

Illinois Institute of Technology

MATH 564FinalProject

Breast Cancer Detection Using Classification Models

Sumanth Donthula (A20519856) and Ram Vaka (A20481446)

Prof. Lulu Kang

NOVEMBER 25th ,2022

# Introduction

Breast Cancer (BC) is one of the most prevalent diseases in women and accounts for the majority of new cancer cases and cancer-related deaths, making it a serious public health issue in today's society. Because it can encourage prompt clinical care for patients, an early diagnosis of BC can considerably enhance the prognosis and likelihood of survival. A more precise categorization of benign tumors might spare people from receiving unneeded medical care. As a result, there is a lot of study on the proper diagnosis of BC and the classification of individuals into benign or malignant categories. Machine learning (ML) is widely acknowledged as the preferred approach in BC pattern classification and forecast modeling due to its distinct advantages in essential features discovery from complicated BC datasets.

The Data set contains the different dependent features to predict the Breast Cancer whether it is Benign or Malignant. The Data set is taken from the Study done by University of California at Irvine from there ML Data Set Repository [1]. The dataset used in this article was developed by Dr. William H. Wolberg, a physician at the University Of Wisconsin Hospital in Madison, Wisconsin, in the United States, and is openly accessible. Dr. Wolberg used fluid samples obtained from patients with solid breast masses and Xcyt, an easy-to-use graphical computer tool that can analyze cytological characteristics based on a digital scan, to construct the dataset. The software computes 10 features from each sample cell using a curve-fitting approach, then computes each feature's mean value, extreme value, and standard error for the picture, returning a 30 real-valued vector.

In this Project we will implement classifying models like K-Nearest Neighbors (KNN) , Support Vector Machine , and Logistic Regression Models and compares the model accuracies and other metrices to get the best model . Data Preparation Analysis has been implemented before creating the models before implementing the models the data is prepared using

# Supervised Machine Learning Overview

Supervised learning involves training a computer algorithm on input data that has been labeled for a specific output. The model is trained until it is able to recognize the underlying relationships and patterns between the input and output labels, allowing it to produce precise labeling results when given with data that the model is not trained on.

Identifying the classification a news article falls under or forecasting the amount of sales for a specific future date are examples of classification and regression problems that supervised learning is good at solving. Making meaning of data within the context of a particular inquiry is the goal of supervised learning.

We have used two supervised machine learning models namely:

**1) Logistic Regression**

A predictive analytic technique based on the idea of probability, logistic regression is a machine learning approach that is used for categorization problems.

The cost function is usually restricted to the range between 0 and 1 according to the logistic regression hypothesis. Because it can have a value larger than 1 or less than 0, which is not feasible according to the logistic regression hypothesis, linear functions cannot accurately describe it.

Hypothesis and expectation for logistic regression

**The Sigmoid Function:**

The Sigmoid function converts any real number to values between 0 and 1. We use the sigmoid function in machine learning to convert predictions to probability values that is in range(0,1).



Fig 1. Sigmoid Function Curve

When using linear regression we use the following hypothesis:

hΘ(x) = β0 + β1X

For logistic regression we will use linear regressuin but the sigmoid function applied over it

σ(Z) = σ(β0 + β1X)

We have expected that our model will give values between 0 and 1.

Z = β0 + β1X

hΘ(x) = sigmoid(Z)

hΘ(x) = 1/(1 + e^-(β0 + β1X)

**2)Support Vector Machine:**

Finding a hyperplane in an N-dimensional space (N is the number of features) that categorizes the data points clearly is the goal of the support vector machine algorithm.

**Hypothetical hyperplanes:**

We should find a plane with huge margin which separates data points from both classes, maximizing the margin distance adds some support, to increase confidence for future data point categorization.

