Inspiring Excellence

CSE422: Artificial Intelligence

Project Name: Breast Cancer Classification

Submitted By:

Group 1

Name	**ID**
Mashrur Safir Shabab	20241037
Hasan al mahmud chowdhury	23141069
Mahfuza Sultana Mim	18101703

Section: 07

Submitted To:

Mostofa Kamal Sagor, Zarin Tahia Hossain

Lecturer

Brac University

Introduction

Breast cancer is a major global public health issue, and early identification is key to enhancing patient outcomes. Machine learning algorithms in particular have shown considerable potential for improving the accuracy of breast cancer classification using Artificial intelligence (AI) techniques. We collect data on personal indicators associated with breast cancer, preprocess and engineer the features and train and evaluate the machine learning models. Thus, This Project provides a thorough analysis of breast cancer classification using AI techniques.

Motivation

This study on breast cancer classification and the purpose of this report's development are driven by a variety of factors, including both practical and scientific considerations. The following are some major causes:

Firstly**,** Breast cancer is a severe health issue, especially for women. The chance of a successful course of medication and positive patient outcomes are significantly increased by early diagnosis and appropriate classification of cancers. By helping doctors identify breast cancer more precisely, the creation of an ML-based classification system can improve healthcare.

Secondly**,** To enhance patient outcomes and cut costs, the healthcare sector is increasingly implementing data-driven strategies. This initiative fits in with the more significant trend of using data analytics and AI in healthcare.

Thirdly, artificial intelligence has advanced remarkably, especially in the areas of machine learning and deep learning. AI algorithms can effectively analyze complicated medical data when used in the exciting and potential field of medical diagnosis.

Then again, the Implementation of machine learning theories acquired in an artificial intelligence course is made in this project. It improves students' understanding of AI principles by giving them practical experience in data preparation, model selection, and evaluation.

Briefly put, the significance of the breast cancer classification effort stems from its potential to enhance the quality of healthcare through AI-driven diagnosis by enabling early detection and intervention, improving patient outcomes and reducing healthcare costs. It addresses the critical need for precise breast cancer diagnosis, takes advantage of AI advances, provides educational value, promotes multidisciplinary collaboration, and is consistent with the more significant trend of data-driven healthcare, making it an essential and crucial achievement.

