MA902 Essay

An Application of Supervised Machine Learning Algorithms for Breast Cancer Diagnostic

by Jayani Dipalkumar Bhatwadiya
2004351
Department:Department Of Mathematical Science
University of Essex

Abstract

This study presents a comparison of six different supervised machine learning(ML) algorithms : Logistic Regression, Support Vector Machines, K-Nearest Neighbors, Random Forest, Gradient Boosting and Ridge Classifier on the **Kaggle dataset** created by **"AI for Social Good: Women Coders' Bootcamp."** For the implementation of Machine Learning Algorithms, the data was clean, pre-process and standardize. The dataset was trained with the 90% training phase and 10% Testing phase. Results show that all the exiting machine learning algorithms performed well (all exceeded 80% test accuracy). The **K-Nearest Neighbors** perform well with testing accuracy of 93.9% among six machine learning models.[2]

KEYWORDS :
machine learning algorithms ; supervised learning; K-Nearest Neighbors; AI for Social Good: Women Coders' Bootcamp.

1 Introduction

Breast cancer is one of the most commonly diagnosed cancers in women all over the world. According to statistics, one in every seven women will be diagnosed with breast cancer. Every year, 55,200 new cases of breast cancer are recorded, with approximately 11,400 women dying as a result.[1] Breast cancer is known to be one of the most deadly diseases, with a mortality rate of 20%.

Mammography, a breast cancer screening method, is an efficient way to diagnose cancer at an early stage and increase the patient's survival rate by treating the disease at an early stage. However, though there is no reason to question the expert radiologists' examination technique for mammograms,but external factors such as exhaustion, distractions, and human error must be minimised. [13]

A mammogram that is prone to errors can lead to decisions that are harmful to patients. If a mammogram reveals a cancer tumour, treatment is normally recommended. On the one hand, patients may be unaware of breast cancer due to a lack of treatment, and on the other hand, an instance of breast cancer may be identified, but there is no cancer, resulting in unnecessary treatment.[8] Computer-Assisted Detection (CAD) software can help to reduce the amount of incorrect interpretations and improve mammography screening accuracy.[14]

The relevance of the same research was recognised in this study, which elaborated on the use of machine learning algorithms for classification to

detect breast cancer using a dataset provided by **"AI for Social Good: Women Coders' Bootcamp"** and eventually got a good result that was better than the radiologist's diagnosis [2]

2 Literature Review

Women should get a mammogram every year, according to the **"American Cancer Society (ACS)"** As a result, there is a fair chance of detecting breast cancer. Inter-observer and intra-observer errors have been discovered in medical image processing, as well as Computer-Aided Diagnosis (CAD), which is focused on mathematical aspects to detect cancer cells.[6]

Another research used artificial algorithms to create and test mammograms in five institutions in South Korea, the United States, and the United Kingdom, and the results were very good when compared to expert radiologists' success.[9].

According to a study conducted in the United Kingdom, the effected cancer cells detected by operating characteristic curve (AUC-ROC)for the AI system had higher result than the AUC-ROC of the average radiologist with margin of 11.5%. This AI system improves the performance and reduced the workload, and it is approved for the clinical trails to improve the accuracy of breast cancer detection.[11]

The technique is to detect cancer cells through deep learning or neural networks gives more accurate result. Study shown that in comparison of mammograms promising result have been achieved through convolutional neural network. Many medical institutions encouraged to use deep learning technique to detect the cancer tumour to increase the chances of performance.[3]

Cancer detection through microwave imaging has been carried out through many researchers. It maps through distribution of electrical property in human body. Cancer tissues related to the property of microwave imaging. Cancerous cells has been detected when microwave property deals with the tissue of body. [4]

A technique called **noninvasive impedance imaging** for breast cancer detection is used to detect the cancerous cells. Analysis has been done with **"vitro impendance measuremnet"** technique with both cancerous tissues and normal tissues, cancerous tissues were first detected. [19]

3 Problem Statement

The use of machine learning methods in the CAD System results high accurate result in mammogram screening for detecting early signs of breast cancer. However, these methods require a large amount of data in order to deal with cancer tumours, and need powerful tools to increase faster performance of the data training process. [15]

