INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 02, February 2024, pp : 488-491

BREAST CANCER PREDICTION USING SUPERVISED MACHINE LEARNING ALGORITHMS

N. Suwathiswari[1], Mr. G. Ramkumar,

[1]Department of Computer Science, Sri Kaliswari College (Autonomous) Sivakasi, India.

[2]M. C. A., M.E., Department of Computer Science, Sri Kaliswari College (Autonomous) Sivakasi, India.

ABSTRACT

Breast Cancer represents one of the disease that make a high number of deaths every year. It is the most common type of all cancers and the main cause of women's deaths worldwide. Classification and data mining methods are an effective way to classify data. Especially in medical field, where those methods are widely used in diagnosis and analysis to make decisions. In this paper, a performance comparison between different machine learning algorithms. Random forest algorithm, Support Vector Machine(SVM), Logistic Regression and Decision Tree on the Breast Cancer Wisconsin (Diagnostic) datasets is conducted. The main objective is to assess the correctness in classifying data with respect to efficiency and effectiveness of each algorithm in terms of accuracy, precision, recall and specificity. The results obtained are very competitive and can be used for predict and treatment.

Keywords: Breast Cancer, Machine Learning, Classification, Accuracy, Precision, SVM.

1. INTRODUCTION

Around the world, Breast cancer is the most widely recognized type of cancer alongside lung and bronchus cancer, prostate cancer, colon cancer, and pancreatic cancer among others. Breast cancer might be a prevalent reason for death, and it's the main kind of malignant growth that is boundless among ladies in the around the world. Breast Cancer causes are multifactorial and include family ancestry, weight hormones, radiation treatment, and even reproductive factors. As indicated by the report of the world health organization every year, 2.1 million ladies are recently affected by breast cancer, and furthermore cause the highest number of can cerrelated deaths among ladies . In 2023, it is assessed that 627,000 ladies died from breast cancer - that is roughly 15% of all cancer deaths among ladies. While breast cancer growth rates are higher among ladies in extra developed areas, rates are expanding in about each locale internationally.

Many imaging techniques are developed for early identification and treatment of breast cancer and to scale back the amount of death and lots of aided breast cancer diagnosis methods are wont to increase the symptomatic precision.

Machine Learning algorithms are widely utilized in intelligent human services frameworks, particularly for breast cancer diagnosis and guess. There are many many machine learning classification and algorithms for prediction of breast cancer outcomes but during this paper, we are comparing various sorts of classification algorithms like Random forest algorithm, Support Vector Machine(SVM), Logistic Regression and Decision Tree. And furthermore, assess and compare the performance of the varied classifiers as far as accuracy, precision, recall, and f1-Score. The outcomes obtained during this paper provide a summary of the condition of modern Machine Learning strategies for breast cancer Prediction.

2. LITERATURE SURVEY

In today's medical world, doctors can use it to quickly and accurately interpret cancer. Because of that we can use machine learning to prevent the death by making an artificial intelligent model that can predict breast cancer and the method that be used is comparison between the ANN and SVM algorithms to see which algorithm suit the best for cancer prediction.

The study concluded by comparing two Artificial Neural Network algorithms and the Support Vector Machine algorithm to predict cancer based on several health attributes in the dataset using supervised machine learning. According to the results of our experiments and evaluating algorithm using Confusion Matrix, the Support Vector Machine algorithm outperforms ANN[1].

In this paper, we propose a diabetes prediction model using data mining techniques. We apply four data mining techniques such as Random Forest, Logistic Regression, and Decision Tree. In logistic regression, the accuracy is high, i.e., 1.0%, in comparison to other data mining techniques[2]. The main data mining algorithms discussed in this paper are algorithm, K means, C4.5 algorithm, Genetic algorithm and logistic regression. It is found that the genetic algorithm gives a better performance over five data mining algorithm[3].

Therefore three machine learning classification algorithms namely Decision Tree, SVM and ANN are used in this experiment to detect cancer at an early stage. Results obtained show SVM outperforms with the highest accuracy of

95% comparatively other algorithms[4]. A set of operation was led to assess this accuracy regarding a set.of data. mining procedures including. Decision.Trees (j48), ANN, and hybrid proposed method of decision-tree and SVM into diabetes disease diagnosis. Results showed that hybrid classification in proposed framework outperforms other classifiers with an accuracy rate of 87%[5].

3. METHODOLOGY

3.1 Support Vector Machine (SVM)

Support Vector Machine is of the Supervised Machine Learning characterization strategies that are broadly applied inside the field of cancer malignant growth determination and guess. Support Vector Machine works by choosing basic examples from all classes referred to as help vectors and isolating the classes by creating a linear function that partitions them as comprehensively as conceivable utilizing these help vectors. In this way, it is regularly said that planning between an input vector to a high dimensionality space is framed utilizing Support Vector Machine that intends to search out the preeminent reasonable hyperplane that separates the data set into classes. This linear classifier intends to expand the space between the decision hyperplane and along these lines the closest data, which is named the minimal distance, by finding the most appropriate hyperplane.

Logistic Regression (LR)

Logistic Regression is a key machine learning classification procedure. It has a place with the gathering of linear classifiers and is fairly practically like polynomial and statistical regression. Logistic regression is quick and similarly simple, and it's helpful for you to decipher the outcomes. In spite of the fact that it's basically a path for binary classification, it additionally can be applied to multiclass issues. This is frequently not the same as statistical regression, as statistical regression contemplates with the forecast of consistent qualities. Logistic regression models the likelihood that reaction falls into a specific classification. A logistic regression model helps us solve, via the Sigmoid function, for situations where the output can take but only two values, 0 or 1.

Decision Tree (DT)

Breast cancer prediction of A decision tree is a flow chart created by a computer algorithm to make decisions or numeric predictions based on information in a digital data set. Collect datasets on breast cancer, Pre-processing data in for performing the Decision Tree data mining approach (Benign, Malignant),Data that has been pre-processed.

Random Forest

The random forest algorithm is used to analyze the medical case diagnosis of breast cancer. The random forest algorithm can combine the characteristics of multiple eigenvalues, and the combined results of multiple decision trees can be used to improve the prediction accuracy.

3.2 Dataset Used

Breast Cancer Wisconsin (Diagnostic) dataset downloaded from Kaggle. This dataset includes the medical information for 768 cases of female patients.

The dataset also includes eight numeric-valued characteristics, where the value of one class is treated as a cancer test result of type 0 and the value of another class is treated as a result of type 1 cancer testing. In this dataset includes 570 Instances and 9 attributes.

The sample was divided into two parts, one with 80% of the data for training and 20% of the data for testing. Python and Jupyter notebook are used to execute the suggested mechanism. Python is a open-source language. In this paper,used the packages such as Numpy, Pandas, Scikit-Learn, Matplotlib, etc.. Python is the language of choice for data processing software

3.3 Preprocessing

Data Imputation: Data imputation is a method for retaining the majority of the dataset's data and information by substituting missing data with a different value.In this paper,there are many zero values in the dataset,so replacing with median values.

Label Encoding:It refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated.

It is an important pre-processing step for the structured dataset in supervised learning.

Feature Scaling or Standardization: It is a step of Data Pre Processing that is applied to independent variables or features of data. It helps to normalize the data within a particular range. Sometimes, it also helps in speeding up the calculations in an algorithm.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 02, February 2024, pp : 488-491

4. RESULT

Accuracy

Accuracy is used in classification problems to tell the percentage of correct predictions made by a model. Accuracy score in machine learning is an evaluation metric that measures the number of correct predictions made by a model in relation to the total number of predictions made.Calculate it by dividing the number of correct predictions by the total number of predictions.

Accuracy=Number of Correct Prediction/ Total Number of Prediction

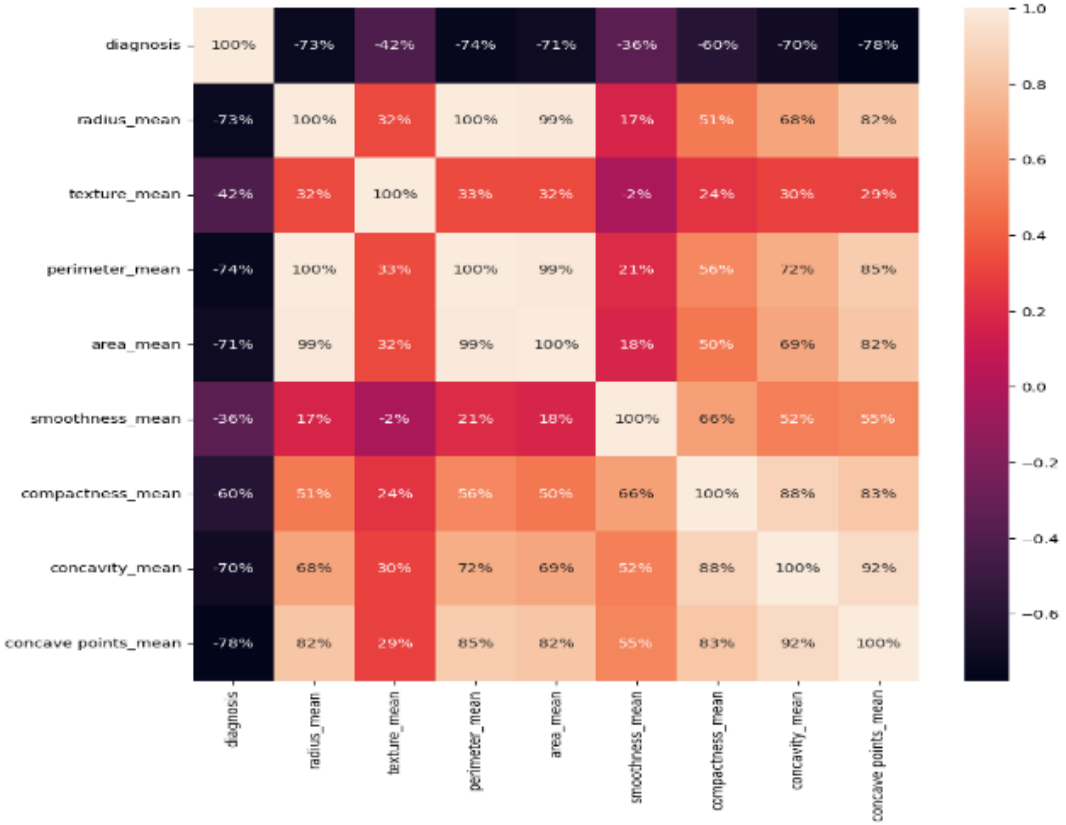

Fig.1: Heatmap for an attributes

Table1. Performance analysis for Proposed Algorithms

Techniques	Accuracy	Precision	Recall
SVM	95%	98%	94%
ANN	87%	90%	88%

From above comparison,.SVM accuracy is high than other techniques.

Table2. Performance analysis for Existing Algorithms

Existing Algorithms	Accuracy
Logistic Regression	99%
Decision Tree	1.0%
Random Forest	1.0%

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor : 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 02, February 2024, pp : 488-491

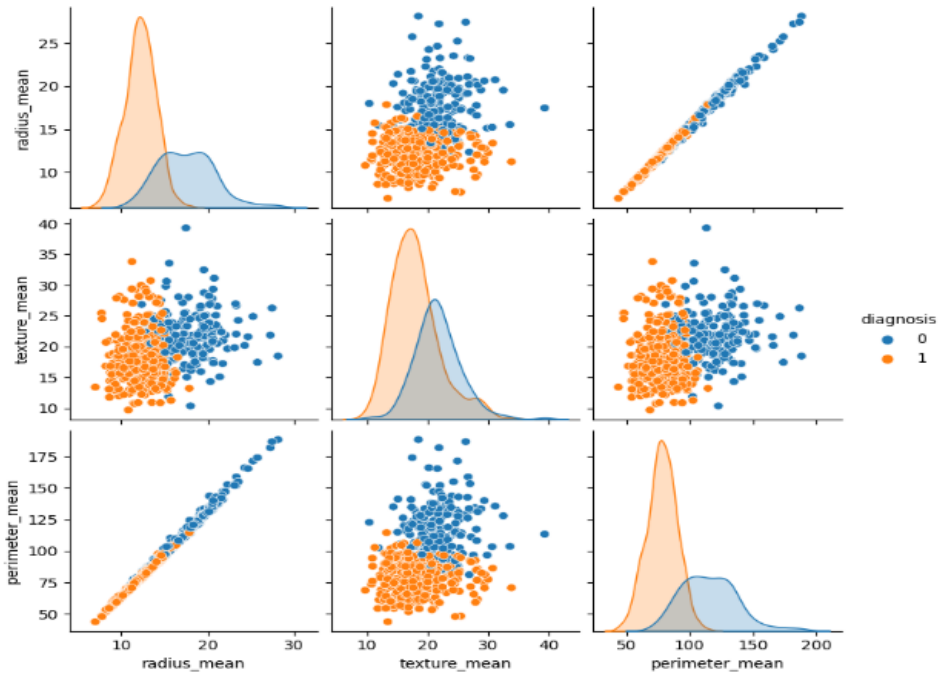

Fig2: pairplot

5. CONCLUSION AND FUTURE WORK

An SVM can make better prediction and achieve better performance. Cancer is a major health challenge in the world. Early prediction of Ccancer will result in improved results. This paper presents a cancer prediction among womens model with the help of Random Forest, Decision Tree, and Logistic Regression techniques to predict cancer . The proposed mechanism is implemented using Python.

To used Breast Cancer Wisconsin (Diagnostic) Dataset ,then preprocess the data,splitted training and testing data,and predicting the accuracy for used SVM algorithm.In the model, the accuracy is high 95% as compared to other models. In the future,large real time dataset will be collected and implemented.

6. REFERENCE

[1] Muhammad Amine Naji, Sanaa El Filali, Kawtar Aarika, EL Habib Benlahumar, Rachida Ait Abdelouhahid, Olivier Debauche, “Machine learning algorithms for Breast cancer prediction and Diagnosis”,Procedia Computer Science ,2021.

[2] Nan Wang,Xueping Du,Kehui Mei,Yuan Zhen, ”Classification Prediction of Breast Cancer Based on Machine Learning”, Procedia Computer Science ,2023.

[3] Varsha Nemade,Vishal Fegade,”Machine Learning Techniques for Breast Cancer Prediction”. Procedia Computer Science ,2018.

[4] Eslam AL Maghayreh,Awais Mohmood,Wail Elkilami, and Mohammad Faisal Nagi,Procedia “Automated Breast Cancer Diagnosis Based on Maxhine Learning Algorithm”,Computer Engineering,2019.

[5] Mamatha Sai Yarabarla,Lakshmi Kaya Ravi,A.Sivasangari,”Breast Cancer Prediction via Machine Learning” Procedia Computer Science ,2019.