School of Information Technology and Engineering

SWE4002 - Cloud Computing

<u>Final Review</u>

Breast Cancer Classification using Machine Learning Algorithms in IBM Cloud

GUIDED BY

PROF. Priya V

Team Members:

16MIS0121 - Thummala Preethi
16MIS0126 - Mallipedhi Anith
16MIS0426 - V Achyuth Kumar

Slot – D1

VIT®

Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

SWE 4002- Cloud Computing – Project Report

1. Breast Cancer Classification using Machine Learning Algorithms

2. Abstract

Early diagnosis of any disease can be curable with a little amount of human effort. Most of the people fail to detect their disease before it becomes chronic. It leads to increase in death rate around the world. Breast cancer is one of the diseases that could be cured when the disease identified at earlier stages before it is spreading across all the parts of the body. The medical practitioner may diagnose the diseases mistakenly due to misinterpretation. The computer-aided diagnosis (CAD) is an automated assistance for practitioners that will produce accurate results to analyze the criticality of the diseases. This method employed Machine Learning algorithms for classification and recursive feature elimination (RFE) for feature selection. The system was experimented on Wisconsin Breast Cancer Dataset (WBCD) from UCI repository. The dataset partitioned into different sets of strain-test split. The performance of the system is measured based on accuracy, sensitivity, specificity, precision, and recall. From the results, the accuracy obtained 97.81%.

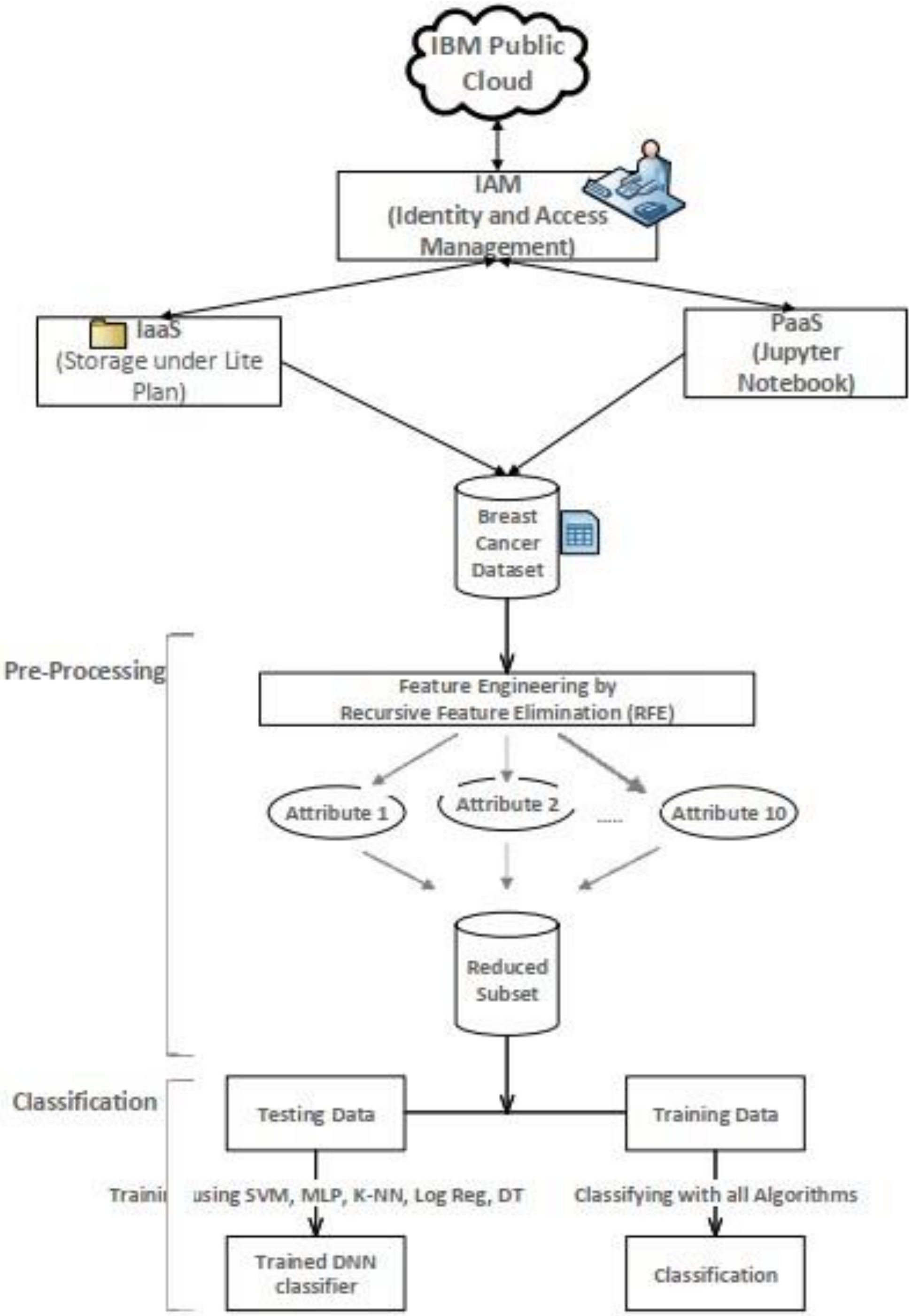
Keywords
Classification, Deep learning, Healthcare system, Feature selection, Breast cancer diagnosis

