**Breast Cancer Diagnosis Using Machine Learning – Project Report**

**1. Introduction**

Breast cancer remains one of the most prevalent and life-threatening diseases affecting women worldwide. Early and precise diagnosis significantly enhances treatment success and patient survival rates. In this project, we apply machine learning techniques to classify tumours as malignant (cancerous) or benign (non-cancerous), based on a set of physical features derived from medical imaging.

The dataset used in this study is sourced from the UCI Machine Learning Repository and includes a comprehensive set of features extracted from digitized images of breast tissue obtained through digital mammography.

**2. About the Dataset**

The dataset contains 569 records, each corresponding to a patient. It consists of 33 features that capture the characteristics of the tumour. Key information includes:

* **Number of Patients:** 569
* **Number of Features:** 33
* **Target Label (Diagnosis):**
  + **M** – Malignant (Cancerous)
  + **B** – Benign (Non-cancerous)
* **Empty Column:** One column (Unnamed: 32) was found to be empty and was excluded from the analysis.

Each record includes the following tumour features:

| **Feature** | **Description** |
| --- | --- |
| radius\_mean | Average distance from the centre to points on the perimeter |
| texture\_mean | Standard deviation of grey-scale values |
| area\_mean | Area of the tumour |
| smoothness\_mean | Local variation in radius lengths |
| compactness\_mean | Perimeter² / Area - 1.0 |
| concavity\_mean | Severity of concave portions of the contour |

**3. Data Preprocessing**

To ensure data quality and compatibility with machine learning algorithms, the following preprocessing steps were performed:

* **Column Removal:** Dropped the empty column (Unnamed: 32)
* **Missing Values:** Verified and ensured no missing data remained
* **Normalization:** Scaled feature values to maintain uniformity
* **Feature Selection:** Selected the most relevant features to improve model performance and reduce noise

The preprocessing pipeline was implemented using the following tools:

* Python
* Pandas
* NumPy
* Scikit-learn

**4. Exploratory Data Analysis (EDA)**

Exploratory analysis was carried out to gain insights into the dataset and detect underlying patterns. Key steps included:

* **Visualization:** Histograms, box plots, and heatmaps to study feature distributions
* **Class Distribution:** Analysis of benign vs. malignant cases
* **Correlation Analysis:** Features such as radius\_mean and concavity\_mean showed strong positive correlation with malignancy

These insights were critical for selecting appropriate models and features.

**5. Model Selection**

Several supervised learning algorithms were explored to build classification models. These included:

| **Model** | **Reason for Selection** |
| --- | --- |
| Logistic Regression | Simple, interpretable, and efficient |
| K-Nearest Neighbours | Non-parametric and easy to implement |
| Support Vector Machine | Effective for high-dimensional spaces |
| Random Forest | Handles complex relationships and highlights key features |

**Performance Evaluation Metrics:**

* Accuracy
* Precision
* Recall
* F1 Score

Among all, **Random Forest** delivered high accuracy and also helped identify the most influential features.

**6. Results and Discussion**

The trained models demonstrated high classification performance. Key results include:

* **Top Predictive Features:**
  + radius\_worst
  + concave points\_worst
  + perimeter\_mean
  + area\_mean
* **Best Performing Models:**
  + **SVM** and **Random Forest** showed high precision and recall across test sets.

**Interpretation:**  
These models exhibit strong potential for aiding medical professionals in early diagnosis. However, they are intended as decision-support tools and not as replacements for clinical judgment.

**7. Conclusion**

This study demonstrates the successful application of machine learning in classifying breast cancer tumours. Through rigorous data preprocessing, feature engineering, and model evaluation, reliable prediction systems were built. The results confirm that such tools can be effectively integrated into diagnostic workflows to enhance decision-making and support timely interventions.

**8. References**

* UCI Machine Learning Repository – Breast Cancer Wisconsin Dataset
* Python Libraries Used:
  + Pandas
  + NumPy
  + Matplotlib
  + Seaborn
  + Scikit-learn

**9. About the Data**

The dataset used in this analysis originates from the **Breast Cancer Wisconsin (Diagnostic) Dataset**, hosted by the **UCI Machine Learning Repository**. It contains diagnostic features calculated from digitized images of fine needle aspirate (FNA) of breast masses. Each image quantifies characteristics of cell nuclei present in the tumour.

