الجامعة الإسلامية العالمية ماليزيا

INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA

يونيبرسيتي اسلام انتارابڠسا مليسيا

Garden of Knowledge and Virtue

KULLIYYAH OF INFORMATION & COMMUNICATION TECHNOLOGY

CSCI 4340 MACHINE LEARNING

SEMESTER 1, 2022/2023

SECTION 1

ASSIGNMENT 1

CLASSIFICATION ALGORITHM

AHMED JOBAER	**1918243**

LECTURER

DR. AMELIA RITAHANI BINTI ISMAIL

Table of Content

1.0 INTRODUCTION

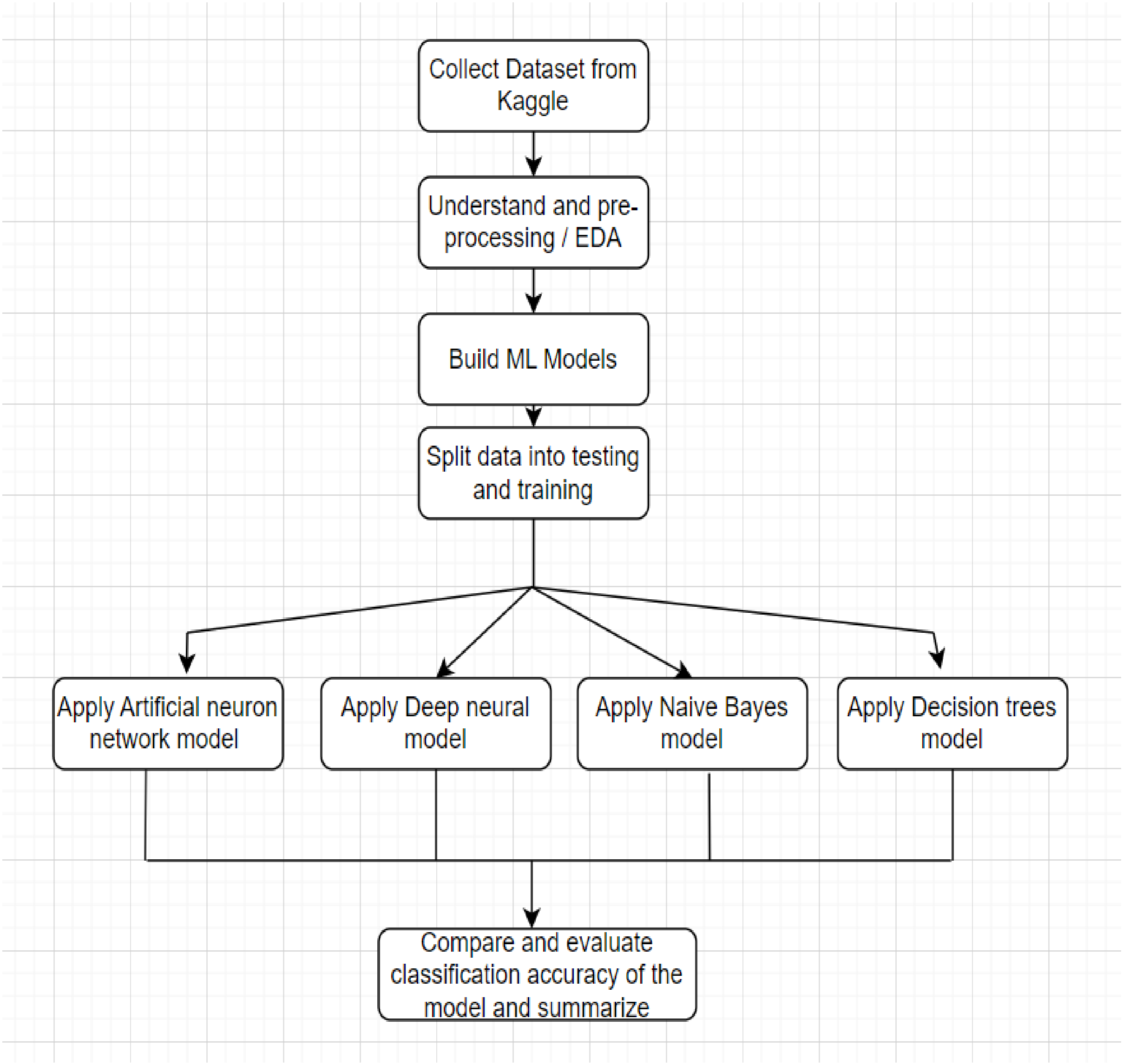
This report is focused on evaluating the performance of various machine learning algorithms in detecting breast cancer, using a dataset collected from Kaggle (https://www.kaggle.com/datasets/merishnasuwal/breast-cancer-prediction-dataset). Breast cancer is one of the most common types of cancer among women worldwide and early detection and diagnosis are crucial for improving the chances of survival. Machine learning models have been widely used in the field of medical imaging to identify and diagnose breast cancer. Among the various machine learning models, Artificial Neural Networks (ANNs), Decision Trees (DT), Naive Bayes (NB) and Deep Neural Networks (DNN) are some of the popular models that have been used to predict breast cancer. ANNs are multi-layered feedforward networks that are capable of learning complex relationships between inputs and outputs. DT are a set of algorithms that can be used for classification and regression tasks, they construct a decision tree from input features to the output. NB is a probabilistic model that is based on the Bayes theorem and is known for its simplicity and efficiency in handling large datasets. DNN are neural networks that have more than one hidden layer, and are able to learn complex representations of the data.