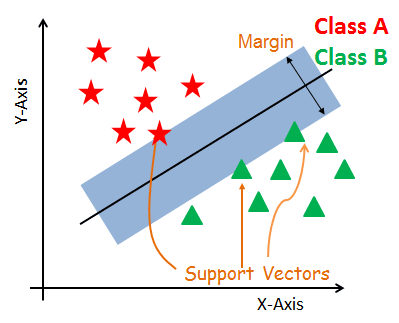


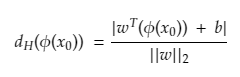
Fig2. SVM Hyperplane Illustration

**SVM Algorithm :**

Finding a hyper plane with great margins is the goal of the Model.

In other words, “The goal is to maximize the minimum distance.”

maximum distance is given by:



If the point from the positive group is substituted in the hyperplane equation while generating predictions on binary training data divided into positive and negative groups, we will obtain a value greater than 0.

wT(Φ(x)) + b > 0

for negative case :

wT(Φ(x)) + b < 0.

# 3 Unsupervised Machine Learning Overview

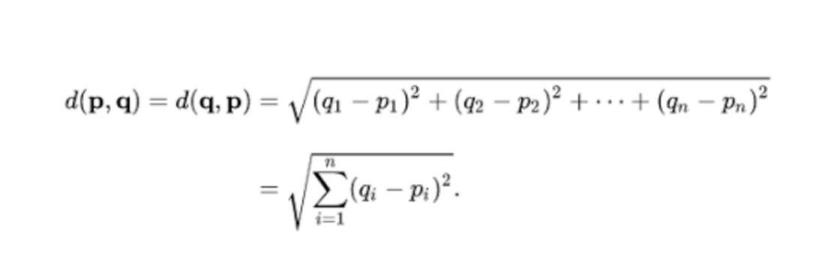
Unsupervised Machine Learning models classify the unlabeled datasets based on the patterns found in the data. These algorithms generally clusters the data to identify the patterns.

We have used KNN Model from Unsupervised learning.

KNN is a non-parametric approach and a lazy learning algorithm. Because it only stores data from the training phase, lazy learning implies that the algorithm learns in almost no time. The saved data will then be used to evaluate a new query point. A non-parametric method is one that does not make any assumptions about distributions. As a result, KNN does not need to find any distribution parameters. During the parametric method, the model discovers new parameters, which are then used for prediction. KNN has only one hyper-parameter (a parameter supplied by the user to the model), which is the number of points that must be considered for comparison purposes.

The query point in the above image is yellow, and we want to know which class it belongs to (red or green). With K=3, the yellow point's three nearest neighbors are considered, and the class is assigned to the query point based on the majority. Similarly, for K=5, the comparison takes into account 5 nearest neighbors, and the majority determines which class the query point belongs to. One thing to keep in mind here is that if the value of K is even, Because the data has an even number of classes, it may cause problems when taking a majority vote.(i.e., 2). As a result, choose K as an odd number when the data has an even number of classes and an even number when the data has an odd number of classes. Working of KNN algorithm

The Euclidean distance between a query point (q) and a training data point (p) is defined as



# Implementation

We have implemented the models using R Studio. Firstly we started with visualizing and understanding the data using different plotting techniques and finding the central tendencies in R. Once the pre processing of data and selected variables are observed we have trained the models. Once SVM, K-NN and Logistic Regression models are build we have evaluated the confusion matrix and accuracy of the models. We have used the R packages like ggplot2, reshape2, ca Tools, e1071, class, Information Value, corrplt for plotting, for reshaping the data frame, for train and test data split, for SVM, for KNN for finding the optimal value in logistic regression and finding the correlation in columns of the dataset. All these packages are installed and called while implementing the specific functionalities in R.

# Results

The below results contains the preprocessing of the data set, checking if the data is distributed uniformly or any outliers are present.

## Data Preparation Analysis:

The histograms and box plots of all the features or dependent variables are plotted. As we have many independent variables and most of them are less correlated with the diagnosis variable which is the target variable we are using for our classification. The independent variables whose correlation with diagnosis are dropped and the remaining columns are used to build the classifiers.

