

BREAST CANCER PREDICTION WITH **MACHINE LEARNING**

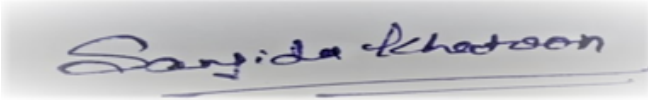


Submitted to the faculty of the Master of Science in Data Science Department of the Maulana Abul Kalam Azad University of Technology (MAKAUT) in partial fulfillment of the requirements of the degree of

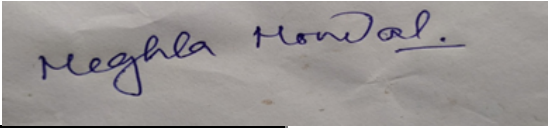
Master of Science
in
Data Science

Declaration

We declare that the work contained in this thesis is our own, except where explicitly stated otherwise. Where material has been used from other sources it has been properly acknowledged. In addition this work has not been submitted to obtain another degree or professional qualification.

Signed: 

Name: Sanjida Khatoon(M.Sc. data science)

Signed: 

Name: Meghla Mondal(M.Sc data science)

ACKNOWLEDGEMENT

First and foremost, thanks to GOD for blessing us with His grace to accomplish this thesis. We want to acknowledge the great encouragement and guidelines provided by our project supervisor Prof Swapan Das, who was very kind to devote his precious time and guide us in every aspect. Also we would like to thank all the Faculty members of the Department of Data Science, from whom we learned a lot and gained a lot of valuable knowledge.

Last but not the least we are also grateful to the authors of all the books and web pages we referred to, during the course of writing this report

BREAST CANCER PREDICTION WITH MACHINE LEARNING

Project advisor: Asst.Prof. Swapan Das
Professor of Computing & Analytics

Project Member: Meghla Mondal

(Reg. No:-23498119003)

Sanjida Khatoon

(Reg. No:-182341810027)



NSHM Knowledge Campus

**124, B. L. Saha Road, Kolkata -
700053, INDIA** establishes a
compulsion in cancer research. Many
research teams from the biomedical and
bioinformatics fields

content

1. Abstraction
2. Introduction
3. Methodology
 - dataset
 - Machine Intelligence Libraries
 - Data pre-processing
 - Data Visualization
 - Methods Used
 - ❖ logistic Regression
 - ❖ Support Vector Classifier
 - ❖ K-Nearest Neighbor Classifier
 - ❖ Random Forest
 - ❖ Adaboost Classifier
 - ❖ Gradient Boosting Classifier
 - ❖ Extreme Gradient Boosting Classifier
 - Parameter used
 - Train and Test
 - confusion Matrix
 - Cross Validation
 - Classification Method
4. Conclusion
5. Future Scope
6. References
7. Appendix

1. Abstraction:

In order to support and supervise patients, the key detection and estimation of cancer type should establish a compulsion in cancer research. Many research teams from the biomedical and bioinformatics fields have been advised to learn and evaluate the use of machine learning (ML) methods because of the relevance of classifying cancer patients into high or low risk clusters. To predict breast cancer, the logistic regression method and many classifiers have been proposed to generate profound predictions about breast cancer data in a new environment. This paper discusses the various approaches to data mining using classification to create deep predictions that can be applied to Breast Cancer data. In addition, by testing datasets on different classifiers, this analysis predicts the best model that delivers high efficiency. In this paper, the UCI machine learning repository has 699 instances with 11 attributes collected from the Breast cancer dataset. First, the data set is pre-processed, visualized and fed to different classifiers such as Logistic Regression, Support Vector Classifier, K-Nearest Neighbour, Decision Tree and Random Forest. 10-fold cross validation is implemented and testing is carried out in order to create and validate new models. Effective analysis shows that Logistic Regression generates the deep predictions of all classifiers and obtains the best model delivering strong and precise outcomes, followed by other methods: Support Vector Classifier, K-Nearest Neighbour, Decision Tree and Random Forest. Most models were less reliable compared to the approach of logistic regression.

2. Introduction

Cancer refers to the development of abnormal cells that divide uncontrollably and destroy normal body tissue.

