MACHINE LEARNING

PROJECT

TOPIC: Breast Cancer Prediction



**Computer Science and Engineering Department TIET, Patiala**

# July-Dec 2023

Submitted to:

Dr. Anjula Mehto

Submission By:

Deeksha Goel

Roll no: 102103681

|  |  |
| --- | --- |
| **INDEX** | |
| Introduction ……………………………………………………...  Problem statement ……………………………………………  Dataset………………………………………………………………  Methodology or workflow…………………………………  Results……………………………………………………………….  Conclusion ………………………………………………………..  Github Link……………………………………………. | **3**  **4 - 5**  **5**  **6**  **7-8**  **8**  **8** |

**INTRODUCTION**

Breast Cancer Prediction using Machine Learning

Breast cancer is a prevalent and life-altering health concern worldwide. Early detection and accurate diagnosis play a pivotal role in improving patient outcomes. This project employs Machine Learning (ML) techniques to predict breast cancer, leveraging a dataset that incorporates various clinical attributes.

The project explores multiple ML models, each offering distinct strengths in classification accuracy. Models are evaluated based on essential metrics such as accuracy, precision, recall, and F1-score. By systematically analyzing these models, the aim is to contribute a reliable and efficient tool for breast cancer prediction.

The significance of this project lies in its potential to aid healthcare professionals in making informed decisions, promoting early detection, and ultimately enhancing patient care. The implementation of ML in healthcare, specifically for cancer prediction, reflects a promising intersection of technology and medicine, emphasizing the project's relevance and potential impact.

PROBLEM STATEMENT:

Breast cancer is a prevalent health concern affecting women globally, and early

detection is crucial for effective treatment. The aim of this project is to develop a

machine learning model that predicts the likelihood of breast cancer based on relevant clinical features.

Objectives:

1. Data Collection: Gather a comprehensive dataset containing clinical features related to breast cancer, including patient demographics, tumor characteristics, and diagnostic results.
2. Data Preprocessing: Perform thorough data cleaning and preprocessing, handling missing values, and ensuring data quality.
3. Feature Selection: Identify the most influential features contributing to breast cancer prediction, enhancing model efficiency and interpretability.
4. Model Development and Evaluation: Implement various machine learning models such as Logistic Regression, Decision Trees, Random Forest to train the predictive model and assess the performance using relevant metrics such as accuracy, precision, recall and F1 score.
5. Hyperparameter Tuning: Optimize the chosen model's hyperparameters to achieve the best predictive accuracy.
6. Validation and Testing: Validate the model using a separate dataset and perform robust testing to ensure its generalization to new, unseen data.

DATASET:

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

Attribute Information:

1) ID number  
2) Diagnosis (M = malignant, B = benign)  
3-32)

Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)  
b) texture (standard deviation of gray-scale values)  
c) perimeter  
d) area  
e) smoothness (local variation in radius lengths)  
f) compactness (perimeter^2 / area - 1.0)  
g) concavity (severity of concave portions of the contour)  
h) concave points (number of concave portions of the contour)  
i) symmetry  
j) fractal dimension ("coastline approximation" - 1)

The mean, standard error and "worst" or largest (mean of the three  
largest values) of these features were computed for each image,  
resulting in 30 features. For instance, field 3 is Mean Radius, field  
13 is Radius SE, field 23 is Worst Radius.

All feature values are recoded with four significant digits.

Missing attribute values: none

Class distribution: 357 benign, 212 malignant

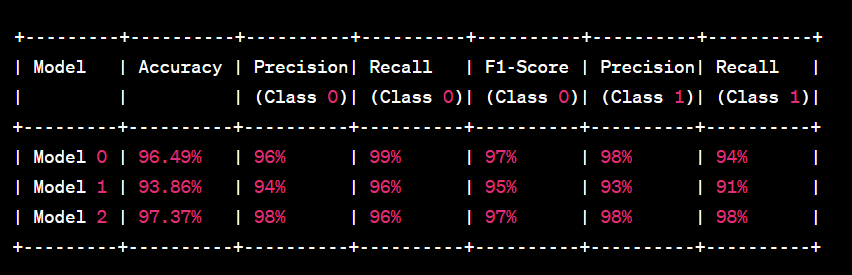
**Download link:** <https://www.kaggle.com/uciml/breast-cancer-wisconsin-data>

METHODOLOGY AND WORKFLOW:

* Problem Definition: Clearly define the problem as predicting breast cancer based on clinical features.
* Data collection: Collect a diverse and comprehensive dataset containing patient demographics, medical history, tumor characteristics, and diagnostic results.
* Data Exploration: Perform exploratory data analysis (EDA) to understand the dataset's characteristics and Identify patterns, trends, and potential correlations between features.
* Data Preprocessing: Handle any missing values, outliers or anomalies and standardize or normalize numerical features. We will also encode categorical values and address class imbalance (if any).
* Feature Selection: Use statistical methods or machine learning algorithms to select the most relevant features.
* Data Splitting: Split the dataset into training and testing sets to assess model generalization.
* Model Selection: Choose machine learning algorithms suitable for binary classification (e.g., Logistic Regression, Decision Trees).
* Model Training: Train the selected models on the training dataset.
* Hyperparameter Tuning: Optimize the chosen model's hyperparameters to achieve the best predictive accuracy.
* Model Evaluation: Evaluate models on the testing dataset using metrics such as accuracy, precision, recall, and F1 score. Generate a confusion matrix for a detailed analysis.

Lastly, we’ll validate and test it using a test dataset to ensure generalization.

RESULT:



Wherein the **Model 0** is: Logistic Regression

**Model 1**: Decision Tree Classifier

**Model 2**: Random Forest

Overall Conclusion-

* All models exhibit high accuracy, with Model 2 achieving the highest at 97.37%.
* The models demonstrate robust performance with high precision, recall, and F1-scores for both classes.
* Model 2 appears to be the most effective in predicting breast cancer based on the provided metrics.
* The project is successful in developing machine learning models that can accurately predict breast cancer based on the given dataset.

# 

# 

# Github link of the Project:

# Signature of Faculty member