* **Total Records:** 569 patients
* **Total Features:** 33 columns
* **Diagnosis Label:**
  + M – Malignant (cancerous)
  + B – Benign (non-cancerous)

**Column Breakdown:**

* **ID column**: A unique identifier for each patient
* **Diagnosis column**: Target variable indicating cancer type
* **30 numeric features**: Describe tumour properties such as radius, texture, smoothness, etc.
* **1 empty column** (Unnamed: 32): Contains no data and was removed

The numerical features are derived from images and describe various dimensions of the tumours, including averages, standard errors, and worst-case (maximum) values.

**10. Project Report**

The primary objective of this project is to **develop a reliable and accurate machine learning model** capable of predicting whether a breast tumour is malignant or benign based solely on its physical characteristics. By analysing the dataset and training various classification models, we aim to:

* Identify the most influential features contributing to diagnosis
* Evaluate the performance of different algorithms
* Interpret model predictions in a medically relevant context
* Propose a model that can assist healthcare professionals in diagnostic decision-making

This project emphasizes both predictive performance and interpretability, ensuring the solution is scientifically sound and clinically useful.

**11. Data Description**

Here is a detailed description of the dataset's columns and what they represent:

| **Feature Name** | **Description** |
| --- | --- |
| id | Unique identifier for each patient |
| diagnosis | Target variable: M (Malignant), B (Benign) |
| radius\_mean | Mean distance from centre to perimeter points |
| texture\_mean | Standard deviation of grayscale values |
| perimeter\_mean | Mean perimeter of the tumour |
| area\_mean | Mean area of the tumour |
| smoothness\_mean | Local variation in radius lengths |
| compactness\_mean | (Perimeter² / Area) - 1 |
| concavity\_mean | Severity of concave portions in the contour |
| concave points\_mean | Number of concave portions in the contour |
| symmetry\_mean | Symmetry of the tumour shape |
| fractal\_dimension\_mean | Fractal dimension - "coastline approximation" - 1 |
| radius\_se | Standard error of radius\_mean |
| texture\_se | Standard error of texture\_mean |
| perimeter\_se | Standard error of perimeter\_mean |
| area\_se | Standard error of area\_mean |
| smoothness\_se | Standard error of smoothness\_mean |
| compactness\_se | Standard error of compactness\_mean |
| concavity\_se | Standard error of concavity\_mean |
| concave points\_se | Standard error of concave points\_mean |
| symmetry\_se | Standard error of symmetry\_mean |
| fractal\_dimension\_se | Standard error of fractal\_dimension\_mean |
| radius\_worst | Maximum value of radius\_mean across measurements |
| texture\_worst | Maximum value of texture\_mean |
| perimeter\_worst | Maximum value of perimeter\_mean |
| area\_worst | Maximum value of area\_mean |
| smoothness\_worst | Maximum value of smoothness\_mean |
| compactness\_worst | Maximum value of compactness\_mean |
| concavity\_worst | Maximum value of concavity\_mean |
| concave points\_worst | Maximum value of concave points\_mean |
| symmetry\_worst | Maximum value of symmetry\_mean |
| fractal\_dimension\_worst | Maximum value of fractal\_dimension\_mean |
| Unnamed: 32 | Empty column with no values (excluded) |

**12. Future Scope**

* Integrate deep learning models like CNNs for better accuracy
* Deploy the model as a web-based diagnostic tool using Flask or Streamlit
* Apply techniques like SMOTE for handling class imbalance (if present)
* Collaborate with medical professionals to validate model predictions in real-world cases

**13. About the Author**

This project was developed by **Sohel Datta**, with a keen interest in building practical AI solutions for healthcare.

* **GitHub:** [soheldatta17](https://github.com/soheldatta17)
* **LinkedIn:** [sohel-datta](https://www.linkedin.com/in/sohel-datta-68927828a/)