There are different types of cancer like :

- lung cancer,
- kidney cancer,
- breast cancer,
- bladder cancer,
- colorectal cancer

and many more. Among these, breast cancer is one of the most widely spread diseases in the world. Breast cancer is the abnormal growth of breast cells in women and rarely to men. The cause of breast cancer is multifactorial. Several risk factors for breast cancer have been known nowadays. The risk factors are classified into non modifiable risk factors: age, sex, genetic factors (5-7%), family history of breast cancer, history of previous breast cancer and proliferative breast disease; modifiable risk factors: menstrual and reproductive factors, radiation exposure, hormone replacement therapy, alcohol and high fat diet; and environmental factors: organochlorine exposure,electromagnetic field and smoking. This tumor can be classified as *benign* and *malignant*. *Benign* is *noncancerous*, it does not invade nearby tissue or spread to other parts of the body but *malignant* is *cancerous* that can invade and kill nearby tissue and spread to other parts of the body. Breast cancer occurs when a malignant tumor(mass of tissue) occurs in the breast. In this project, benign is denoted by 0 and malignant is denoted by 1.It is very hard and time taking task for the doctors to diagnose breast cancer in a patient at commencing

stage. But it becomes easy by using application of artificial intelligence, machine learning helps in prediction as well as detection of breast cancer effectively and accurately. Machine learning gives the system the capacity to learn automatically and improve for experience. It uses different algorithms for prediction and computation of accuracy. By using effective models like logistic regression, SVC, KNN, decision tree and random forest which give high accuracy assists in prediction of breast cancer. The highly effective model is judged on 10 fold cross validation of testing. The validation is done on the bases of these parameters: accuracy, RMSE Error, sensitivity, specificity, FMeasure, ROC Curve Area and Kappa statistics and time taken to build the model .

3. Methodology

A. Dataset

A dataset is an intrinsic requirement to produce a robust method for the detection of breast cancer. It is very difficult to collect dataset due to unavailability of samples and privilege of the patients. In this project, we have collected the dataset from the UCI machine learning repository . The data is retrieved from Wisconsin Breast Cancer Database, source - University of Wisconsin Hospital Madison, Wisconsin USA. The data contains 699 instances and 11 attributes (as of 15th July, 1992). The dataset contains 458- Benign and 241- Malignant, benign is denoted by 0 and malignant is denoted by 1. The data features are as follows: sample code number, clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses, class.

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	conc
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	

Fig 1 : sample of the dataset

B. Machine Intelligence Libraries

The libraries used in this article are: matplotlib and seaborn for data visualization, pandas for data manipulation or analysis and numpy for numerical computation. To implement all machine learning algorithms scikit-learn is used.

```
#import libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings("ignore")
plt.style.use('ggplot')
```

Fig 2 : Importing libraries

C. Data Pre-processing

To make a viable analysis, the raw data is cleaned such that more than one machine learning algorithm is executed in one dataset to pick the right one out of them. By deleting comma, large space and unused attributes (sample 'id' is removed), the raw data is washed. If the dataset includes null values, the mean value of the row or column is replaced, or if several null values exist, the row or column is deleted. The dependent attribute is binarized, which transforms the values of the attribute type to binary values, so that it is possible for breast cancer diagnosis.

D. Data Visualization

Data visualization provides an important suite of tools for gaining a qualitative understanding. This helps in exploring and getting to know the dataset and helps in identifying patterns, corrupt data, outliers and much more. Data visualization is used to express and demonstrate key relationships in plots and charts. In this article, data visualization is used to know about patients having benign and malignant in the breast cancer dataset. Fig. 3 is a bar graph between class and count which shows that benign patients are more than malignant patients, there are 458 benign patients and 241 malignant patients. Fig. 4 is a scatter plot graph between perimeter mean and area mean which shows that perimeter mean of malignant cells is increasing with the increased area mean.