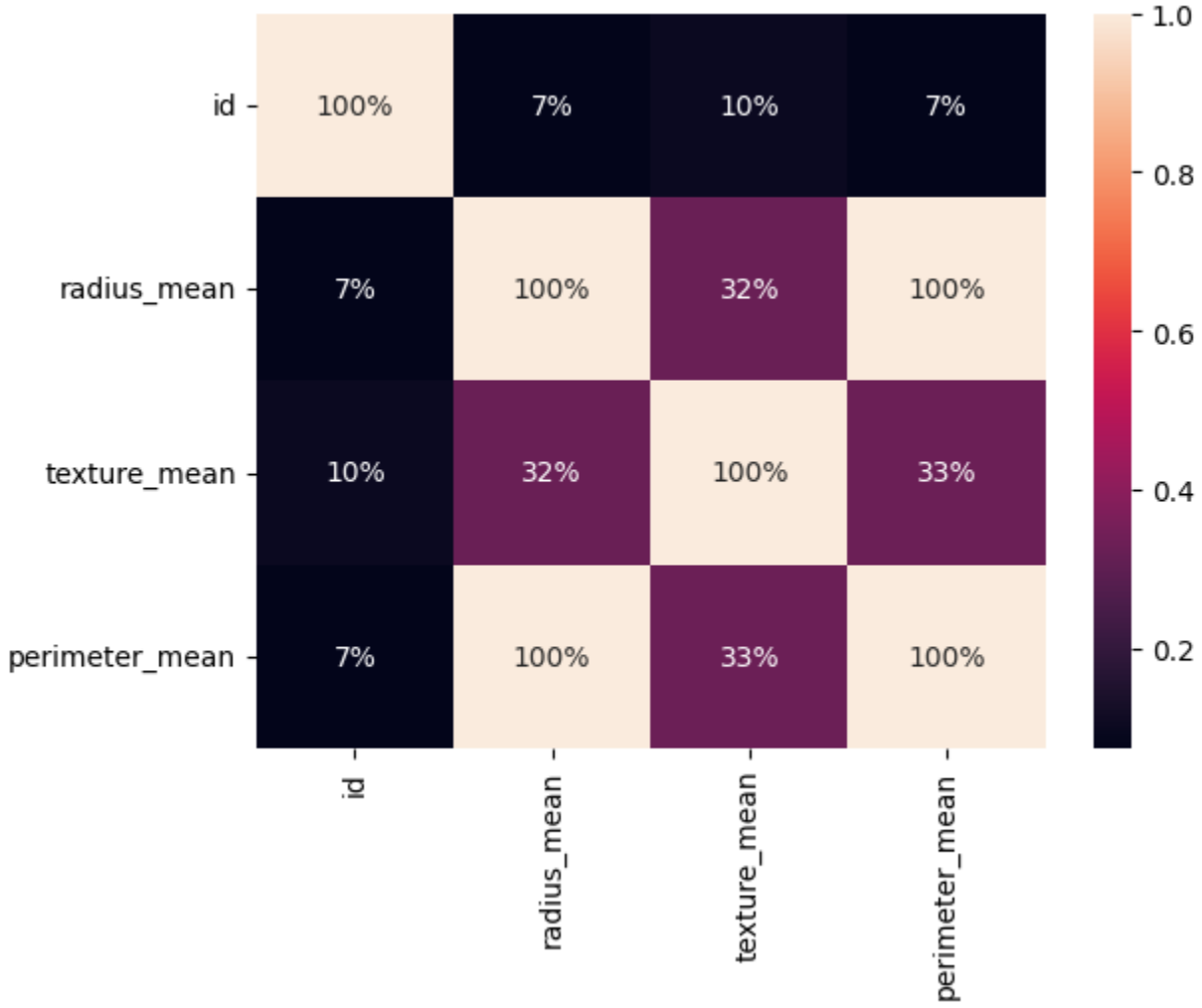
Dataset Description

Link:
https://www.kaggle.com/code/niteshyadav3103/breast-cancer-classification/input?fbclid=IwAR22RKVDEwDXHXbUWPVQUJIV7yYWdsASh1KItQbzy_05lQubNsqbcz9eMyE

Number of columns: 32
Number of features: 32

This is a dataset with 569 observations and 33 factors pertaining to breast cancer diagnosis is used in this study. The dataset consists of several different attributes that capture different elements of tumor characteristics. The mean and standard error of the radius, texture, perimeter, and area measurements are among these characteristics, along with others like smoothness, compactness, and concavity. Additionally, a diagnosis of either benign or malignant is included next to each entry, giving a categorical target variable for classification tasks. Notably, the dataset is largely complete except for the Unnamed: 32 column, which has null values and is therefore left out of the study. This dataset's numerous and varied properties provide a great basis for using machine learning.

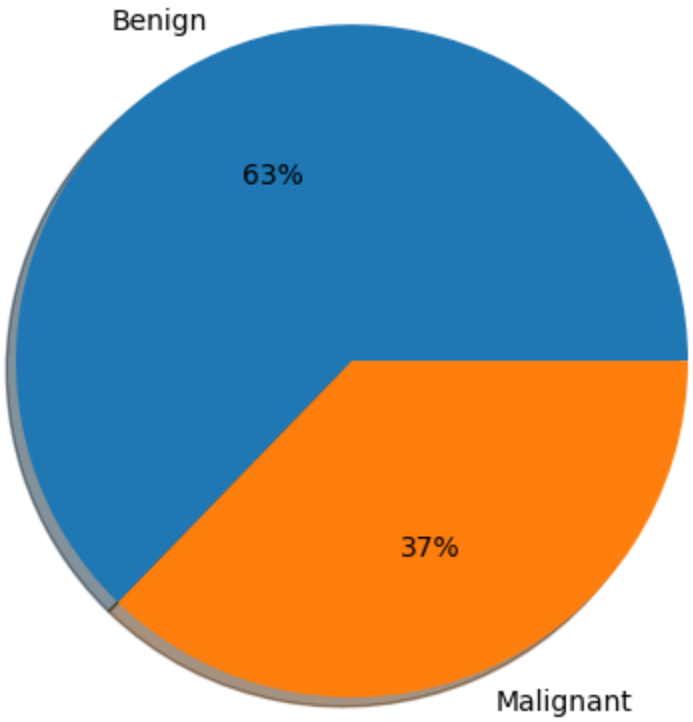
Correlation of the features along with the label/class:



Here, the image's text provides additional data information, such as the percentage of squares in the dataset. For instance, "1.0" indicates 100% of squares have the same ID, while "0.6" indicates 60% have the same mean radius. This visual representation helps understand data distribution and identify unusual patterns, allowing for a better understanding of the data.

Biased/Balanced:

As we can see in this pie chart, the data set is biased towards one class over the other.



Data Preprocessing:

Problem 1: As there are 569 null values in 1 column.

Solution: Drop the Column

Problem 2: There is a column named ID which had no impact on the decision of cancer being Malignant or Benign

Solution: While training the model that particular feature was not used

id	0
diagnosis	0
radius_mean	0
texture_mean	0
perimeter_mean	0
area_mean	0
smoothness_mean	0
compactness_mean	0
concavity_mean	0
concave points_mean	0
symmetry_mean	0
fractal_dimension_mean	0
radius_se	0
texture_se	0
perimeter_se	0
area_se	0
smoothness_se	0
compactness_se	0
concavity_se	0
concave points_se	0
symmetry_se	0
fractal_dimension_se	0
radius_worst	0
texture_worst	0
perimeter_worst	0
area_worst	0
smoothness_worst	0
compactness_worst	0
concavity_worst	0
concave points_worst	0
symmetry_worst	0
fractal_dimension_worst	0
Unnamed: 32	569
dtype: int64	

id	False
diagnosis	False
radius_mean	False
texture_mean	False
perimeter_mean	False
area_mean	False
smoothness_mean	False
compactness_mean	False
concavity_mean	False
concave points_mean	False
symmetry_mean	False
fractal_dimension_mean	False
radius_se	False
texture_se	False
perimeter_se	False
area_se	False
smoothness_se	False
compactness_se	False
concavity_se	False
concave points_se	False
symmetry_se	False
fractal_dimension_se	False
radius_worst	False
texture_worst	False
perimeter_worst	False
area_worst	False
smoothness_worst	False
compactness_worst	False
concavity_worst	False
concave points_worst	False
symmetry_worst	False
fractal_dimension_worst	False
Unnamed: 32	True
dtype: bool	

Before pre-processing

```
id                         0
diagnosis                  0
radius_mean                0
texture_mean               0
perimeter_mean             0
area_mean                  0
smoothness_mean            0
compactness_mean           0
concavity_mean             0
concave points_mean        0
symmetry_mean              0
fractal_dimension_mean     0
radius_se                  0
texture_se                 0
perimeter_se               0
area_se                    0
smoothness_se              0
compactness_se             0
concavity_se               0
concave points_se          0
symmetry_se                0
fractal_dimension_se       0
radius_worst               0
texture_worst              0
perimeter_worst            0
area_worst                 0
smoothness_worst           0
compactness_worst          0
concavity_worst            0
concave points_worst       0
symmetry_worst             0
fractal_dimension_worst    0
dtype: int64
```

```
id                         False
diagnosis                  False
radius_mean                False
texture_mean               False
perimeter_mean             False
area_mean                  False
smoothness_mean            False
compactness_mean           False
concavity_mean             False
concave points_mean        False
symmetry_mean              False
fractal_dimension_mean     False
radius_se                  False
texture_se                 False
perimeter_se               False
area_se                    False
smoothness_se              False
compactness_se             False
concavity_se               False
concave points_se          False
symmetry_se                False
fractal_dimension_se       False
radius_worst               False
texture_worst              False
perimeter_worst            False
area_worst                 False
smoothness_worst           False
compactness_worst          False
concavity_worst            False
concave points_worst       False
symmetry_worst             False
fractal_dimension_worst    False
dtype: bool
```

