**Building a Simple Machine Learning Model on Breast Cancer Prediction**

**Jemimah Beulah Panga**

Dept.of Computer Science and Engineering,

Karunya Institute of Technology and Sciences,

Coimbatore - 641114 , India

**Abstract**

Breast Cancer is the most leading malignancy affecting 2.1 million women each year which leads to the greatest number of deaths among women. Early treatment not only helps to cure cancer but also helps in prevention of its recurrence. And hence this system mainly focuses on prediction of breast cancer where it uses different machine learning algorithms for creating models like decision tree, logistic regression, random forest and svm which are applied on pre-processed data which suspects greater accuracy for prediction. Amongst all the models, Random Forest Classification leads to best accuracy with 98.57%.These techniques are coded in python by importing numpy, pandas, sklearn and seaborn libraries.

**Keywords** : Breast cancer prediction , Cancer dataset, Logistic regression, Decision Tree, Random Forest Classification, Support vector machine, Machine learning, Numpy, Pandas.

**Introduction**

According to the World health organization, Breast cancer is the most frequent cancer among women and it is the second dangerous cancer after lung cancer. In 2018, from the research it is estimated that a total 627,000 women lost their life due to breast cancer that is 15% of all cancer deaths among women. In case of any symptom, people can visit an oncologist. Doctors can easily identify breast cancer by using Breast ultrasound, Diagnostic mammogram, Magnetic resonance imaging (MRI), Biopsy. Based on these test results, doctors may recommend further tests or therapy. Early detection is very crucial in breast cancer. If chances of cancer are predicted at an early stage the survivability chances of the patient may increase. An alternate way to identify breast cancer is using machine learning algorithms for prediction of abnormal tumors. Thus,the research is carried out for the proper diagnosis and categorization of patients into malignant and benign groups.

Three types of tumors are as follows:-

a. Benign tumors are not cancerous, they cannot spread or they can grow very slowly. And if doctors remove them, then they cannot cause any harm to the human body.

b. In Premalignant tumors the cells are not cancerous but they have potential to become malignant.

c. Malignant cells are cancerous and they can spread rapidly in the body.

In machine learning, cancer classification can be done using benign or malignant cells and could significantly improve our ability for prognosis in cancer patients, poor progress has been made for their application in the clinics. However, before gene expression profiling can be used in clinical practice, studies with larger data samples and more adequate validation are needed. And thus, our aim is to develop a prediction system that can predict chances of breast cancer on a huge amount of data.