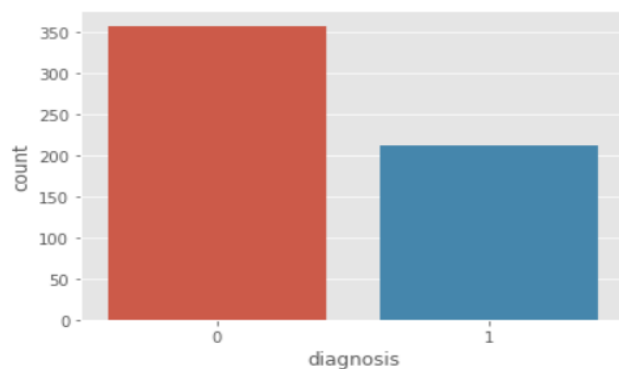


Fig. 3. Bar graph to count cell

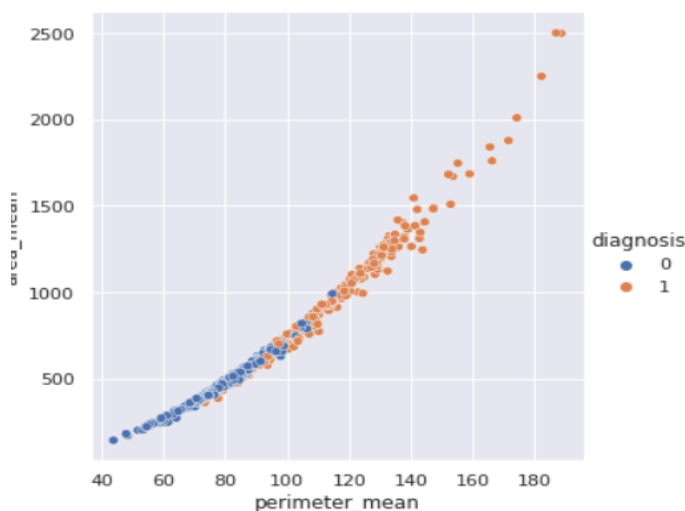


Fig. 4. Scatter plot

E. Methods Used

1) Logistic Regression:

It is a supervised learning algorithm used to predict the probability of a target variable. The target variable or dependent variable should have two possible classes like in this article, there are two classes:

Benign and Malignant.

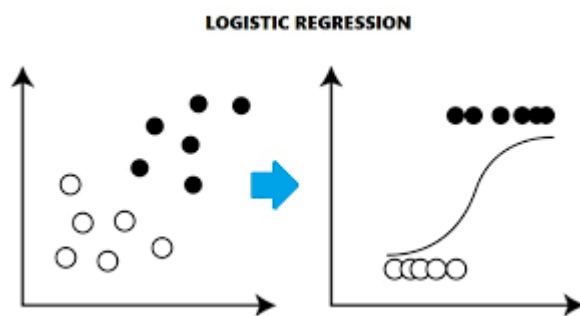


Fig. 5 This figure shows how logistic regression classifies two classes

2) Support Vector Classifier:

The aim of SVC (Support Vector Classifier) is to respond to the information that we have, returning our data to a best match hyperplane that separates or categorizes it. From there we can then feed in any functionality to the classifier after getting the hyperplane to see what the predicted class is.

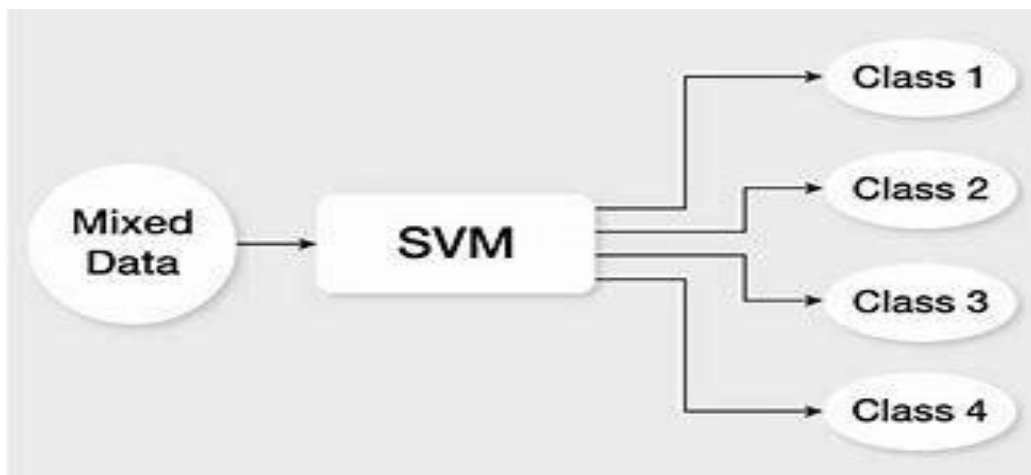


Fig. 6 this image shows how svm classifies.