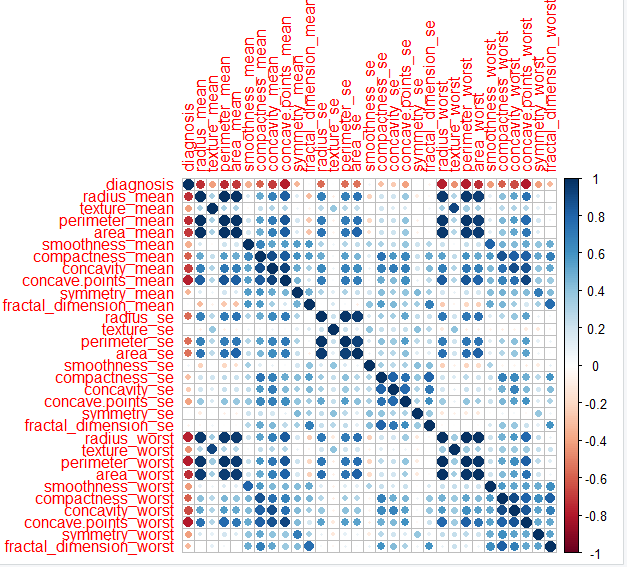


Fig 3. Correlation plot of Data frame

The following features from the dataset is highly correlated with the Diagnosis :

* concave.points\_worst
* perimeter\_worst
* concave.points\_mean
* radius\_worst
* area\_mean
* perimeter\_mean
* area\_worst
* radius\_mean

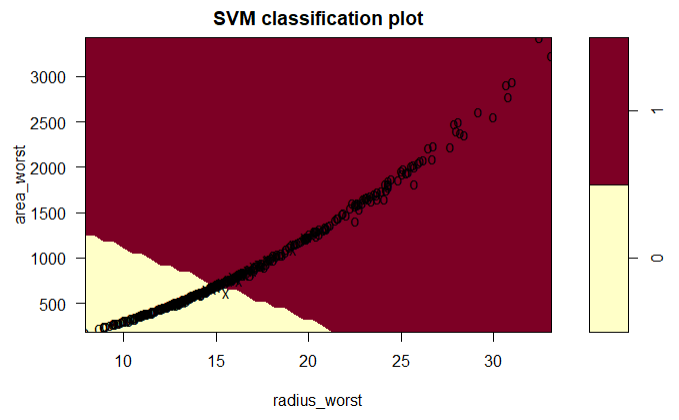
These 8 independent variables are used to build the models. We can observe from the correlation plot that there is no severity of multi co linearity in the above variables.

**Supervised Models Implementation:**

**Support Vector Machine:**

When trained by Support Vector Machine the accuracy of the model is 95.34 %. The below figure shows the decision boundaries created by SVM for classifying whether the diagnosis is benign(0) or malignant(1). The points which falls in the red region has the prediction as malignant and the points which falls in the yellow region are categorized as benign. The boundary was taken with the interaction of area\_worst and radius\_worst features.

Fig4. SVM Decision Boundary



**Logistic Regression Model:**

The Logistic Model gave the highest accuracy score of 98.4 % . The below plot is the visualization of logistic regression, the dependent variable diagnosis with one of the independent variable (radius\_mean) we can see that the sigmoid function has been applied and all the values of the linear model are converted into the ranges of 1 and 0 which represents whether the cancer is Malignant or Benign.

