IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

MACHINE LEARNING MODELS FOR EARLY DETECTION OF BREAST CANCER

[1]Uzma Nazir, [2]Er.Mandeep Kaur

[1]M.Tech Student, [2]Assistant Professor,

[1]Computer Science Engineering,

[1]Desh Bhagat University, Mandi Gobindgarh, Fatehgarh Sahib, Punjab- 147301, India

Abstract: This research is justified from the scientific field since through the development of an intelligent system will support mammographic analysis in timely detection of breast tumors thus generating an innovative idea that contributes to the science of health. From the financial field it is justified that this system is not very expensive compared to other systems that are in the technological market, it will also reduce the waiting time for a timely and early diagnosis of breast cancer, thus obtaining more patients attended that will bring high prestige to the institution. Continuing with the social sphere, it is justified since it intends to support the mammographic analysis, where there will be a minimum margin of error and in addition to helping the oncology area by having a timely diagnosis. Also this system indirectly benefit the students of the medical career, as well as the specialists in charge of this area. Finally, this present investigation is justified technologically since present technologies will be used as much as software and hardware that will allow this intelligent system to be carried out in order to support the radiologist with the mammographic analysis. However, this research on breast cancer, which successfully carried out the calculation analysis, using various models of Machine Learning like logistic regression, support vector machine, decision tree, and random forest Consequently, Logistic Regression with dependent and independent variables results in the test data accuracy value of 96%. Thereafter, SVM vide linear hyperplane determining the affected patients with breast cancer and number of predictions achieved 97% of accuracy. Subsequently, the Decision Tree modeling with Random Forest achieved 96% of accuracy collectively.

Index Terms **– Breast Cancer Detection, Machine Learning, Random Forest, Logistic Regression, Decision Tree, Support Vector Machine.**

I. INTRODUCTION

Machine learning is a field of computer science that gives computers the ability to learn without being explicitly programmed. Machine Learning (ML) is one of the applications of Artificial Intelligence (AI) which focuses on developing a system that is able to learn on its own without having to be programmed repeatedly. ML requires a data (data training) as a learning process before producing a result. So, in simple terms it can be explained that Machine Learning is computer programming to achieve certain criteria/performance by using a set of training data or past experience [1]. Several studies have been conducted, concluding that ML can be used in the medical field to predict disease [2] [3] [4][5].

Until now there are several ML algorithms that can be used and developed for various purposes. A study was conducted to compare the effectiveness of several algorithms in ML, including Naïve Bayes (NB) , Radial Basis Function (RBF) and Support Vector Machine (SVM) [4]. Based on the research results , it was obtained that SVM is the algorithm with the highest level of accuracy . Evidenced by the algorithm test which shows the accuracy value of the SVM algorithm reaches 93.75% while for the NB algorithm it is only 71.67% and RBF is 70.01% [4].

The support vector machine (SVM) is a relatively new classification or prediction method developed by Cortes and Vapnik in the 1990s as a result of the collaboration between the statistical and the machine learning research community[3]. SVM is a classification technique for nonlinear problems. SVMis most suitable for working accurately and efficiently with high dimensionality feature spaces in addition to that SVMis based on strong mathematical foundations and results in a simple way and very powerful algorithms [6]. Breast cancer or Carcinoma Mammae is a condition when cancer cells form in the breast tissue. Cancer can form in the glands that produce milk (lobules), or in the ducts (ducts) that carry milk from the gland to the nipple.

Cancer can also form in the fatty tissue or connective tissue in the breast. Breast cancer is the second leading cause of death rate in women. Breast cancer represents about 12% of all new cancer cases and 25% of all cancers in women [7].The world health organization, the International Agency for Research on Cancer (IARC) estimates that more than 400,000 women die each year from breast cancer [6]. There are still many people who assume that cancer is the same as a tumor, in fact tumors that appear are not always cancerous. Examination with biopsies and mammography can be used to detect the type of breast cancer.

The results of a biopsy examination with Fine Needle Aspiration (FNA) can determine whether the type of breast cancer cell is malignant or benign. The growth of breast cancer that starts from a tumor is grouped into several stages, ranging from stage 0 to IV. The delay in detecting symptoms of breast cancer causes many sufferers to find out about their condition after entering a high stage (the average is at stages III and IV). In this condition, the risk of death is much higher.

This study aims to build an ML application using the SVM, Decision Tree, Random Forest and Logistic Regression algorithm that can be used to diagnose breast cancer based on data patterns taken from biopsies. The results of the diagnosis can produce a prediction for determining which type of breast cancer is malignant or benign.

II. LITERATURE REVIEW

- **Logistic Regression**

A better understanding of the factors involved in pentest would help to make a better decision and the reason is that it provides us with a better analytical capacity that allows us to choose the best possible path, for this research we will make use of the logistic regression because it is a versatile tool and that with dichotomous variables it is the most used [8][9][10]. For this, we must ask ourselves, what is logistic regression?

ML regression analysis deals with the study of the dependencies of a variable of interest concerning one or more explanatory variables through techniques for their analysis and modeling. It helps us estimate the uncertain expectation of the dependent variable given the independent variables. The objective of the estimation is a purpose of the independent variables called the regression function and it describes the variation of the variable of interest based on a probability distribution [8][9][10].

Many methods have been developed to carry out regression analysis and their use in practice depends on the form of the data generation process which depends on making assumptions about this process. The best-known regression method is linear regression, which uses the conventional regression analysis; the method is limited since it is only applicable if the dependent variable is continuous, independent, and identically distributed when the dependent variable is categorical, the conventional regression analysis is not the most appropriate because it fails to comply with the fact that it must be continuous, it can take negative values, it must be distributed normally and in terms of error it must be independent and identically distributed, these do not hold in the cases and where the dependent variable is dichotomous.

$$\boldsymbol{\sigma(z) = \frac{1}{1 + e^{-x}}} \ \boldsymbol{(eq.1)}$$

Where x is a real number. From this equation, we can see that as x approaches minus infinity, the quotient approaches zero and as x approaches infinity the quotient approaches one, as shown in figure 1.