In this report, we will evaluate the performance of these models using the dataset collected from Kaggle, and compare their accuracy, precision and recall. By doing so, we hope to determine which algorithm is most effective in identifying breast cancer. Furthermore, we will provide insights into the potential use of these models in clinical settings, and how they can be used to improve the early detection and diagnosis of breast cancer.

2.0 Experimental Setup

2.0.1 Flowchart

The below shows the flow of experiment throughout the learning:


Collect Dataset from Kaggle
Understand and pre-processing / EDA
Build ML Models
Split data into testing and training
Apply Artificial neuron network model
Apply Deep neural model
Apply Naive Bayes model
Apply Decision trees model
Compare and evaluate classification accuracy of the model and summarize

2.1.0 import the dataset and explore

2.1 Dataset

Breast_cancer_data.csv is collected from Kaggle. It is a platform for data science competition with a web-based interface for data exploration, analysis, modelling and submission of predictions for evaluation.

This Dataset contains information about lumps from x-ray. It has 6 columns, 5 of the columns have features to predict if the patient should go for diagnosis or not.

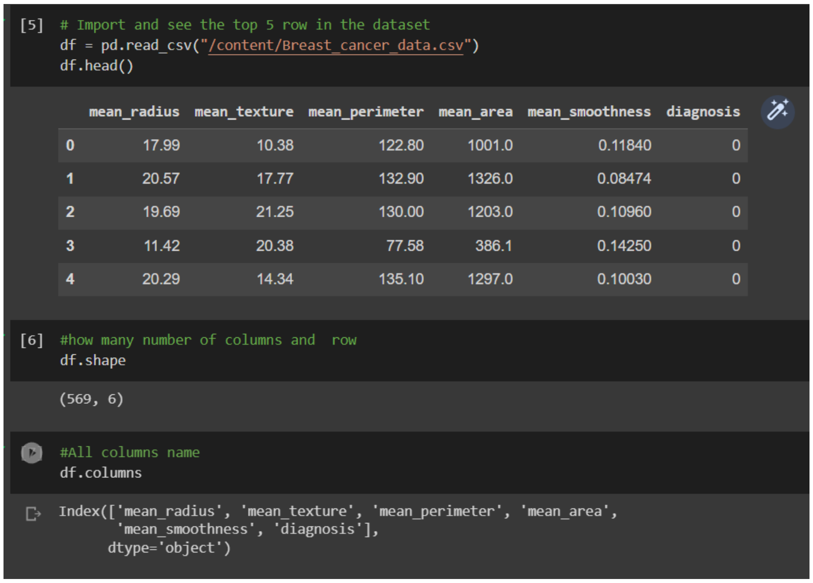
1. Mean_radius (float)

2. Mean_texture (float)
3. Mean_perimeter (float)
4. Mean_area (float)
5. Mean_smoothness(float)
6. Diagnosis (binary)

2.2.0 EDA

Exploratory Data Analysis (EDA) is an essential step in understanding and preparing data for modelling. The process typically involves Data Cleaning,Data Visualization, Data Transformation, and Data Analysis.

