Anticipating The Severity of Mammographic Mass Using Data Mining Technique

Lokesh K[1], Sreedevi M[2]

[1]Dept of Computer Science, S V University, Tirupati
[2]Assistant Professor, Dept of Computer Science, S V University, Tirupati

Abstract— *Mammography is viewed as the least expensive and most productive strategy to distinguish malignancy in a preclinical stage and bosom screening programs were made accurately with the goal of recognizing disease in prior stages. The bosom screening programs typically produce a tremendous measure of information, explained by the Breast Imaging Reporting and Data System (BI-RADS) made by the American College of Radiology. The BI-RADS framework decides a standard vocabulary to be utilized by radiologists when concentrating each finding. The primary objective of this work is to deliver AI models that anticipate the result of a mammography from a diminished arrangement of explained mammography discoveries. Nonetheless, the low sure prescient worth of bosom biopsy coming about because of mammogram understanding prompts roughly 70% pointless biopsies with considerate results. In this exploration paper information mining order calculations; Artificial Neural Network (ANN) and Support Vector Machine (SVM) are investigated on mammographic masses informational collection. Exactness of ANN and SVM are 80.3% and 81.9% of test tests separately. Our examination shows that out of these three arrangement models SVM predicts seriousness of bosom disease with least blunder rate and most noteworthy exactness.*

I. INTRODUCTION

Breast Cancer is quite possibly the most noticeable infections predominant in females. In 2016 alone it is being assessed that almost 246 thousand new instances of intrusive bosom malignant growth will be determined along to have 61 thousand non-obtrusive cases [1]. It's anything but a hard excursion for any malignancy patient, and a guardian all through. It gets imperative to analyze bosom malignant growth early, given its high death rate in the later stages. Mammography is the most dependable strategy utilized in this day and age for diagnosing bosom malignancy. Bosom Image Reporting and Data System (BI-RADS), a brand name of the American College of Radiology was acquainted with characterize the results of mammograms into four classifications, which was later on expanded to six. Mammography is viewed as the most economical and most capable procedure to distinguish danger in a preclinical stage and chest screening programs were made accurately determined to perceive illness in earlier stages.

Analytic evaluation of a patient in type of BI-RADS scale may require a further biopsy before the specialist articulates their last finding about a mammogram. The tumor biopsy may result either in threatening or kind tumor. On the off chance that the tumor was amiable, we might have kept away from the biopsy however the need of this biopsy was just when the specialist wasn't certain in a patient's BIRADS evaluation of the mammogram. Almost 70% of the biopsies done, prompt kind outcomes which is an exceptionally large number of patients and might have been forestalled [3]. In writing, radiologists show extensive variety in deciphering a mammography. In such cases, Fine Needle Aspiration Cytology (FNAC) is received. Yet, the normal right distinguishing proof pace of FNAC is just 90% [5]. The objective of BI-RADS to recognizing evidence is to give out a patient to either a liberal that doesn't have chest illness or an unsafe who has strong verification of having chest harmful development [7]. The inspiration driving this examination is to assemble the limit of specialist to choose the earnestness of a mammographic mass injury from BI-RADS properties of pointless bosom biopsies and the patient's age.

II. DATA MINING

Data mining is the way toward removing legitimate, already obscure, and at last understandable data from enormous data sets and utilizing it to settle on essential business choices. The separated data can be utilized to shape an expectation or order model, or to recognize relations between data set records [4]. Data Mining includes a coordination of strategies from various teaches, for example, data set and information distribution center innovation, insights, AI, elite registering, design acknowledgment, neural organizations, information perception, picture and sign handling, and spatial or fleeting information examination. By performing information mining, intriguing information, consistencies, or significant level data can be extricated from data sets and saw or perused from various points. The found information can be applied to dynamic, measure control, data the executives and inquiry preparing [2].