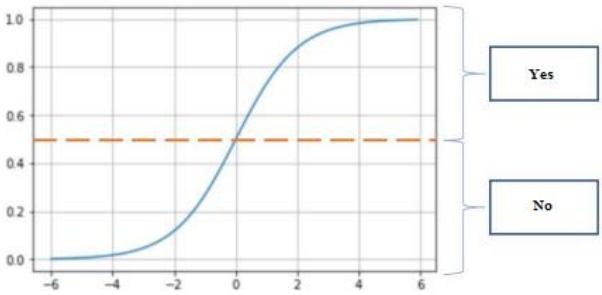


Figure 1: Graph of the sigmoid function and its threshold

To used to approximate the dependency relationship between a dependent variable Y, the independent variable X, and a random term ε. This model can be expressed as:

$$\boldsymbol{Y = \beta_{j0} + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i} \quad \boldsymbol{(eq.2)}$$

To predict a $\boldsymbol{Y}$ value greater than 1. The specific form of the logistic regression model is:

$$\boldsymbol{\pi(x) = \frac{e^{\beta_{j0} + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i}}{1 + e^{\beta_{j0} + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i}}} \boldsymbol{(eq.3)}$$

The previous defaults drive the evolution towards generalized linear models, within which the so-called log-linear and the so-called LOGIT models can be included as :-

$$\boldsymbol{logistic(p) = log\left(\frac{p}{1-p}\right) = log(p) - log(1-p)} \quad \boldsymbol{(eq.4)}$$

Therefore, the logistic model expresses the dependent variable as the occurrence or not of an event in terms of probability. Among the positive aspects of this methodology, it is found that it is a simple and easy to interpret model, it is a light process from the point of view of computational resources, in addition to allowing the use of multiple variables even with few variables. cases for each one of them, and allows to obtain consistent estimates of the probability of default, identify the risk factors that determine said probabilities, as well as the influence or relative weight of these on them .

On the other hand, being a linear methodology, it does not allow direct solving of non-linear problems, for example, if the probability is U-shaped, that is, it is initially reduced by increasing a characteristic, and later the probability increases by continuing increasing the characteristic, a logistic model cannot reflect this behavior directly, this forces to transform this characteristic previously so that the model can register this non-linear behaviour.

- **Decision Tree**

The decision tree methodology is a widely used data mining method to establish classification systems based on multiple covariates or to develop prediction algorithms for an objective variable [11][12][13][14] The Tree Models behave like a cluster analysis (creation of homogeneous groups and different from each other) and at the same time as a predictive method. Given a data set, we will obtain diagrams of logical constructions that serve to represent and categorize a series of conditions that occur successively, for the resolution of a problem both classification and regression. The process of creating the tree is laborious due to the complicated casuistry and combinatorial that is generated each time it is necessary to divide a node[11][12][13][14].

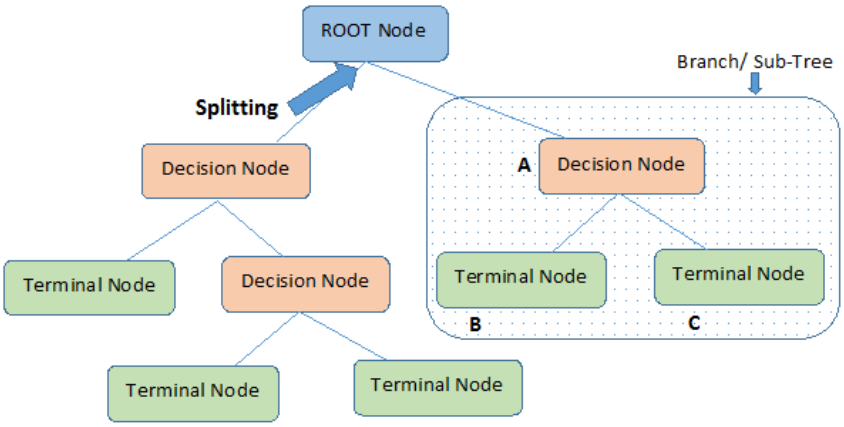


Figure 2: Model Decision Tree

The construction of trees must take into account the following aspects: criteria for the selection of independent variables for the division of each node, criteria for choosing the optimal cut within each independent variable or node, criteria for choosing cut groups for independent variables nominals with more than one category, the minimum number of observations to build each node, stopping criteria, limits on the number of nodes and splits, criteria for handling missing values, and whether validation data will be used to control the construction process [11][12][13][14].

Regarding the criteria for selecting the optimal cut-off point for each variable, since we find ourselves in our specific case with a problem in which both the dependent variable and the predictors are qualitative, some are:

Chi-square: the cut-off is a selected or grouping of categories of the independent variable with the highest value of the associated statistic, crossing the dependent variable with the independent one. Normally a significance test is applied.

$$x^2 = \sum_{i=1}^{m}\sum_{j=1}^{k}\left(\frac{(n_{i,j} - \frac{n,n_j}{N})^2}{\frac{n,n_j}{N}}\right) (eq.5)$$

Gini index: the smaller the better, since it indicates a greater distance between classes (also called impurity). Subsequently, the division of the independent variable that improves the Gini index is selected compared to not using it.

$$I(Gini) = \left(1 - \sum_{i=1}^{k}\left(\frac{n_i}{n}\right)^2\right)^2 (eq.6)$$

Entropy: measures the information gained in divisions. The less, the better, the more information. Subsequently, the division that improves the entropy index concerning the base entropy of the parent node is selected.

$$I(Entropy) = -\log\sum_{i=1}^{k} p_i \; logp_i (eq.7)$$

Regardless of the selected criterion, for simplicity, the possibility that each node is only divided into two parts at each step will be considered. The algorithm ends when some of the following stopping criteria are met: there are not enough observations in the final sheets to consider splitting, the maximum depth has been reached, or there is no improvement in the splitting criteria at any node. The final tree model may be overfitting if the data is complex.

After obtaining the final tree, you can act as in the cluster methods: choose the subtree that seems most stable. This process is known as pruning. Finally, regarding the advantages of these models, it is a technique that has great descriptive power, allows non-linear relationships, there is no theoretical assumption of the data, provides measures of the importance of the variables, etc. However, among the disadvantages, we find low reliability (poor generalization) and low predictive efficiency.

- **Random Forest**

The Random Forest algorithm is a supervised learning technique that generates multiple decision trees on a training data set: the obtained results are combined to obtain a single model that is more robust compared to the results of each tree separately.

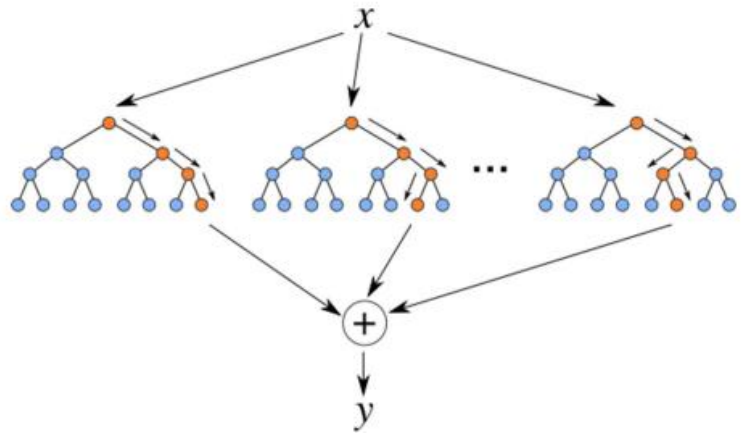


Figure 3: Random Forest Diagram

The Random Forest algorithm goes one step further in avoiding the problem of variable selection, avoiding rigidly deciding on a set of variables, and taking advantage of bagging at the same time. It is a question of incorporating two sources of variability (resampling of observations and variables) to gain generalization capacity and reduce overfitting while conserving the ability to adjust well particular relationships in the data (interactions, nonlinearity, cuts, problems). extrapolation, etc.). Random Forest also avoids the problem of very dominant predictor variables. With only Bagging, in the case of a very dominant pair of variables, the trees would be similar. By adding randomness to the variables used, different trees are obtained, which reduces the variance of the model [15][16][17].

$$\boldsymbol{\varphi(T) = \sum_{t \in \widetilde{T}} p(t)\, \varphi\,(t)\ \ (eq.8)}$$

where p (t) is the probability that a given example corresponds to leaf t and φ(t) is the impurity of the terminal node t.