Import the dataset and understand the overall dataset. What are the columns and what it contains and many more.



```
[5] # Import and see the top 5 row in the dataset
    df = pd.read_csv("/content/Breast_cancer_data.csv")
    df.head()
```

	mean_radius	mean_texture	mean_perimeter	mean_area	mean_smoothness	diagnosis
0	17.99	10.38	122.80	1001.0	0.11840	0
1	20.57	17.77	132.90	1326.0	0.08474	0
2	19.69	21.25	130.00	1203.0	0.10960	0
3	11.42	20.38	77.58	386.1	0.14250	0
4	20.29	14.34	135.10	1297.0	0.10030	0

```
[6] #how many number of columns and  row
    df.shape
```

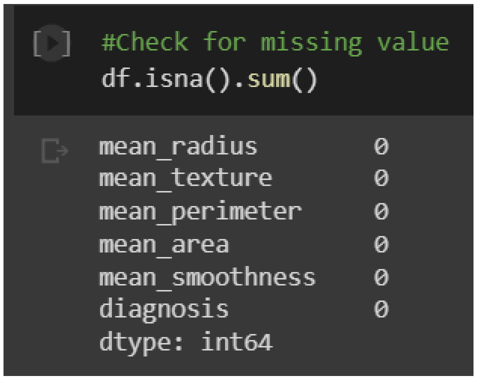
```
(569, 6)
```

```
#All columns name
df.columns
```

```
Index(['mean_radius', 'mean_texture', 'mean_perimeter', 'mean_area',
       'mean_smoothness', 'diagnosis'],
      dtype='object')
```

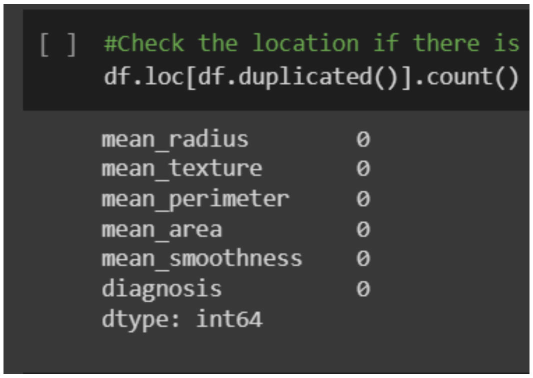
2.2.1 Data Cleaning

This section identifies and treats missing, duplicate or inconsistent data. But this dataset is already clean as it dont have any duplicate or missing value.



```
#Check for missing value
df.isna().sum()
```

```
mean_radius         0
mean_texture        0
mean_perimeter      0
mean_area           0
mean_smoothness     0
diagnosis           0
dtype: int64
```

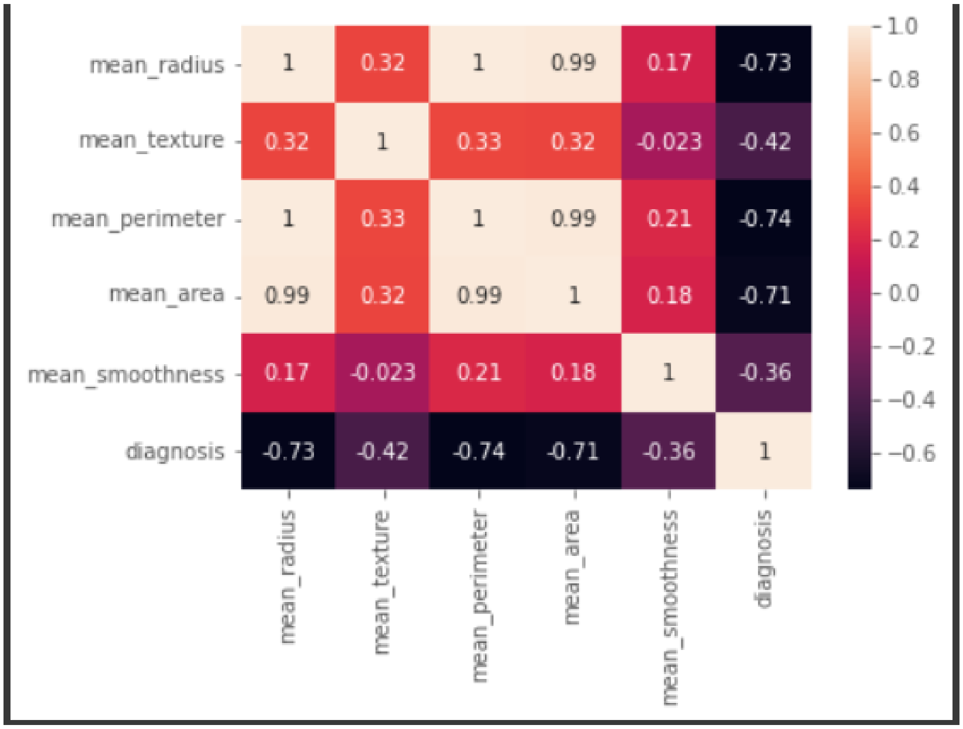


```
#Check the location if there is
df.loc[df.duplicated()].count()
```

```
mean_radius         0
mean_texture        0
mean_perimeter      0
mean_area           0
mean_smoothness     0
diagnosis           0
dtype: int64
```