4 Methodology

4.1 Machine Intelligence Library

In this study, machine learning libraries is used which is very important to predict the result such as, **numpy** for working with arrays , Pandas to create dataframe to fit the data, **matplotlib** for creating graphical representation, **scikit-learn** to use various machine learning algorithms including classification, clustering and regression.[2]

4.2 The Dataset

To detect breast cancer, the machine learning algorithms were trained on the **Kaggle Dataset** created by **"AI for Social Good: Women Coders' Bootcamp"**. It is very small and contains **.csv** file having five columns named **mean radius, mean texture, mean perimeter, mean area, mean smoothness, and diagnosis**. From the dataset **diagnosis** column considered for Testing data and rest of columns considered for Training purpose.[2]

4.3 Machine Learning (ML) Algorithms

This section represents all the machine learning algorithms which is carried out in this study, Machine Learning Algorithms are mainly divided in two categories :

- Supervised Algorithms
- UnSupervised Algorithms

On the one hand, in Supervised Learning the classification happens with respect to same label. The model has trained with respect to input-output pairs. The techniques fall under supervised learning are : **Classification** and **Regression**

On the other hand, Unsupervised Learning uses Unlabelled data, it allows the model to select the pattern. The type of Unsupervised learning mainly includes **Clustering**

In this study, the experiments have done with all Supervised Learning Algorithms as below :

1. Logistic Regression
2. Support Vector Machines (SVM)
3. K-Nearest Neighbors
4. Random Forest
5. Gradient Boosting
6. Ridge Classifier

4.3.1 Logistic Regression

Logistic Regression algorithm is mainly used for classification problem. It is like the Linear Regression but the difference between this two regression is Logistic Regression uses a cost function which is very complex, its called **Sigmoid Function**. [12]

The limit of this cost function is between 0 to 1, therefore linear regression model fails to achieve this goal, because Linear Regression model have values greater than 1 or less than 0, and its not possible in the case of Logistic Regression. [Equation (1)] represents Logistic Regression hypothesis expectation. [12]

$$0 \leq h_\theta(x) \leq 1 \tag{1}$$

Sigmoid Function contains value between 0 to 1. Sigmoid plays critical role in order to achieve prediction. [Equation (2)] is the formula to calculate sigmoid function.

$$f(x) = \frac{1}{1 + e^{-(x)}} \tag{2}$$

4.3.2 Support Vector Machines (SVM)

Support Vector Machine is used for both classification and regression technique. However, It is widely used to solve classification problems. In SVM the each data is plotted in n-dimensional space. The classification is done by finding hyper-plane that distinguish two different classes.

The Kernel Trick is used to mapped training data with higher dimension where the data is linearly distributed, which was carried out by dot product which maps the input features.[7] To separate two different classes data hyper plane is used. There are many possible hyper-plane can be chosen. But the hyper-plane with maximum margin distance is the best to distribute both data points from two classes. Maximum distance gives more assurance to classify the data points for future predictions.[5]

In Support Vector Machine the data points which are more closer to the hyper-plane that can change position and orientation of the hyper-plane. To change the position of the hyper-plane support vectors used.[5] The (Figure 1) shows the data points of two classes were plotted through hyper-plane.

