DSE315/615: Data Science in Practice

Breast Cancer Prediction using Machine Learning

Akshat Singh
20031

Varun Nair
20303

November 18, 2023

Indian Institute of Science Education and Research (IISER) Bhopal.
India

1 Introduction

[1]

Machine learning is a powerful tool that enhances efficiency by enabling computers to learn from data and make predictions or decisions without explicit programming. Its practical applications range from streamlining business processes and automating tasks to improving personalized recommendations in areas like entertainment and e-commerce. In healthcare, it aids in diagnostics and treatment planning by extracting valuable insights from complex medical data, showcasing its broad utility in optimizing various facets of our daily lives.

Breast cancer poses a significant global health challenge, necessitating innovative approaches for early detection and improved prognostic assessments. This project tries to use machine learning to predict the nature of breast cancer—whether benign or malignant.

2 Objectives

The primary goals of this project include developing a machine learning model that accurately classifies breast cancer cases as benign or malignant and assessing the model's performance against established metrics.

3 Data Collection

Data for this project were sourced from the UCI ML repository, a widely recognized platform for Machine Learning Data, accessible through the following link: UCI ML repository. This repository is renowned for hosting diverse datasets, providing a valuable resource for machine learning practitioners and researchers.

4 Preprocessing

The provided data exhibited no missing values; therefore, no processing was required. Principal component analysis was conducted, and the top 15 components were selected based on the analysis of the scree plot shown in fig 1 Subsequently, features were standardized for each model.

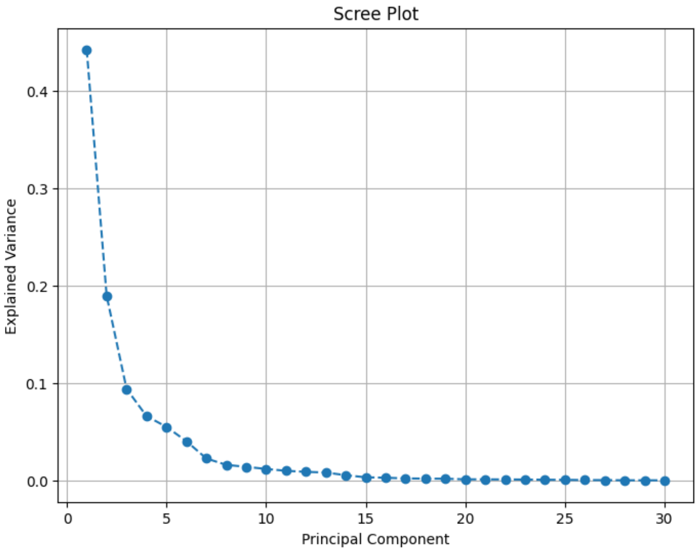


Figure 1: Scree Plot for PCA analysis

[1] Codes available at: https://github.com/Borun2002/Breast-Cancer-Classification

5 Exploratory Data Analysis

The dataset has 30 features, including things like radius and texture, each with three values. We noticed strong connections, especially between radius, perimeter, and area. The three radius values (radius1, radius2, and radius3) were pretty similar, so we're thinking about making things simpler by either removing similar features or using something called Principal Component Analysis (PCA).

On the analysis of each feature in people with benign and malignant tumors. It was observed that people with malignant tumors generally had higher average values for feaures with same names(radius1, radius2, and radius3) as shown in fig 2 compared to those with benign tumors. This suggests these features might be important in telling apart benign and malignant tumors.

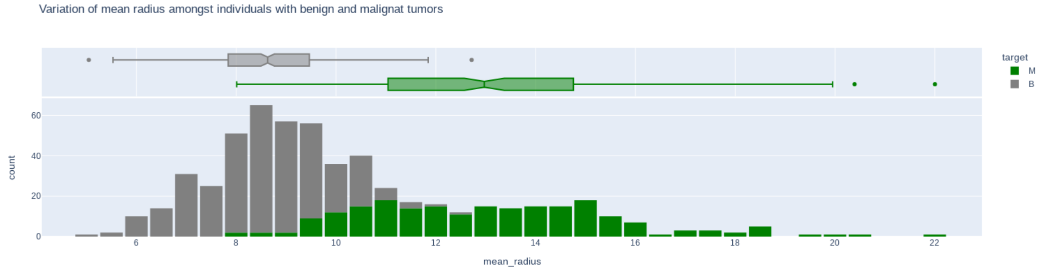


Figure 2: Variation of mean radius

6 Model Selection

The following classification models were used for training and on each of these models, hyper-parameter tuning was done and the parameters with the best accuracy scores were chosen.:

- **Baseline Classifier:** The baseline classifier used here is a majority class classifier which predicts the most frequent class in the training data for every instance. No hyper-parameter tuning was used as it is the reference model.
- **Decision Tree Classifier:** A decision tree classifier is a model that recursively splits the dataset based on features, making decisions at each node to classify instances into different classes.
- **Support Vector Machine Classifier:** A Support Vector Machine (SVM) classifier is a model that aims to find a hyperplane that best separates classes in the feature space, making it effective for both linear and non-linear classification tasks.
- **K-Nearest Neighbours Classifier:** A K-Nearest Neighbors (KNN) classifier is a model that classifies instances based on the majority class of their k-nearest neighbors in the feature space.
- **Logistic Regression Classifier:** A Logistic Regression classifier is a linear model that uses the logistic function to model the probability of belonging to a particular class, making it suitable for binary and multi-class classification tasks.
- **Random Forest Classifier:** A Random Forest Classifier is an ensemble model that consists of a collection of decision trees, where each tree is trained on a random subset of the data. It combines their predictions to improve overall performance and robustness.
- **Bagging Classifier:** A Bagging Classifier is an ensemble model that combines multiple base classifiers, typically decision trees, trained on random subsets of the training data. It averages or takes a majority vote of the predictions to improve overall model performance and reduce overfitting. Here decision tree was used.
- **Voting Classifier:** A Voting Classifier is an ensemble model that combines the predictions of multiple individual classifiers, often of different types, using a majority vote (hard voting) or weighted average (soft voting) to make the final prediction. Here, Decision tree, SVM, KNN,

Logistic Regression, Random Forest and Bagging Classifiers were used for both hard and soft voting.