After pre-processing

Dataset splitting:

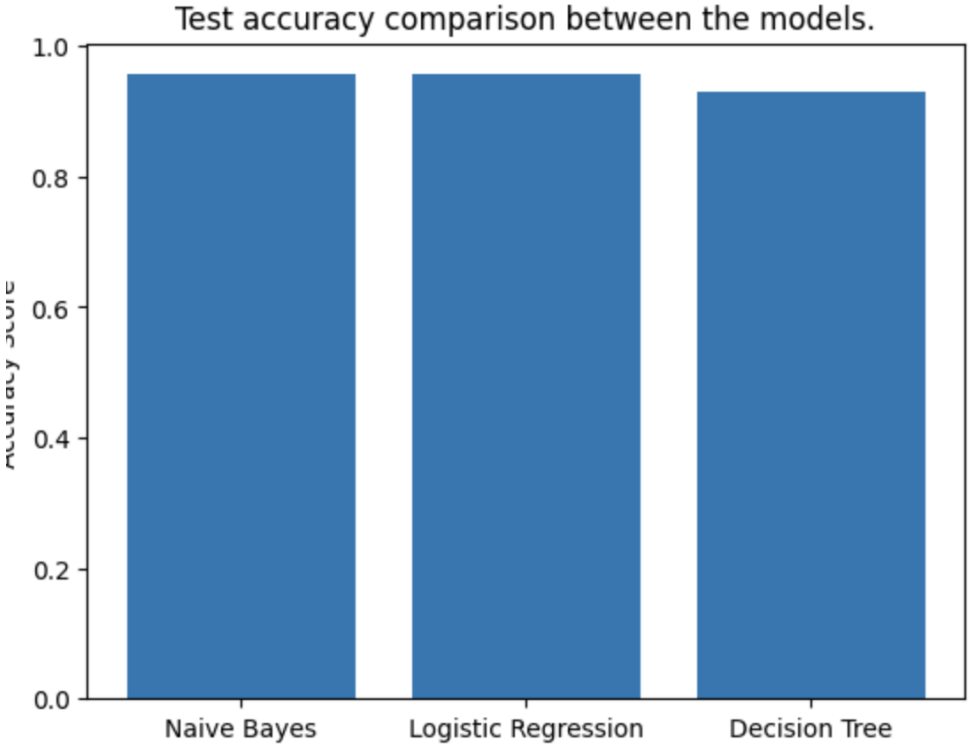
To train the model data splitting was done as 80% for training model and 20% for testing. On this basis- Training data set - 455 and Testing data set - 114

Model Training:

Model Name	Accuracy (%)	Error(%)
Naive Bayes classifier	95.61%	4.39%
Logistic Regression Classifier	95.61%	4.39%
Decision Tree Classifier	92.98%	7.02%

From the table, we can see that the Logistic Regression model and Naive Bayes Classifier model showed the best performance with 95.61% accuracy and only 4.39% error. On the 92.98% accuracy, the error was double that of the LogisticRegression model and Naive Bayes Classifier Model.

Model selection/Comparison analysis:



From the Bar chart comparison we can see that Logistic Regression and Naive Bayes Classifier have better performance than the other models.

Model Testing:

To taste model, the data was used-

1. Single instance- array([[1], [0], [0], [1],[0], [1], [1], [1], [0], [0], [0], [0], [1], [0], [1], [1], [0] [0], [1] , [0], [1], [0], [0], [1], [1], [0], [0], [0], [0], [0], [0], [0], [0], [0], [0], [0], [0], [1], [0], [0], [1], [1], [0], [0], [0], [0], [1], [1], [1], [0], [1], [1], [1], [0], [0], [0], [1], [1], [0], [0], [0], [1], [1], [0], [0], [1], [1], [0], [0], [0], [0], [0], [1], [1], [1], [1], [0], [0], [1], [0], [0], [0], [0], [0], [0], [1], [0], [0], [1], [1], [1], [0], [1], [1], [0], [0], [0], [0], [0], [0], [0], [0], [0], [1], [0], [0], [1], [0], [0],[1], [1], [0], [0], [0], [0], [0], [1]])

Result for Naive Bayes classifier:

```
[1 0 0 1 0 1 1 1 0 0 0 0 1 0 1 0 0 0 1 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 1 1 0 0 0 0 0 1 1 1 0 1 1 1 0 0 0 0 1 0 0 0 1 1 0 0 1 1 0 0 0 0 0 1 0
 1 1 0 0 1 0 0 0 0 1 0 1 0 0 1 1 1 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 0 1 1 0 0 0
 0 0 1]
```

Here, the accuracy is about 95.61%

Result for Logistic Regression

```
[1 0 0 1 0 1 1 1 0 0 0 0 1 0 1 0 0 0 1 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0
 0 0 1 1 1 0 0 0 0 1 1 1 0 1 1 1 0 0 0 0 1 0 0 0 1 1 0 0 1 1 0 0 0 0 0 1 0
 1 1 0 0 1 0 0 0 0 0 0 1 0 0 1 1 1 0 1 1 0 0 0 0 0 0 0 0 0 1 0 0 1 1 0 0 0
 0 0 1]
```

Here, the Accuracy is about 95.61%

Result for Decision Tree Classifier:

```
[1 0 0 1 0 1 1 1 0 0 0 0 1 0 1 1 0 0 1 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0
 0 0 1 1 1 0 0 0 0 1 1 1 0 1 0 1 0 0 1 1 1 0 0 0 1 1 0 0 1 0 0 0 0 0 1 1 0
 1 1 0 0 1 0 0 0 0 1 0 1 0 0 1 1 1 0 1 1 0 0 0 0 0 0 0 0 0 1 0 0 1 1 0 0 0
 0 0 1]
```

Here, the Accuracy is about 92.98% which is a bit lower than the other two models.

Conclusion:

For the classification of breast cancer in this project, we successfully created and assessed machine learning models. By applying various machine learning algorithms such as Logistic Regression, Decision tree, and Naive Bayes classifiers we obtained an accuracy ranging from 92.98% to 95.61% which shows that these models demonstrate outstanding performance in predicting breast cancer.

With an accuracy rate of 95.61% respectively, the Logistic Regression and Naive Bayes Model performed most favorably among the evaluated models, demonstrating their potential as beneficial tools for breast cancer prediction. With an accuracy of 92.98%, the Decision tree also displayed good results but it still needed to be improved in order to be more predictive.

These results demonstrate the possibility of exploiting crucial personal key indicators to predict breast cancer using machine learning methodologies. To ensure the accuracy and generalizability of these models in actual clinical settings, more investigation and validation are required leveraging larger datasets and additional assessment criteria.

However, the findings of this study offer a useful basis for further investigation and real-world applications of machine learning in the prediction of breast cancer, which has the potential to dramatically advance early identification and prevention of this serious health issue.

Future Work:

Machine learning models can improve their generalization to new data by using larger and more diverse datasets, which are representative of the real patient population. Moreover, new feature extraction methods can significantly impact the accuracy of the classification model, capturing subtle differences between benign and malignant tumors. Advanced machine learning methods, such as deep learning, are effective for image classification but can be computationally expensive and require large datasets. Ensemble learning methods combine predictions from multiple models to improve overall accuracy, making them an effective way to improve breast cancer classification models.

Clinical decision support systems can be developed using machine learning models to help doctors make better decisions about patient care. These systems can identify patients at high risk of developing breast cancer, recommend the best treatment course, and monitor patient progress during treatment. Overall, the development of machine learning models can significantly enhance the accuracy of breast cancer classification models.