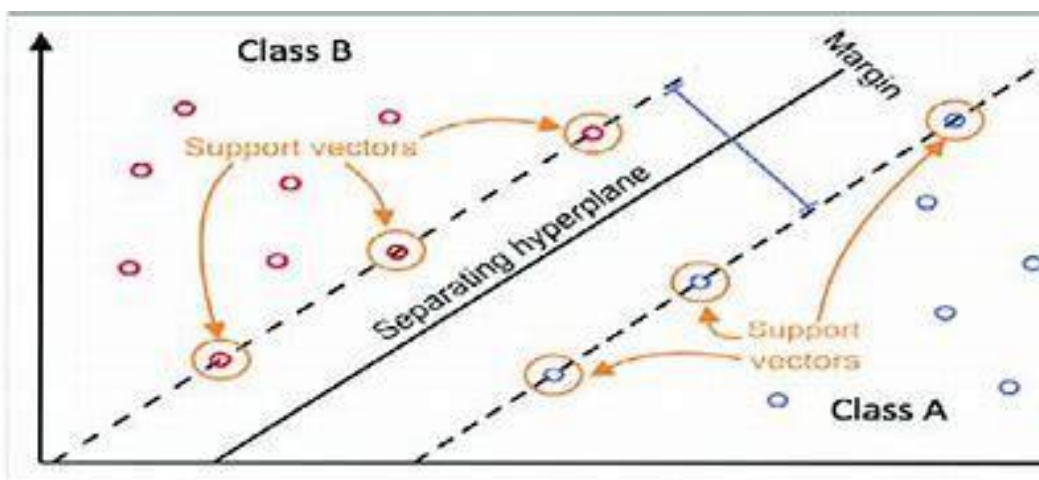


Fig.7 Graphical representation of SVC

3) K-Nearest Neighbor:

Also known as KNN, K-Nearest Neighbor is a supervised learning algorithm that can be used both for regression and classification issues. But in machine learning, KNN is widely used for classification problems. KNN operates under a theory that implies that each data point that falls next to each other is of the same class.