- **Support Vector Machine**

Support Vector Machines (SVM) have become very popular due to their great classification potential [34]. The SVM transforms the input vectors to a space of higher dimension through a nonlinear transformation. Given a space of features with samples of two different classes, the goal of the SVM is to find an equation plane:

$$\boldsymbol{w^T x + b = 0\ (eq.9)}$$

that allows them to be separated [18] [19], as shown in Figure 4.

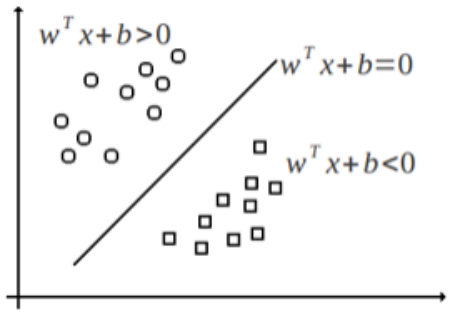


Figure 4: Plane separator of two classes using SVM.

The objective of SVM is that this plane is obtained in such a way that its distance with each one of the classes is maximal. For this, the support vectors are defined, which are those that characterize the limit of the class and that is obtained from the examples closest to the separating plane [18][19]. The problem then becomes one of maximizing the distance or margin ρ between the support vectors of the different classes, as shown in Figure 5.

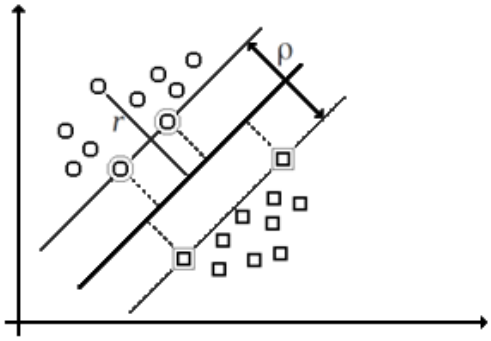


Figure 5: Example of a separable problem in two dimensions. They are called support vectors to the points that make up the two lines parallel to the hyperplane, being the distance ρ between them the greatest possible.

If we define the notation $y_i = -1$ if it is below the plane and $y_i = 1$ if it is above the plane, the samples x_i of the classes satisfy the following equation:

$$\boldsymbol{y_i\,(wx_i + b) - 1 \geq 0\,(eq.10)}$$

In particular for the support vectors , the above equation is equality. What the distance from a point in space to a plane corresponds to the projection, can be expressed in the margin as:

$$\boldsymbol{L(w.b.a) = \frac{1}{2}\,||w||^2 - \sum_{i=1}^{l} a_i\left(y_i\left((x_i.w + b) - 1\right)\right)\,(eq.11)}$$

With N the number of samples used to train the classifier. This problem is usually written in its dual form and solved using Lagrange multipliers, which is finally expressed as follows:

$$\boldsymbol{Max \sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i,j=1}^{l} a_i a_j y_i y_j\, x_i^T x_j\,(eq.12)}$$

In the case that the classes are not linearly separated in the space of characteristics, the SVM allows the introduction of a parameter C, corresponding to the weight or punishment of misclassifying the samples. So the optimization problem consists of a balance between separating classes and minimizing errors, given by the following function:

$$\boldsymbol{K(x_i, j_i) = \left(\left(x_i, x_j^T\right) + \ 1\right)^n\ (eq.13)}$$

Linear kernel functions take the form of equations:

$$\boldsymbol{K(x_i, j_i) = x_i, x_j^T \quad (eq.14)}$$

The rbf kernel function has the form of the equation:

$$\boldsymbol{K(x_i, j_i) = exp\frac{||x_i, j_i||^2}{2\sigma^2}\ (eq.15)}$$

- **Confusion Matrix**

The confusion matrix is an evaluation measure to assess quality classifier. The confusion matrix states the correct amount of test data classified and the amount of misclassified test data (Windrawati, 2020). There are several sizes that can be used in assessing the data evaluating the classification model such as accuracy or recognition rate, error rate or level error, recall or sensitivity or true positive rate, specificity or true negative rate, precision, F measure or F1 or F-score or harmonic average of precision and recall [20].

Confusion Matrix		**The Actual Class**	
		1	**2**
Prediction Class	**1**	**True Positive**	**False Negative**
	2	**False Positive**	**True Negative**

Table 1: Confusion Matrix Table

$$\boldsymbol{Accuracy = \frac{TP + FN}{TP + FP + TN + FN}\quad (eq.16)}$$

$$\boldsymbol{Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}\quad (eq.17)}$$

$$\boldsymbol{Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}\quad (eq.18)}$$

III. PROPOSED MODEL

- **Diagnosis of Breast Cancer**

Breast cancer is formed when cells in the breast grow abnormally and out of control. These cells generally form tumors. Breast cancer can be detected by doing a biopsy and mammography. Biopsy is an examination technique that is carried out by taking fluid from the breast using FNA, then the results of the FNA biopsy will be re-examined in the laboratory to obtain a diagnosis. Mammography is an examination technique using low-grade X-rays to assess breast tissue. Biopsy is the most accurate examination for detecting breast cancer, because it can show the type of cancer cells and their stage. Breast cancer can be assessed by three main factors, namely tumor (primary tumor), lymph nodes (regional lymph nodes), and spread (distant metastases). Figure 6 describes indicators for breast cancer diagnosis.

Primary Tumor (T)	
T0	No evidence of primary tumor
Tis	Carcinoma in situ
T1, T2, T3, T4	Increasing size and/or local extension of the primary tumor
TX	Primary tumor cannot be assessed (use of TX should be minimized)
Regional Lymph Nodes (N)	
N0	No regional lymph node metastases
N1, N2, N3	Increasing number or extend of regional lymph node involvement
NX	Regional lymph nodes cannot be assessed (use of NX should be minimized)
Distant Metastasis (M)	
M0	No distant metastases
M1	Distant metastases present

Figure 6: Indicators for Breast Cancer Diagnosis.

