

# Predictive Analysis Of Breast Cancer Using Machine Learning Techniques

*Análisis predictivo del cáncer de mama utilizando técnicas de aprendizaje automático*

*Análise preditiva de câncer de mama com o uso de técnicas de aprendizado de máquina*

Rashmi Agrawal<sup>1</sup>

**Received:** May 5<sup>th</sup>, 2019

**Accepted:** July 24<sup>th</sup>, 2019

**Available:** September 16<sup>th</sup>, 2019

**How to cite this article:**

A. Raschmi, "Predictive Analysis Of Breast Cancer Using Machine Learning Techniques,"  
*Revista Ingeniería Solidaria*, vol. 15, no. 3, 2019.  
doi: <https://doi.org/10.16925/2357-6014.2019.03.01>

---

Artículo de investigación. <https://doi.org/10.16925/2357-6014.2019.03.01>

<sup>1</sup> International Institute of Research and Studies, Faridabad, India.

**ORCID:** <https://orcid.org/0000-0003-2095-5069>

E-mail: [drrashmiagrwal78@gmail.com](mailto:drrashmiagrwal78@gmail.com), [rashmi.fca@mriu.edu.in](mailto:rashmi.fca@mriu.edu.in)

## Abstract

This paper is a product of the research Project "Predictive Analysis Of Breast Cancer Using Machine Learning Techniques" performed in Manav Rachna International Institute of Research and Studies, Faridabad in the year 2018.

*Introduction:* The present article is part of the effort to predict breast cancer which is a serious concern for women's health.

*Problem:* Breast cancer is the most common type of cancer and has always been a threat to women's lives. Early diagnosis requires an effective method to predict cancer to allow physicians to distinguish benign and malicious cancer. Researchers and scientists have been trying hard to find innovative methods to predict cancer.

*Objective:* The objective of this paper will be predictive analysis of breast cancer using various machine learning techniques like Naïve Bayes method, Linear Discriminant Analysis, K-Nearest Neighbors and Support Vector Machine method.

*Methodology:* Predictive data mining has become an instrument for scientists and researchers in the medical field. Predicting breast cancer at an early stage helps in better cure and treatment. KDD (Knowledge Discovery in Databases) is one of the most popular data mining methods used by medical researchers to identify the patterns and the relationship between variables and also helps in predicting the outcome of the disease based upon historical data of datasets.

*Results:* To select the best model for cancer prediction, accuracy of all models will be estimated and the best model will be selected.

*Conclusion:* This work seeks to predict the best technique with highest accuracy for breast cancer.

*Originality:* This research has been performed using R and the dataset taken from UCI machine learning repository.

*Limitations:* The lack of exact information provided by data.

**Keywords:** Naïve Bayes, Linear Discriminant, Support Vector, K-Nearest Neighbors, Breast Cancer, Predictive Analysis

## Resumen

Este artículo es producto del proyecto de investigación "Análisis predictivo del cáncer de mama utilizando técnicas de aprendizaje automático" realizado en el Instituto Internacional de Investigación y Estudios Manav Rachna, Faridabad, en el año 2018.

*Introducción:* el presente artículo es parte de un esfuerzo para predecir el cáncer de seno, lo cual es una preocupación seria para la salud de las mujeres.

*Problema:* el cáncer de mama es el tipo más común de cáncer y siempre ha sido una amenaza para la vida de las mujeres. El diagnóstico precoz requiere un método efectivo para predecir el cáncer que permita a los médicos distinguir el cáncer benigno y el maligno. Investigadores y científicos han estado tratando de encontrar métodos innovadores para predecir el cáncer.

*Objetivo:* el objetivo de esta investigación es el análisis predictivo del cáncer de seno utilizando diversas técnicas de aprendizaje automático, como el método Naïve Bayes, el análisis discriminante lineal, K-Nearest Neighbors y el método de máquina de vectores de apoyo.

*Metodología:* la minería de datos predictivos se ha convertido en un instrumento para científicos e investigadores en el campo de la medicina. La predicción del cáncer de mama en una etapa temprana ayuda a una mejor cura y tratamiento. KDD (Knowledge Discovery in Databases) es uno de los métodos de minería de datos más populares utilizados por los investigadores médicos para identificar los patrones y la relación entre las variables y también ayuda a predecir el resultado de la enfermedad en función de los datos históricos de los conjuntos de datos.

**Resultados:** para seleccionar el mejor modelo para la predicción del cáncer, se estimará la precisión de todos los modelos y se seleccionará el mejor modelo.

**Conclusión:** este trabajo busca predecir la mejor técnica con la mayor precisión para el cáncer de seno.

**Originalidad:** esta investigación se realizó utilizando R y el conjunto de datos tomado del repositorio de aprendizaje automático UCI.

**Limitaciones:** la falta de información exacta proporcionada por los datos.

**Palabras clave:** Naïve Bayes, discriminante lineal, vector de apoyo, K-Nearest Neighbors, cáncer de mama, análisis predictivo.