Data mining is the critical advance in the information revelation measure. The fundamental assignments of Data mining are by and large isolated into two classifications: Predictive and Descriptive. The goal of the prescient undertakings is to foresee the worth of a specific quality dependent on the upsides of different characteristics, while for the spellbinding ones, the goal is to separate beforehand obscure and valuable data like examples, affiliations, changes, irregularities and huge designs, from enormous information bases. There are a few methods fulfilling these destinations of information mining [6]. A portion of these can be arranged into the accompanying classifications: grouping, characterization, affiliation rule mining, successive example disclosure and examination.

The advancement of information mining frameworks has gotten a lot of consideration lately. It's anything but a vital job in cutthroat organizations in a wide assortment of business conditions. It has been broadly applied to a wide assortment of errands like deals investigation, medical services, E-trade, producing, and so on Various examinations have been made on proficient Data mining strategies and the important applications.

In this exploration paper we thought about Classification Rule Mining for information disclosure and produced the guidelines by applying our created approach on mammographic clinical data set.

III. METHODOLOGY

3.1 Support Vector Machine

The SVM is a new type of machine learning methods based on statistical learning theory. Because of good promotion and a higher accuracy, SVM has become the research focus of the machine learning community. SVMs are set of related supervised learning methods used for classification and regression [9]. Several recent studies have reported that the SVM generally are capable of delivering higher performance in terms of classification accuracy than the other data classification algorithms. SVM is on the basis of statistical learning theory by Vapnik et al proposed a new learning method, which is built on the basis of a limited number of samples in the information contained in the existing training text to get the best classification results [10].

A special property of SVM is, SVM simultaneously minimize the empirical classification error and maximize the geometric margin. So SVM called Maximum Margin Classifiers. SVM is based on the Structural risk Minimization. SVM map input vector to a higher dimensional space where a maximal separating hyperplane is constructed. Two parallel hyperplanes are constructed on each side of the hyperplane that separate the data. The separating hyperplane is the hyperplane that maximize the distance between the two parallel hyperplanes. An assumption is made that the larger the margin or distance between these parallel hyperplanes the better the generalization error of the classifier [9].

3.2 Artificial neural organization (ANN)

An ANN is an information getting ready perspective that is impelled by the way where a characteristic tangible framework in human brain works. ANNs are used comprehensively for the course of action of different issues, including portrayal, vision, talk, plan affirmation, control structures, etc. A colossal number of neurons present in the human frontal cortex structures the vital part of the neural framework perspective and go probably as simple taking care of segments [4]. A fake neuron is a little getting ready unit and plays out a clear computation that is fundamental to the action of a neural framework. The model of a neuron contains the essential parts like wellsprings of data, synaptic burdens, inclination, adding crossing point, and incitation work.

3.2.1 Multilayer Perceptron (MLP)

A MLP is a champion among the most generally perceived Neural Network plan that has been used for various applications. The MLP organize is commonly made out of different centers or dealing with units, and it is figured out into a movement of no less than two layers [6]. The essential layer (or the most diminished layer) is named as an information layer where it gets the external information while the last layer (or the most dumbfounding layer) is a yield layer where the response for the issue is gotten. The disguised layer is the widely appealing layer in the data layer and the yield layer, and may frame with somewhere around one layers. The arrangement of MLP could be communicated as a nonlinear improvement issue. The objective of MLP learning is to find the best loads that limit the differentiation between the information and the yield. The most predominant getting ready estimation used in NN is Back propagation (BP), and it has been used in dealing with various issues in model affirmation and portrayal. This computation depends on a couple of boundaries, for instance, different covered center points at the hid layers learning rate, energy rate, enactment work and the quantity of preparing to happen. Besides, these boundaries could change the exhibition on the gaining from awful to great exactness [2].

IV. EXPERIMENTAL RESULTS

The analyses have been directed by utilizing Python programming dialect. The Python Scikit-learn is a bundle for information characterization, grouping and representation. We have considered the Mammography mass data from the UCI Machine Learning Repository [8] dataset for experimentation. The Mammography mass data having 961 instances and 6 attributes. In this dataset, 516 instances classified as benign and 445 instances as malignant. There are 162 missing values of different attributes. The values of ordinal attribute represent categories with some intrinsic ranking while they nominal attribute represent categories with no intrinsic ranking in nominal type.

