CASE STUDY

Breast Cancer Detection

Author: Sameer Hate

Date: 04-05-2025

Table of Contents

Introduction

General

Breast cancer is one of the leading causes of cancer-related deaths in women worldwide. Early and accurate diagnosis is critical for improving survival rates. Traditional diagnostic methods such as biopsies are effective but invasive, time-consuming, and resource-intensive. Leveraging machine learning techniques can aid healthcare professionals in making quicker and more accurate predictions, thus supporting timely medical intervention.

Author’s Note

This case study presented in this work has been independently completed by me. This project serves as a medium to deepen my understanding of Data Science concepts and enhance my practical skills in applying machine learning and data analysis techniques to real-world problems.It may be prone to errors but I try my best to eliminate them as much as I can. In case you identify any such error or have a more optimal technique to solve this problem, your feedback and suggestions are always welcome as this will eventually help me strengthen my concepts.

Problem Statement & Expected Results

Objective

To develop a machine learning-based classification model that can accurately predict whether a tumor is malignant or benign based on features derived from breast cancer diagnostic data.

Goals

- Perform exploratory data analysis to understand patterns and correlations.
- Preprocess and normalize the data wherever necessary.
- Train and evaluate multiple machine learning models (e.g., SVM, KNN, CatBoost, ANN).
- Compare models using metrics such as accuracy, standard deviation, confusion matrix and R^2.
- Select the best-performing model for deployment or further development.

Expected Outcomes

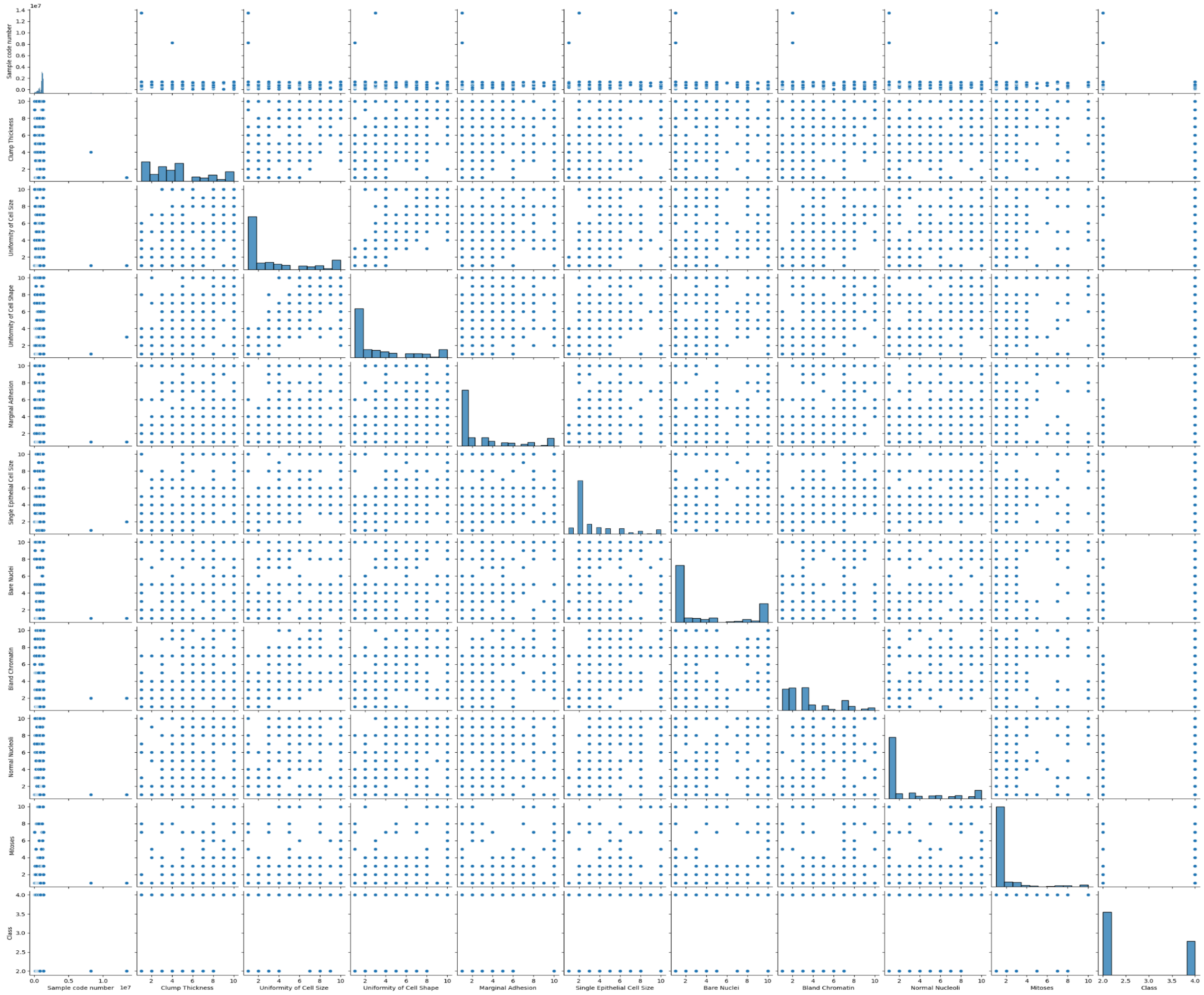
A robust classification model capable of accurately identifying malignant tumors, thereby assisting clinicians in early diagnosis and potentially saving lives.