The results of the biopsy are then taken to the laboratory for a more detailed examination. The FNA will show several indicators related to the development of breast cancer cells. Table 2 describes the description of the indicators from the FNA results.

#	INDICATORS	DESCRIPTION
1	Clump Thickness	Assess whether the cell is mono or multi-layer
2	Uniformity of cells sizes	Evaluate the consistency of cell size in the sample
3	Uniformity of cells shape	Evaluate the consistency of the shape of the cells in the sample
4	Marginal Adhesion	Calculates the proportion of cells that are fused together
5	Single Epithelial Cell size	Measures the enlargement of epithelial cell size
6	Bare Nuclei	The proportion of nuclei that are surrounded by cytoplasm versus those that are not
7	Bland Chromatin	Assess the similarity of the "texture" of the nucleus in the fine to coarse range
8	Normal Nucleoli	Determines whether the nuleoli are small and barely visible or more visible
9	Mitoses	Describes the level of mitotic activity

Table 2.Description of FNA outcome indicators in Breast Cancer Wisconsin

- **WN Street, OL Mangasarian, and WH Wolberg. Breast Cancer Wisconsin (Prognostic) Dataset. UCI**

The data used in this study uses the Wisconsin Breast Cancer dataset(Diagnostic), made by Dr. William H. Wolberg, W. Nick Street, and Olvi L. Mangasarian. The database is in the form of dataset obtained from

the results of digital image analysis of breast masses using the FNA method, they analyzed the development of abnormal living cells in digital images.

#	Attribute	Domain
H1	Sample code number	id number
H2	Clump Thickness	1 -10
H3	Uniformity of Cell Size	1 -10
H4	Uniformity of Cell Shape	1 -10
H5	Marginal Adhesion	1 -10
H6	Single Epithelial Cell Size	1 -10
H7	Bare Nuclei	1 -10
H8	Bland Chromatin	1 -10
H9	Normal Nucleoli	1 -10
H10	Mitoses	1 -10
H11	Class:	(2 for benign, 4 for malignant)

Figure 7: Wisconsin Breast Cancer dataset information

Figure 8, describes the information from the Wisconsin Breast Cancer dataset. Breast Cancer Wisconsin have 699instances(benign: 458 and malignant: 241), 2 classes (65.5% malignant and 34.5% benign), and 11 attribute values integrates (Asri et al., 2016). Datasets it is also an indicator that can be seen in living cells to detect the presence of breast cancer, each record has nine attributes indicators besides Sample Code Number And class the nine attributes are assessed on an interval scale of 1 to 10, with a scale of 10 being the most abnormal situation assessment, so that the closer the value of each attribute is to 10, the more likely it is to be detected malignant(malignant).

H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11
1000025	5	1	1	1	2	1	3	1	1	12
1002945	5	4	4	5	7	10	3	2	1	2
1015425	3	1	1	1	2	2	3	1	1	2
1016277	6	8	8	1	3	4	3	7	1	2
1017023	4	1	1	3	2	1	3	1	1	2
1017122	8	10	10	8	7	10	9	7	1	4
1018099	1	1	1	1	2	10	3	1	1	2
1018561	2	1	2	1	2	1	3	1	1	2
1033078	2	1	1	1	2	1	1	1	5	2
1033078	4	2	1	1	2	1	2	1	1	2

Figure 8: Illustrate sample datasets from Breast Cancer Wisconsin

- **Research Framework**

The flow chart depicted in Figure 9 illustrates the research's framework. The flowchart depicts the phases of This study involves normalizing the data as a preprocessing step, training and testing the model using machine learning techniques based on support vector machines, decision trees, logistic regression, and random forests, and evaluating the model using confusion matrices.

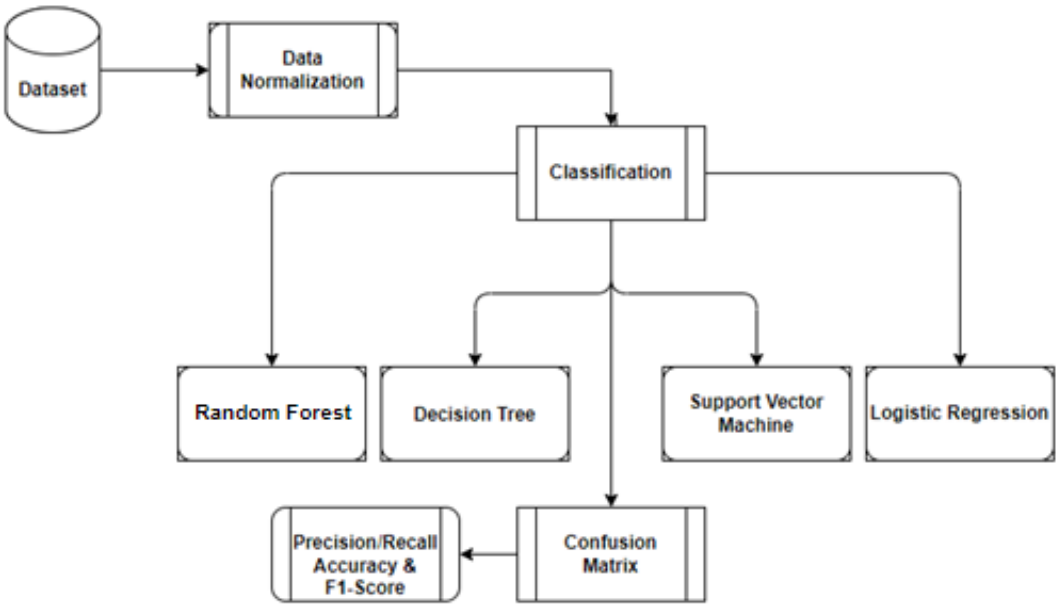


Figure 9: Proposed Methodology

IV. RESULTS

- **Algorithm Support Vector Machine**

Since we have the information of the two distributions, we apply the algorithm of SVM using the Probability Product Kernel so that the weight of the hyperplane k is given by the value of π_k. The scheme splits the entire dataset into the two classes (Benign and Malignant) into which they are to be separated, this is done to extract the hyper plane of each class, if we didn't split everything first the data set in classes we would not know to which hyper plane each class corresponds.

Since to work with values that are representative of each class instead of working with all the data, we extract the hyper plane of each class, in this way for each class we have a set of vectors representing the means and a set of matrices that represent the covariance's associated with each mixture of hyper plane of each class.

In order to extract the hyper plane, the expectation maximization method is applied that in addition to generating the information of the hyper plane also allows to calculate the probability with which each of these hyper plane generates the data. Instead of removing points from the original data set or identifying the points most representative at the end of this step, we have the information of the averages, the covariance's and the probability with which each Gaussian generates the data set original data, this is the data compression step.