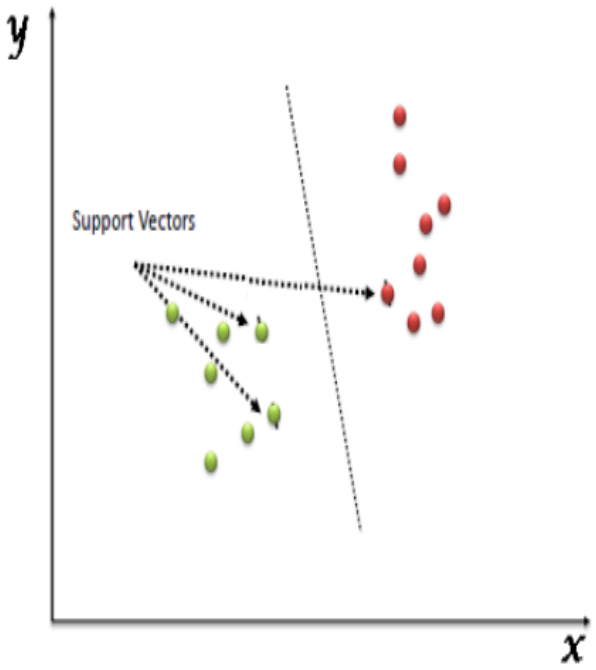

Figure 1: Hyper-plane with two classes

4.3.3 K-Nearest Neighbors

Its widely used algorithm in classification problem. In this method prediction is done based on few similarities between old and new data points. It uses **feature similarities** to assume the value of new data points, which further assign the value based on the how closely the new data point related to training data. [17]

K-Nearest Neighbors is also called **non-parametric algorithm** because it does not make prediction on underlying data, it assumes the new data with respect to the training data points.It stores the data at the training time and whenever new data will come in the model at that time it will going to find some similarities between new coming data and the data which was previously stored, and this same steps will perform up to how many value of K has been given.(Figure 2) shows the prediction with K=3.

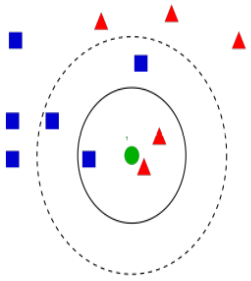

Figure 2: Example of a kNN classifier

The Algorithm is based on distance, in which the data points are belongs to the nearest one with the specified k value. To find the distance between two data points Euclidean Distance is used. [Equation.(3)] shows the distance between two data points.

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{3}$$

4.3.4 Random Forest

Random Forest is one of the type of Classifier Algorithm. The term **Forest** stands for jungle, which is made from multiple trees. Likewise, Random Forest creates samples of decision tree and get assumption of these decision tree samples and finally decide best solution. Model accuracy is depend on total number of decision tree samples, if there is more samples

were trained then highest accuracy will occur, and it prevents to make model over-fitting.

The Algorithm works in two steps, in first step N-number of decision tree were created that needs to combine to make one random forest, In second step the prediction will happen on basis of each number of decision tree, and when the data points occur that time the model has make final decision based on majority. (Figure 3) shows the Random Forest is divided in two decision tree samples.

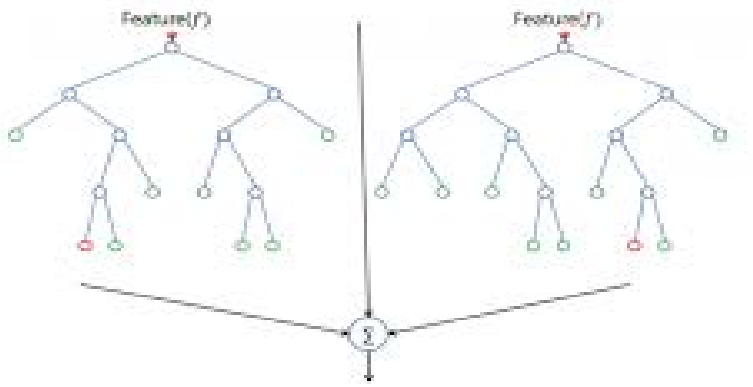

Figure 3: Random Forest in two Samples

4.3.5 Gradient Boosting

The Gradient Boosting Algorithm introduced from **AdaBoost Algorithm**.In AdaBoost Algorithm in each observations, the decision tree will be assigned an equal weight. After evaluating the first decision tree the lower weight assigned to those observations which are easy to classify and higher weight assigned to those observation which are difficult to classify.[16]

Many models trains through Gradient Boosting Algorithm with gradual, sequential additive and manner. The main difference between two algorithm is how these algorithm can identify the decision trees. Decision trees identifies by AdaBoost Model with high weight data while gradient boosting does the same thing using loss function. shows the formula for the loss function. Through this Loss function model's coefficients measured, in which situation our model is over-fitting or under-fitting.[16]

4.3.6 Ridge Classifier

The Ridge Classifier model comes from family of Ridge Regression. It is simple variation of Linear Regression. It solve the problem using Regression Method by converting the labeled data into [-1,1]. Target Class accepted highest value and multi-output regression is applied for multi class data. The Ridge Algorithm works as below :