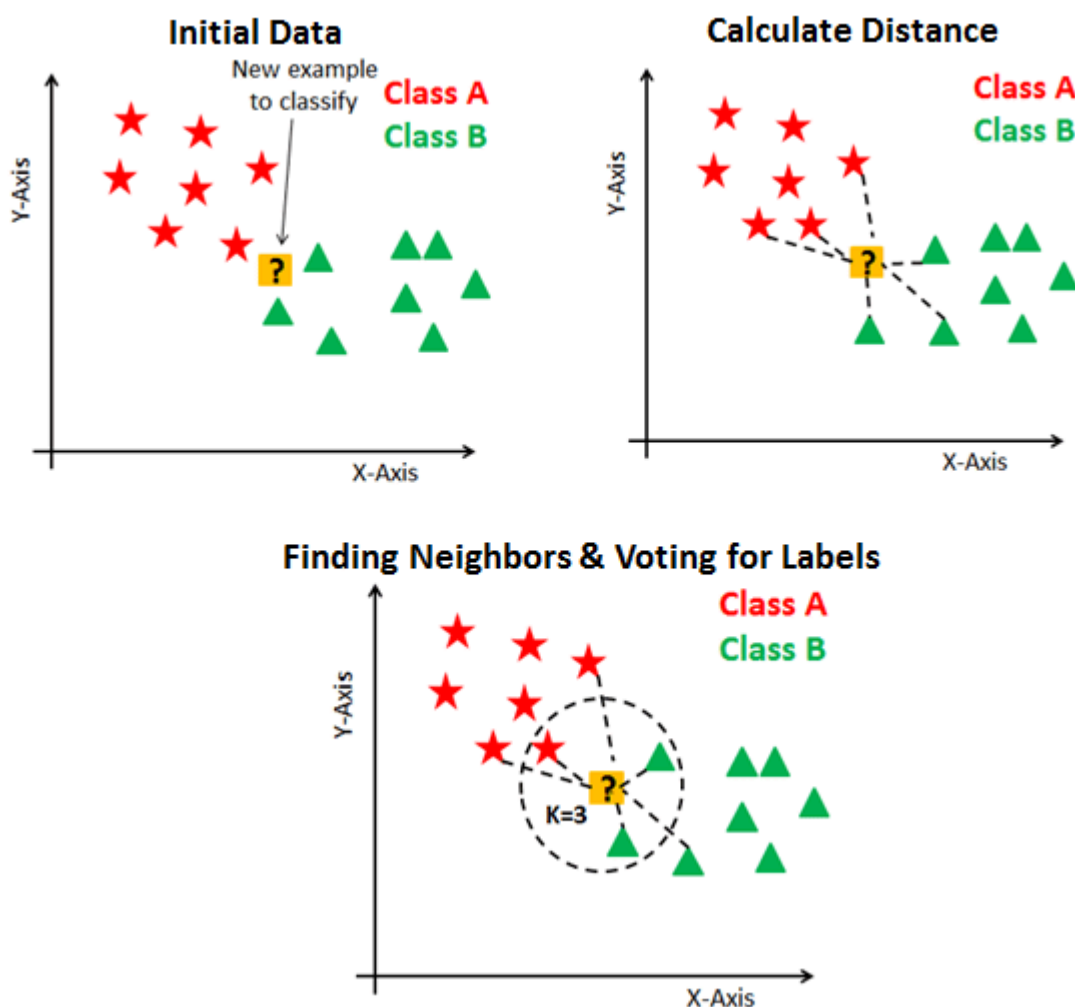


Fig.8. KNN classification pictorial representation

4) Random Forest:

Random forests or random decision forests are an ensemble learning tool for classification, regression and other tasks that function by creating a number of decision trees at training time and generating the class that is the class mode (classification) or mean/average predictor (regression) of the individual trees.

5) Adaboost Classifier

Ada-boost or Adaptive Boosting is an ensemble boosting classifier. AdaBoost is an iterative ensemble method.. The basic concept behind Adaboost is to set the weights of classifiers and train the data sample in each iteration such that it ensures the accurate predictions of unusual observations. Any machine learning algorithm can be used as a base classifier if it accepts weights on the training set. Adaboost should meet two conditions:

1. The classifier should be trained interactively on various weighted training examples.
2. In each iteration, it tries to provide an excellent fit for these examples by minimizing training error.

6) Gradient Boosting Classifier

The Gradient Boosting Classifier is an additive ensemble of a base model whose error is corrected in successive iterations (or stages) by the addition of Regression Trees which correct the residuals (the error of the previous stage)

7) Extreme Gradient Boosting Classifier

The XGBoost stands for extreme Gradient Boosting, which is a boosting algorithm based on gradient boosted decision trees algorithm. XGBoost applies a better regularization technique to reduce overfitting, and it is one of the differences from the gradient boosting.

F. Parameters used

Cancer is extracted by analyzing the results obtained and are evaluated by considering various parameters and are

Explained in detail here.

1) The percentage of test instances that are correctly classified on a given test set is determined as the accuracy of a classifier.

Accuracy = (Number of correctly classified instances by rules ÷ Total number of instances by rules) * 100

2) Sensitivity and Specificity is calculated from the Confusion Matrix obtained in the model.

3) Sensitivity is the proposition of the positive instances that are correctly identified

$TP = (TP / TP + FN) * 100$

4) Specificity is the proposition of the negative instances that are correctly identified

$TN = (TN / TN + FP) * 100$

5) RMSE Root Mean Square Error is a measure of the difference between values predicted by a model and the values actually observed.

6) F-measure, ROC curve area and Kappa Statistics are also calculated using Confusion Matrix with the help of WEKA tool.

G. Train and Test:

It is a technique for evaluating the performance of a machine learning algorithm. It is the method of measuring the accuracy of the models. Here firstly, the dataset is split into two parts: features (except class) is stored in X and class is stored in Y. Then the data is divided to train and test: 80% of data is trained and 20% of data is tested. The 80% of data is fed to the machine learning models and accuracy is calculated using the 20% of the data. In this project, 80% of the data that is 558 instances with 9 attributes from X and 558 instances with 1 attribute from Y is fed to the models: logistic regression, support vector classifier, KNN, decision tree and random forest. Then it is tested to calculate the accuracy with 20% of data that is 140 instances and 9 attributes from X and 140 instances with 1 attribute from Y.

Support Vector Classifiers gave the highest accuracy among the models, with 98.2% accuracy.

	Model	Score
1	Logistic regression	0.9824561403508771
2	Support Vector Classifier	0.9766081871345029
3	K-NN classifier	93.672514619883
4	Random Forest	95.90643274853801
5	Adaboost Classifier	93.5672514619883
6	Gradient Boosting Classifier	97.66081871345029
7	Extreme Gradient Boosting Classifier	98.24561403508771

TABLE I - shows the accuracy obtained by the models.

H. Confusion Matrix

A comparison is drawn between the actual class labels and the predicted class labels based on the class labels by the classifiers. The following describes the case when we deal with two-class classification problems. The generated confusion matrix is 2×2 matrixes . The following figure shows the confusion matrix of the logistic regression model in heatmap.

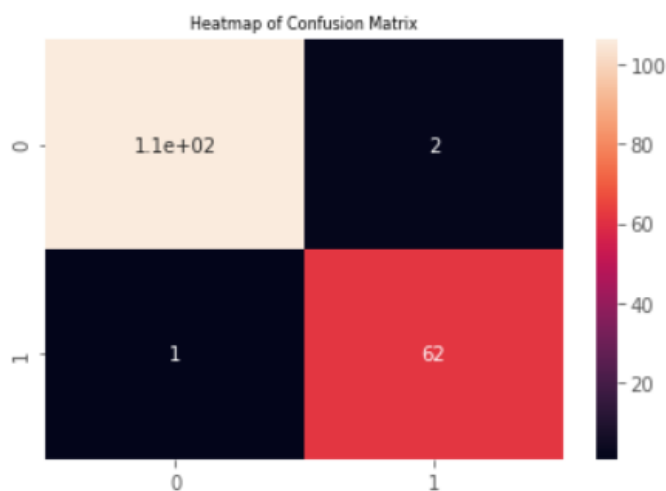


Fig 9: confusion matrix of logistic regression

I . Cross validation:

Based on the models performance on unseen data we can say whether our model is Under-fitting/Over-fitting/Well generalised. Cross validation (CV) is one of the technique used to test the effectiveness of a machine learning models

J. Classification Report

There are four ways to check if the predictions are right or wrong:

1. **TN / True Negative:** the case was negative and predicted negative
2. **TP / True Positive:** the case was positive and predicted positive
3. **FN / False Negative:** the case was positive but predicted negative
4. **FP / False Positive:** the case was negative but predicted positive

Precision — *The percent of our predictions were correct.*

Precision:- Accuracy of positive predictions.

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

Recall — *The percent of the positive cases we caught.*

Recall:- Fraction of positives that were correctly identified.

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

F1 score — *The percent of positive predictions were correct.*

$$\text{F1 Score} = 2 * (\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision})$$

Support

Support is the number of actual occurrences of the class in the specified dataset. Imbalanced support in the training data may indicate structural weaknesses in the reported scores of the classifier and could indicate the need for stratified sampling or rebalancing. Support doesn't change between models but instead diagnoses the evaluation process.

```
print(classification_report(y_test,y_pred))
```

	precision	recall	f1-score	support
0	0.96	0.99	0.97	108
1	0.98	0.92	0.95	63
accuracy			0.96	171
macro avg	0.97	0.96	0.96	171
weighted avg	0.97	0.96	0.96	171

Fig. 10. classification report of Random Forest

Conclusion:

Machine learning is an easy and effective way to predict the kind of tumor the patient is suffering from, as there is an increase in the number of women suffering from breast cancer. We need to predict the cancer class to which a patient will be classified by extracting

the hidden information of different attributes that could be used to maximize performance in general by leveraging the best available tools. In this project, comparing the accuracy of different models, namely: Logistic Regression, Support Vector Classifier, K-Nearest Neighbour, and Random Forest. The result concludes that Support Vector Classifier obtains the best model with 98.2% accuracy to predict breast cancer. Support Vector Classifier gives best performance in comparison with other models in terms of parameters: accuracy, RMSE Error, sensitivity, specificity, F-Measure, ROC Curve Area, Kappa statistics and time taken to build the model .

Scope of machine Learning

Machine learning is an application of **artificial intelligence** (AI). The system provided by ML has the ability to **automatically learn** and **improve** from past experiences. So, they can perform without being **explicitly programmed**. It focuses on the development of computer programs which can **access data** and use it to learn for themselves.

In simple terms, in this field provides computer the **ability** to learn without being explicitly programmed. It provides algorithms which can be trained to perform a task.

ML is also being used for **data analysis**, such as detection of regularities in the data by appropriately dealing with imperfect data, interpretation of continuous data used in the Intensive Care Unit, and for intelligent alarming resulting in **effective** and **efficient monitoring**.

It is argued that the successful implementation of ML methods can help the integration of **computer-based** systems in the healthcare environment providing opportunities to **facilitate** and **enhance the work** of medical experts and ultimately to **improve the efficiency** and **quality of medical care**.

In medical diagnosis, the main interest is in establishing the existence of a disease followed by its **accurate identification**.

There is a separate category for each disease under consideration and one category for cases where no disease is present.

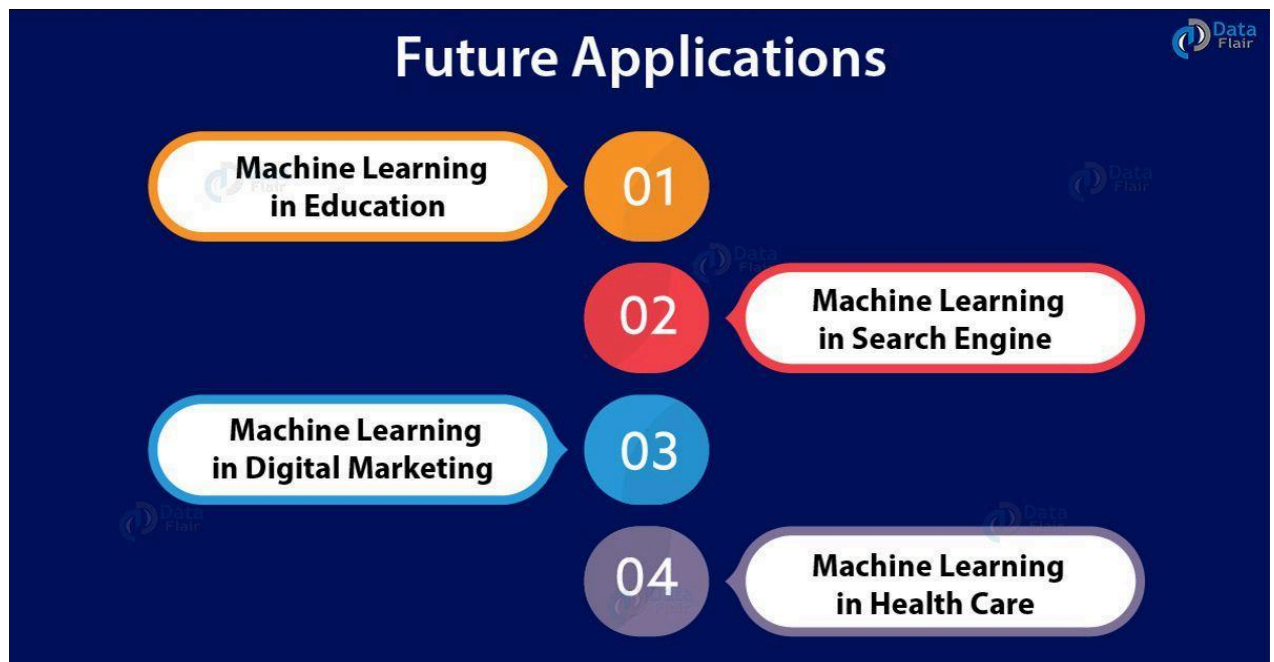
Here, machine learning improves the accuracy of medical diagnosis by **analyzing data** of patients.

The measurements in this Machine Learning applications are typically the results of certain medical tests (**example blood pressure, temperature and various blood tests**). This can also be **medical diagnostics (such as medical images)**,

presence/absence/intensity of various symptoms and basic physical information about the **patient(age, sex, weight etc.)**.

On the basis of the results of these **measurements**, the doctors narrow down on the **disease** inflicting the patient.

So we can conclude that machine learning has a very vast future in further development or we can say it helps to next generation to find something in digital platform so it have versatile scope in future.



References

1. Predicting Breast Cancer Using Logistic Regression and Multi-Class Classifier. Sultana, Jabeen and Jilani, Abdul Khader. 2018, International Journal of Engineering & Technology, pp. 22-26.
2. Using deep learning to enhance cancer diagnosis and classification. Fakoor, Rasool, et al. Arlington : s.n., 2013, Journal of Machine Learning Research, Vol. 28.
3. Breast cancer detection by leveraging Machine Learning. Vaka, Anji Reddy, Soni, Badal and K., Sudheer Reddy. s.l. : Elsevier B.V., 2020, pp. 1-5. ICT Express (2020).
4. On Breast Cancer Detection: An Application of Machine Learning Algorithms on the Wisconsin Diagnostic Dataset. M. Agarap, Ambien Fred. Viet Nam : Association of Computing Machinery, 2018, ICMLSC.
5. Performance Comparison of Machine Learning Techniques for Breast Cancer Detection. Gbenga, Dada Emmanuel, Christopher, Ngene and Yetunde, Daramola Comfort. s.l. : Nova Explore, 2017, Nova Journal Of Engineering and Applied Science, pp. 1-8.
6. Machine Learning Classification Techniques for Breast Cancer Diagnosis. Omondiagbe, David A., Veeramani, Shanmugam and Sidhu, Amandeep S. s.l. : IOP Publishing, 2019, pp. 1-16.
7. Breast Cancer Detection Using Machine Learning. Chithrakaran, R., et al. 11, s.l. : Blue Eyes Intelligence Engineering & Science Publication, September 2019, International Journal of Innovative Technology and Exploring Engineering, Vol. 8, pp. 3123-3126.
8. Using Machine Learning Algorithm for Breast Cancer Risk Prediction and Diagnosis. Asri, Hiba, et al. s.l. : Elsevier B.V., 2016, Procedia Computer Science, pp. 1064-1069.
9. Artificial intelligence methods for the diagnosis of breast cancer by image processing: a review. Farahnaz Sadoughi, Zahra Kazemy, Farahnaz Hamedan, Leila Owji, Meysam Rahmanikati, Tahere Talebi Azadboni. s.l. : Dovepress, 2018, Breast Cancer- Targets and Therapy, pp. 219-230.
10. Prediction of benign and malignant breast cancer using data mining techniques. Vikas Chaurasia, Saurabh Pal, BB Tiwari. Jaunpur, UP, India : SAGE, 2018, Journal of Algorithms & Computational Technology, pp. 119-126.
11. A support vector machine-based ensemble algorithm for breast cancer diagnosis. Haifeng Wang, Bichen Zheng, Sang Won Yoon, Hoo Sang Ko. s.l. : Elsevier B.V., 2017, European Journal of Operational Research, pp. 687-699.
12. Breast Cancer Histopathological Image classification: Deep Learning Approach. Mehdi Habibzadeh Motlagh,

Appendix

<https://colab.research.google.com/drive/1WC7UGUAhx607aQkuo8N-MHLe6HUWp64M?usp=sharing>