2.2.2 Data Visualization

See the correlation for every column. All 5 feature columns are negatively correlated with the diagnosis column.



3.0 Evaluating our model

The Figure below shows the confusion matrix that displays the summary results of the heart disease classification that is classified according to 4 classes: True Negative (TN), False Negative (FN), False Positive (FP), and True Positive (TP).

		Predicted values	
		0	1
Actual values	0	True Negative (TN)	False Position (FP)
	1	False Negative (FN)	True Positive (TP)

3.1 Evaluation Metrics

A TP. FP. and FN. were used to assess the performance of this supervised machine learning task with multiple classes as output. The following is the formula for calculating these values:

3.1.1 Accuracy

Accuracy (ACC) will show how many of our predictions were correct. Accuracy should be as high as possible for us to know that the model is working properly.

ACC= $\frac{(TP+TN)}{(TP+TN+FP+FN)} X\,100$

3.1.2 Precision

The accuracy of a model's positive prediction is known as precision, also known as Positive Predictive Value (PPV).

PPV = $\frac{TP}{(TP+FP)} X\,100$

3.1.3 Sensitivity

By dividing the total number of positives by the total number of positive predictions made, the sensitivity, recall, or true positive rate (TPR) is calculated. The ideal recall value is 1.0.

R= $\frac{TP}{TP+FN}$

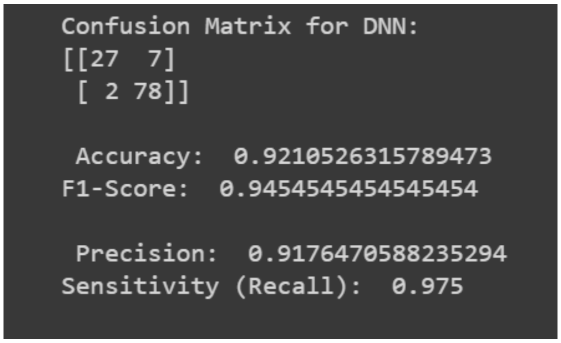
F-measure : harmonic mean of precision and recall

F = $\frac{2pr}{p+r}$

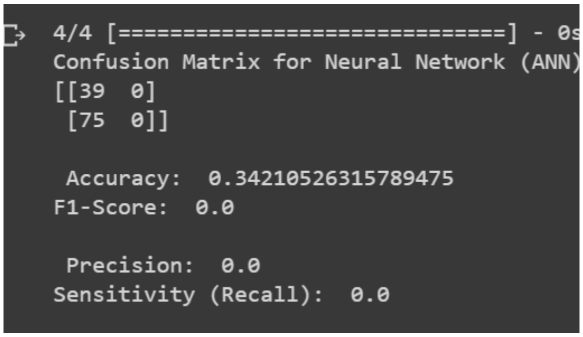
Evaluation using Cross-Validation Accuracy				
Parameter	**ANN**	**DT**	**NB**	**DNN**
Cross	0.3421	0.9649	0.9298	0.9210