1. Binary Classification
    - The target value will be classified into +1 or -1 which is based on class from the data belongs.
    - For predicting the target variable, builds **Ridge()** model and its loss function is denoted by MSE +12.
    - The **decision function()** is used to predict the value, the function decides the class, if the value is above 0 then it will be consider as Positive class otherwise it will be consider in Negative class.
2. Multi-Class Classification
    - To create multi output regression model **LabelBinarizer()** is used.it will create individual ridge for each class and then each of the class will be trained through the model.
    - After training the Ridge model the prediction was measured for each class by using argmax.

5 Result

Experiments done in this study were conducted on a laptop with Intel Core(TM) i3-6300HQ CPU @2.30GHz x 4, 8GB of DDR3 RAM as hardware and for the coding purpose and graphical representation of the result Jupiter Notebook is used.

Various machine learning algorithms have trained on the Dataset and it showing great result with considering both Training and Testing accuracy. Table shows the result of machine learning models which is performed on **"Women Coders Dataset"**

Machine Learning Models	Training Accuracy(%)	Testing Accuracy(%)
Logistic Regression	87	88.6
Support Vector Machines	88.4	90.4
K-Nearest Neighbors	1.00	93.9
Random Forest	89.9	93.00
Gradient Boosting	92.3	92.1
Ridge Classifier	88.8	89.5

All machine learning algorithm training accuracy measured :(1) Logistic Regression finished its training with average of 87% accuracy, (2) Support Vector Machines (SVM) finished with the average training accuracy of 88.4%, (3) K-Nearest Neighbors gives best result to achieve training accuracy which is 1.00%, (4)Random Forest trained and finished with the training accuracy of 89.9%, (5) Gradient Boosting finished with the average accuracy of 92.3%, and finally (6) Ridge Classifier has trained and gave accuracy of 88.8% respectively.

Testing accuracy of all Machine Learning algorithms also measured : (1) Logistic Regression finished its testing with average of 88.6% accuracy, (2) Support Vector Machines (SVM) finished with the average testing accuracy of 90.4%, (3) K-Nearest Neighbors gives good result having testing accuracy which is 93.9%, (4)Random Forest trained and finished with the testing accuracy of 93.0%, (5) Gradient Boosting finished with the average testing accuracy of 92.1%, and finally (6)Ridge Classifier has trained and gave testing accuracy of 89.5% respectively.

The experiment shows, **K-Nearest Neighbors** performed well on described dataset with Training accuracy 1.00% and Testing accuracy 93.9% accordingly.

Using **matlpotlib** library the all machine algorithms training and testing accuracy is graphically plotted, the (Figure 4) (Figure 5) shows the graphical representation of both accuracy.

6 Conclusion

This study presents an application of different supervised machine learning algorithms including (1) Logistic Regression, (2) Support Vector Machines, (3) K-Nearest Neighbors, (4) Random Forest, (5) Gradient Boosting and (6) Ridge Classifier for detection of breast cancer. All mentioned algorithms well trained and given high performance.[2]

For future study, deep learning models should be applied, deep learning or neural network includes hidden layer to train model, input data