## Resumo

Este artigo é produto do projeto de pesquisa "Predictive Analysis of Breast Cancer Using Machine Learning Techniques" (Análise preditiva de câncer de mama com uso de técnicas de aprendizado de máquina) realizado no Manav Rachna International Institute of Research and Studies, em Faridabad, em 2018.

**Introdução:** o presente artigo é parte de um esforço para prognosticar o câncer de mama, que é uma séria questão para a saúde da mulher.

**Problema:** o câncer de mama é o tipo mais comum de câncer e sempre foi uma ameaça à vida das mulheres. Um diagnóstico precoce requer um método efetivo para prognosticar o câncer e permitir que os médicos distingam se é benigno ou maligno. Os pesquisadores e cientistas têm tentado encontrar métodos inovadores para prognosticar o câncer.

**Objetivo:** o objetivo deste artigo é a análise preditiva do câncer de mama a partir do uso de várias técnicas de aprendizado de máquina, como o método Naïve Bayes, a Análise Discriminante Linear, o Método dos Vizinhos mais Próximos (k-NN) e o método de Máquina de Vetores de Suporte.

**Metodologia:** a coleta de dados preditiva se tornou um instrumento para cientistas e pesquisadores na área médica. Prognosticar o câncer de mama em estágio inicial ajuda na cura e no tratamento. A extração de conhecimentos de bases de dados (*Knowledge Discovery in Databases* ou KDD) é um dos métodos mais populares de extração de dados usado pelos médicos pesquisadores para identificar os padrões e as relações entre variáveis e ajuda na previsão do efeito da doença com base no histórico de informações de bases de dados.

**Resultados:** para escolher o melhor modelo para o prognóstico de câncer, a precisão de todos os modelos será estimada e o melhor modelo será selecionado.

**Conclusão:** o presente trabalho busca prever a melhor técnica com a precisão mais alta para o câncer de mama.

**Originalidade:** esta pesquisa foi realizada com R e o conjunto de dados recolhido do repositório da máquina de aprendizado da UCI (Universidade da Califórnia em Irvine).

**Limitações:** a falta de informação exata fornecida pelos dados.

**Palavras-chave:** Naïve Bates, linear discriminante, vetor de suporte, vizinhos mais próximos, câncer de mama, análise preditiva.

## 1. INTRODUCTION

Breast cancer is a type of tumor which is malignant in nature. It is the most common type of cancer found in women, affecting almost 10 % of all women at some stage of life. The only way to prevent the spread of this cancer is early diagnosis and timely treatment. Breast cancer starts its development in lobules or duct of the cells. The

main reasons for breast cancer are either because of the abnormality inherited from the parents which is about 10 % of the cases and the remaining 90 % of the cases are due to genetic abnormalities that are caused by the ageing process. Breast cancer can be treated using local and systematic treatments. Surgery and radiation are categorized under local treatments whereas chemotherapy and hormone therapy are categorized under systematic treatments. Local and systematic treatments are often recommended collectively by doctors for best results.

Many data mining techniques are extensively used to diagnose various stages of breast cancer. Prediction of cancer, whether it is malignant or benign, can be diagnosed using various data mining techniques. Classification and clustering are some of the most widely used data mining methods that are used to classify and cluster the data. In the medical field, the aforementioned methods are widely used in diagnosis and in analyzing the data to reduce the number of false positive and false negative decisions which will further help in taking accurate decisions. The paper is organized as follows: In section 1 we provide the introduction to the breast cancer. In section 2 we discuss how we can use data mining techniques for breast cancer prediction. Section 3 describes the data set. In section 4 various classification algorithms are discussed. The remaining sections are used to evaluate and then select and apply the best model for predicting breast cancer.

## 2. BREAST CANCER PREDICTION

### 2.1 BACKGROUND

Statistics demonstrate that the leading cause of death among women across the world is breast cancer. Predicting breast cancer at an early stage helps in improving the cure rate and treatment. KDD (Knowledge Discovery in Databases) is one of the most popular data mining methods used by medical researchers to identify the patterns and the relationship between variables and also helps in predicting the outcome of the disease based upon historical data of datasets.

For the said purpose, various data mining techniques are used to help and support doctors in better decision making. There are two major techniques in data mining which are used in various applications; clustering and classification. Classification, which is also known as supervised learning, requires a set of stored observations known as training data. A classification model is prepared and fit onto that data to predict the future which is applied in another set of stored records known as testing data. Each data set has specified input data and class labels, hence named as

supervised learning. Famous classification techniques are Decision trees, k-Nearest Neighbors, SVM, neural networks etc. On the other hand, clustering is known as unsupervised learning in which data is presented at once to the algorithm and it is the role of the algorithm to divide the data into different clusters. The division of each data point in a cluster is done on the basis of distance measured; which is used in the selection process. One of the most widely used measured distances is Euclidean distance which is easy to use and effective as well. Other measured distances are cosine distance and Manhattan distance. Famous clustering techniques are k-means clustering, k-mediod, DBSCAN and hierarchical clustering. Prediction of breast cancer is possible through both of the data mining techniques - i.e. clustering and classification - but previous available breast cancer data prediction has been conducted using various classification algorithms like naïve Bayes, random forest and support-vector machines to analyze the symptoms and to predict the disease at its earliest stages. In this research paper we apply suitable data mining classification algorithms that can be used to predict the early detection & recurrence of breast cancer in both males and females and will also help in improving the accuracy of the algorithm.

## 2.2 LITERATURE REVIEW

Research on breast cancer is working towards predicting the recurrence and occurrence of breast cancer at various stages [1,2,3,4]. For women older than 40 years of age, Nulliparity is an increased risk factor. Kelsey et al [5] discussed the reproductive factors and breast cancer in detail.

Premenopausal breast cancer also needs significant medical attention. Francis et al [6] studied Premenopausal breast cancer in relation to adjuvant ovarian suppression. This study was performed with the background that suppression of ovarian estrogen production reduces the recurrence of hormone-receptor-positive early breast cancer in premenopausal women, but its value when added to tamoxifen is uncertain.

An interesting study of breast cancer in black and white women was done by DeSantis et al [7] in which breast cancer statistics on convergence of incidence rates between black and white women were analyzed. They concluded that widening racial disparities in breast cancer mortality are likely to continue, at least in the short term, in view of the increasing trends in breast cancer incidence rates in black women. Another similar study was performed by DeSantis et al in 2017 [8] in which breast cancer statistics on racial disparity in mortality by state was performed. Whelan et al [9] studied regional nodal irradiation in early-stage breast cancer and concluded that patients in the nodal-irradiation group had higher rates of grade 2 or greater acute pneumonitis.

Oeffinger et al [10] analyzed the 2015 guideline update from the American cancer society [17,18] for breast cancer screening for women at average risk and concluded that these recommendations should be considered by physicians and women in discussions about breast cancer screening.

### 3. DESCRIPTION OF DATASET

For the experiments from the UCI machine learning repository, we used the Breast Cancer Wisconsin Data Set with 569 instances and 31 attributes. This is a multivariate dataset with real type attributes. Null value for any attribute is not present in the dataset. Features have been computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. Then these ten real-valued features are figured for each cell nucleus which defines distinctiveness of the cell nuclei present in the image. These features are:

Smoothness (S), Compactness (C), Radius (R), Perimeter (P), Area (A), Concavity (Co), Texture (T), Concave points (CP), Fractal Dimension (FD) and Symmetry (Sy).

For simplicity we have used the abbreviations of these features as mentioned above. This dataset is classified into (M or B) Malignant or Benign tumor.

#### 3.1 A VIEW OF THE DATASET

	<b>Diagnosis</b>	<b>T_mean</b>	<b>R_mean</b>	<b>P_mean</b>	<b>A_mean</b>
1	M	10.38	17.99	122.80	1001.0
2	M	17.77	20.57	132.90	1326.0
3	M	21.25	19.69	130.00	1203.0
4	M	20.38	11.42	77.58	386.1
5	M	14.34	20.29	135.10	1297.0
6	M	15.70	2.45	82.57	477.1

	<b>S_mean</b>	<b>C_mean</b>	<b>Co_mean</b>	<b>CP_mean</b>
1	0.11840	0.27760	0.3001	0.14710
2	0.08474	0.07864	0.0869	0.07017
3	0.10960	0.15990	0.1974	0.12790
4	0.14250	0.28390	0.2414	0.10520
5	0.10030	0.13280	0.1980	0.10430
6	0.12780	0.17000	0.1578	0.08089

	<b>Sy_mean</b>	<b>FD_mean</b>	<b>R_se</b>	<b>T_se</b>
1	0.2419	0.07871	1.0950	0.9053
2	0.1812	0.05667	0.5435	0.7339
3	0.2069	0.05999	0.7456	0.7869
4	0.2597	0.09744	0.4956	1.1560
5	0.1809	0.05883	0.7572	0.7813
6	0.2087	0.07613	0.3345	0.8902

	<b>P_se</b>	<b>A_se</b>	<b>S_se</b>	<b>C_se</b>	<b>Co_se</b>
1	8.589	153.40	0.006399	0.04904	0.05373
2	3.398	74.08	0.005225	0.01308	0.01860
3	4.585	94.03	0.006150	0.04006	0.03832
4	3.445	27.23	0.009110	0.07458	0.05661
5	5.438	94.44	0.011490	0.02461	0.05688
6	2.217	27.19	0.007510	0.03345	0.03672

	<b>CP_se</b>	<b>Sy_se</b>	<b>FD_se</b>	<b>R_worst</b>
1	0.01587	0.03003	0.006193	25.38
2	0.01340	0.01389	0.003532	24.99
3	0.02058	0.02250	0.004571	23.57
4	0.01867	0.05963	0.009208	14.91
5	0.01885	0.01756	0.005115	22.54
6	0.01137	0.02165	0.005082	15.47