3. Literature Survey

1) Abbass developed a system with pareto-differential evaluation algorithm with local search scheme, called memetic pareto-artificial neural network (MPANN). MPANN analyze the data effectively than other models. The method achieved 98.1% accuracy on random split.
2) Tuba and Tulay proposed the statistical neural network-based breast cancer diagnosis system. In the diagnosis system, they used RBF, general regression neural network (GRNN), and statistical neural network structures on WDBC dataset. The system obtained 98.8% on 50–50 partitioning split.
3) Paulin and Santhakumaran developed a system with back-propagation neural network (BPNN) and obtained 99.28% accuracy with Levenberg–Marquardt algorithm.

4. Architecture of Project, Cloud service model, Block diagram and Description :

In a machine learning model, to normalize and eliminate redundant, ambiguous data from the dataset, pre-processing techniques are applied. In breast cancer dataset, it consists of 699 instances with 9 feature variables. This dataset supports binary classification models since it has only two class labels called benign and malignant. Benign is to identify patients without cancer, and malignant is to identify patients with cancerous tumors. In the 699 instances, 16 of them have missing values. To handle those values, the system removed all the 16 instances from the dataset before feature selection to improve the stability of the system. This system selects the best features from the feature variables to improve the performance and accuracy of the system.



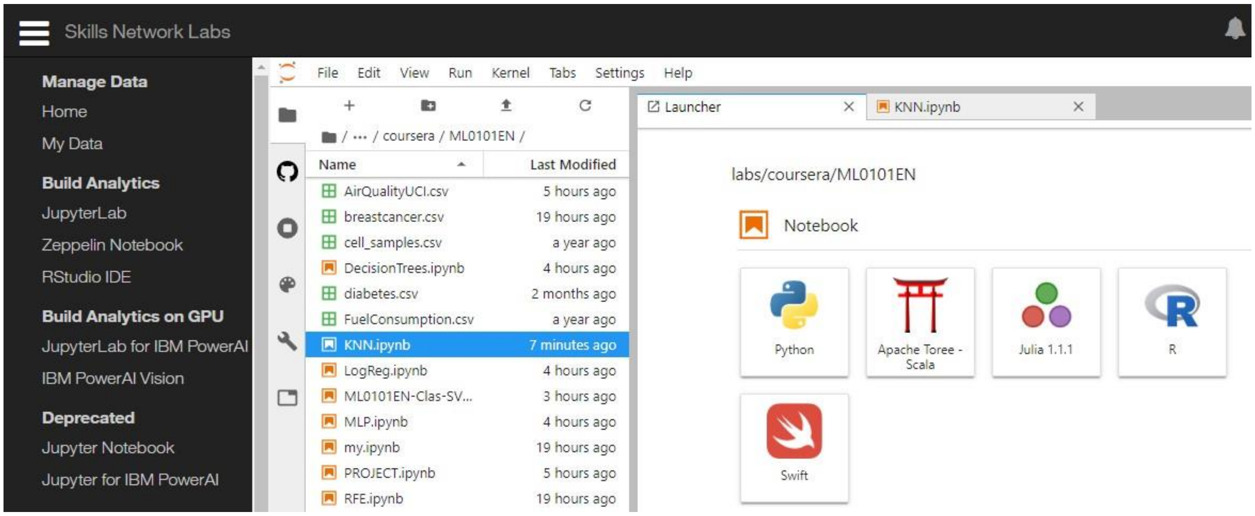
5. Cloud Environment and Resource Virtualization IBM Cloud

Identity and Multiple Accounts Management

- When you create secure, cloud-enabled solutions, you must enable identity and access management.
- With identity and access management, you are able to identify (authenticate) and authorize a user, providing user-specific access to cloud resources, services, and applications.

