UNIVERSITY OF ULSTER
FACULTY OF COMPUTING, ENGINEERING AND BUILT ENVIRONMENT
COURSEWORK SUBMISSION SHEET

This sheet must be completed in full and attached to the front of your project paper before submission to Module Coordinator/Instructor, Dr. Xuemei Ding, via Blackboard Learn.

Student’s Name ………Surya Patnaik…………………………………………

Registration No ………B00749292………………………………………………

Course Title ………………MSc Data Science……………………………………

Module Code/Title ……COM737/Machine Learning and Data Modelling………

Lecturer …………………………Dr. Xuemei Ding………………………………..

Date Due …………………………28th April 2019……………….………………

(NB: Latest hand-in time is 23:59 pm on the due date unless otherwise advised)

Submitted work is subject to the following assessment policies:

1. Coursework must be submitted by the specified date.
2. Students may seek prior consent from the Course Director to submit coursework after the official deadline; such requests must be accompanied by a satisfactory explanation, and in the case of illness by a medical certificate.
3. Coursework submitted without consent after the deadline will not normally be accepted and will therefore receive a mark of zero.

I declare that this is all my own work and that any material I have referred to has been accurately referenced. I have read the University’s policy on plagiarism and understand the definition of plagiarism. If it is shown that material has been plagiarised, or I have otherwise attempted to obtain an unfair advantage for myself or others, I understand that I may face sanctions in accordance with the policies and procedures of the University. A mark of zero may be awarded and the reason for that mark will be recorded on my file.

Comparative Study of classification techniques using Breast Cancer Wisconsin Diagnostic Data Set

Surya Patnaik
MSc (PT) Data Science 2018-19
Faculty of Computing, Engineering and Built Environment
Ulster University
Jordanstown, Northern Ireland
Patnaik-s@ulster.ac.uk

Abstract**—**The objective of this paper is to undertake analysis of the Breast Cancer Wisconsin (Diagnostic) Data Set and do a comparative study of machine learning techniques followed by predicting whether the cancer is benign or malignant.

Keywords - Machine Learning, Data Classification, Breast Cancer Wisconsin Dataset, Comparative Study

I. Introduction

Breast cancer is the 4th most common cause of cancer death in the UK, accounting for 7% of all cancer deaths (2016). Breast Cancer can affect both males and females. In females in the UK, breast cancer is the 2nd most common cause of cancer death, with around 11,500 deaths in 2016. About one in eight women are diagnosed with breast cancer during their lifetime[1]. Worldwide, breast cancer is the leading type of cancer in women, accounting for 25% of all cases[2]. There's a good chance of recovery if it's detected in its early stages.

There is a growing trend to apply Machine Learning techniques on cancer datasets for analysis and provide insights. In this paper we follow a sequence of steps to prepare the data and apply machine learning algorithm and compare their performance to predict whether the cancer is benign and malignant.

II. Dataset

There are three available methods of diagnosing Breast cancer: Mammography, FNA with visual interpretation and surgical biopsy. The Wisconsin Breast Cancer Data[3] grew out of the desire by Dr. Wolberg to accurately diagnose breast masses based solely on a Fine Needle Aspiration (FNA). He identified nine visually assessed characteristics of an FNA sample which he considered relevant to diagnosis[4].

Ten real-valued features are computed for each cell nucleus [4]:

a) radius (mean of distances from center to points on the perimeter)

b) texture (standard deviation of gray-scale values)

c) perimeter

d) area

e) smoothness (local variation in radius lengths)

f) compactness (perimeter^2 / area - 1.0)

g) concavity (severity of concave portions of the contour)

h) concave points (number of concave portions of the contour)

i) symmetry

j) fractal dimension ("coastline approximation" - 1)

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.

All feature values are recoded with four significant digits.