- **Stacking Classifier:** A Stacking Classifier is an ensemble model that combines multiple base classifiers by training a meta-classifier to make predictions based on the outputs of the individual classifiers. It involves stacking the predictions of the base classifiers as additional features for training the meta-classifier. Here, Decision tree, SVM, KNN, Logistic Regression, Random Forest and Bagging Classifiers were used for the training.

Other than these classification models, 4 clustering models were used to check how good they are compared to the classification models if no labels are there, for all the models the number of clusters were set to 2. The models were:

- **K-Means**: K-Means is an unsupervised machine learning algorithm used for clustering. It partitions the input data into k clusters by iteratively assigning each data point to the cluster whose centroid is closest and updating the centroids based on the mean of the points in each cluster.

- **Agglomerative Clustering:** Agglomerative Clustering is an unsupervised machine learning algorithm used for hierarchical clustering. It starts with each data point as a singleton cluster and iteratively merges the closest pairs of clusters until only one cluster remains. The linkage criterion (e.g., ward, complete, average) determines how the distance between clusters is calculated during merging.

- **BIRCH:** BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) is an unsupervised machine learning algorithm for hierarchical clustering. It is designed for large datasets and incrementally builds a tree structure to represent the clusters. BIRCH is memory-efficient and suitable for streaming data, making it well-suited for scenarios with high-dimensional datasets.

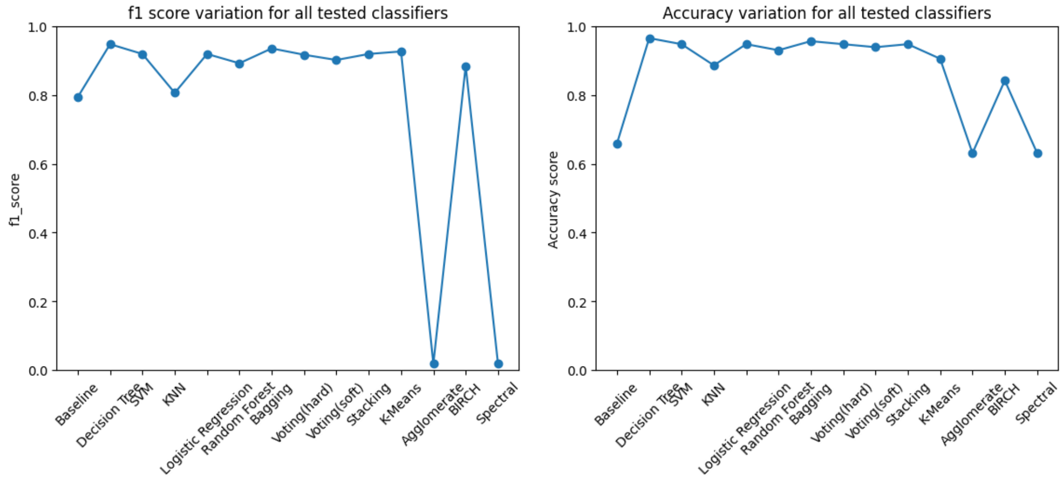
- **Spectral Clustering:** Spectral Clustering is an unsupervised machine learning algorithm used for clustering. It works by transforming the data into a spectral domain, where clusters are identified based on the eigenvalues of the similarity matrix. Spectral clustering can be particularly effective for datasets with complex structures or non-convex shapes.

7 Results

After hyper-parameter tuning, the final accuracy and f1-score of the models are as follows(for clustering, the clusters are compared with labels):

Model	Accuracy	f1-score
Baseline Classifier	0.64	0.79
Decision Tree Classifier	0.96	0.94
SVM Classifier	0.94	0.91
KNN Classifier	0.88	0.80
Logistic Regression	0.94	0.91
Random Forest Classifier	0.92	0.89
Bagging Classifier	0.95	0.93
Voting Classifier(Hard)	0.94	0.91
Voting Classifier(Soft)	0.93	0.90
Stacking Classifier	0.94	0.91
K-Means Cluster	0.90	0.92
Agglomerative cluster	0.63	0.01
BIRCH Cluster	0.84	0.88
Spectral Cluster	0.63	0.01

Table 1: Different Models and their accuracy and f1-score



(a) f1 scores of all models

(b) Accuracy of all models

8 Conclusion

After a thorough analysis and comparison of various models, it has been determined that **Decision Tree** Model had the best accuracy and f1-score among the models used. Hence this is selected as the final model.

This is a practical selection as the data is not balanced and Decision Trees excel with imbalanced datasets, they adeptly model non-linear boundaries without being influenced by class distribution. Their flexibility in handling diverse data patterns, insensitivity to skewed class proportions, interpretability, and natural treatment of missing values make them a fitting choice.

Additionally, noteworthy insights emerged regarding the performance of clustering algorithms. Although not surpassing the efficacy of classification algorithms, clustering demonstrated decent results. Despite limited exploration in this area, it is acknowledged that clustering algorithms could serve as viable options in scenarios where labeled training data is scarce or unavailable.