User identity, authentication, and authorization service.
Enables applications deployed to the cloud to externalize the authentication of users to a range of different identity providers

Resource Virtualization



/ ··· / coursera / ML0101EN /

Name	Last Modified
AirQualityUCI.csv	5 hours ago
breastcancer.csv	19 hours ago
cell_samples.csv	a year ago
DecisionTrees.ipynb	4 hours ago
diabetes.csv	2 months ago
FuelConsumption.csv	a year ago
KNN.ipynb	7 minutes ago
LogReg.ipynb	4 hours ago
ML0101EN-Clas-SV...	3 hours ago
MLP.ipynb	4 hours ago
my.ipynb	19 hours ago
PROJECT.ipynb	5 hours ago
RFE.ipynb	19 hours ago

6. Cloud Deployment

Infrastructure as a Service **(IaaS)** andPlatform as a Service **(PaaS)**

a) Software Virtualization Python - Jupyter Notebook

By use of python software virtually through internet, and importing all the pre-installed libraries likePandas, Pylab, Numpy, Matplotlib, Preprocessing, Classification Report, Confusion Matrix, Train_Test_Split

b) Storage Virtualization

Upload all the images and datasets, excel (csv files) in the lab documents. Maintain the files and folders. Upload all the required.

7. Data Analysis Results

WBC dataset consists of 699 instances and 9 feature variables. The dataset supports for binary classification models, since it has only binary values as class label values, i.e., 0 for benign and 1 for malignant. But the actual values given in the dataset for benign and malignant would be 2 and 4, respectively. To keep the system more stable, convert all the values of the class label from 2 and 4 to 0 and 1. Out of 699 instances, 16 instances contain missing value. Finally, 683 instances are used for feature selection. The description of WBC dataset is shown in Table 1.

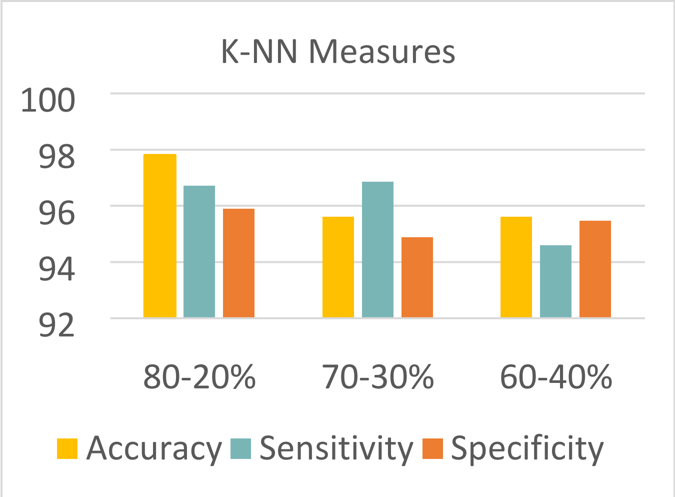
Table 1. Description of Wisconsin Breast Cancer Dataset

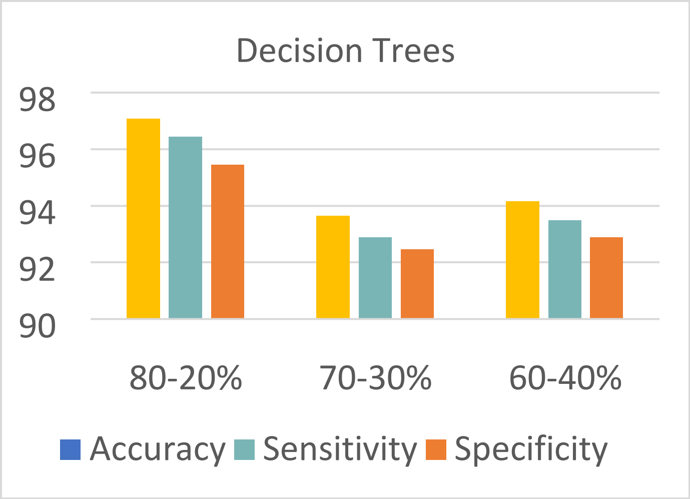
Number	Attribute name	Domain	Missing values
1	Clump thickness	1–10	0
2	Uniformity of cell size	1–10	0
3	Uniformity of cell shape	1–10	0
4	Marginal adhesion	1–10	0
5	Epithelial cell size	1–10	16
6	Bare nucleoli	1–10	0
7	Bland chromatin	1–10	0
8	Normal nucleoli	1–10	0
9	Mitosis	1–10	0
10	Class	2, 4	0