	<b>T_worst</b>	<b>P_worst</b>	<b>A_worst</b>	<b>S_worst</b>
1	17.33	184.60	2019.0	0.1622
2	23.41	158.80	1956.0	0.1238
3	25.53	152.50	1709.0	0.1444
4	26.50	98.87	567.7	0.2098
5	16.67	152.20	1575.0	0.1374
6	23.75	103.40	741.6	0.1791

	<b>C_worst</b>	<b>CO_worst</b>	<b>CP_worst</b>
1	0.6656	0.7119	0.2654
2	0.1866	0.2416	0.1860
3	0.4245	0.4504	0.2430
4	0.8663	0.6869	0.2575

5	0.2050	0.4000	0.1625
6	0.5249	0.5355	0.1741

	<b>Sy_worst</b>	<b>FD_worst</b>
1	0.4601	0.11890
2	0.2750	0.08902
3	0.3613	0.08758
4	0.6638	0.17300
5	0.2364	0.07678
6	0.3985	0.12440

## 3.2 ATTRIBUTE INFORMATION

**Table 1.** Attribute Information

<b>Attribute Name</b>	<b>Attribute Type</b>
<b>Diagnosis</b>	"factor"
<b>R_mean</b>	"numeric"
<b>T_mean</b>	"numeric"
<b>P_mean</b>	"numeric"
<b>A_mean</b>	"numeric"
<b>S_mean</b>	"numeric"
<b>C_mean</b>	"numeric"
<b>Co_mean</b>	"numeric"
<b>CP_mean</b>	"numeric"
<b>Sy_mean</b>	"numeric"
<b>FD_mean</b>	"numeric"
<b>R_se</b>	"numeric"
<b>T_se</b>	"numeric"
<b>P_se</b>	"numeric"
<b>A_se</b>	"numeric"
<b>S_se</b>	"numeric"
<b>C_se</b>	"numeric"
<b>Co_se</b>	"numeric"
<b>CP_se</b>	"numeric"
<b>Sy_se</b>	"numeric"
<b>FD_se</b>	"numeric"

(continúa)



(viene)

Attribute Name	Attribute Type
R_worst	"numeric"
T_worst	"numeric"
P_worst	"numeric"
A_worst	"numeric"
S_worst	"numeric"
C_worst	"numeric"
CO_worst	"numeric"
CP_worst	"numeric"
Sy_worst	"numeric"
FD_worst	"numeric"

Source: own work

### 3.3 CLASS DISTRIBUTION

Class distribution represents the instances belonging to each class. The following table presents this information with percentages and with an absolute count.

	freq	percentage
<b>B</b>	357	62.74165
<b>M</b>	212	37.25835

A Barplot of the predicting class variable is drawn as follows in Fig 1 for graphical representation of the class distribution.



**Figure 1.** Class Distribution


Source: own work

To get the idea of distribution of each attribute we have drawn probability density plots for the distribution of attributes.



$$Z = \beta_1\chi_1 + \beta_2\chi_2 + \dots + \beta_a\chi_a$$

$$S(\beta) = \frac{\beta^T \mu_1 - \beta^T \mu_2}{\beta^T C \beta} \quad \text{Score function}$$



$$S(\beta) = \frac{\bar{Z}_1 - \bar{Z}_2}{\text{Variance of } Z \text{ within groups}}$$

Linear Discriminant analysis is a classification (and dimension reduction) method [11]. It finds the linear combination of the variables that separates the target variable classes. The target can be a binary or multiclass variable. As an input, the LDA algorithm uses a data set of cases. For each case, a categorical variable and several predictor variables are used which are used to define the class. These predictor variables are numeric in nature. The input data is often visualized in the form of a matrix where each variable is a column and each row is a case. This data is used by the algorithm to divide the space into regions. These divided regions are labeled by categories and also have linear boundaries; hence the "L" in LDA. This model predicts the category of a new unseen case in accordance to which region it lies in. Generally, the model predicts the cases which lie in a region well.

## 4.2 CART

Classification and Regression Tree (CART) is one of the commonly used Decision Tree algorithms [12]. In general, it is known that decision trees follow a recursive partitioning approach. CART follows an extension of this approach and splits each of the input nodes into two child nodes. Therefore, the CART decision tree is a Binary Decision Tree in which the algorithm identifies a condition at each level of the decision tree where variables and levels are used for splitting the single input node (data sample) into two child nodes.

A few simple steps for building Decision Tree algorithms are:

1. Labelled input data is taken with a target variable and a list of independent variable. If independent variable .(typo)
2. Best split is found for each of the independent variables
3. For the split, the best variable is chosen

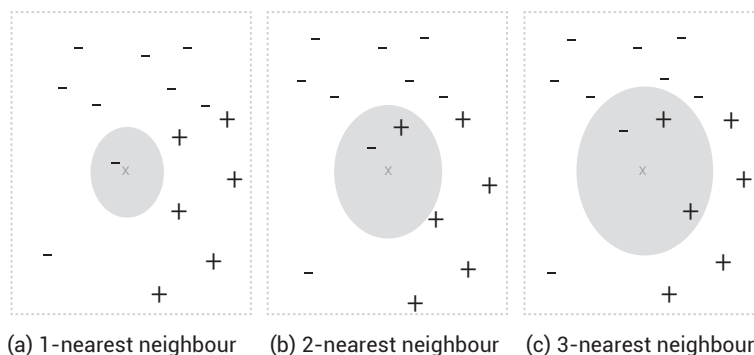
4. Input data is split into left and right nodes
5. The previous steps are continued until each node meets its stopping criteria
6. The Decision tree is built