Dataset

Description

The breast cancer dataset consists of 683 patient records, each described by 10 numeric attributes related to characteristics observed in fine-needle aspirate (FNA) tests of breast masses. These attributes include features such as clump thickness, uniformity of cell size and shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitoses. Each feature is assigned an integer value ranging from 1 to 10, indicating severity or abnormality. The dataset also includes a target variable labeled as “Class,” where a value of 2 indicates a benign tumor and 4 represents a malignant tumor. This dataset is commonly used in binary classification tasks for developing and evaluating machine learning models that aid in early breast cancer detection.

Link to the Dataset: Click Here



Approach

- #IMPORTING LIBRARIES : This cell imports the necessary Python libraries for data manipulation (pandas, numpy), model evaluation (scikit-learn), visualization (matplotlib.pyplot, seaborn), and model selection (train-test split and cross-validation).

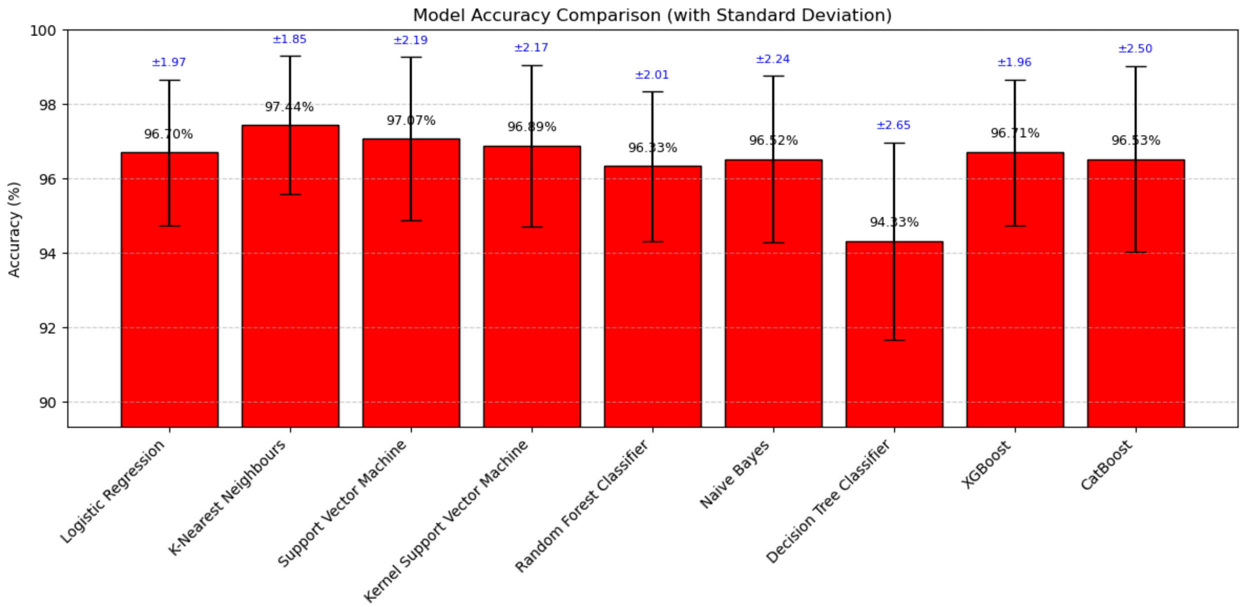
- #PREPARING THE DATASET : This cell loads the dataset and separates the features (X) and the target variable (y). It converts the class labels into binary format: 0 for malignant (originally labeled as 2) and 1 for benign (originally labeled as 4). Lists are also initialized to store model names, accuracies, and standard deviations for performance comparison.