**Related Works**

With the evolution of medical research, numerous new systems have been developed for the detection of breast cancer. The research associated with this area is outlined in brief as follows. Sakri et al. [[1](https://ieeexplore.ieee.org/document/8371664)] focused on the enhancement of the accuracy value using a feature selection algorithm named as particle swarm optimization (PSO) along with machine learning algorithms K-NNs, Naive Bayes (NB) and reduced error pruning (REP) tree. Their work perspective holds the Saudi Arabian women’s breast cancer problem, and according to their report, it is one of the major problems in Saudi Arabia. Their reports suggest that women with age range greater than 46 are the main victim of this malicious disease. Holding this sentiment, authors of [[1](https://ieeexplore.ieee.org/document/8371664)] implemented five phase-based data analysis techniques on the WBCD dataset. They reported a comparative analysis between classification without feature selection method and classification with a feature selection method. They have acquired 70%, 76.3%, and 66.3% accuracy for NB, RepTree, and K-NNs, respectively. They used the Weka tool for their data analysis purpose. With PSO implemented, they have found four features that are best for this classification task. For NB, RepTree, and K-NNs with PSO, they obtained 81.3%, 80%, and 75% accuracy values, respectively. Banu and Subramanian [[2](https://www.researchgate.net/publication/344035493_Breast_Cancer_Prediction_A_Comparative_Study_Using_Machine_Learning_Techniques)] have emphasized Naive Bayes techniques on breast cancer prediction and described a comparison study on Tree Augmented Naive Bayes (TAN), Boosted Augmented Naive Bayes (BAN) and Bayes Belief Network (BBN). They used SAS-EM (Statistical Analytical Software Enterprise Miner) for the implementation of the models. The same popular WBCD dataset is used in their work. According to their findings with the help of gradi-ent boosting 91.7%, 91.7%, and 94.11% accuracy have been achieved for BBN, BAN, and TAN, respectively. Hence, their research suggests that TAN is the best classifier among Naive Bayes techniques for this dataset. Chaurasia et al. [[3](https://journals.sagepub.com/doi/abs/10.1177/1748301818756225)] implemented Naive Bayes, RBF network, and J48 Decision Tree techniques on WBCD dataset. For their purpose of research, they used the Waikato Environment for Knowledge Analysis (WEKA) version 3.6.9 as a tool of analysis. For Naive Bayes, they obtained 97.36% accuracy which is greater than 96.77% and 93.41% accuracy values resulted from the RBF network and J48 Decision Tree, respectively. Azar et al. [[4](https://www.springerprofessional.de/en/decision-tree-classifiers-for-automated-medical-diagnosis/11664048)] introduced a method for the prediction of breast cancer using the variants of the decision tree. The modalities used in this technique are the single decision tree (SDT), boosted decision tree (BDT), and decision tree forest (DTF). The decision is taken by training the data set and after that testing. The outcomes presented show that the accuracy obtained by SDT and BDT is 95.75% and 97.07%, respectively, in the training phase which clarifies that BDT performed better than SDT. BDT showed the best performance in terms of sensitivity, and SDT was the best only considering speed. In [[5](https://scholar.google.com/citations?user=vwSPIPQAAAAJ&hl=ja)], the authors demonstrated a procedure for the detec-tion of breast cancer. The experiments that have been done for detecting the disease are discussed here using local linear wavelet neural network (LLWNN), and recursive least square (RLS) to enhance the performance of the system. The LLWNN-RLS is providing the maximum values of average Correct Classification Rate (CCR) 0.897 and 0.972 for 2 and 3 predictors, respectively, with a few calculation times. It also provides the lowest value of minimum description length (MDL) and average squared classification error (ASCE) with much lesser time. Senapati et al. [[6](https://www.researchgate.net/publication/281264041_A_Breast_Cancer_Classifier_Using_a_Neuron_Model_with_Dendritic_Nonlinearity)] proposed a hybrid system for the detection of breast cancer using KPSO and RLS for RBFNN. The centers, as well as variances of RBFNN, are adjusted using K-particle swarm optimization and adjusted using back propagation. The classification accuracy achieved by RBFNN-KPSO and RBFNN-extended Kalman filter is 97.85% and 96.4235%, respectively, whereas the coverage time is 8.38 s and 4.27 s, respectively. Hasan et al. [[7](https://www.semanticscholar.org/paper/Mathematical-model-development-to-detect-breast-Hasan-Islam/742c29c704817adff5ac839fedf82a3e68ec8d45)] developed a mathematical model for the prediction of breast cancer based on the symbolic regression of Multigene Genetic Programming. The ten-fold technique is used to avoid overfitting here. A com-parative study is also illustrated. The stopping criteria for the model were generated but the generation level did not reach zero. The highest accuracy obtained by ANNs is 98.57% whereas the lowest accuracy derived from the RFs and LR is 95.7%. A variant of SVM [[8](https://www.semanticscholar.org/paper/Performance-analysis-of-support-vector-machines-in-Azar-El-Said/6ab057f0467400e2613b72b960ab80f2e84dbb75)] is introduced for the diagnosis of breast cancer. Here six kinds of SVM are explained and used for performance evaluation. The stand-ard SVM results are compared with other types of SVM. Four-fold cross-validation is used for training and testing. The highest accuracy, specificity, and sensitivity achieved by St-SVM are 97.71%, 97.9%, and 97.08%, respectively, in the training phase. The highest accuracy, sensitivity, and specificity obtained by NSVM, LPSVM, SSVM, and LPSVM are 96.5517% 98.2456%, 96.5517%, and 97.1429% individually in the testing phase. The authors in [[9](https://www.dcc.fc.up.pt/~pedroferreira/)] presented an efficient method for the detection of breast cancer by categorizing the features of breast cancer data utilizing the inductive logic programming technique. A comparison study with a propositional classifier is also drawn. Kappa statistics, F-measure, area under the ROC curve, true positive rate, etc. are calculated as a performance measure. The system simulated two platforms named Aleph and WEKA. Jhajharia et al. [[10](http://ieeexplore.ieee.org/abstract/document/7830107/)] appraised variants of decision tree algorithms for the diagnosis of breast cancer. The system used the most common decision tree algorithms named CART and C4.5 which are simulated in the WEKA platform using Matlab and Python. The CART implemented in Python achieved the highest accuracy 97.4% and the highest sensi-tivity 98.9% is obtained in the CART which is implemented in Matlab, and 95.3% specificity is acquired by CART and C4.5, respectively, which are simulated in WEKA. [11] The accuracy obtained by SVM (97.13%) is better than the accuracy obtained by C4.5, Naïve Bayes and k-NN that have an accuracy that varies between 95.12 % and 95.28 %. It can also be easily seen that SVM has the highest value of correctly classified instances and the lower value of incorrectly classified instances than the other classifiers. [[12](https://www.semanticscholar.org/paper/Feature-Selection-with-Fast-Correlation-Based-for-Khourdifi-Bahaj/9f41033b3d9be311c2ce5bdb4eb425697093163d)]The experimental results show that SVM gives the highest accuracy of 97.9% without FCBF but if we apply this method we find that the SVM and MLP show the best results in comparison with other algorithms. The results will help to choose the best learning algorithm machine classification for breast cancer prediction.