### 4.3 k-NN

k- Nearest Neighbor classifier (k-NN) is a classification algorithm which delays the modeling procedure of the training data until it is necessary to label and classify the examples [13,14]. Therefore k-NN is also known as the "lazy learner". The k-NN algorithm follows a non-parametric technique as it does not make any assumptions on data distribution passed to the algorithm. Despite its simplicity, the algorithm performs very well with the following prerequisites:

- a) Set of stored records
- b) A distance metric to compute the distance between records
- c) Value of k i.e. number of nearest Neighbors

To classify an unknown record, the distance between other records are computed and based on this distance k-Nearest Neighbors are selected to decide the class label of the unknown record. Figure 3 represents the 1-nearest, 2-nearest and 3-nearest Neighbors. The class label of the data point is determined based on the class labels of its neighbors.

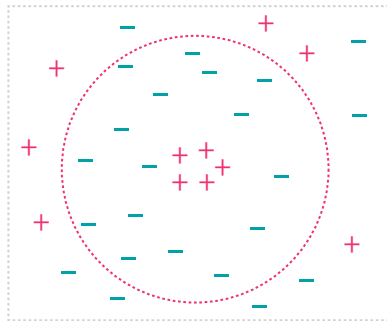


**Figure 3.** 1, 2 and 3 - Nearest Neighbor

Source: own work

When neighbors have more than one class label, the data point is classified as per the majority class of its nearest neighbors. As shown in figure 3 (a), the 1- nearest neighbor of the data point is a negative example and so the data point will be classified

as a negative class. In figure 3 (c), the number of nearest neighbors is three and the neighborhood contains two positive examples and one negative examples. Data point will be classified as per the majority vote. Hence the data point will be assigned the positive class label [19,20]. This discussion highlights the importance of choosing the right value of  $k$ . If  $k$  is too small, then the nearest neighbor classifier may be susceptible to overfitting because of noise in the training data. If  $k$  is too large, the classifier may not correctly classify the test instance as the list of nearest neighbors may include data points which are located very far from its neighborhood as shown in figure 4.



**Figure 4.** k-Nearest Neighbour Classification with large  $k$

Source: own work

The most common distance function, when computing the distance between two records 'p' and 'q', is Euclidean distance  $d(p,q)$  defined by

$$d(p, q) = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$$

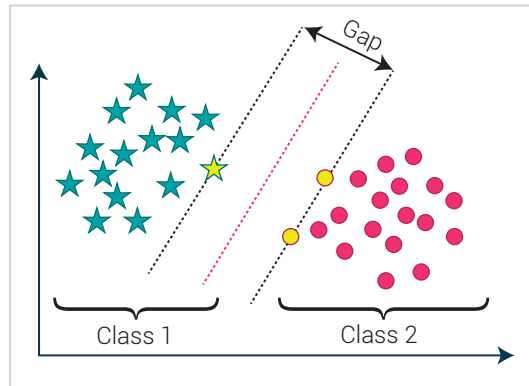
where  $p=(p_1, p_2, \dots, p_n)$  and  $q=(q_1, q_2, \dots, q_n)$ .

The other popular distance functions which are also used by  $k$ -NN are Manhattan, Minkowski and cosine distance metrics.

## 4.4 SVM

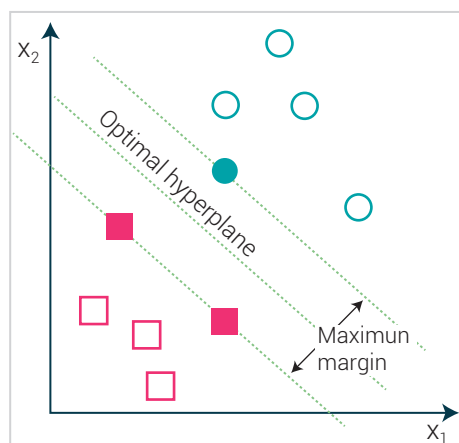
Support Vector Machine (SVM) is a classification technique which has its roots in statistical learning. This technique gives promising results in many practical applications like handwritten digit recognition and categorization of text. SVM is also popular with high dimensional data as it avoids the curse of dimensionality. However, SVMs do not directly provide probability estimates but these are calculated with an expensive five-fold cross validation [15].

Figure 5 shows a plot of a dataset which contains examples belonging to two different classes. The dataset is linearly separable as we can find a hyperplane so that all objects of one class reside on one side of hyperplane and objects of other class reside on other side.



**Figure 5. Linearly Separable Dataset**  
Source: own work

However, as shown in the figure 5, many hyperplanes are possible. The classifier must choose one of these hyperplanes to best represent its decision boundary. If the margin is small, the decision boundary can have a significant impact on the classification and are more susceptible to model overfitting. Figure 6 shows the optimal hyperplane for a decision boundary.