III. Methodology

Analysis was drawn through a sequence of steps as described below

1. Exploratory Data Analysis

a. Number of instances: 569
b. Missing attribute values: none
c. Check for NA/-Inf/Inf values - None
d. Number of attributes: 32 (ID, diagnosis, 30 real-valued input features)
e. Class distribution: 357 benign, 212 malignant

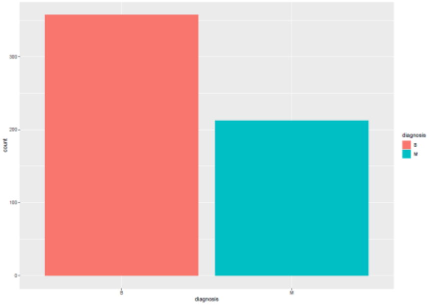

Figure 1 – Split of Diagnosis

f. Correlation Matrix - Correlation among variables of the cancer dataset was captured using corrplot library in R [5]. The correlation coefficients were computed using spearman method and were ordered using AOE. The highlighted boxes in Figure 2 suggest a high correlation exists between some of the attributes like perimeter, radius and area.

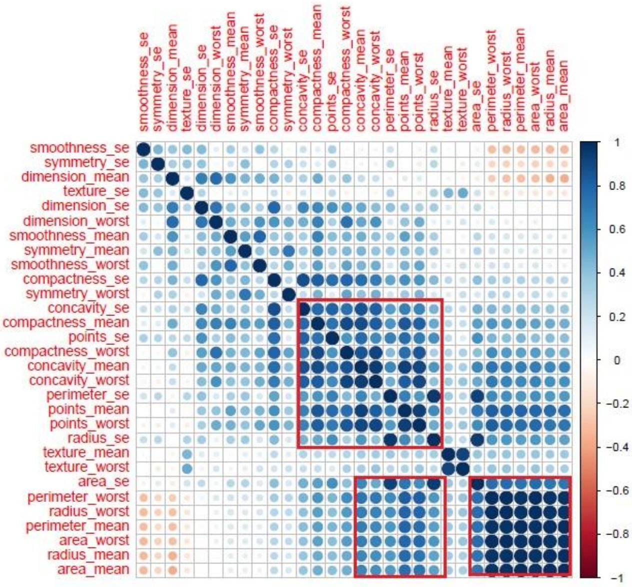

Figure 2 – Correlation of variables

2. Test and Train datasets –

The following sequence of steps were followed to create test and train datasets

i. Split the dataset into Benign and Malignant Dataset
ii. Split each dataset (Benign and Malignant Dataset) into development subset 80% (including training and validation subset) and testing subset 20%
iii. Benign and Malignant development Dataset were combined to create a Dev dataset
iv. Benign and Malignant Testing Dataset were combined to create a Test dataset
v. Development subset had 454 observations and Test subset had 115 observations with 72 Benign and 43 Malignant

3. Machine learning and Data Modelling

The dataset has 2 classes Benign and Malignant and hence is a binary classification problem. The modelling was done using machine learning techniques. Each model was trained using the development subset followed by the trained model fit to the test subset. As part of this research project we have tested 6 machine learning techniques and compared their performance in terms of accuracy.

a) k-Nearest Neighbor –

KNN is a non-parametric, lazy classification model (it takes little or no training time.) that predicts a class based on the features of known observations that are close to it. The K-Nearest Neighbour Algorithm is the simplest of all machine learning algorithms. It is based on the principle that the samples that are similar, generally lies in close vicinity[6]. K nearest neighbour was trained with k set to 20 nearest neighbours with 10-fold cross validation. The optimum K value was found to be 8 using the following formula

Max Balanced accuracy (K neighbours (Mean of accuracy for each of the 10-fold validation))

KNN performed very well on the train dataset with 98.83 % accuracy. There was only one false negative.

	Benign	Malignant	Total
Benign	72	0	72
Malignant	1	42	43
Total	72	42	

Table 1 – KNN Confusion Matrix

Advantages:[7]

a) Easy to understand and implement.
b) Can be used both for Classification and Regression

Disadvantages:

a) Becomes significantly slower as the volume of data increases

b) Support Vector Machines – Linear Kernel

SVM is a discriminative classifier formally defined by a separating hyperplane. The algorithm outputs an optimal hyperplane which categorizes new examples. In two dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side. [8].

SVM linear kernel finds the largest possible linear margin that separates these two regions.

SVM linear Kernel performed very well on the train dataset with 98.83 % accuracy. There was only one false negative.

	Benign	Malignant	Total
Benign	71	0	71
Malignant	1	43	44
Total	72	43	

Table 2 – SVM Linear Kernel Confusion Matrix

Advantages:[9]

a) Performs well, If greater number of features and lesser than training sample
b) Performs well, If a smaller number of features and large training sample
c) Has built in multi class functionality

Disadvantages:

a) Memory Intensive

c) Support Vector Machines - Gaussian Kernel

SVM algorithm was applied using the Gaussian Kernel otherwise called Radial Basis kernel. It is generally used for non-linear data.

SVM Gaussian Kernel had full accuracy on the train dataset.

	Benign	Malignant	Total
Benign	72	0	72
Malignant	0	43	43
Total	72	43	

Table3 – SVM Gaussian Kernel Confusion Matrix

Advantages: [9]

a. Performs well If a smaller number of features and intermediate training sample

b. Has built in multi class functionality

d) Naïve Bayes Classification

Naive Bayes is a statistical classification technique based on Bayes Theorem. It is called Naive as it assumes that all variables contribute towards classification and are mutually correlated. Naive Bayes classifier assumes that the effect of a feature in a class is independent of other features[10][11].

Naïve Bayes had 95.13 % accuracy. There were 2 False positive and 3 true negative conditions.

	Benign	Malignant	Total
Benign	69	2	71
Malignant	3	41	44
Total	72	43	

Table 4 Naïve Bayes Confusion Matrix

Advantages:[7]

a) Simple supervised learning algorithms.
b) Fast
c) High accuracy and speed on large datasets.

Disadvantages: [7]

a) Less accurate on small datasets

e) Neural Networks

Deep learning is a specific subfield of machine learning: a new take on learning representations from data that puts an emphasis on learning successive layers of increasingly meaningful representations[12]. This was implemented using Keras and the parameter loss was set as "binary_crossentropy" as it is a binary classification problem. The number of iterations i.e. epochs was set to 20 with a batch_size of 20.

The model performed very well and predicted 100%.

	Benign	Malignant	Total
Benign	72	0	72
Malignant	0	43	43
Total	72	43	

Table 5 Neural Networks Confusion Matrix

Advantages:

a) High Accuracy
b) Suitable for complex non-linear problems

Disadvantages:

a) Slow to Train

f) Decision Trees

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions. A decision tree is drawn upside down with its root at the top**.** In the image, the top text represents a condition/internal node, based on which the tree splits into branches/ edges. The end of the branch that doesn't split anymore is the decision/leaf [13].

Decision Trees can be used in both regression and classification problem. Library rpart.plot was used to implement the decision tree.

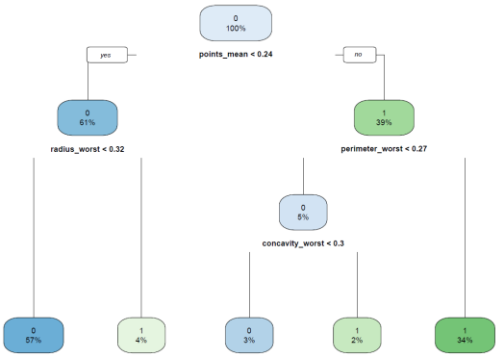

Figure 3 – Decision Tree

Decision Trees had 96.73 % accuracy. There were 3 true negative conditions.

	Benign	Malignant	Total
Benign	69	0	69
Malignant	3	43	46
Total	72	43	

Table 6 – Decision Trees Confusion Matrix

Advantages:[7]

a) Easy to interpret and explain.
b) Simple and Fast

Disadvantages: [7]

a) Can overfit
b) Long Training Time

g) Logistic Regression

Logistic regression is a method for fitting a regression curve, y = f(x), when y is a categorical variable. The typical use of this model is predicting y given a set of predictors x. The predictors can be continuous, categorical or a mix of both [14]. This was implemented using the GLM i.e. Generalized linear model function. In this project we call it a "binomial logistic regression" model as the variable to predict is binary, however, logistic regression can also be used to predict a dependent variable which can assume more than 2 values.

The GLM function is used to fit generalized linear models, specified by giving a symbolic description of the linear predictor and a description of the error distribution [15][16].

It had an accuracy of 0.9514

Advantages: [9]

a) Many ways to regularize model
b) Performs well, If a greater number of features than training samples

c) Suitable for Non-Linear decision Boundaries

IV. RESULTS

Table 7 and Table 8 capture the prediction results of each of the machine learning technique in both absolute and percentage.

Classification Technique	Agreement	
	FALSE	TRUE
kNN Nearest Neighbour	1	114
Support Vector Machines – Linear Kernel	1	114
Support Vector Machines - Gaussian Kernel	0	115
Naive Bayes Classification	5	110
Neural Networks	0	110
Decision Trees	5	110

Table 7 Agreement Table

Classification Technique	Agreement Percentage	
	FALSE	TRUE
kNN Nearest Neighbour	0.0869	0.9913
Support Vector Machines – Linear Kernel	0.0869	0.9913
Support Vector Machines - Gaussian Kernel		1
Naive Bayes Classification	0.0434	0.9565
Neural Networks	0	1
Decision Trees	0.0434	0.9565

Table 8 Agreement Table in percentage

CONCLUSIONS

This project focuses on Comparative Study of classification techniques.

a) All the models performed very well both in terms of accuracy and execution speed due to a small test dataset.

b) Support Vector Machines Gaussian Kernel and Neural Networks had the highest accuracy

c) Logistical Regression had less accuracy in comparison to others

d) The optimum K for k Nearest Neighbours was 8

Classification Technique	Accuracy
kNN Nearest Neighbour	0.9883
Support Vector Machines – Linear Kernel	0.9886
Support Vector Machines - Gaussian Kernel	1
Naive Bayes Classification	0.9518
Neural Networks	1
Decision Trees	0.9673
Logistic Regression	0.9514

Table 9 Accuracy of Machine learning techniques

FURTHER DEVELEOPMENTS

a) Add more statistical features within the dataset and perform analysis.
b) Test this model on bigger datasets and see how they behave.
c) Apply Random forest ensemble technique
d) Apply Gradient Boosting Technique

REFERENCES

[1] "Breast cancer statistics | Cancer Research UK." [Online]. Available: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer. [Accessed: 28-Apr-2019].

[2] "WHO | World Cancer Report 2014," *WHO*, 2015.

[3] "Index of /ml/machine-learning-databases/breast-cancer-wisconsin." [Online]. Available: https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/. [Accessed: 28-Apr-2019].

[4] O. L. Mangasarian, W. N. Street, and W. H. Wolberg, "Breast Cancer Diagnosis and Prognosis Via Linear Programming," *Oper. Res.*, 2008.

[5] "An Introduction to corrplot Package." [Online]. Available: https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html. [Accessed: 28-Apr-2019].

[6] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification. IEEE Trans Inf Theory IT-13(1):21-27," *IEEE Trans. Inf. Theory*, 1967.

[7] S. D. Jadhav and H. Channe, "Comparative Study of K-NN , Naive Bayes and Decision Tree Classification Techniques." 2016.

[8] "Chapter 2 : SVM (Support Vector Machine) — Theory – Machine Learning 101 – Medium." [Online]. Available: https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72. [Accessed: 28-Apr-2019].

[9] "Lecture7 SVM =nc." .

[10] "Naive Bayes Classification using Scikit-learn (article) - DataCamp." [Online]. Available: https://www.datacamp.com/community/tutorials/naive-bayes-scikit-learn. [Accessed: 15-Apr-2019].

[11] N. Friedman, D. Geiger, and M. Goldszmit, "Bayesian Network ClassifiersOverfitting and Underfitting With Machine Learning Algorithms (no date). Available at: https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/ (Accessed: 1 July 2018).," *Mach. Learn.*, 1997.

[12] "Chapter 1. What is deep learning? - Deep Learning with R." [Online]. Available: https://livebook.manning.com/#!/book/deep-learning-with-r/chapter-1/38. [Accessed: 28-Apr-2019].

[13] “Decision Trees in Machine Learning – Towards Data Science.” [Online]. Available: https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052. [Accessed: 28-Apr-2019].

[14] “How to Perform a Logistic Regression in R | DataScience+.” [Online]. Available: https://datascienceplus.com/perform-logistic-regression-in-r/. [Accessed: 28-Apr-2019].

[15] “Logit Regression | R Data Analysis Examples.” [Online]. Available: https://stats.idre.ucla.edu/r/dae/logit-regression/. [Accessed: 28-Apr-2019].

[16] “R: Fitting Generalized Linear Models.” [Online]. Available: https://stat.ethz.ch/R-manual/R-patched/library/stats/html/glm.html. [Accessed: 28-Apr-2019].