics across the dataset.
- df.shape returns (569, 33) indicating 33 columns including ID and target.
- df.isnull().sum() confirms there are **no null values**.
- df['diagnosis'].value_counts() shows the distribution of benign and malignant cases.
- **Encoded** the target column diagnosis as binary labels:
    - 'M' → 1 (Malignant)
    - 'B' → 0 (Benign)
    This encoding is needed for correlation analysis and modeling.
- df.corr() was used to compute correlations between numeric features and the target.

📈 **Plot: Diagnosis Distribution**



This bar chart visualizes the number of malignant vs. benign diagnoses in the dataset.

🔥 **Plot: Top 10 Correlated Features Heatmap**



This bar helps identify which features are most relevant for the prediction task.

⚙️ Data Preprocessing

Steps taken:

1. **Dropped the id column** as it holds no predictive value.
2. **Separated features and target variable** (X, y)
3. **Train-test split** using train_test_split() with 80/20 ratio.
4. **Standardized features** using StandardScaler() to normalize the data before model training:
   - fit_transform() was applied to X_train
   - transform() was used on X_test

🧪 Modeling with Logistic Regression

The logistic regression model was trained using:

```
from sklearn.linear_model import LogisticRegression
```

- The model was fitted to the training data and used to predict test values.

📊 Evaluation Metrics

After training the logistic regression model, the following evaluation metrics were computed on the test set:

- **Accuracy: 0.9737**
  → The model correctly classified **97.37%** of the cases.
- **Precision, Recall, F1-Score:**

	precision	recall	f1-score
0	0.97	0.99	0.98
1	0.98	0.95	0.96

(The exact numbers will be taken from the classification_report() output)
These metrics provide insight into how well the model performs on both benign (0) and malignant (1) classes, especially in imbalanced datasets.

- **Confusion Matrix:**

```
Confusion Matrix:
 [[70  1]
 [ 2 41]]
```

Gives a breakdown of true positives, true negatives, false positives, and false negatives — useful for visualizing the model's performance.

📋 Classification Report

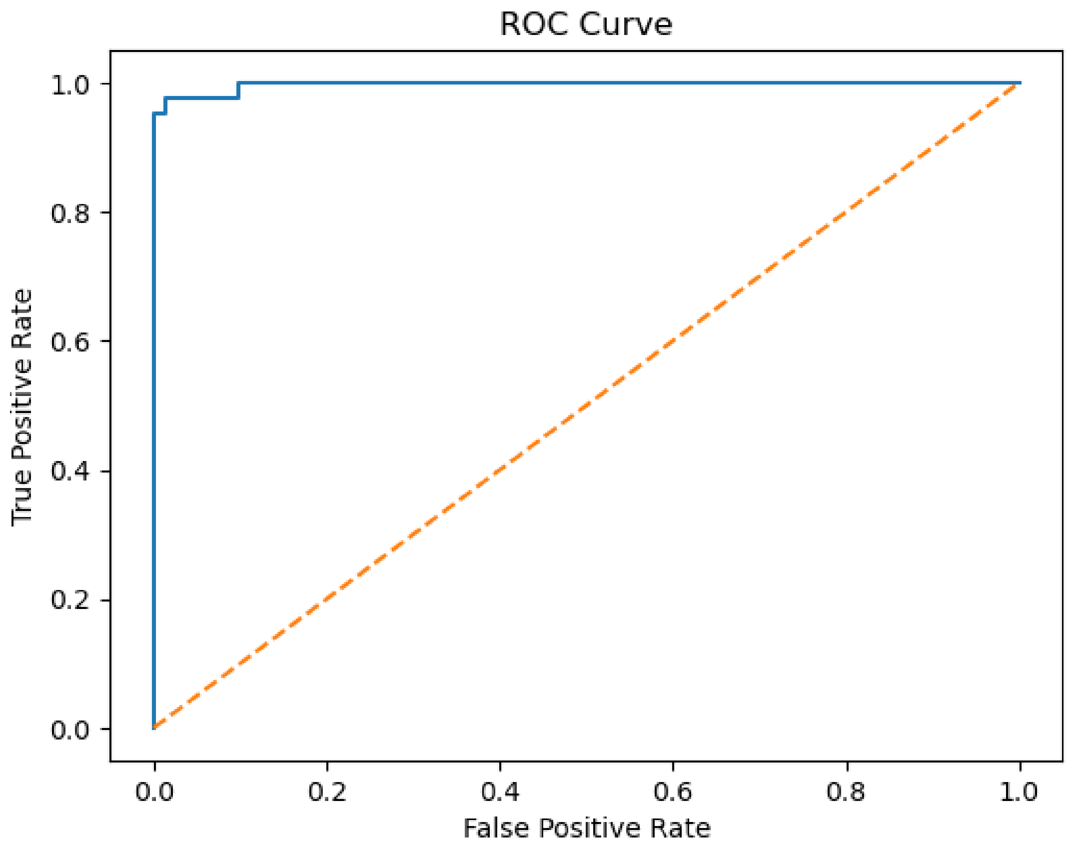
```
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.99      0.98        71
           1       0.98      0.95      0.96        43

    accuracy                           0.97       114
   macro avg       0.97      0.97      0.97       114
weighted avg       0.97      0.97      0.97       114
```

- **Precision**: Proportion of positive identifications that were actually correct.
- **Recall**: Proportion of actual positives that were correctly identified.
- **F1 Score**: Harmonic mean of precision and recall — balances both.

📈 ROC Curve & AUC Score



The **AUC score** is **0.9974**, which indicates excellent model performance in distinguishing between classes across all thresholds.

🧾 Conclusion

- The logistic regression model performs strongly on this dataset with **high accuracy**, **balanced precision/recall**, and **excellent AUC**.
- Only minor misclassifications occurred (1 false positive, 2 false negatives).
- This model can serve as a solid baseline for breast cancer prediction tasks.