**Figure 6. Margin of a Decision Boundary**  
Source: own work

A linear SVM is a classifier which searches for a hyperplane with the largest margin. Hence it is also known as maximal margin classifier.

## 4.5 RANDOM FOREST

In the classification technique algorithm "Random Forest" many decision trees are combined into a single mode [16]. Generally, the predicted output of a decision tree may not be accurate but when multiple decision trees are combined together, the efficiency of the classifier increases. Random forest is similar to bootstrapping an algorithm with the Decision tree model; CART in particular. For example: If we have 500 observations in a population with 20 variables, the Random forest algorithm will try to build multiple CART models with diverse illustrations using a different initial variable for each model. The final output will be a function of each model prediction which can be calculated as a mean of each prediction.

## 5. EVALUATION OF CLASSIFICATION ALGORITHMS

To evaluate the algorithms, we split our dataset into 10 parts and a 10 fold cross validation was performed during each experiment. We implemented classification algorithms in the R open source tool. The output of each algorithm is given below:

### 5.1 OUTPUT OF LINEAR DISCRIMANT ANALYSIS (LDA)

#### *Linear Discriminant Analysis*

569 samples were taken using 30 predictor and two classes i.e. 'B', and 'M'. 10 fold cross validation was performed and no pre-processing was done

**Table 2.** Results of Resampling

<b>Kappa</b>	<b>Accuracy</b>
0.9077115	0.9579531

**Source:** own work

## 5.2 OUTPUT OF CART

### *CART*

569 samples were taken using 30 predictor and 2 classes: 'B', 'M'. 10 fold cross validation was performed and no pre-processing was done

**Table 3.** Results of Resampling across tuning parameters (CART)

<b>cp</b>	<b>Kappa</b>	<b>Accuracy</b>
0.004716981	0.8277594	0.9191340
0.049528302	0.7995353	0.9067896
0.792452830	0.3619641	0.7502107

**Source:** own work

Accuracy was calculated so as to assist in choosing the optimal model by means of the largest value and here the final value cp = 0.004716981 was selected for the model.

## 5.3 OUTPUT OF KNN

### *k-Nearest Neighbors*

569 samples were taken using 30 predictors and 2 classes: 'B', 'M'. 10 fold cross validation was performed and no pre-processing was done.

**Table 4.** Results of Resampling across tuning parameters (KNN)

<b>k</b>	<b>Kappa</b>	<b>Accuracy</b>
5	0.8543771	0.9330784
7	0.8545504	0.9330784
9	0.8623644	0.9366185

**Source:** own work

Accuracy was calculated so as to assist in choosing the optimal model by means of the largest value and here the final value k=9 was selected for the model.



## 5.4 OUTPUT OF SVM

### *Support Vector Machines with Radial Basis Function Kernel*

569 samples were taken using 30 predictors and 2 classes: 'B', 'M'. 10 fold cross validation was performed and no pre-processing was done.

**Table 5.** Results of Resampling across tuning parameters (SVM)

<b>c</b>	<b>Kappa</b>	<b>Accuracy</b>
0.25	0.9097508	0.9577673
0.50	0.9438755	0.9735891
1.00	0.9511668	0.9771292

**Source:** own work

Tuning parameter 'sigma' was held constant at a value of 0.04992736. We calculated accuracy to choose the optimal model by means of the largest value and here the final value for sigma as 0.04992736 and c=1 was selected for the model.

## 5.5 OUTPUT OF RANDOM FOREST

### *Random Forest*

569 samples were taken using 30 predictor and 2 classes: 'B', 'M'. 10 fold cross validation was performed and no pre-processing was done.

**Table 6.** Output of Random Forest

<b>Mtry</b>	<b>Kappa</b>	<b>Accuracy</b>
2	0.9288987	0.9666299
16	0.9287452	0.9665684
30	0.9323033	0.9683530

**Source:** own work

The largest value was taken when selecting the optimal model. The final value used for the model was mtry = 30.