The data set of size N became information of its clusters that in total are defined by parameters. Where K in the number total classes and D the dimension of the data. For very large values of N this greatly reduces the number of data to work with the distance formula as follows:

$$\boldsymbol{D = \frac{|Ax \ + \ By \ + \ C|}{\sqrt{A^2 + B^2}}.}$$

Equation above is converted into a dot product in a vector so that it becomes:

$$\boldsymbol{[AB] \begin{bmatrix} x \\ y \end{bmatrix} + C = 0}$$

Suppose $\boldsymbol{w} = [A \ B]$ and $\boldsymbol{x} = \begin{bmatrix} x \\ y \end{bmatrix}$ and b = C, then we get:

$$D = \frac{|Ax + By + C|}{\sqrt{A^2+B^2}} = \frac{|w \cdot x + b|}{\sqrt{w^2+C^2}} = \frac{|w \cdot x + b|}{\sqrt{w^2}} = \frac{|w \cdot x + b|}{||w||}$$

The margin value can be found using the middle value between the two classes as follows:

$$margin = \frac{1}{2} \ (d^+ - d^-)$$

$$K=\frac{1}{2} \ \left(\frac{|w \cdot x_1 + b|}{||w||} - \frac{|w \cdot x_2 + b|}{||w||}\right)$$

$$K=\frac{1}{2} \ \left(\frac{1}{||w||} - \frac{(-1)}{||w||}\right)$$

$$K=\left(\frac{1}{||w||}, \ ||w|| \neq 0, \right)$$

Where:

D +: the distance between the hyperplane against class +1, distance between hyperplane and class -1. Each class must add restrictions on the data from each class so that it does not enter into the margins, the limitations are as follows:

$$w \cdot x_i + b \leq -1, \text{if } y = -1,$$
$$w \cdot x_i + b \geq +1, \text{if } = +1,$$

or it can be written as follows:

$$yi \ (w \cdot xi + b) - 1 \geq 0, \ \forall 1 \leq i \leq n, i \in N.$$

Maximizing the equivalent margin value by minimizing $\|w\|2$. Then the search for the best hyperplane with the largest margin value can be formulated into a quadratic programming optimization problem as follows:

$$max \ margin = min \ \frac{1}{2} \ ||w||^2$$

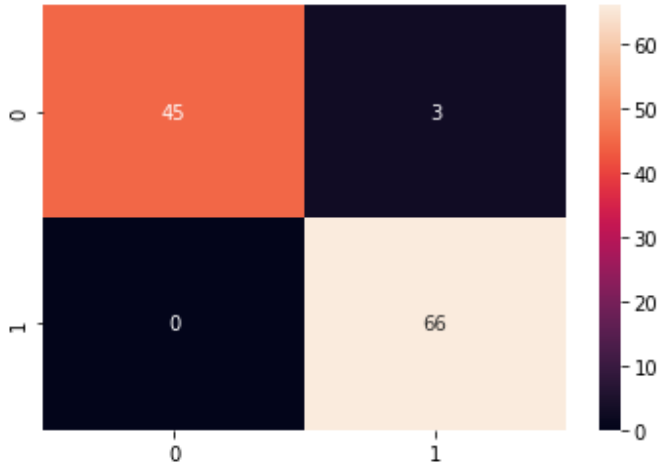
with constraints:

$$yi \ (w \cdot xi + b) - 1 \geq 0, \ \forall 1 \leq i \leq n, i \in N.$$

This problem can be solved by converting the equation into a function lagrange:

$$min \ L_p(w.b.a) = \frac{1}{2} \ ||w||^2 - \sum_{i=1}^{l} a_i \left(y_i((x_i.w + b) - 1)\right)$$

One consequence of using SVM is that the classification model takes as input probability functions and not data vectors. To use the classification model on a vector x a Gaussian with mean equal to the desired vector is constructed classify and with covariance equal to the identity matrix.




```
print(classification_report(y_test,y_predict))
```

	precision	recall	f1-score	support
0.0	1.00	0.94	0.97	48
1.0	0.96	1.00	0.98	66
accuracy			0.97	114
macro avg	0.98	0.97	0.97	114
weighted avg	0.97	0.97	0.97	114

Figure 10: Breast Cancer Prediction using Support Vector Machine

- **Algorithm Decision Tree**

A set of rules for dividing a large (heterogeneous) population in this decision tree model into smaller (homogeneous) populations by considering the target variable. Usually, the target variable is classified precisely and is intended to calculate the probability of each record related to its category from the decision tree model or to group the same records into one class. Making a decision tree can use one of several decision tree algorithms, which aim to model data sets that have not been grouped by class.

One of the most popular and widely used learning methods is the decision tree. This method is an attempt to get the approximation function through discrete values. The concept of the decision tree itself is to transform data in the decision table into a decision tree and decision rules known as rules.

Data in a decision tree is usually represented in the form of a table containing attributes and records. Attributes declare the parameters that will be used as the basis for forming the Tree. For example, to determine the level of human risk of stroke, it is necessary to consider criteria such as the history of heart disease, glucose levels, lifestyle, type of work, smoking status, and gender. One of the attributes used to specify Data in a per-Data solution is termed the target attribute. While the attribute itself has a value that is termed an instance.

The decision tree goes through a process in the form of converting table data into a tree model, converting the form of the tree model into rules, then simplifying the rules. The first step in building a decision tree is to calculate the total entropy value of the number of data samples, then group the variables for the gain value for each attribute. When the calculation is complete, the attribute with the highest gain value becomes the root, the other attributes become the branch, the branch is then recalculated to see which other attribute has the highest gain value.

The calculation step is repeated continuously so that all attributes are executed. The main benefit of using a decision tree is that it can break down complex decision-making processes into simpler ones. This allows the decision-making process to interpret solutions, not problems. Decision Trees also help to explore data and find hidden relationships between input variables and target variables. Decision trees combine

data exploration and modeling and are the first step in the modeling process even when used as the final model for other techniques.

A decision tree, also known as a flowchart, is shaped like a tree structure, with each internal node representing an attribute test, each branch representing the output of the test results, and the leaf nodes representing the class distribution. The top node is referred to as the root node. The decision tree is used to classify data samples whose class is not yet known into existing classes. In the data testing path, all data must first go through the root node and finally through the leaf node. This will conclude class predictions for the data.