**Methods and Materials**

**1.Data Set Description**

The breast cancer dataset was retrieved from the UCI machine learning repository [[13](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic))]. There are 699 instances in this dataset, where the cases are either benign or malignant. For such cases, 458 (65.50%) are benign, and 241 (34.50%) are malignant. The class in the dataset is partitioned into 2 or 4, wherever 2 corresponds to the benign case, and 4 corresponds to the malignant case. The attributes consist in the dataset which is in Table 1 excluding sample code number and class level.The benign cases are identified as a negative class, and the malignant cases are identified as a positive class in our research.

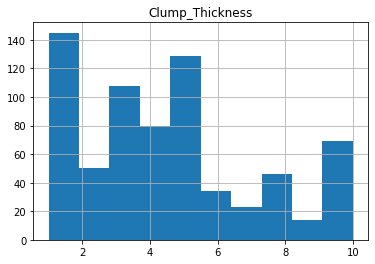
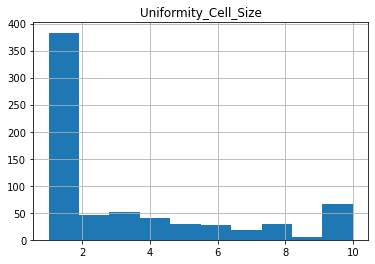
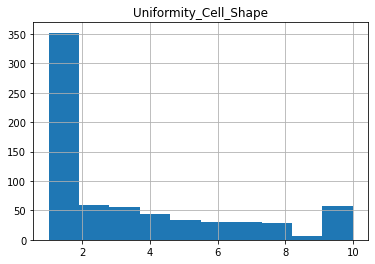
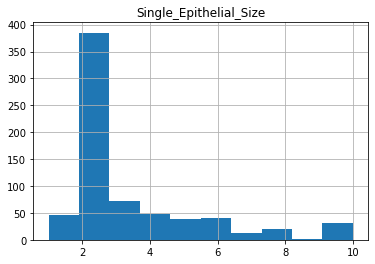
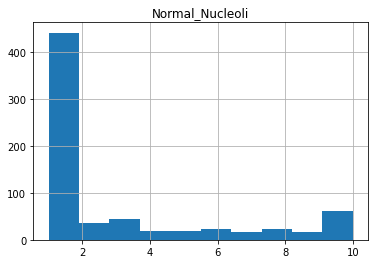
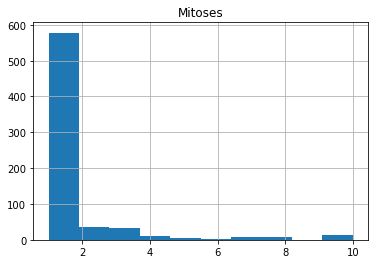
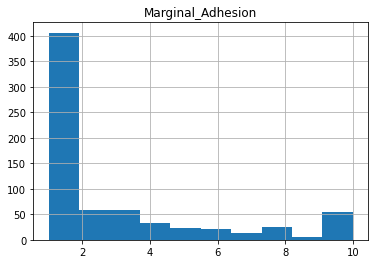
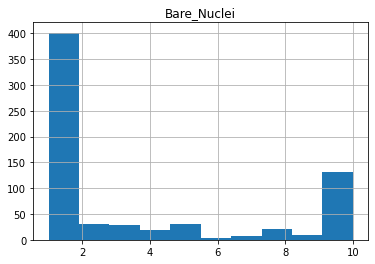
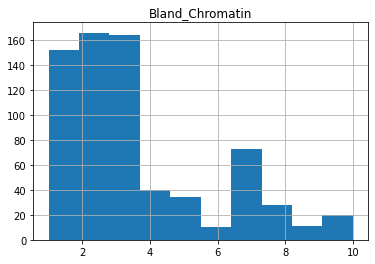
|  |  |  |
| --- | --- | --- |
| Table 1 Attributes of the dataset |  | |
| Attributes | Domain | Symbol |
| Clump thickness | 1–10 | x1 |
| Uniformity of cell size | 1–10 | x2 |
| Uniformity of cell shape | 1–10 | x3 |
| Marginal adhesion | 1–10 | x4 |
| Single epithelial cell size | 1–10 | x5 |
| Bare nuclei | 1–10 | x6 |
| Bland chromatin | 1–10 | x7 |
| Normal nuclei | 1–10 | x8 |
| Mitoses | 1–10 | x9 |

**2.Data Preprocessing**

Data preprocessing is used to complement missing values, identify or remove outliers, and solve self-contradiction. The sample code number is reduced from the dataset as it has no impact on diseases. There are 16 absent values of traits in the dataset. The mean replaces the absent traits for that class. Additionally, the dataset employs random selection to confirm the proper circulation of the data.