## 6. SELECTING BEST MODEL

Models used: Random Forest, LDA, KNN, SVM, CART

Number of resamples: 10

**Table 7. Accuracy using Best Model**

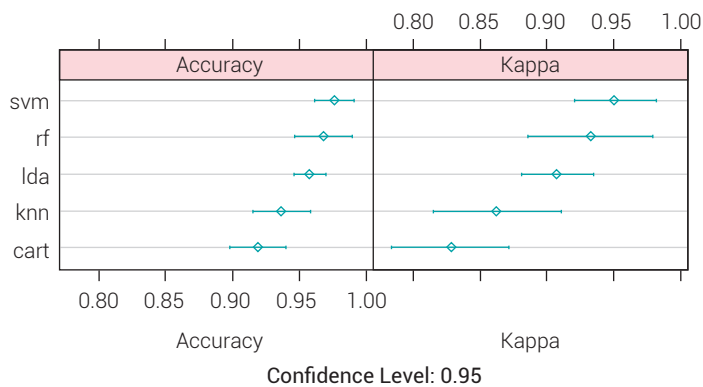
	Min.	Max.	1st Qu.	3rd Qu.	Median	Mean	NA's
<b>LDA</b>	0.9310345	0.9821429	0.9473684	0.9649123	0.9561404	0.9579531	0
<b>CART</b>	0.8793103	0.9649123	0.8977130	0.9298246	0.9122807	0.9191340	0
<b>k-NN</b>	0.8928571	0.9824561	0.9154919	0.9605263	0.9392015	0.9366185	0
<b>SVM</b>	0.9298246	1.0000000	0.9692199	0.9826830	0.9824561	0.9771292	0
<b>RF</b>	0.9137931	1.0000000	0.9510495	0.9956140	0.9736842	0.9683530	0 (yes..ok)

Source: own work

**Table 8. Kappa Matrix**

	Min	Max	1st Qu.	3rd Qu.	Median	Mean	NA's
<b>LDA</b>	0.8481675	0.9615385	0.8834356	0.9252498	0.9032563	0.9077115	0
<b>CART</b>	0.7503075	0.9246032	0.7832072	0.8512626	0.8150176	0.8277594	0
<b>k-NN</b>	0.7623762	0.9626719	0.8180708	0.9143116	0.8683356	0.8623644	0
<b>SVM</b>	0.8521401	1.0000000	0.9338370	0.9629610	0.9619238	0.9511668	0
<b>RF</b>	0.8185232	1.0000000	0.8942308	0.9906680	0.9432635	0.9323033	0

Source: own work



**Figure 7. Dotplot of Results**

Source: own work

Results show that for the Wisconsin breast cancer dataset, SVM and RF give the best results.

## 7. PREDICTIONS USING BEST MODEL

### 7.1 PREDICTIONS USING SVM

Confusion Matrix and Statistics

Reference	Prediction	<b>B</b>	<b>M</b>
B		71	2
M		0	40

**Table 9.** Predictions using SVM

<b>Accuracy :</b>	0.9823
<b>Kappa :</b>	0.9617
<b>Specificity :</b>	0.9524
<b>P-Value [Acc &gt; NIR]</b>	: <2e-16
<b>No Information Rate :</b>	0.6283
<b>Sensitivity :</b>	1.0000
<b>Mcnemar's Test P-Value</b>	0.4795
<b>Detection Rate :</b>	0.6283
<b>Pos Pred Value :</b>	0.9726
<b>Neg Pred Value :</b>	1.0000
<b>Prevalence :</b>	0.6283
<b>Balanced Accuracy :</b>	0.9762
<b>95% CI :</b>	(0.9375, 0.9978)
<b>Detection Prevalence :</b>	0.6460
<b>'Positive' Class :</b>	B

**Source:** own work

## 7.2 PREDICTIONS USING RF

Confusion Matrix and Statistics

Reference	Prediction	B	M
B		71	0
M		0	42

**Table 10.** Predictions using Random Forest

<b>Accuracy :</b>	1
<b>Kappa :</b>	1
<b>Prevalence :</b>	0.6283
<b>P-Value [Acc &gt; NIR] :</b>	< 2.2e-16
<b>Mcnemar's Test P-Value :</b>	NA
<b>Sensitivity :</b>	1.0000
<b>Neg Pred Value :</b>	1.0000
<b>No Information Rate (NIR) :</b>	0.6283
<b>Balanced Accuracy :</b>	1.0000
<b>Pos Pred Value :</b>	1.0000
<b>Detection Rate :</b>	0.6283
<b>Specificity :</b>	1.0000
<b>95% CI :</b>	(0.9679, 1)
<b>Detection Prevalence :</b>	0.6283
<b>'Positive' Class :</b>	B

**Source:** own work

## 8. CONCLUSION AND FUTURE SCOPE

Breast cancer is the most common cancer in women worldwide. Broadly speaking, there are two types of breast cancer: malignant and benign. Cancer prediction at an earlier stage could save many women's lives by early detection and treatment accordingly. In this paper we studied breast cancer and its types. To predict the breast cancer, we applied 5 machine learning techniques: linear Discriminant analysis, support vector machine, k-nearest neighbor, random forest and CART. We found that SVM and random forest gives the best results for breast cancer prediction. This work can be enhanced in the future by modifying the existing machine learning techniques or developing new algorithms to get better accuracy. Machine learning is the

future of breast cancer prediction and machine learning models are getting better day by day with the help of the extensive research performed in this domain by the researchers. In the coming decades, Artificial Intelligence and Machine Learning are set to change the medical industry. With the incorporation of advanced techniques like convolutional neural networks, prediction of breast cancer will outperform traditional pathological tests.