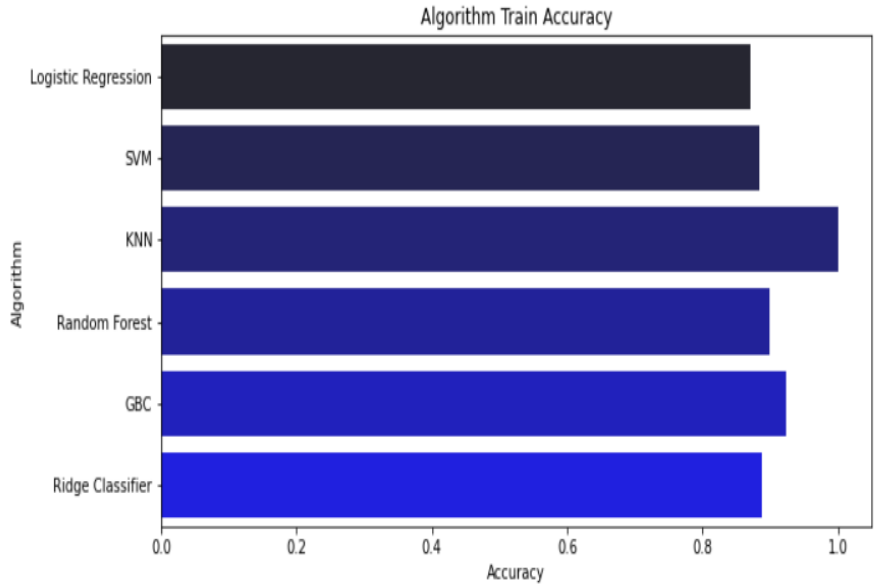

Figure 4: Training Accuracy

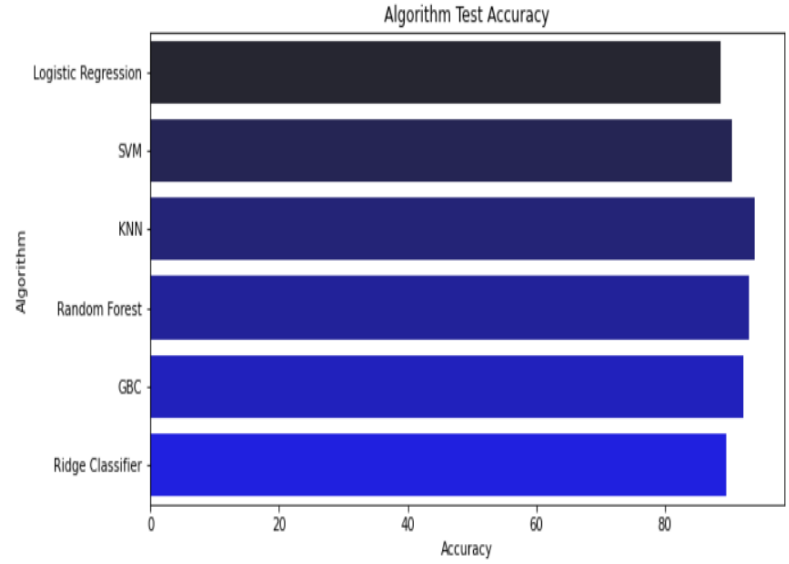

Figure 5: Testing Accuracy

passes through multiple layers, model then extract best features and it gives best result with good accuracy and efficiency. Biggest advantage of deep learning model over machine learning model is it tries to extract high level features from the data. Deep Learning models use to train large scaled unstructured data such as image, Video, Sound and text.[10]

Another study should be applied with the use of **K-Fold Cross Validation** technique. Its an computer vision method which uses k-number of folds. By increasing k value model trains multiple times which will give best prediction result. These type of techniques are very accurate and efficient, However, it also need hyper parameter tuning for the Machine Learning Algorithms.[18]

References

[1] Cancer Research UK (2020). Breast cancer statistics, 2020.

[2] Abien Fred Agarap. On breast cancer detection: An application of machine learning algorithms on the wisconsin diagnostic dataset. *CoRR*, abs/1711.07831, 2017.

[3] Saira Charan, Muhammad Jaleed Khan, and Khurram Khurshid. Breast cancer detection in mammograms using convolutional neural network. In *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, pages 1–5, 2018.

[4] E.C. Fear, P.M. Meaney, and M.A. Stuchly. Microwaves for breast cancer detection? *IEEE Potentials*, 22(1):12–18, 2003.

[5] Rohith Gandhi. Support vector machine — introduction to machine learning algorithms, 2018.

[6] Karthikeyan Ganesan, U. Rajendra Acharya, Chua Kuang Chua, Lim Choo Min, K. Thomas Abraham, and Kwan-Hoong Ng. Computer-aided breast cancer detection using mammograms: A review. *IEEE Reviews in Biomedical Engineering*, 6:77–98, 2013.

[7] Aurelien Geron. Hands-on machine learning with scikit-learn, keras tensorflow, 2019.

[8] Matthias Elter Alexander Horsch. Cadx of mammographic masses and clustered microcalcifications: a review. *Medical Physics*, 36:2052–2068, 2009.

