Name: Eshwari Shrike
GitHub link: https://github.com/Eshwari00/Happy-Monk-Assignment

DEEP LEARNING IN WISCONSIN BREAST CANCER DETECTION AND CLASSIFICATION

INTRODUCTION

- Breast cancer is a type of cancer that forms in the cells of the breasts.
- According to global statistics, Breast cancer (BC) is one of the most common cancers among women worldwide representing most new cancer cases and cancer-related deaths making it a significant public health problem in today's society.
- The correct diagnosis of BC and classification of patients into malignant or benign groups is the subject of much research.
- Because of its unique advantages in critical features detection from complex BC datasets, machine learning (ML) is widely recognized as the methodology of choice in BC pattern classification.

OBJECTIVE

- This analysis aims to observe which features are most helpful in predicting malignant or benign cancer and to see general trends that may aid us in model selection and hyper parameter selection.
- The goal is to classify whether the breast cancer is benign or malignant. To achieve this, I have used Deep learning Artificial Neural Network methods to fit a function that can predict the discrete class of new input.

DATASET SPECIFICATION

- Source dataset is in csv format.
- Dataset contain 569 rows and 32 columns.
- Attribute Information:
    1) ID number
    2) Diagnosis (M = malignant, B = benign)
- Ten real-valued features are computed for each cell nucleus:
    1) radius (mean of distances from centre to points on the perimeter)
    2) texture (standard deviation of gray-scale values)
    3) perimeter
    4) area
    5) smoothness (local variation in radius lengths)
    6) compactness (perimeter² / area — 1.0)
    7) concavity (severity of concave portions of the contour)
    8) concave points (number of concave portions of the contour)
    9) symmetry
    10) fractal dimension (“coastline approximation”)
- There are no missing values in dataset.
- The output of the dataset is “diagnosis” which notifies whether the cancer is Malignant(cancerous) or Benign(non-cancerous).
- The mean, standard error and “worst” or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

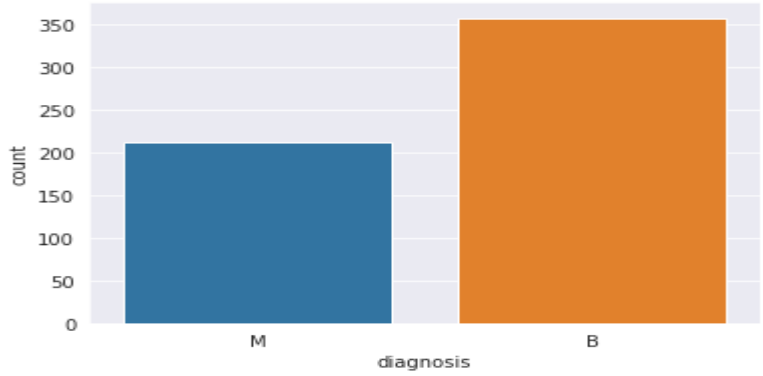
DATASET OVERVIEW & PREPROCESSING

- After going through the dataset I have noticed that data is clean and no null values are present.
- Therefore, only thing we have to do in preprocessing is Standardization.
- Standardization is must in Deep Learning, because while creating the Artificial Neural Network, we have to scale the data into smaller numbers because the deep learning algorithms multiplies the weights and input data of the nodes and takes lots of time, so for reducing that time we have to scale the data. For scaling I have used standard scaler, we scale the training and testing dataset.

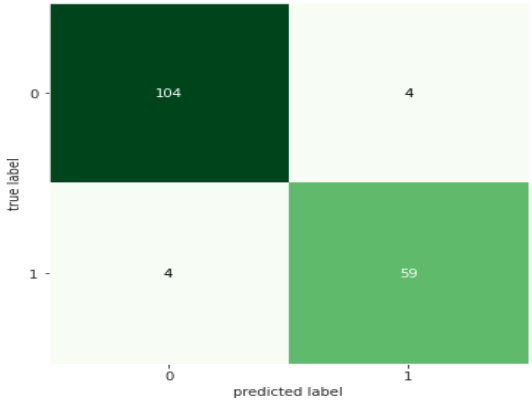
VISUALIZATION

- Data is unevenly distributed

```
B    357
M    212
Name: diagnosis, dtype: int64
```



- Confusion matrix:



- Most correlated features to target column.

```
smoothness_se            -0.067016
fractal_dimension_mean   -0.012838
texture_se               -0.008303
symmetry_se              -0.006522
fractal_dimension_se      0.077972
concavity_se              0.253730
compactness_se            0.292999
fractal_dimension_worst   0.323872
symmetry_mean             0.330499
smoothness_mean           0.358560
Name: diagnosis, dtype: float64
```

ALGORITHM USED FOR PREDICTION

- To solve this classification dataset we have used below algorithms after dividing in training and testing dataset.
  - Artificial Neural Network

PROCEDURE

Step 1: In the first step, Input units are passed i.e data is passed with some weights attached it to the hidden layer. Here we have created input layer with 9 nodes.
Step 2: Then inputs are multiplied by their weights. Weight is the gradient or coefficient of each variable, it shows the strength of the particular input. After assigning the weights, a bias variable is added.
Step 3: Then Activation Function is applied to the linear equation. The importance of the activation function is to inculcate nonlinearity in the model.
Step 4: Here we have used **ReLU AF** for input and hidden layers because it is faster to compute and its derivative is faster to compute. **Sigmoid** AF for output layer because we must predict the probability as an output. Since probability of anything exists only between the range of 0 and 1, sigmoid is the right choice.
Step 5: While compiling I have used Adaptive Moment Estimation (Adams) optimizer. Because data is large and have so many parameters. So Adam is best choice.
Step 6: After passing through every hidden layer, we move to the last layer i.e. our output layer which gives us the final output. The output layer will predict the output i.e., cancer is malignant or benign on the basis of given input parameters.

```
from tensorflow.keras.layers import Dropout   #Dropout library is imported to overcome the overfitting of model
ann=Sequential()
ann.add(Dense(units=9,activation="relu")) #Input layers with 9 neurons
ann.add(Dropout(0.3))  #Dropout rate
ann.add(Dense(units=9,activation="relu"))  #Hidden layer with 9 neuron
ann.add(Dense(units=1,activation="sigmoid")) #output layer
ann.compile(optimizer='adam',loss="binary_crossentropy")
ann.fit(xtrain,ytrain, epochs=95,validation_data=(xtest,ytest))
```

CLASSIFICATION REPORT & CONFUSION MATRIX

CLASSIFICATION REPORT:

- The classification report is key metrics in a classification problem.

	precision	recall	f1-score	support
0	0.97	0.96	0.97	108
1	0.94	0.95	0.94	63
accuracy			0.96	171
macro avg	0.95	0.96	0.96	171
weighted avg	0.96	0.96	0.96	171

This above figure shows the main classification metrics precision, recall and f1-score on a per-class basis.

1. Precision

- It represents that out of all the data points which have been identified as positive(malignant) by our model and how many are actual positive(malignant) datapoints.
- when our model predicts that a patient has malignant or Benign, it is correct around 94% and 97% respectively of the time.

- Precision also gives us a measure of the relevant data points. It is important that we don't start treating a patient who actually doesn't have a malignant, but our model predicted as having it.
- Hence, we will try to reduce the precision as much as possible.

2.Recall

- It represent that out of all actual positive datapoints (Malignant), how many have been truly identified as positive(malignant) by our model.
- For all the patients who actually have Malignant, recall tells us how many we correctly identified as having a Malignant.

3.F1_score

- F1 score is the harmonic mean of recall and precision.
- When we want that recall and precision should be balanced, we use f1 score.

4.Accuracy

- It represents the percentage of correctly predicted values out of all data points.
- Here accuracy is 96%.

CONFUSION MATRIX:

- Confusion Matrix is the visual representation of the Actual VS Predicted values.
- It represents the different combinations of Actual VS Predicted values.

```
[[104   4]
 [  3  60]]
```

- True Positive(TP) : TP represents , out of actual positive(malignant) data points how many are predicted correctly by model. Here TP is 60
- True Negative(TN): TN represents, out of actual negative(Benign) data points how many are falsely predicted by model. Here TN is 104.
- False Negative (FN) / Type II Error : FN represents, out of actual positive(malignant) data points how many are falsely predicted by model. Here FN is 3. Our objective is to minimize the FN because it will detect negative (Benign) patient as a positive(malignant).
- False Positive (FP) / Type I Error: FP represents, out of actual negative(Benign) data points how many are predicted correctly by model. Here FP is 4.

CONCLUSION

- It will be helpful for prediction of model, even if the patterns or symptoms gets evolved in future.
- The accuracy is 96%.
- From below figure we can say that with decreasing epochs v training loss also gets decreased but the validation loss remains constant.