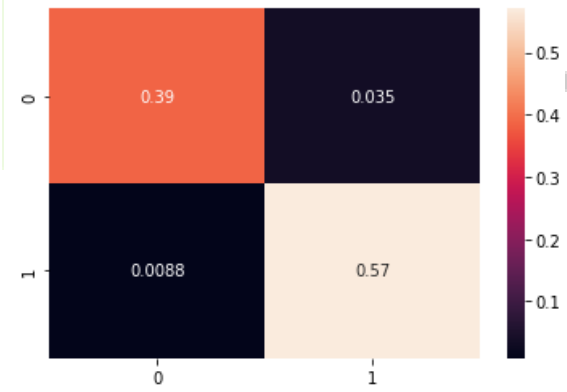
Data attributes must be categorical data. If it is continuous, the attribute must be discretized first. The Decision Tree method has several advantages over other methods for large databases, namely:

1. Has a relatively faster speed.
2. Can be turned into a classification rule easily and simply.
3. Can use SQL queries to access the database.
4. High accuracy compared to other methods.

- ***The algorithm is as under***: –

Classify function (attribute, tree)
Input: attribute: the attribute to be classified;
Tree: the decision tree used to classify;
Output: class: the class of the attribute.
Class ← take-value (root (tree), instance);
if leaf (root (tree)) then return class
else classify(instance, sub-tree(tree, class));
end if
end function

```
from sklearn import metrics #Import scikit-learn

cm=confusion_matrix(y_test,y_pred_Df)
sns.heatmap(cm/np.sum(cm),annot=True)
plt.ioff()
```




```
print(classification_report(y_test,y_pred_Df))
print("Decision Tree Accuracy:",metrics.accuracy_score(y_test, y_pred_Df))
```

```
              precision    recall  f1-score   support

         0.0       0.98      0.92      0.95        48
         1.0       0.94      0.98      0.96        66

    accuracy                           0.96       114
   macro avg       0.96      0.95      0.95       114
weighted avg       0.96      0.96      0.96       114

Decision Tree Accuracy: 0.956140350877193
```

Figure 11: Breast Cancer Prediction using Decision Tree

- **Algorithm Random Forest**

Random Forest is a method used to create decision trees using the concept of information entropy. ID3 algorithm can be execute using a recursive function. The Random Forest algorithm tries to build a top-down decision tree, starting with which attribute must be checked first and placed as the root. The trick is to evaluate all existing attributes using a statistical measure (which is widely used in information gain) to measure the effectiveness of an attribute in classifying a collection of sample data, the steps for Random Forest work as follows.

1. Calculation of information gained from each attribute using:

$$\boldsymbol{Gain(S,A) = Entropy(S) - \sum_{v=condition(A)} \frac{|S_v|}{S} Entropy(S_v)}$$

$$\boldsymbol{Entropy(S) = -P^{+}log_2P^{+} - P^{-}log_2}$$

2. Selection of the attribute that has the greatest information gain value.
3. Establishment of a node that contains these attributes.
4. The information gain calculation process will continue to be repeated and carried out until all data is included in the same class. The selected attribute is no longer included in the calculation of the information gain value.

- **Entropy**

An object classified in a tree must be tested for its entropy value. Entropy is a measure of information theory that can determine the characteristics of impurity and homogeneity of a data set. From this entropy value, the information gain value for each attribute is then calculated. where : the sample (data) space used for training $p+$: the number of positive solutions in the sample data for certain criteria $p-$: the number of negative solutions in the sample data for certain criteria From the above equation it can be concluded that the definition of entropy (S) is the amount bits that are estimated to be needed to be able to extract a $class$ $(+ or -)$ from several random data in a sample space S. Entropy can be said to be the bits needed to declare a class. The smaller the entropy value, the better it is used in extracting a class. The code length for optimally representing information is log_2P^{+} bits for messages that have probability p. So the number of bits expected to extract in a class is $-P^{+}log_2P^{+} - P^{-}log_2P^{-}$.

- **Information Gain**

In Random Forest, information gain is used to measure how well an attribute separates the training example into the target class. The attribute with the highest information will be selected. To define gain, this scheme first employs an idea from information theory called entropy. Entropy measures the amount of information contained in an attribute. The purpose of measuring the value of information gain is to select attributes that will be used as branches in the formation of a decision tree. Choose the attribute that has the greatest information gain value.

The Random Forest algorithm stops if the attributes or variables that are considered perfect classify training sets or recursively operate on n values, where n is the number of possible values of something to get the best attribute. Mathematically, the information gain of an attribute A is written as follows.

$Information\ Gain = \sum value(A)\ |Sv|\ |S|\ Gain(S,A) = Entropy(S) - Entropy(SV)$ where:

A = attribute

V = a possible value for attribute A

Value (A) = possible set for attribute A

|Sv| = number of samples for the value v

|S| = total number of data samples Entropy

(Sv) = entropy for samples that have a value of v

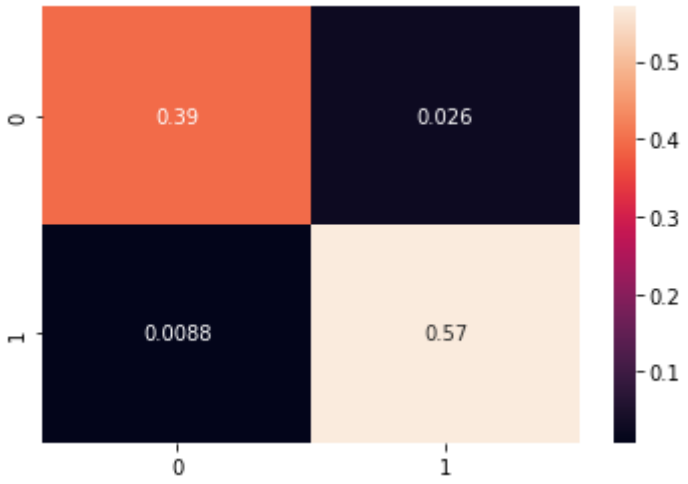
- ***Random Forest algorithm can be summarized as follows***:

1. *Select the best attribute.*
2. *Place a branch for each value of the attribute.*
3. *Divide the instances into subsets, one for each value.*
4. *Repeat the process for each branch using the appropriate subset.*
5. *If the instances of a branch are of the same class, the process terminates for that branch.*