**3.Data Visualization**

Visualization of data is an imperative aspect of data science. It helps to understand data and also to explain the data to another person. Here is the visualization of each feature from the dataset.



**4.Training and Testing Phase**

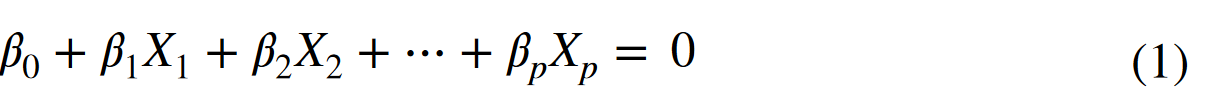
The training phase extracts the features from the dataset and the testing phase is used to determine how the appropriate model behaves for prediction. The dataset is divided into two sections those are the training and testing phases. We have trained 90% of data and tested 10% of data.

**5.Theoretical Considerations**

In the machine learning strategies, the learning procedure can be parted into two principal classifications such as supervised and unsupervised learning. In supervised learning, an arrangement of information cases is utilized to prepare the machine and is marked to give the right outcome. But in case of unsupervised learning, there are no pre-decided informational indexes, no idea of the usual result, which implies that the objective is harder to accomplish.

* **Support Vector Machine**

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they're able to categorize new text. So you're working on a text classification problem The hyperplane that is used in p-dimensions is as follows:



where X1, X2,..., and Xp are the data points in the sample space of p-dimension and β0, β1, β2,..., and βp are the hypo-thetical values.

* **Random Forests**

Random forest classifier is a powerful supervised classification tool. The RF classification is an ensemble method that can be studied as a form of the nearest neighbor predictor. Ensemble learning is the method by which statistical methods like class-sifiers or experts are strategically developed and incorporated to solve a specific problem of computational intelligence. RF generates a forest of classification trees from a given data-set, rather than a single classification tree.The workflow of random forest is given below:

(i) From the training set, pick K data points randomly.

(ii) From these K data points, generate the decision trees.

(iii) From generated trees, choose the number of N-tree and repeat steps (i) and (ii).

(iv) Form the N-tree that predicts the category to which the data points relate for a new data point, and assign the new data point via the category with the highest probability

* **Decision Tree**

Decision tree is the most powerful and popular tool for classification and prediction in machine learning.The Decision trees algorithm consists of two parts: nodes and rules (tests).A Decision tree is like tree structure, where node denotes a test on an attribute,Branch represents an outcome of the test, and each leaf node holds a class label.For detecting breast cancer, it’s leaf nodes are categorised as benign and malignant. And then certain rules are established to check if the tumor is benign or malignant.

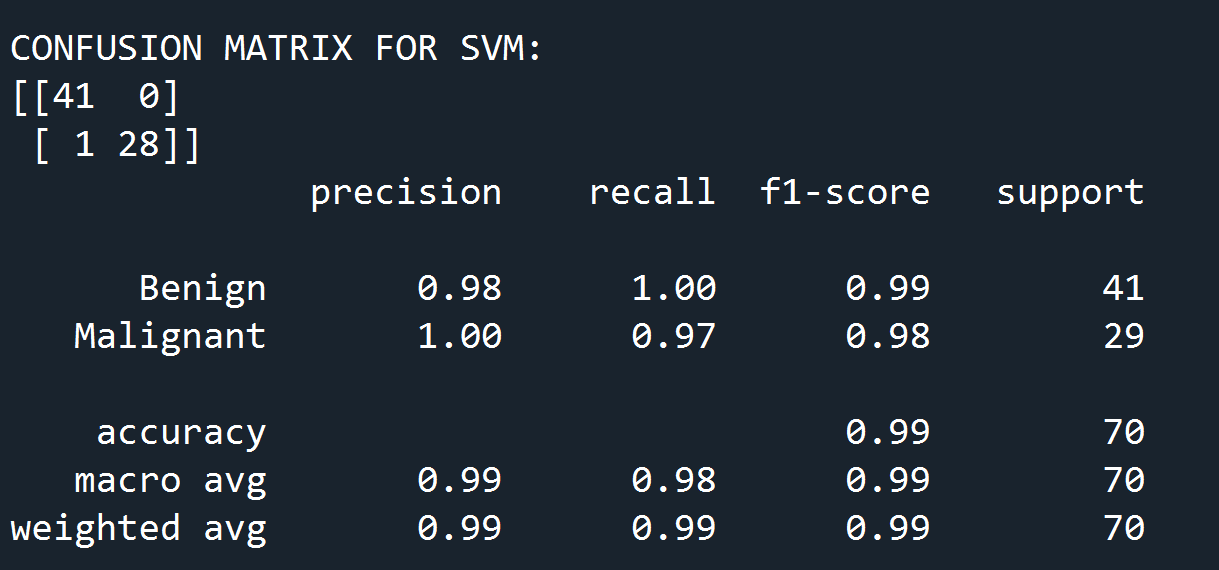
* **Logistic Regression**:

A logistic regression model predicts dependent data variables by analyzing the relationship between one or more existing independent variables.It is generally used when the dependent variable is binary and provides a constant output. Estimated efficiency of the Logistic model will be the 95% which is obviously lower than Random Forest.

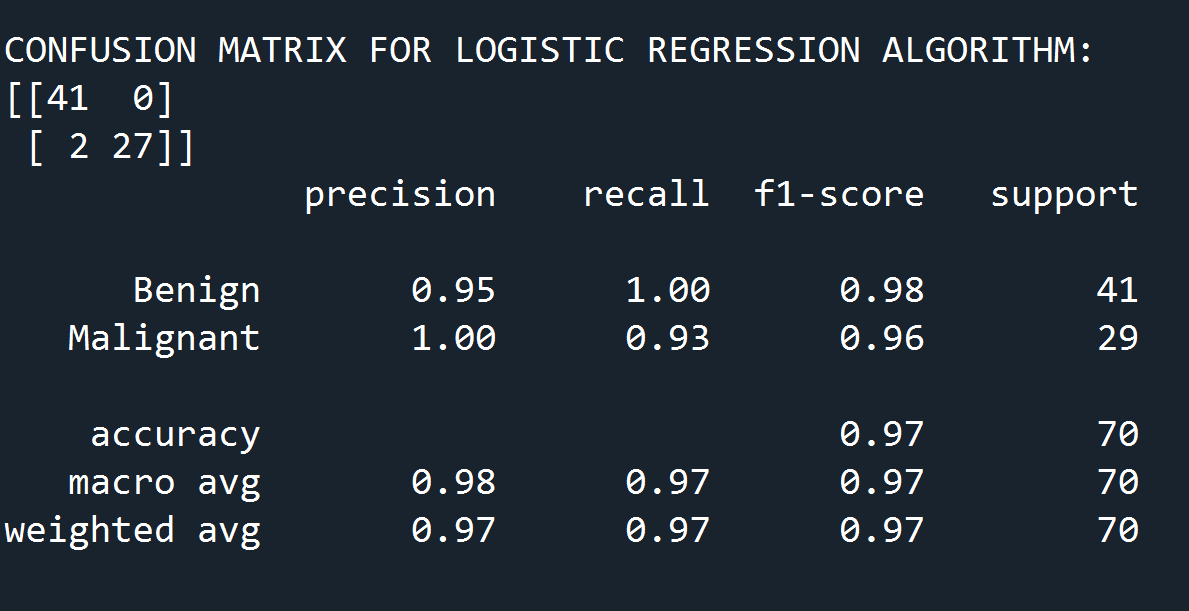
**6. Performance Measure Parameters**

The performance of machine learning techniques is meas-ured with respect to a few performance measure parameters. A confusion matrix including TP, FP, TN, and FN for actual data and predict data is formed to evaluate the parameters. The output of the implication on the dataset is given below:

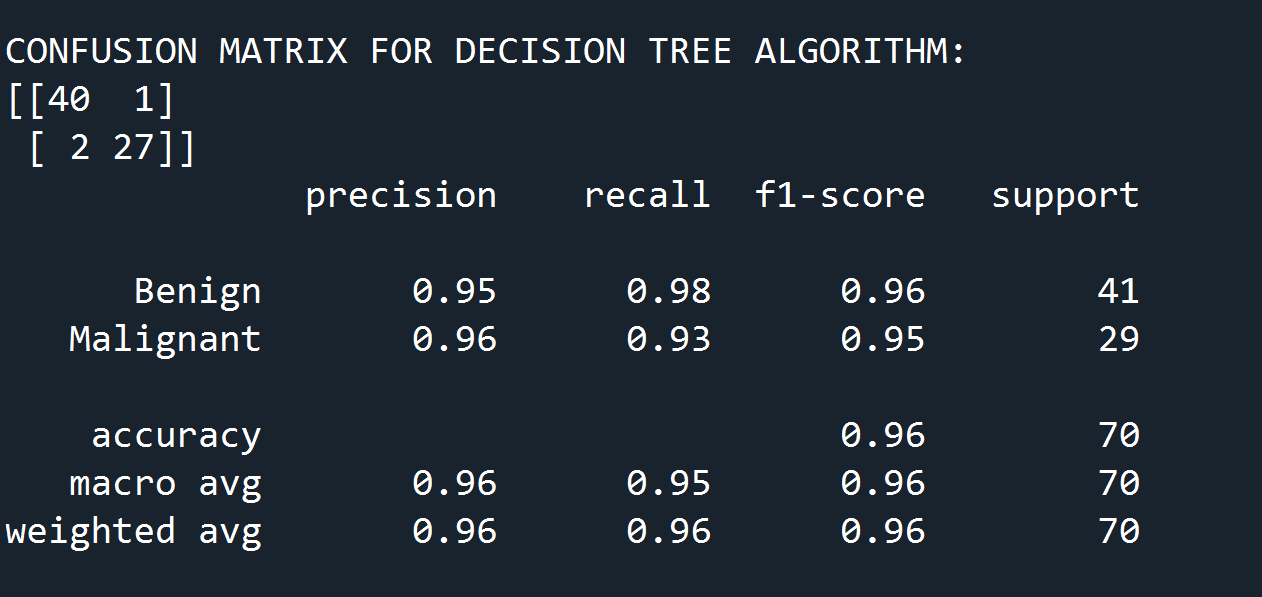
* **SVM**



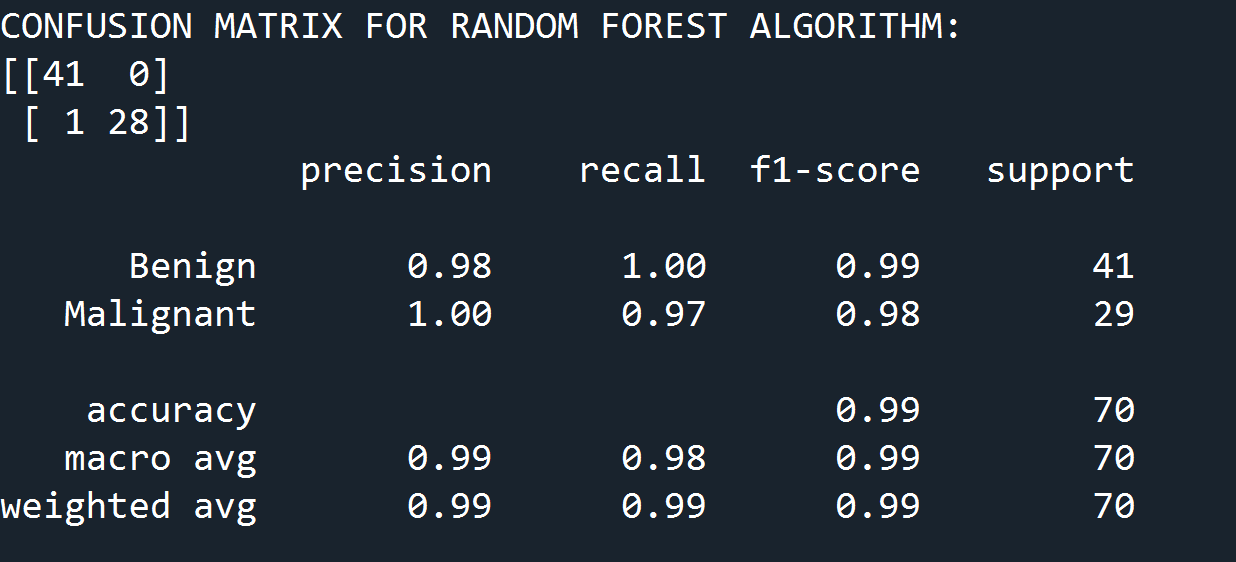
* **LOGISTIC REGRESSION**



* **DECISION TREE**

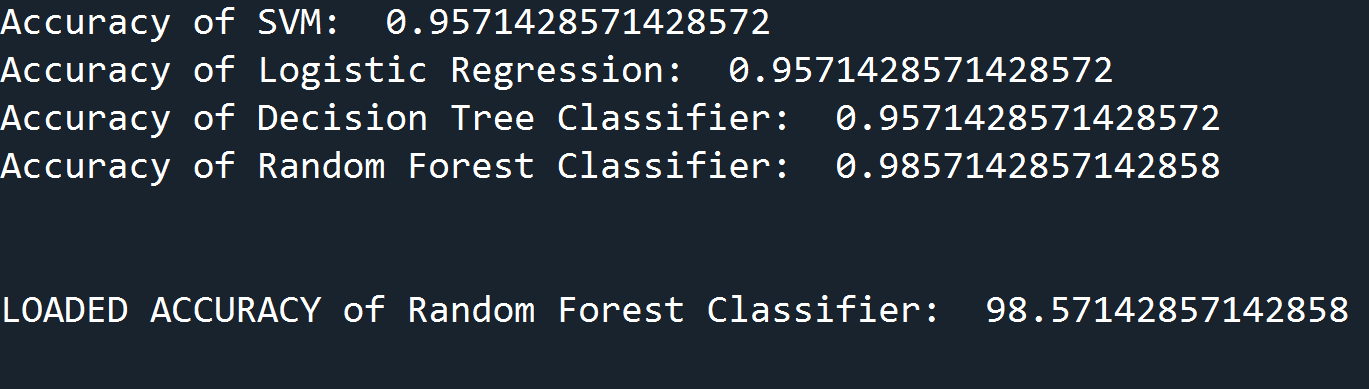


* **RANDOM FOREST**



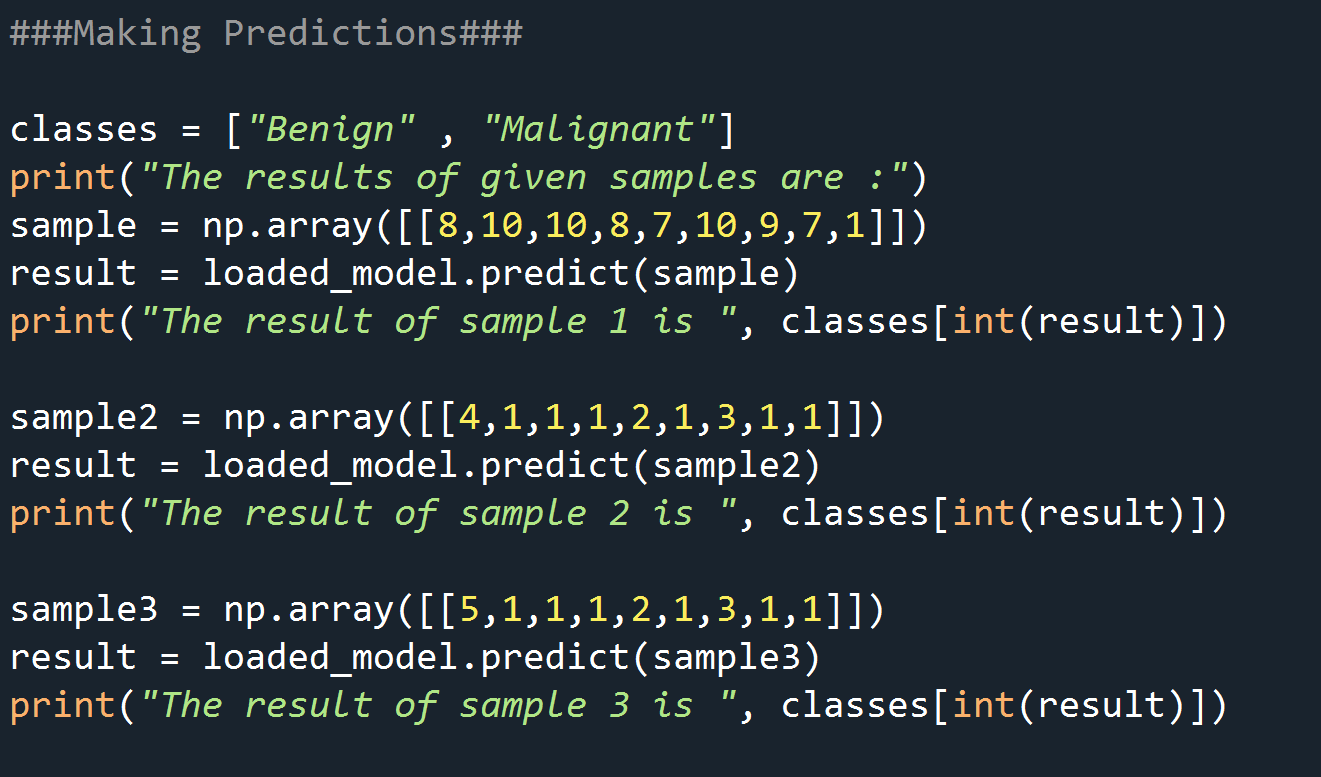
**7. Implementation**

To predict whether a cell is benign or malignant, we have used five machine learning techniques such as SVM, Decision tree, Random Forest Classification, SVM and Logistic Regression individually. Scikit-learn, an open-source machine learning library in Python programming language, is used. Spyder is an open-source web application that permits to develop and share reports that include live code, visualizations, equations, and narrated text which is utilized to fulfill our goal. The confusion matrix is calculated for each technique. From the dataset of 699 instances, we used 629 instances which are 90% of the total data to train for all five techniques. We used 70 instances to test both our trained models. Random Forest Classification outperformed all other machine learning techniques so far we have studied with the highest accuracy of 98.57% whereas SVM, LR and Decision tree achieved the next highest accuracy of 95.7%. In summary, Random Forest Classification was able to show its power in terms of effectiveness and efficiency based on accuracy and recall. The output of the algorithms is shown below:

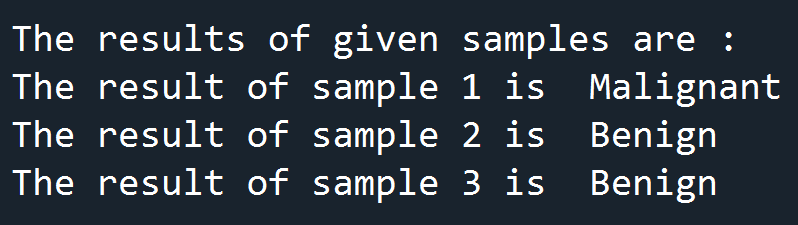


We have given a sample data to check working of the model:

**Sample code:**



**Output:**



So if the given input is larger, it's classified as Malignant cells, else it's classified as Benign cells.

**CONCLUSION**

In this paper, different types of models are reviewed and their accuracies are computed and compared with each other, so that the best cancer prediction model can be used by doctors in real life to identify breast cancer relatively faster than previous methods. Above written study proposed that the Random Forest Classification algorithm is proficiently utilized and efficient for detection of breast cancer as compared to Decision tree, SVM and Logistic Regression algorithms. The analysis of the results signifies that the integration of multidimensional data along with different classification, feature selection and dimensionality reduction techniques can provide auspicious tools for inference in this domain. Here, our main aim is to reduce the error rates with maximum accuracy.

**References**

[1] Sakri SB, Rashid NBA, Zain ZM. Particle swarm optimization feature selection for breast cancer recurrence prediction. IEEE Access. 2018;6:29637–47.

[2] Banu AB, Subramanian PT. Comparison of Bayes classifiers for breast cancer classification. Asian Pac J Cancer Prev (APJCP). 2018;19(10):2917–20.

[3] Chaurasia V, Pal S, Tiwari B. Prediction of benign and malig-nant breast cancer using data mining techniques. J Algorithms Comput Technol. 2018;12(2):119–26

[4] Azar AT, El-Metwally SM. Decision tree classifiers for automated medical diagnosis. Neural Comput Appl. 2012;23(7–8):2387–403

[5] Senapati MR, Mohanty AK, Dash S, Dash PK. Local linear wavelet neural network for breast cancer recognition. Neural Comput Appl. 2013;22(1):125–31

[6] Senapati MR, Panda G, Dash PK. Hybrid approach using KPSO and RLS for RBFNN design for breast cancer detection. Neural Comput Appl. 2014;24(3–4):745–53.

[7] Hasan MK, Islam MM, Hashem MMA (2016) Mathematical model development to detect breast cancer using multigene genetic programming. In: Proc. 5th International Conference on Informatics, Electronics and Vision (ICIEV), Dhaka, 2016, pp. 574–579.

[8] Azar AT, El-Said SA. Performance analysis of support vector machines classifiers in breast cancer mammography recognition. Neural Comput Appl. 2013;24(5):1163–77.

[9] Ferreira P, Dutra I, Salvini R, Burnside E. Interpretable models to predict Breast Cancer. In: Proc. IEEE International Confer-ence on Bioinformatics and Biomedicine (BIBM), Shenzhen, 2016, pp. 1507–1511.

[10] Jhajharia S, Verma S, Kumar R. A cross-platform evaluation of various decision tree algorithms for prognostic analysis of breast cancer data. In: Proc. International Conference on Inven-tive Computation Technologies (ICICT), Coimbatore, 2016, pp. 1–7

[11] R. M.Mohana, R. Delhi Howsalya Devi, Anita Bai, “Lung Cancer Detection using Nearest Neighbour Classifier”, International Journal of Recent Technology and Engineering (IJRTE), Volume-8, Issue-2S11, September 2019

[12] Y. Khourdifiand M. Bahaj,"Feature Selection with Fast Correlation-Based Filter for Breast Cancer Prediction and Classification Using Machine Learning Algorithms,"2018 International Symposium on Advanced Electrical and Communication Technologies (ISAECT), Rabat, Morocco, 2018

[13] Breast Cancer Wisconsin (Original) Data Set : [Breast Cancer Wisconsin (Diagnostic) Data Set](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic))