Predictive Analytics

Data Classification Assignment

Report

Breast Cancer Classification

"Developing Machine Learning Models for Early Diagnosis of Breast Cancer: Utilizing the Wisconsin Breast Cancer Diagnosis Dataset to Predict Malignancy"

In this problem statement, the goal is to create a predictive model that can accurately distinguish between benign and malignant breast tumors using the features provided in the WBCD dataset. This study aims to assist in early diagnosis, improving treatment outcomes for breast cancer patients.

Data Source:

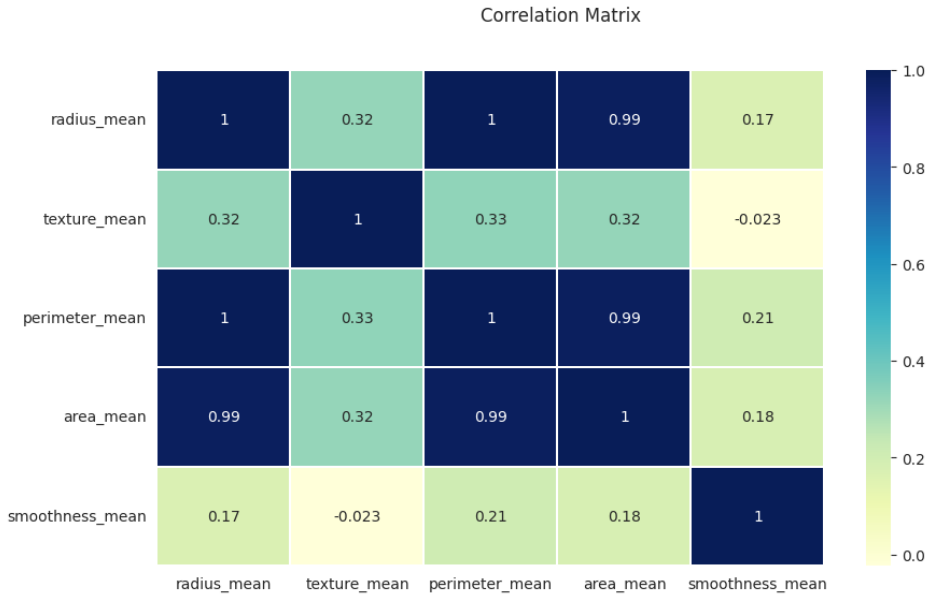
Breast Cancer Wisconsin (Diagnostic) Dataset is publicly available to all.
I took the dataset from UCI Machine Learning Repository.
Link: https://doi.org/10.24432/C5DW2B.

About the data set:
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.

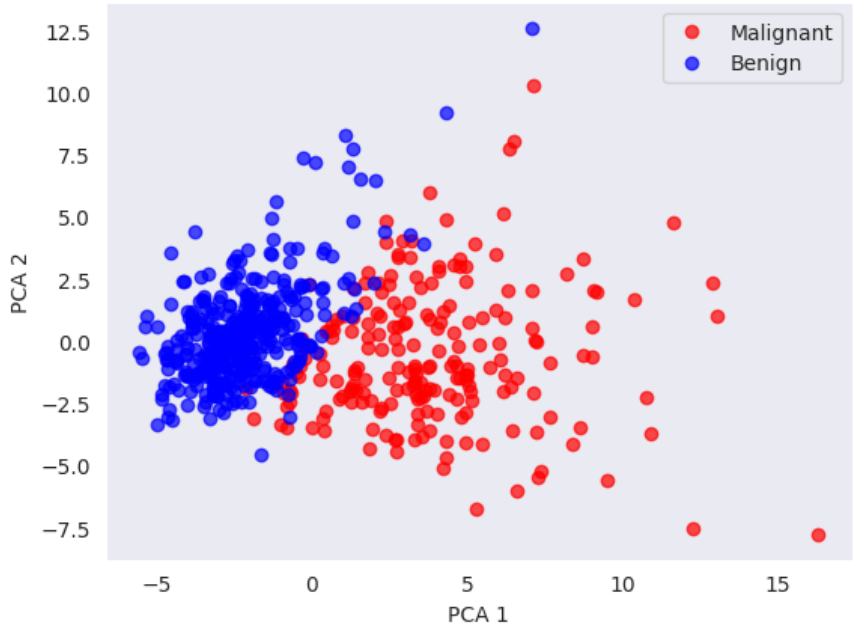
It contains:
The dataset has 569 records.
There are 32 Columns/ features:

Preprocessing:

- Dataset is checked for Null and duplicate values.
- Creating a correlation matrix for the Wisconsin Breast Cancer Diagnosis (WBCD) dataset is a valid approach in the data analysis process. A correlation matrix helps in understanding the relationships between different features in the dataset.



- Applying Principal Component Analysis (PCA) to the Wisconsin Breast Cancer Diagnosis (WBCD) dataset can be a suitable approach, especially if the goal is to reduce the dimensionality of the dataset while retaining most of the important information. PCA can help in identifying the most significant features that contribute to the variance in the data, which can be particularly useful if the dataset has a large number of features.



Models Used:

Decision Trees:
A decision tree classifier can forecast the label that will be assigned to new input data. Based on the input feature values, it separates the input data into subsets. Recursively dividing the input data based on the most instructive feature creates the decision tree.
Advantages and Disadvantages:

- the ability to handle both categorical and continuous input features
- the ability to interpret the resulting model and make decisions based on the learned rules
- the ability to handle missing values in the input data.

However, they can be prone to overfitting and may not perform well on datasets with a large number of features or complex relationships between features and labels.

Random Forests:

A Random Forest Classifier is a machine learning algorithm that forecasts the output label for a fresh input data point using a variety of decision trees. The input data is randomly sampled, and several decision trees are built using various subsets of the data to create a Random Forest Classifier. Every decision tree in the forest is consulted when a new input data point is classified.
Advantages and disadvantages:

- the ability to handle both categorical and continuous input feature
- the ability to reduce overfitting
- the ability to handle missing values in the input data
- the ability to provide a measure of feature importance

However, it can be computationally expensive and may not be as interpretable as single decision trees.

KNN:

It assigns an output label based on the majority class among the k closest data points to a new input point. The algorithm creates a tree-based data structure during training to quickly find the k closest neighbors of any new input data point.
Advantages and disadvantages:

- the ability to handle both categorical and continuous input features

- the ability to learn complex decision boundaries
- the ability to generalize well to new data

However, it may be sensitive to the choice of k and the distance metric used, and may be prone to overfitting if the training data is noisy or imbalanced.

Gradient Boosting:

It predicts the output label for each input data point using a decision tree-based algorithm. Using this method, decision trees are gradually added to the model, with each tree forecasting the residual error of the ones that came before it.
Advantages and disadvantages:

- Faster training time and lower memory usage, especially on large datasets
- It is also less prone to overfitting and can handle both categorical and continuous input features

However, it may not perform well on datasets with many irrelevant features or noise.

SVM:

The main idea of SVM is to find the optimal hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points. It aims to maximize the margin between data points of different classes.
Advantages and disadvantages:

- effective in high dimensional spaces and is relatively memory efficient
- It can be used for both classification and regression tasks
- It can handle non-linear data, making it adaptable to various types of datasets

However, The performance of SVM is heavily dependent on the choice of kernel and regularization parameters. Finding the right settings can be challenging.

Results:

This table has all the results for the models before and after hyperparameter tuning. Hyperparameter tuning was done for Random Forest and Decision Tree. GridSearchCV was used for tuning RF and RandomSearch was used for Decision Tree. The accuracy was less for Tuned DT and Tuned RF because there was another hyperparameter tuning done (manually) by having for loops that checked for some values from the list.

Table 1: Metrics of the Models

	Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	ROC
0	Random Forest	0.947368	0.930233	0.930233	0.957746	0.930233	0.943990
1	Tuned Random Forest	0.929825	0.926829	0.883721	0.957746	0.904762	0.920734
2	SVM	0.964912	0.953488	0.953488	0.971831	0.953488	0.962660
3	KNN	0.947368	0.930233	0.930233	0.957746	0.930233	0.943990
4	KNN From Scratch	0.956140	0.952381	0.930233	0.971831	0.941176	0.951032
5	MLP	0.973684	0.954545	0.976744	0.971831	0.965517	0.974288
6	Gradient Boosting	0.973684	0.954545	0.976744	0.971831	0.965517	0.974288
7	Decision Tree	0.964912	0.953488	0.953488	0.971831	0.953488	0.962660
8	Tuned DT	0.947368	0.911111	0.953488	0.943662	0.931818	0.948575

Conclusion:

Successful implementation of WBCD dataset for prediction of breast cancer. From the results we can see that the highest accuracy was achieved by two models that are Gradient Boosting and MLP i.e. 97.36% on this dataset.

We achieved all the objectives of this assignment.

GitHub Repo:
https://github.com/TheAmanGupta20/DataClassification.git