4.1 Results and discussion

The whole dataset is divided for training the models and test them by the ratio of 70:30% respectively. The training set is used to estimate each model parameters, while the test set is used to independently assess the individual models.

In this step the mammography dataset has to go through a cleaning process to remove duplicate records and fill missing data. In this data set 162 instances having missing values. The performance of a learning model is dependent on the quality features. Data preparation is an important step when building a model. This phase consists of replace missing data. The proposed stream imputes the missing values then trains and optimizes the two models. So, in this step, we replace missing values using Missing imputation strategy as mean was selected. The missing data results are shown in the screen shots of shown in the figure-1 and figure-2.

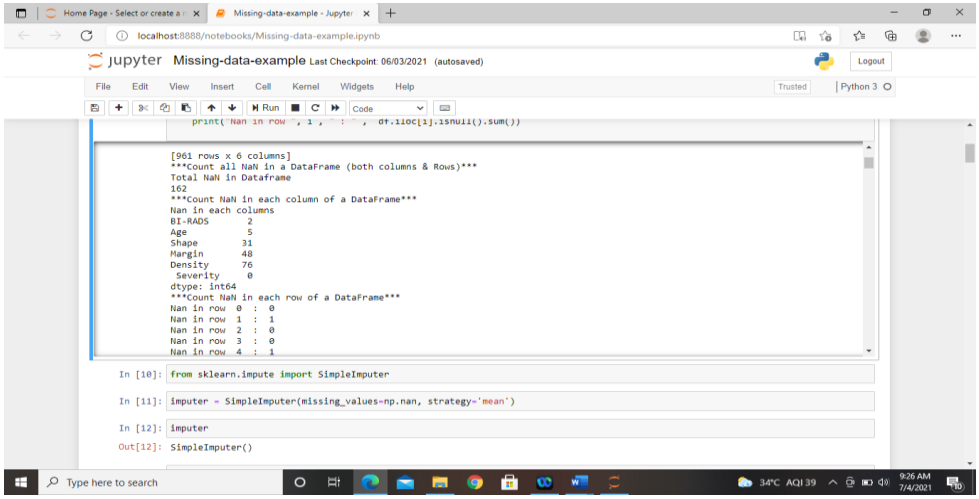