Validation Accuracy				

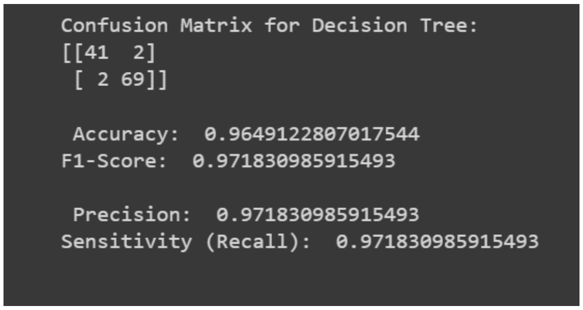
Confusion Matrices Comparison				
Parameter	**ANN**	**DT**	**NB**	**DNN**
TN	39	41	40	27
FP	75	5	5	7
FN	0	3	3	2
TP	0	69	66	78



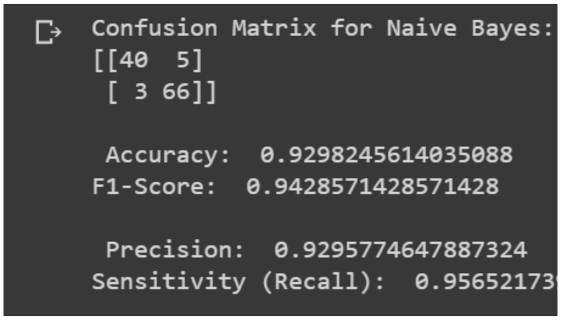
```
Confusion Matrix for DNN:
[[27  7]
 [ 2 78]]

 Accuracy:  0.9210526315789473
F1-Score:  0.9454545454545454

 Precision:  0.9176470588235294
Sensitivity (Recall):  0.975
```



```
4/4 [==============================] - 0s
Confusion Matrix for Neural Network (ANN)
[[39  0]
 [75  0]]

 Accuracy:  0.34210526315789475
F1-Score:  0.0

 Precision:  0.0
Sensitivity (Recall):  0.0
```



```
Confusion Matrix for Decision Tree:
[[41  2]
 [ 2 69]]

 Accuracy:  0.9649122807017544
F1-Score:  0.971830985915493

 Precision:  0.971830985915493
Sensitivity (Recall):  0.971830985915493
```



```
Confusion Matrix for Naive Bayes:
[[40  5]
 [ 3 66]]

 Accuracy:  0.9298245614035088
F1-Score:  0.9428571428571428

 Precision:  0.9295774647887324
Sensitivity (Recall):  0.956521739
```

5.0 Result And Discussion

We present the results of our evaluation of various machine learning algorithms for detecting breast cancer using a Kaggle dataset in this section. We used accuracy, precision, and recall as evaluation metrics to compare the performance of Artificial Neural Networks (ANNs), Decision Trees (DT), Naive Bayes (NB), and Deep Neural Networks (DNN).

Based on our evaluation table, we discovered that Deep Neural Networks, Naive Bayes, and Decision Trees were more accurate than Artificial Neural Networks. This suggests that these models may be better at identifying breast cancer and are more suited to our dataset.

The confusion matrices of the four models were also examined. Decision Trees, Naive Bayes, and Deep Neural Networks all achieved satisfactory results in their confusion matrices, showing that they were able to correctly classify the majority of the dataset's cases. The Artificial Neural Networks model's confusion matrix suggested it was less successful in identifying breast cancer on this dataset.

5.0 Conclusion

To conclude, using a Kaggle dataset, our group was able to assess the performance of various machine learning algorithms in detecting breast cancer. All group members contributed equally to the assignment, and we discovered that Decision Trees, Naive Bayes, and Deep Neural Networks outperformed Artificial Neural Networks in terms of accuracy, precision, and recall. These findings suggest that these models are better suited to our dataset and have the potential to improve breast cancer detection and diagnosis in clinical settings. Furthermore, by analysing the confusion matrices, we discovered that the three classifiers performed well, but the Artificial Neural Networks performed differently. Overall, we learned the value of evaluating different models and their performance on specific datasets to determine the most effective algorithm for a given task, as well as the value of equal participation in a team project.