- **Detailed Algorithm**

$RF(Examples,\ Class,\ Attributes).$
$Examples: learning\ \ examples.$
$Class:\ \ Attribute\ \ to\ \ predict\ \ by\ the\ \ tree.$
$Attributes: list\ \ of\ \ attributes\ \ to\ \ check\ \ for\ \ the\ \ tree.$
$Begin$
$Create\ \ a\ \ root\ for\ \ the\ \ tree$
$if\ \ attributes\ \ =\ \ empty$
$return\ \ a\ simple\ \ node\ \ with\ an\ \ error\ \ value.$
$if\ attributes\ \ consist\ of\ \ records\ with\ \ the\ same\ \ rank$
$return\ \ a\ node\ \ with\ the\ \ same\ \ rank.$
$if\ \ examples\ \ =\ \ empty$
$return\ \ a\ node\ \ with\ the\ \ majority\ \ rank\ value.$
$else:$
$A\ is\ \ the\ attribute\ \ of\ \ \ attributes\ \ that\ best\ \ \ classifies\ \ \ examples$
$The\ \ decision\ \ \ attribute\ \ for\ \ root\ is\ A$
$For\ \ \ each\ \ \ possible\ \ \ value\ \ \ vi\ of\ A\{$
$Add\ \ \ a\ branch\ \ \ to\ root\ \ \ with\ the\ \ \ test\ A\ =\ \ vi$
$Examples\ \ \ vi\ \ \ is\ \ \ the\ subset\ \ \ of\ \ \ \ examples\ \ \ with\ value\ \ vi\ for\ A$
$if\ \ examples\ vi\ \ =\ \ empty$
$then\ \ \ add\ \ \ a\ node\ \ \ (n,l)\ from\ \ \ the\ \ \ created\ branch.$
$else\ add\ \ subtree\ \ to\ \ \ created\ branch$
$RF\ \ (examples\ \ vi,\ \ \ class,\ \ \ Attributes - \{A\})$

```
from sklearn.ensemble import RandomForestClassifier

#Create a Gaussian Classifier
clf=RandomForestClassifier(n_estimators=50)

#Train the model using the training sets y_pred=clf.predict(X_test)
clf.fit(x_train,y_train)

# prediction on test set
y_pred_RF=clf.predict(x_test)

#Import scikit-learn metrics module for accuracy calculation
from sklearn import metrics
# Model Accuracy, how often is the classifier correct?
from sklearn import metrics #Import scikit-learn metrics module for accuracy calculation

cm=confusion_matrix(y_test,y_pred_RF)
sns.heatmap(cm/np.sum(cm),annot=True)
plt.ioff()
print("Accuracy:",metrics.accuracy_score(y_test, y_pred_RF))

print(classification_report(y_test,y_pred_RF))
```



Accuracy: 0.9649122807017544

	precision	recall	f1-score	support
0.0	0.98	0.94	0.96	48
1.0	0.96	0.98	0.97	66
accuracy			0.96	114
macro avg	0.97	0.96	0.96	114
weighted avg	0.97	0.96	0.96	114

Figure 13: Breast Cancer Prediction using Random Forest

- **Logistic Regression Algorithm:**

Logistic Regression is used to answer the question of the probability of a predictable variable with its independent variable. Case in point for logistic regression, for example, the author wants to determine the probability of pollution by looking at some time series fetched from the meteorological department, the pollution level (high or low). In the case of problems the independent variable is a mixture of continuous variables and categories, so it does not need to be considered normality of data on the independent variable. So logistic regression is used if accepting a multivariate normal distribution is not approved. Logistic regression. Opportunities and probabilities provide the same information but in different forms. From these two different forms, opportunities can be changed into probability or vice versa, namely in the following ways:

$$P(S|B) = \frac{odds(S|B)}{1 + odds\ (S|B)}\ and\ odds(S|B) = \frac{P(S|B)}{1 + P\ (S|B)}$$

Calculation of odds above can be its natural log value is calculated as follows:

$$Ln[odds(S|B) = Ln\left[\sum odds(S|B)\right)]\ \ and\ Ln[odds(S|K) = Ln\left[odds(S|K)\right)]^2$$

These two equations can be combined into the equation below to give log odds as a function of class quota size study scheme as (benign and malignant):

$$Ln[odds(S|\text{benign}) = Ln[odds(S|K) = Ln\left[odds(S|B)\right)]\text{malignant}^2$$

Where benign = 1 if the class A and malignant = 0 is class B, if the class benign is less. So it is clear that the log of odds is a linear function of the benign independent variable and can be interpreted as a coefficient on regression analysis. A sign of a positive benign coefficient means the log of odds will increase if malignant increases, where the log of the Maximum Likelihood estimations of successful higher than the less malignant class. The logistic regression equation for k independent variables can be stated as follows:

1. The probability that the study benign class will be successful is P (S) = 5/14 = 0.36 2. The probability that the study class malignant will be successful and the class Less (B) is: P (S | B) = 2 / 5 = 0.400 3.
2. The probability that the study program class benign will be successful and the small study program class benign (K) is: P (S | K) = 3/9 = 0.333 Probability is sometimes expressed in odds terms can be calculated as follows:
3. 1. Odds for a class quota of the study program will be successful is odds (S) = 5/5 = 1 which means the odds of a class benign will be successful or unsuccessful are the same or odds 1 against 1 2. Odds for a class quota of study program will be successful and large class quota is odds (S | B) = 2/3 = 0.667 which means the odds of large class quota that will be successful are 2 to 3 or 0.667 to 1 3. Odds for a class quota the study program will be successful and the Less class quota is odds (S | K) = 3/6 = 0.5 which means the odds of the small class quota that will be successful are 3 to 6 or 0.5 to 1 Odds and the probability of giving the same information, but in the same form different. From these two different forms, the odds can be changed into probabilities or vice versa, namely in the following ways to calculate the maximum likelihood estimates as under:-

$P(S|B) = \frac{odds(S|B)}{1-P(S|B)} = \frac{0.667}{1+0.6667} = 0.40$ and $odds(S|B) = \frac{P(S|B)}{1+P(S|B)} = \frac{0.40}{1-0.40}$ 0.667 The odds calculation above can be calculated as its natural log value as follows:

$Ln[odds(S|\text{NO2}) = Ln[odds(0.667) = -0.405$ and $Ln[odds(S|\text{NO2}) = Ln(0.5) = -0.693$ These two equations can be combined into the equation below to provide log odds as a function of the study program class quota measure (SIZE): $Ln[\sum odds(S|\text{NO2}) = -0.693 + 0.405$ (MLE – Maximum Likelihood Estimates).

- **Pseudo Code of Logistic Regression**

1. Collect Data.
2. Prepare: Numeric values are needed for a distance calculation. A structured data format is best.
3. Analyze MLE (As Defined in #REF3).
4. Train: We'll spend most of the time training, where we try to find optimal coefficients to classify our data.
5. Test: Classification is quick and easy once the training step is done.

This application needs to get some input data and output structured numeric values. Next, the application applies the simple regression calculation on this input data and determines which class the input data should belong to. The application then takes some action on the calculated class example as below for the method.