[9] Hyo-Eun Kim, Hak Hee Kim, Boo-Kyung Han, Ki Hwan Kim, Kyunghwa Han, Hyeonseob Nam, Eun Hye Lee, and Eun-Kyung Kim. Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. *The Lancet Digital Health*, 2(3):e138–e148, 2020.

[10] Sambit Mahapatra. Why deep learning over traditional machine learning? *Nature*, 2018.

[11] Sieniek M. Godbole V. McKinney, S.M. International evaluation of an ai system for breast cancer screening. *Nature*, 577:89–94, 2020.

[12] Ayush Pant. Introduction to logistic regression, 2019.

[13] K. Polat and S. Günes. Breast cancer diagnosis using least square support vector machine. *Digit. Signal Process.*, 17:694–701, 2007.

[14] Margolies L. R. Rothstein J. H. Fluder E. McBride R. B. Shen, L. and W. Sieh. Deep learning to improve breast cancer early detection on screening mammography. *Scientific Reports*, 2017.

[15] Margolies L.R. Rothstein J.H. et al. Shen, L. Deep learning to improve breast cancer detection on screening mammography. *Medical Physics*, 9, 2019.

[16] Harshdeep Singh. Understanding gradient boosting machines, 2018.

[17] TAVISH SRIVASTAVA. Introduction to k-nearest neighbors: A powerful machine learning algorithm (with implementation in python r), 2018.

[18] Ian J Goodfellow Yoshua Bengio and Aaron Courville. Deep learning. *Nature*, 521:436–444, 2015.

[19] Y Zou and Z Guo. A review of electrical impedance techniques for breast cancer detection. *Medical Engineering Physics*, 25(2):79–90, 2003.

Appendices

To load the dataset **pandas** library used and read, only first five data has been showed. **mean radius, mean texture, mean perimeter, mean area, mean smoothness** columns were used to train the model and to test the model only, and **diagnosis** column required.Pre-processing task was performed using described libraries

```
#Feature Selection
bestfeatures = SelectKBest(score_func=f_classif, k=5)
fit = bestfeatures.fit(X_feat,y_feat)
dfscores = pd.DataFrame(fit.scores_)
dfcolumns = pd.DataFrame(X_feat.columns)
#concat two dataframes for better visualization
featureScores = pd.concat([dfcolumns,dfscores],axis=1)
featureScores.columns = ['Assembly','Score']  #naming the dataframe columns
print(featureScores.nlargest(12,'Score'))  #print 10 best features
```

```
          Assembly       Score
2   mean_perimeter  697.235272
0      mean_radius  646.981021
3        mean_area  573.060747
1     mean_texture  118.096059
4  mean_smoothness   83.651123
```

Train the data with all six algorithms and created confusion matrix. I have attached Logistic model code, As per the same way rest of other algorithms code was done and created confusion matrix for it.

```
#Logisitc Regression

Log_Reg=LogisticRegression(C=1, class_weight='balanced', dual=False,
                   fit_intercept=True, intercept_scaling=1, l1_ratio=None,
                   max_iter=1000, multi_class='auto', n_jobs=None, penalty='l2',
                   random_state=0, solver='lbfgs', tol=0.0001, verbose=0,
                   warm_start=False)
Log_Reg.fit(X_train_scaled, y_train)
y_reg=Log_Reg.predict(X_test_scaled)
print("Train Accuracy {0:.3f}".format(Log_Reg.score(X_train_scaled, y_train)))
print('Test Accuracy' "{0:.3f}".format(metrics.accuracy_score(y_test, y_reg)))
cm = metrics.confusion_matrix(y_test, y_reg)
np.set_printoptions(precision=2)
plt.figure()
plot_confusion_matrix(cm, classes=['Benign', 'Malignant'],
                          title='Logistic Regression')
accuracy_list.append(metrics.accuracy_score(y_test, y_reg)*100)
train_accuracy.append(Log_Reg.score(X_train_scaled, y_train))
algorithm.append('Logistic Regression')
```

```
Train Accuracy 0.870
Test Accuracy0.886
[[43  4]
 [ 9 58]]
```