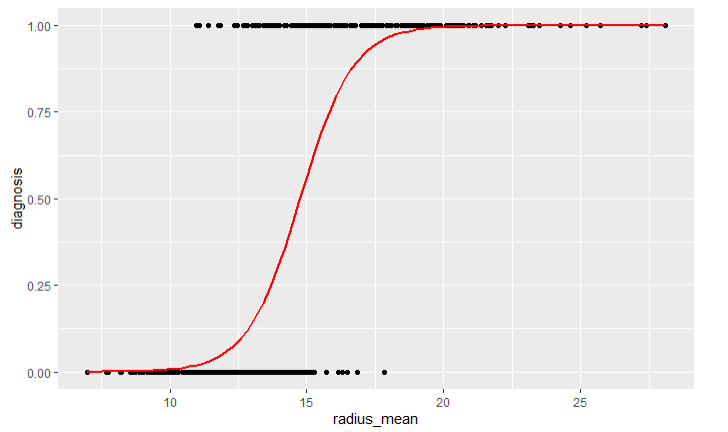


Fig 5. Sigmoid function for radius\_mean with Diagnosis

**Unsupervised Models:**

**Knn Model:**

The KNN Model underperformed when compared to the rest of the Models with an accuracy of 95.34%. The below figure shows the visualization of the Knn classifier between boundaries of two independent variables radius\_mean and area\_mean. The model cluster the data into sets based on the nearest Euclidean points and classifies whether the prediction is Malignant(1) or Benignant(0).

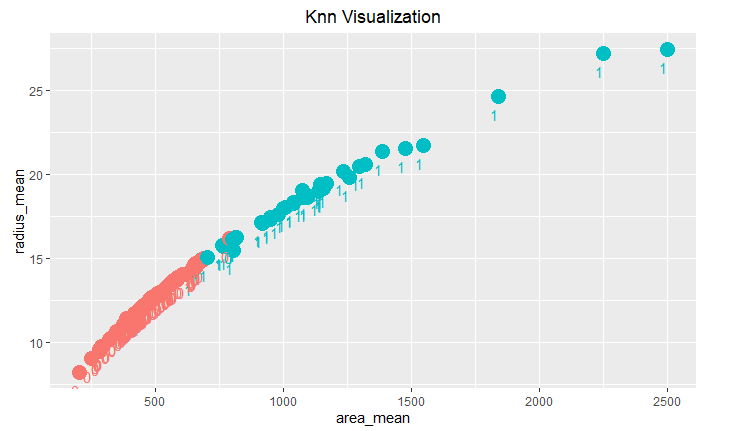
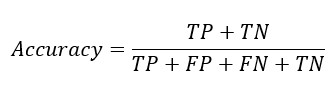


Fig 6. Knn Clusters of radius\_mean and area\_mean boundaries

The Confusion Matrix for all the Models is built and all the True Positive, True Negative, False Positive and False Negative predictions are displayed and the following measure where calculated.

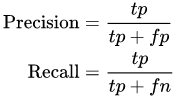
Accuracy, Precision and recall are generally widely used measures to compare classification problems.

• Accuracy: This metric indicates how close the predicted values are to the actual value.



• Precision: It demonstrates the model's accuracy in terms of predicted values.

• Recall : It demonstrates how accurate the model is in terms of known values, or how many actual positives our model has captured by



|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **TP** | **FP** | **TN** | **FN** | **Accuracy(%)** | **Precision(%)** | **Recall(%)** |
| SVM | 44 | 2 | 81 | 2 | 96.89922481 | 95.6522 | 95.6521 |
| KNN | 41 | 5 | 82 | 1 | 95.34883721 | 89.1304 | 97.6190 |
| Logistic Regression | 44 | 2 | 83 | 0 | 98.4496124 | 95.6522 | 1 |

Table.1 . Confusion Matrix and Prediction Measures

We can see that Logistic Regression outperforms other models in terms of Accuracy, precision and recall.

# Conclusion

Classification Models play a key role in the health care related prediction problems as we can see that the predictions can be either 0 or 1, like in our case Benign or Malignant. The predictions can result into serious implications if False Negative cases and False Positive cases are more. For the Breast Cancer Dataset taken from the University of California Irvine, we have identified the important features which can predict the diagnosis whether its Benign or Malignant. We have build some of the Supervised and Unsupervised Models to compare accuracies and performance. So, we can conclude that for the Data Set we used Logistic Regression performs better classification.

# References

1. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29
2. An Introduction to Logistic Regression Analysis and Reporting- CHAO-YING JOANNE PENG
3. WORKSHOP ON SUPPORT VECTOR MACHINES: THEORY AND APPLICATIONS - Theodoros Evgeniou and Massimiliano Pontil
4. KNN Model-Based Approach in Classification Gongde Guo1 , Hui Wang 1 , David Bell 2 , Yaxin Bi 2 , and Kieran Greer