FIGURE-1: Screen shot of attributes missing records

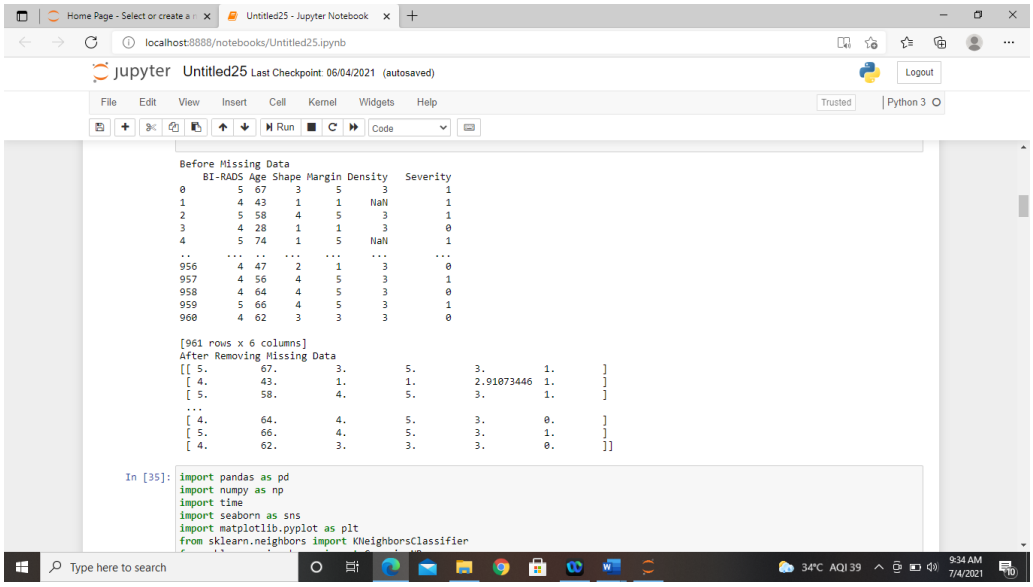

FIGURE-2: Screen shot of before missing and after filling imputation strategy

In the second stage we implement a SVM and MLP algorithms for prediction of Severity (benign and malignant) of mammographic dataset. The results that we got for MLP and SVM as shown in the figure-3 with their corresponding values.

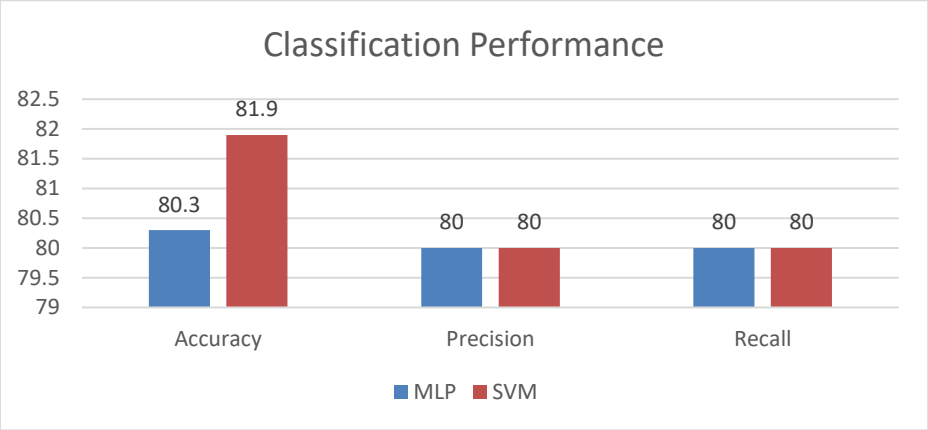

FIGURE 3: Classification Results

From the figure-3, we observe the performance of MLP accuracy has got 80.3%, whereas the performance of SVM accuracy has achieved 81.9%. However, there is an improvement in the accuracy of SVM over MLP model. The SVM accuracy rate is increased 1.6% over the MLP algorithm. In our experimental result the SVM algorithm shows the highest accuracy compared with MLP. The Experimental screen shots of two models are shown in the figure-4 and figure-5.

