Synopsis

Breast Cancer Detection: Using Machine Learning

GUIDED BY:

Sri Nath Dwivedi
Department of CSE

SUBMITTED BY:

Swapnil Siddhartha
Ashutosh Singh
Umang Raj Chaurasia

Abstract

Breast cancer is one of the leading causes for the death of women. In women Breast cancer is treated as the most significant issue. According to statistics released by the International Agency for Research on Cancer (IARC) in December 2020, Breast cancer has now overtaken lung cancer as the most commonly diagnosed cancer in women worldwide. Early diagnosis of this helps to prevent the cancer. If breast cancer is detected in early stage, then Survival rate is very high. Machine Learning methods are effective ways to classify data. Especially in the medical field, where those methods are widely used in diagnosis and analysis for decision making.

Objective

The aim of this project is to predict breast cancer tumors by using different machine learning approaches.

Methodology

1)Dataset collection

We have obtained Breast Cancer Wisconsin (Diagnostic) Dataset from Kaggle. Here 569 Patient's Data Was used for analysis, each instances have 32 Attributes with Diagnosis and Features. Each instance has a parameter of the cancerous non-cancerous cells and we will predict the cancer just by the input of features.

The values of features are in Numeric Format. The 'Target' means the patient Who is having Whether 'Benign' or 'Malignant' Cancer state. Benign means the patient is not having Cancer and Malignant means the patient is having Cancer.

2)Data Visualization

Visualizing the dataset with respect to two catagories

- Begnin: represented as 1
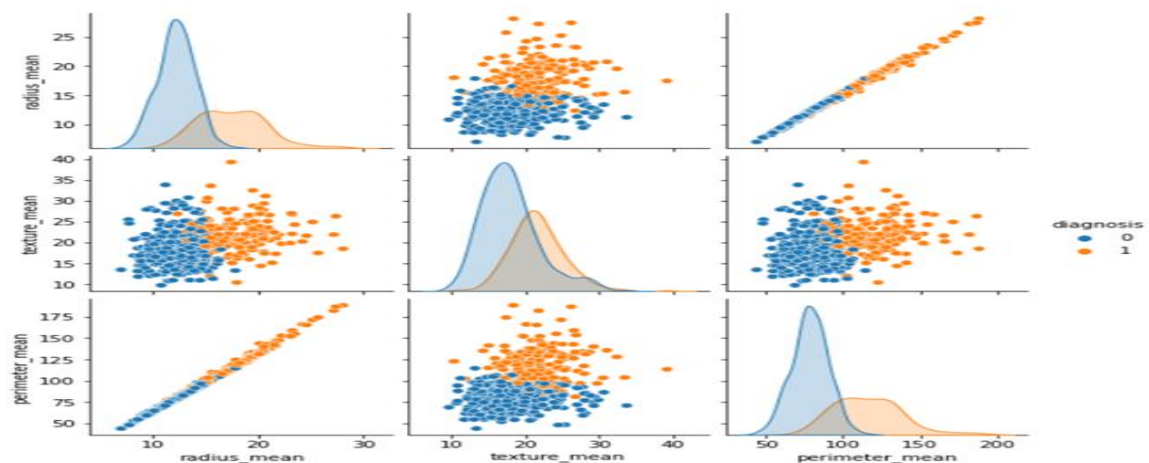- Malignant: represented as 0

Fig. 1 Pairplot of Features

3)Section Heading

We used Google colaboratory as our platform to write our code and get a prediction value. Model selection is the most important step in machine learning. Machine Learning algorithms can be classified as: supervised learning and unsupervised learning. For Our project, we only need supervised learning. We used all Methodologies to Predict the result and Noted their Accuracy.

4)Data Preparation

Separating the data into training data and testing data

```
[ ]  #creating 4 arrays
     X_train, X_test, Y_train,Y_test=train_test_split(X, Y, test_size=0.2, random_state=2)
```

```
▶  print(X.shape, X_train.shape, X_test.shape)
```

Preparing the data to train and test the model. Data is split into training and testing dataset. Eighty percentage of the is used to train the model and then remaining data is used to test the data. Then model is trained using training dataset and then it is tested .

5)Confusion Matrix and Accuracy

Confusion Matrix is used for evaluating the performance of a classification model. The Matrix compares the actual target values with predicted values by the machine learning model. It shows the ways in which your classification model gets confused when it makes predictions.

```
#accuracy on the training data
X_train_prediction=model.predict(X_train)
training_data_accuracy=accuracy_score(Y_train, X_train_prediction)

print('Accuracy on training data: ', training_data_accuracy)

Accuracy on training data:  0.9472527472527472
```

```
#accuracy on the test data
X_test_prediction=model.predict(X_test)
test_data_accuracy=accuracy_score(Y_test, X_test_prediction)

print('Accuracy on test data: ', test_data_accuracy)

Accuracy on test data:  0.9210526315789473
```

Conclusion

The early detection and classification of breast cancer help to prevent the disease's spread. This paper examined different machine learning techniques for detection of breast cancer. The objectives of our study were to analyse the Wisconsin breast cancer dataset by visualizing and evaluating Machine Learning Predictions. However, it is required that before running the algorithm, the dataset must be pre-processed. In future, we like to add larger dataset and check the efficiency and scalability of algorithm.

References

Kaggle, Google search