## REFERENCES

- [1] C. E. Fear, *et al.*, "Confocal microwave imaging for breast cancer detection: Localization of tumors in three dimensions," *IEEE Transactions on biomedical engineering*, vol. 49, no. 8, pp. 812-822, 2002. [Online]. doi: 10.1109/TBME.2002.800759
- [2] N. K. Nikolova, "Microwave imaging for breast cancer," *IEEE microwave magazine*, vol. 12, no. 7, pp. 78-94, 2011. [Online]. doi: 10.1109/MMM.2011.942702
- [3] Xie, Yao, *et al.*, "Multistatic adaptive microwave imaging for early breast cancer detection," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 8, pp. 1647-1657, 2006. [Online]. doi: 10.1109/TBME.2006.878058
- [4] E. J. Bond, *et al.*, "Microwave imaging via space-time beamforming for early detection of breast cancer," *IEEE Transactions on Antennas and Propagation*, vol. 51, no. 8, pp. 1690-1705, 2003. [Online]. doi: 10.1109/TAP.2003.815446
- [5] J. L. Kelsey, D. G. Marilie, and M. J. Esther, "Reproductive factors and breast cancer," *Epidemiologic reviews*, vol. 15, no.1, p. 36, 1993. [Online]. doi: 10.1093/oxfordjournals.epirev.a036115
- [6] P. A. Francis, *et al.*, "Adjuvant ovarian suppression in premenopausal breast cancer," *New England Journal of Medicine*, vol. 372, no. 5, pp. 436-446, 2015. [Online]. doi: 10.1056/NEJMoa1412379
- [7] C. E. De Santis, *et al.*, "Breast cancer statistics, 2015: Convergence of incidence rates between black and white women," *CA: a cancer journal for clinicians*, vol. 66, no. 1, pp. 31-42, 2016. [Online]. doi: 10.3322/caac.21320
- [8] C. E. De Santis, *et al.*, "Breast cancer statistics, 2017, racial disparity in mortality by state," *CA: a cancer journal for clinicians*, vol. 67, no. 6, pp. 439-448, 2017. [Online]. doi: 10.3322/caac.21412

- [9] T. J. Whelan, *et al.*, "Regional nodal irradiation in early-stage breast cancer," *New England Journal of Medicine*, vol. 373, no. 4, pp. 307-316, 2015. [Online]. doi: 10.1056/NEJMoa1415340.
- [10] K. C. Oeffinger, *et al.*, "Breast cancer screening for women at average risk: 2015 guideline update from the American Cancer Society," *Jama*, vol. 314, no. 15, pp. 1599-1614, 2015. [Online]. doi: 10.1001/jama.2015.12783
- [11] M. Kan, *et al.*, "Multi-view discriminant analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 188-194, 2015. [Online]. doi: 10.1109/TPAMI.2015.2435740
- [12] M. Batra, and R. Agrawal, "Comparative Analysis of Decision Tree Algorithms," in B. Panigrahi, M. Hoda, V. Sharma, and S. Goel (eds), *Nature Inspired Computing. Advances in Intelligent Systems and Computing*, vol 652, pp. 1-4 Springer, Singapore, 2018. [Online]. doi: [https://doi.org/10.1007/978-981-10-6747-1\\_4](https://doi.org/10.1007/978-981-10-6747-1_4)
- [13] R. Agrawal, "Design and development of data classification methodology for uncertain data," *Indian Journal of Science and Technology*, vol. 9, no. 3, pp. 1-12, 2016. [Online]. doi: 10.17485/ijst/2016/v9i3/72262
- [14] R. Agrawal, "Integrated Parallel K-Nearest Neighbor Algorithm," *Smart Intelligent Computing and Applications*, Springer, Singapore, pp. 479-486, 2019. [Online]. doi: [https://doi.org/10.1007/978-981-13-1921-1\\_47](https://doi.org/10.1007/978-981-13-1921-1_47)
- [15] R. Agrawal, [Online]. Available: <https://shodhganga.inflibnet.ac.in/handle/10603/169657>
- [16] G. Biau, and E. Scornet. "A random forest guided tour." *Test* 25.2 , 197-227.,2016, 10.1007/s11749-016-0481-7. [Online]. Available: <http://www.lsta.upmc.fr/BIAU/test-bs.pdf>
- [17] American Cancer Society. (2019). Cancer.org. Accessed 11 July 2019. [Online]. Available: <https://www.cancer.org/cancer/breast-cancer/non-cancerous-breast-conditions.html>
- [18] Nihgov. (2019). PubMed Central (PMC). Accessed 11 July 2019. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4115707/>
- [19] R. Agrawal, "Integrated Effect of Nearest Neighbors and Distance Measures in k-NN Algorithm," in V. Aggarwal, V. Bhatnagar, and D. Mishra (eds), *Big Data Analytics. Advances in Intelligent Systems and Computing*, vol 654, pp. 1-6 Springer, Singapore, 2018. [Online]. doi: [https://doi.org/10.1007/978-981-10-6620-7\\_74](https://doi.org/10.1007/978-981-10-6620-7_74)

- [20] R. Agrawal, "A modified K-nearest neighbor algorithm using feature optimization," *Int. J. Eng. Technol.*, vol. 8, no. 1, pp. 28–37, 2016. [Online]. Available: <http://www.enggjournals.com/ijet/vol8issue1.html>