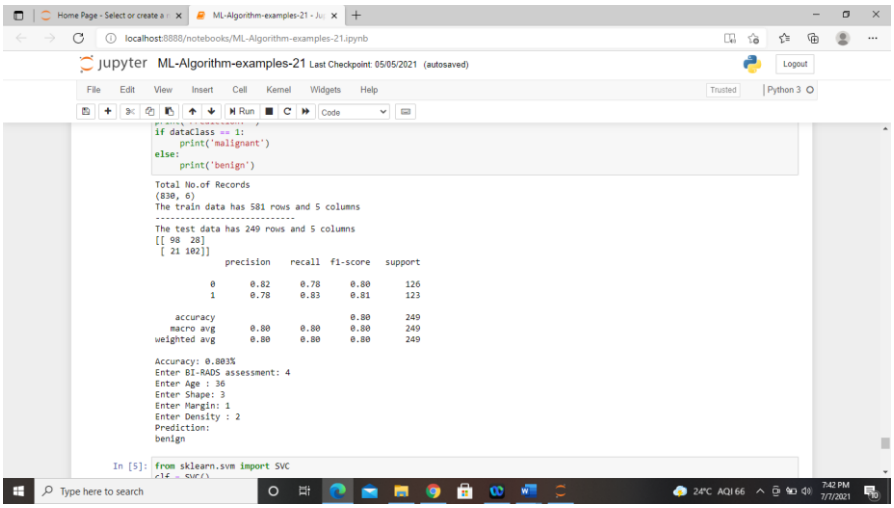

FIGURE 4: Screen Shot of MLP Algorithm

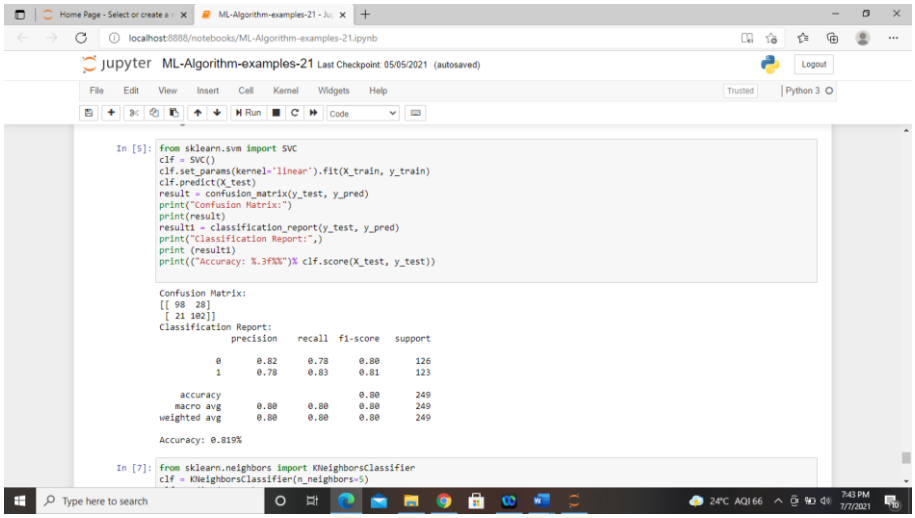

FIGURE 5: Screen Shot of SVM Algorithm

V. CONCLUSION

In this paper, two different classification models have been analyzed for the prediction of the severity of breast masses. These models are namely artificial neural network and support vector machine. The proposed stream imputes the missing values then trains and optimizes the two models. In this paper mainly focused on to establish an accurate classification model for mammographic mass medical diagnosis. The empirical results reveal that the SVM model does outperform the MLP method in terms of learning accuracy and complexity.

REFERENCES

[1] Elmore, J., M. Wells, M. Carol, H. Lee, D. Howard and A. Feinstein, 1994. Variability in radiologists‟ interpretation of mammograms. N. Engl. J. Med., 331:1493-1499.

[2] H. Witten and E. Frank, “Data mining: practical machine learning tools and techniques with Java implementations”, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., (2006)

[3] http://www.breastcancer.org/symptoms/understand_bc/statistics

[4] J.Han and M.Kamber,”Data Mining concepts and Techniques”, the Morgan Kaufmann series in Data Management Systems, 2nd ed.San Mateo, CA; Morgan Kaufmann, 2006.

[5] M. Margaret, Eberl, C.H. Fox, MD, S.B. Edge, C.A. Carter, and M.C. Mahoney, BI-RADS Classification for Management of Abnormal Mammograms, The Journal of the American Board of Family Medicine19, 2006, pp.161-164.

[6] N. Michael, “Artificial Intelligence - A Guide to Intelligent Systems”, 2nd edition, Addison Wesley, 2005.

[7] Simone A. Ludwig. Prediction of breast cancer biopsy outcomes using a distributed genetic programming approach. In ACM International Health Informatics Symposium, IHI 2010, Arlington, VA, USA, November 11 - 12, 2010, Proceedings, pages 694–699, 2010.

[8] UCI machine learning repository. http://archive.ics.uci.edu/ml/.

[9] Vapnik V.N, “Statistical learning Theory”, John Wiley and Sons, New York, USA, 1998.

[10] Vapnik, V.N. The Natural of Statistical Learning theory. Springer–Verleg, New York, USA1995.