#REF3 : logistic_regression MLE <- function(X, Y, alpha, eta, seed = 100, threshold = 100)
# initializations+set.seed(seed)+X$intercept <- 1
theta <- rnorm(ncol(X), sd = 0.5)
delta_theta <- rep(10000, ncol(X))
numIterations <- 15
XY <- apply(X,2,function(x) X * Y)
XY <- sweep(X, 1,Y, "*")
while (any(abs(alpha * delta_theta/(theta + 0.0001)) > eta) & numIterations < threshold
delta_theta <- gradient_descent(X,Y, XY,theta)
theta <- theta - alpha * delta_theta+numIterations <- numIterations

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report , confusion_matrix
```

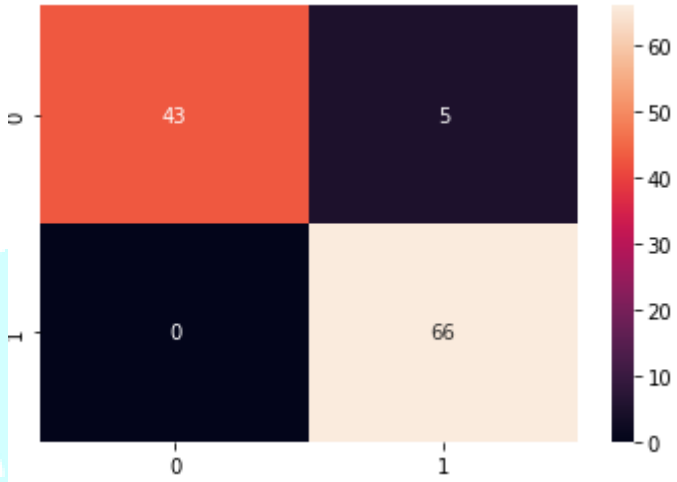
```
logistic_model = LogisticRegression(random_state = 0)
logistic_model.fit(x_train, y_train)
```

```
y_predict =logistic_model.predict(x_test)
y_predict
```

```
array([0., 1., 1., 1., 1., 0., 1., 1., 1., 1., 1., 1., 0., 1., 1., 1., 1.,
       1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 1., 0., 1., 0., 0., 0., 0., 0.,
       1., 1., 0., 1., 1., 0., 1., 1., 1., 0., 1., 1., 0., 0., 1., 0., 1.,
       1., 1., 1., 1., 0., 0., 0., 1., 0., 0., 0., 1., 1., 1., 1., 1., 1.,
       1., 0., 1., 0., 1., 1., 1., 1., 0., 1., 0., 1., 0., 1., 0., 0., 0.,
       1., 0., 1., 0., 0., 0., 0., 1., 1., 0., 0., 1., 1., 1., 1., 1., 0.,
       1., 1., 0., 0., 1., 0., 1., 0., 1., 0., 1., 0.])
```

```
cm = confusion_matrix(y_test,y_predict)
```

```
sns.heatmap(cm ,annot=True)
```




```
print(classification_report(y_test,y_predict))
```

	precision	recall	f1-score	support
0.0	1.00	0.90	0.95	48
1.0	0.93	1.00	0.96	66
accuracy			0.96	114
macro avg	0.96	0.95	0.95	114
weighted avg	0.96	0.96	0.96	114

Figure 14: Breast Cancer Prediction using Logistic Regression

V. CONCLUSION

The results showed that the applications of ML to diagnose breast cancer can produce predictive decisions based on two possibilities, namely living cells in malignant or benign conditions. This application can be used by laboratory personnel to make a diagnosis based on the FNA results obtained from a biopsy examination. The use of the SVM, Logistic Regression Decision Tree and Random Forest algorithm ins ML applications can increase the accuracy of the diagnostic results to a higher level, because the more data that is complete and precise, the more accurate the results of the diagnosis will be. Based on the research that has been done, several important things that need to be considered in building ML applications are as follows, (1) datasets and (2) algorithms for finding data patterns and determining predictions. A complete and precise dataset greatly influences the prediction results. Choosing the right algorithm can also support the accuracy of the prediction results. In further research, it can be developed using algorithm models and other datasets as trial material to determine the best model for building ML applications.

REFERENCES

[1] Kaur, Gagandeep. (2023). Insights on Machine Learning. 10.55083/Isbn.978-93-94435-08-7.ccedtb062332.

[2] Zhao, CY & Zhang, Haixia & Zhang, Xiaoyun & Liu, M & Hu, Z & Fan, B. (2006). Application of Support Vector Machine (SVM) for Prediction Toxic Activity of Different Data Sets. Toxicology. 217. 105-19. 10.1016/j.tox.2005.08.019.

[3] Jarrah, Mohamad & Khader, Yousef & Alkouri, Osama & Al-Bashaireh, Ahmad & Alhalaiqa, Fadwa & Marzouqi, Ameena & Qaladi, Omar & Alharbi, Abdulhafith & Alshahrani, Yousef & Alqarni, Aidah & Oweis, Arwa. (2023). Medication Adherence and Its Influencing Factors among Patients with Heart Failure: A Cross Sectional Study. Medicina. 59. 960. 10.3390/medicina59050960.

[4] Janardhanan, Padmavathi & Heena, L. & Sabika, Fathima. (2015). Effectiveness of Support Vector Machines in Medical Data mining. Journal of Communications Software and Systems. 11. 25-30. 10.24138/jcomss.v11i1.114.

[5] Poornajaf, Maryam & Yosefi, Sajad. (2023). Improvement of the Performance of Machine Learning Algorithms in Predicting Breast Cancer. Frontiers in Health Informatics. 12. 132. 10.30699/fhi.v12i0.400.

[6] Edriss, E., Ali, E., & Feng, WZ (2016). Breast Cancer Classification using Support Vector Machines and Neural Networks.International Journal of Science and Research (IJSR),5(3), 1–6. https://doi.org/10.21275/v5i3.nov161719

[7] Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis.Procedia Computer Science,83(Fams), 1064–1069. https://doi.org/10.1016/j.procs.2016.04.224

[8] George, Darren & Mallery, Paul. (2021). Logistic Regression. 10.4324/9781003205333-27.

[9] Leonard, Thomas. (2020). Logistic Regression. 10.1201/9781003073109-8.

[10] Zhou, Hong. (2020). Logistic Regression. 10.1007/978-1-4842-5982-5_6.

[11] Zhou, Zhi-Hua. (2021). Decision Trees. 10.1007/978-981-15-1967-3_4.

[12] Jo, Taeho. (2021). Decision Tree. 10.1007/978-3-030-65900-4_7.

[13] Suzuki, Joe. (2020). Decision Trees. 10.1007/978-981-15-7568-6_8.

[14] Sundiman, Didi. (2022). Decision tree regression.

[15] Calhoun, Peter & Su, Xiaogang & Spoon, Kelly & Levine, Richard & Fan, Juanjuan. (2021). Random Forest. 10.1002/9781118445112.stat08287.

[16] Sundiman, Didi. (2022). Random Forest Regression.

[17] Berk, Richard. (2020). Random Forests. 10.1007/978-3-030-40189-4_5.

[18] Chopra, Deepti & Khurana, Roopal. (2023). Support Vector Machine. 10.2174/9789815124422123010006.

[19] Machine, Vector & Fan, Tan. (2023). A Tutorial on Support Vector Machine.

[20] Murphy, Andrew & Moore, Candace. (2019). Confusion matrix. 10.53347/rID-68080.