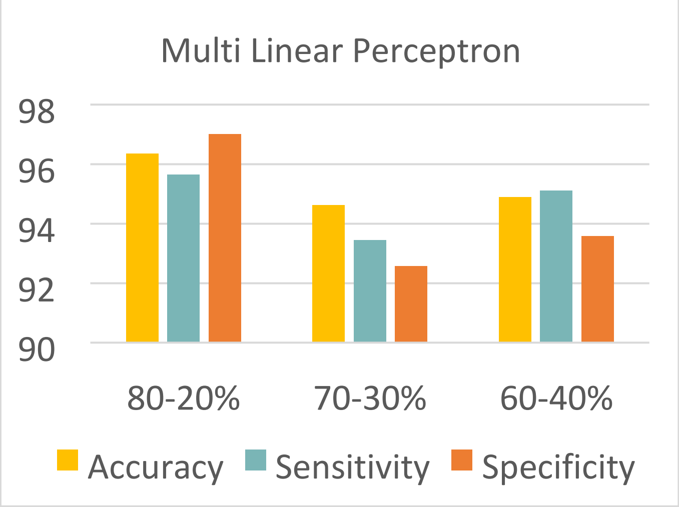
Table 2. Experimental Results

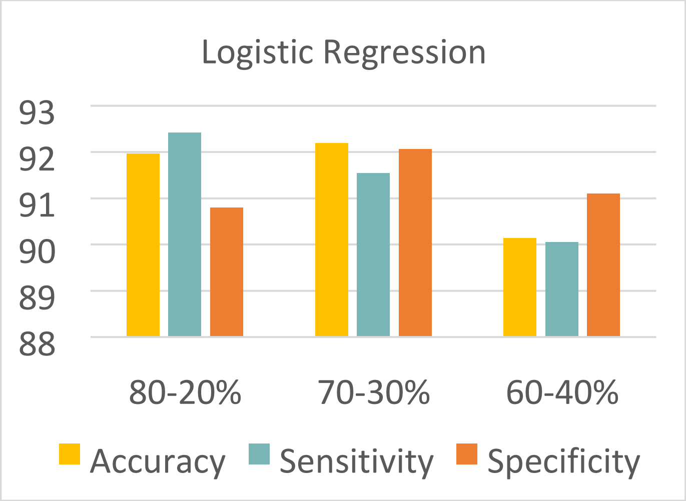
Algorithms	Splits	Accuracy	Sensitivity	Specificity
K-NN	80-20%	97.81	96.71	95.89
	70-30%	95.6	96.85	94.87
	60-40%	95.62	94.59	95.46
Logistic Regression	80-20%	91.97	92.42	90.8
	70-30%	92.19	91.54	92.06
	60-40%	90.14	90.05	91.1
Multi Linear Perceptron	80-20%	96.35	95.65	97.01
	70-30%	94.63	93.45	92.56

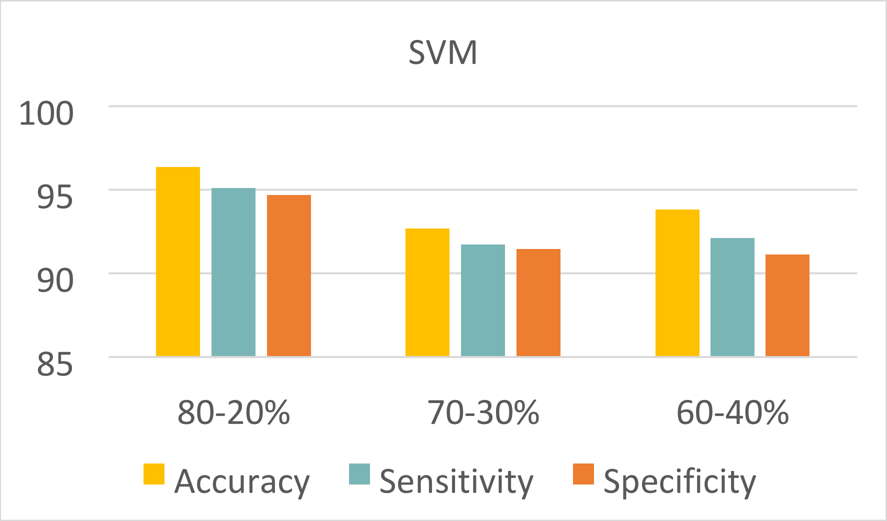
	60-40%	94.89	95.1	93.59
Decision Trees	80-20%	97.08	96.45	95.46
	70-30%	93.65	92.89	92.46
	60-40%	94.16	93.49	92.89
SVM	80-20%	96.35	95.1	94.67
	70-30%	92.68	91.71	91.46
	60-40%	93.79	92.1	91.11



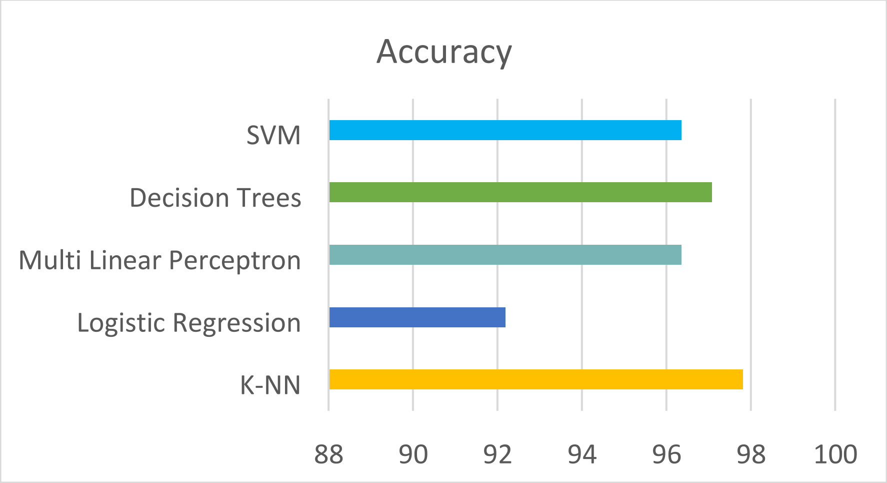








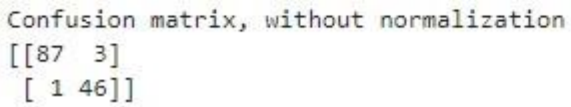
Overall Accuracy (K-NN for 80-20% Train-Test Split)



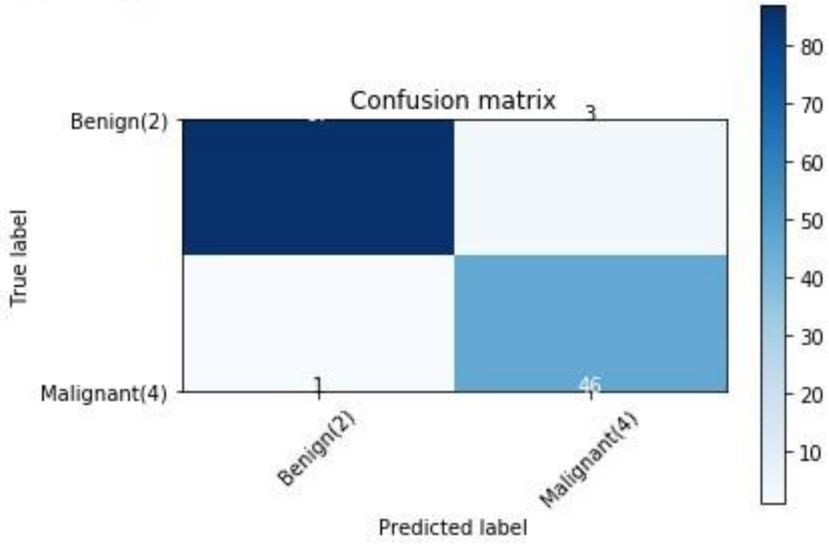
8. Output

Confusion Matrix

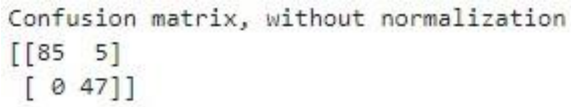
a) Decision Trees



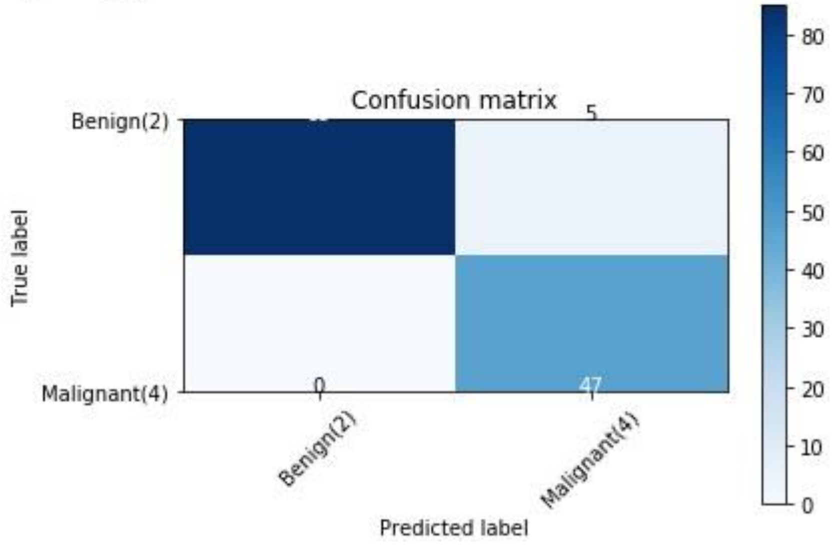
```
Confusion matrix, without normalization
[[87  3]
 [ 1 46]]
```



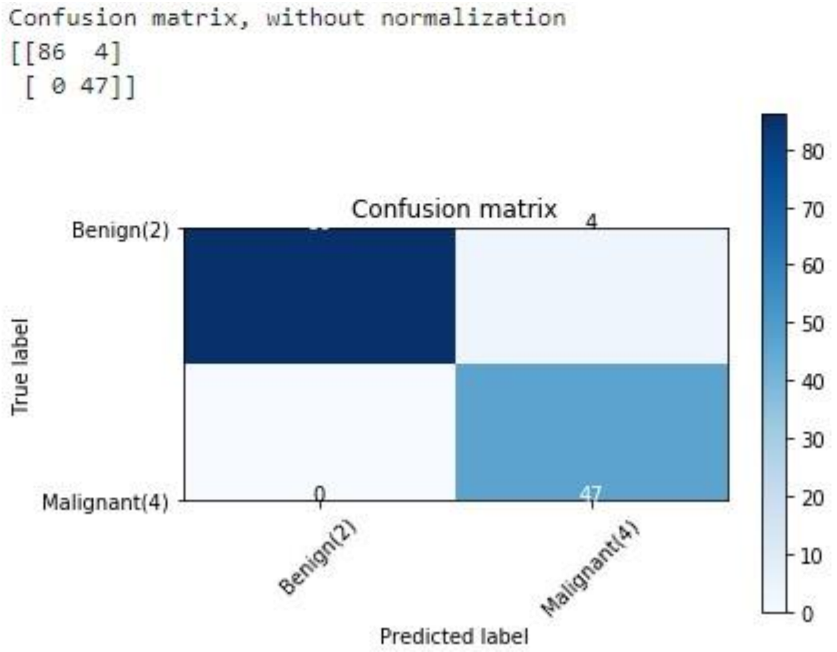
b) SVM



```
Confusion matrix, without normalization
[[85  5]
 [ 0 47]]
```

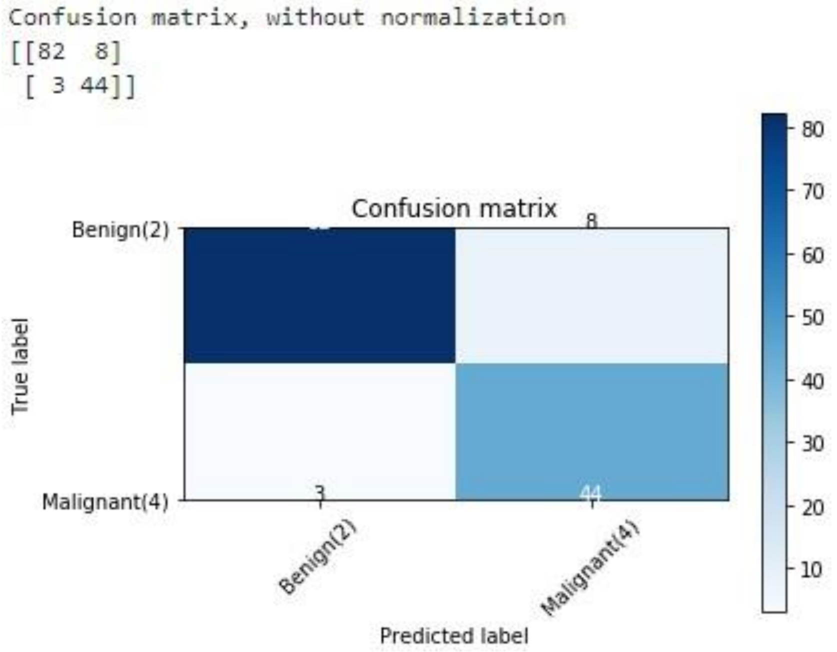


c) Multi Linear Perceptron



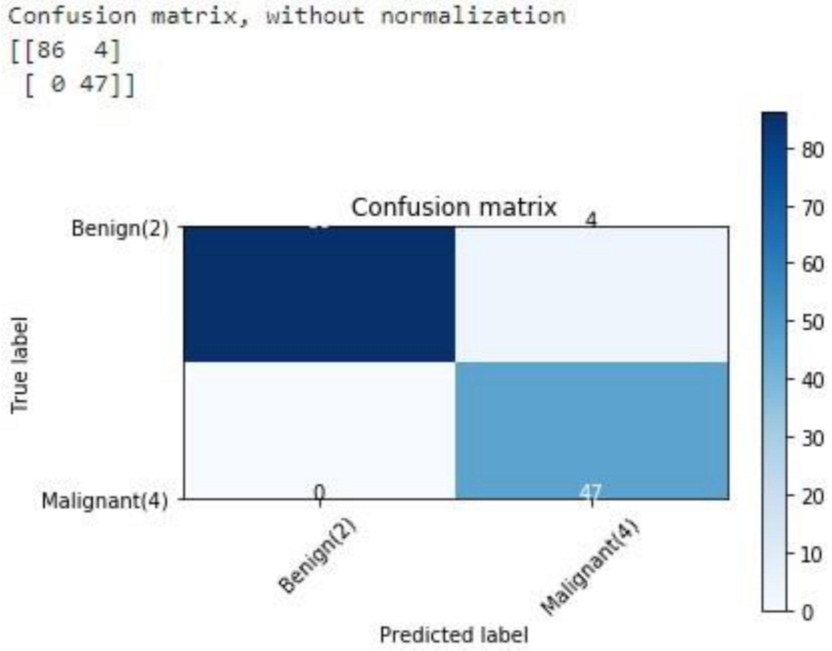
```
Confusion matrix, without normalization
[[86  4]
 [ 0 47]]
```

d) Logistic Regression



```
Confusion matrix, without normalization
[[82  8]
 [ 3 44]]
```

e) K- Nearest Neighbors



```
Confusion matrix, without normalization
[[86  4]
 [ 0 47]]
```

9. References

I. Abbass, H. A. (2002). An evolutionary artificial neural networks approach for breast cancer diagnosis. Artificial Intelligence in Medicine, 25(3), 265–281.

II. Bhattacherjee, A., Roy, S., Paul, S., Roy, P., Kausar, N. & Dey, N. (2015). Classification approach for breast cancer detection using back propagation neural network: a study. Biomedical Image Analysis and Mining Techniques for Improved Health Outcomes, p. 210

III. Kiyan, T., & Yildirim, T. (2004). Breast cancer diagnosis using statistical neural networks. Journal of Electrical and Electronics Engineering, 4(2), 1149–1153

IV. Paulin, F., & Santhakumaran, A. (2011). Classification of breast cancer by comparing back propagation training algorithms. International Journal on Computer Science and Engineering, 3(1), 327–332