- #SPLITTING THE DATASET INTO TRAINING AND TEST SET : This cell splits the dataset into training and testing sets, with 80% of the data used for training and 20% for testing. A fixed random_state ensures reproducibility of results.

- #TESTING VARIOUS CLASSIFICATION MODELS : The different models tested are, Logistic Regression, K-Nearest Neighbours(KNN), Support Vector Machine(SVM), Kernel Support Vector Machine(K-SVM), Random Forest Classifier, Naive Bayes, Decision Tree, XGBoost and CatBoost. We are using the K-Folds cross validation technique here itself for more accurate comparison and model selection.

- #PRINTING ACCURACY AND STANDARD DEVIATION : Prints the average accuracy and standard deviation for multiple classification models which were used. This comparison helps in selecting one model so that we can enhance it.

- #PLOTTING THE ACCURACY AND STANDARD DEVIATION : Creating a bar graph to compare the accuracies and standard deviations of every model. Based on the results obtained K-Nearest Neighbours displayed the best Accuracy with the least Standard Deviation hence we shall be proceeding with that.



- #APPLYING K-FOLDS CROSS VALIDATION & GRID SEARCH : Conducting the K-Folds Cross Validation helps in evaluating the performance of my chosen model.It helps ensure that the model generalizes well to unseen data by using different portions of the dataset for training and testing in multiple iterations.

  Grid Search technique is used for Hyper-parameter testing which again enhances your model. After selecting the suggested hyper-parameters the accuracy of our models increases which is a good sign.

- #USING KNN WITH BEST PARAMETERS : Training the KNN model with the suggested hyper-parameters.

- #CONFUSION MATRIX AND ACCURACY SCORE : Calculating the Accuracy score and Confusion Matrix. The accuracy score obtained is not that far off from our peak model accuracy hence it signifies that our model is not over-fitted.

- #PREDICTING A NEW RESULT : This code cell uses the updated K-Nearest Neighbors (KNN) classifier (classifier_knn_updated) to predict whether a tumor is benign or malignant based on a new input data sample with 9 feature values.

Results

Overall

With the updated KNN Model, we were able to achieve a 96.37% accuracy which mean the model correctly predicted 96.37% of test samples, reflecting high performance with very few misclassifications.

Confusion Matrix:

```
[[84  3]
 [ 2 48]]
```

- 84: True Negatives (correctly predicted as class 0)
- 3: False Positives (incorrectly predicted as class 1)
- 2: False Negatives (incorrectly predicted as class 0)
- 48: True Positives (correctly predicted as class 1)

Specific Prediction

The following values were submitted to the model in order to make a prediction,

- Clump Thickness - 8
- Uniformity of Cell Size - 7
- Uniformity of Cell Shape - 5
- Marginal Adhesion - 10
- Single Epithelial Cell Size - 7
- Bare Nuclei - 9
- Bland Chromatin - 5
- Normal Nucleoli - 5
- Mitoses - 4

Model Prediction : The Tumor is **Benign**

References

GitHub Repository - Click Here

****