JOURNAL OF NORTHEASTERN UNIVERSITY

Volume 25 Issue 04, 2022 ISSN: 1005-3026 https://dbdxxb.cn/ Original Research Paper

EXECUTION AND EVALUATION OF K-NEAREST NEIGHBOR FOR IDENTIFICATION AND VISUALIZATION OF BREAST MALIGNANT GROWTH

Gaurav D Saxena[1], Dr. Shaik Jumlesha[2], K Susmitha[3], Morukurthi Sreenivasu[4], Uppalapu Vinod Kumar[5], Gottala Surendra Kumar[6]

[1]Department of Computer Science, Kamla Nehru Mahavidyalaya, Nagpur, Maharashtra, India

[2]Professor, Department of Computer Science and Engineering, Annamacharya Institute of Technology and Sciences, Tirupati, India

[3]Assistant Professor, Department of Computer Science and Engineering, Annamacharya Institute of Technology and Sciences, Tirupati, India

[4]Associate Professor, Department of Information Technology, GIET Engineering College, JNTUK, Kakinada, Andhra Pradesh, India

[5]Assistant Professor, Department of Computer Science and Engineering, GIET College of Engineering, JNTUK, Kakinada, Andhra Pradesh, India

[6]Assistant Professor, Department of Computer Science and Engineering, Shri Vishnu Engineering College for Women, Bhimavaram, India

gauravsaxena@kamlanehrucollege.ac.in[1], ahmedsadhiq@gmail.com[2], susmitha.karanam10@gmail.com[3], msreenivasucse@giet.ac.in[4], uvinodkumarcse@gmail.com[5], gsurendrakumarcse@svecw.edu.in[6]

ABSTRACT

Breast malignant growth starts as a sluggish developing irregularity or cancer that begins from the milk channel cell lining. Breast malignant growth can either be obtrusive or not. Harmless breast malignant growths can't attack other breast tissues; however obtrusive breast diseases can go from the milk conduit or lobule to other breast tissues. The thickness and mass of the breast are consistent in size and structure because of their heterogeneity. In this paper, the K-Nearest Neighbor technique is utilized to break down the Breast Malignant growth Wisconsin (Diagnostic) dataset from the UCI machine learning repository. For the distinguishing proof of Breast threatening development, support vector machines, K-Nearest Neighbors, random forest, calculated relapse, and a combination of various philosophies can be used. Regardless, for a surprisingly long time, we just used the K-Nearest Neighbor approach for getting ready and assurance, which is a controlled AI calculation. The most restricted way between the model point and the arrangement discernments in the dataset is all not completely permanently established in that frame of mind of social event the data (dataset). Directly following executing all of the cycles as indicated by K-Nearest Neighbor computation on the foreordained dataset, we acquired results with 95.21 percent Accuracy.

KEYWORDS: Breast Cancer, Breast Cancer Diagnostic Dataset, Confusion Matrix, K-Nearest Neighbor, Machine Learning.

Submitted: 16/11/2022 **Accepted**: 05/12/2022

1 INTRODUCTION

In the present situation, the computerized reasoning and AI is progressing is ending up being continuously critical in disorder or sickness distinguishing proof [1]. The usage of a characterization structure to clinical examination is filling in pervasiveness. The evaluation and social event of data from patients, as well as master examination, are fundamental pieces of a respectable investigation. Nevertheless, the usage of a man-made cognizance in a cunning characterization approach has genuinely committed to both decreasing potential missteps made by the lacking people and reviewing clinical data in a more restricted and more complete way.[2] We can say that, the sickness or sickness distinguishing proof is like one of the application where modernized thinking is utilized to assist the specialists in contamination or illness assurance with more precision, subsequently vanquishing difficulties related with specialists owing to a shortfall of fitness or stress, which makes the end problematic.

AI methodologies are expecting a basic part in assurance and representation of breast dangerous development by applying some consistently utilized characterization procedures or strategies to perceive people or patients with breast malignant growth sickness as innocuous from compromising disease and to expect perception. Breast cells threatening development could spread from the one tissue and organs to individual in the body; in these circumstances, the dangerous development is suggested as metastatic breast cells harmful development. The kind of threatening development contrast considering when they occur and how outrageous they are.

Around 100 distinct types of malignant growth exist, each having its own name contingent upon the organ or tissue where in it shows. Fundamentally, understanding the direction of malignant growth, the organ by which it made, and the instrument of disease movement helps in the treatment revelation.[3] Early identification of this infection or fatal disease and its characterization into cases is significant[4].

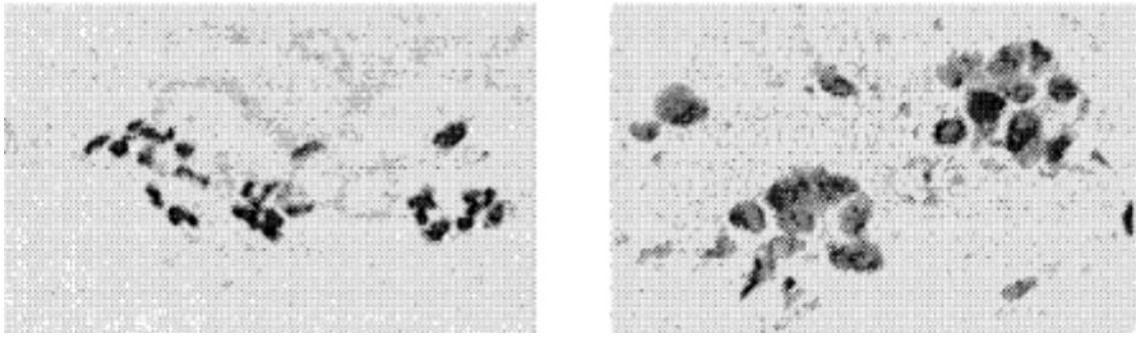
As a matter of fact, enormous information has reformed the size of information and furthermore making esteem from it Huge information has rolled out a major improvement in BI overwhelmingly of unstructured, heterogeneous, non-standard and fragmented medical services information. It doesn't just conjecture yet additionally helps in navigation and is progressively seen as forward leap in continuous headway with the objective is to work on the nature of patient consideration and diminishes the medical services cost. Information mining calculations applied in medical services industry assume a critical part because of their superior presentation in foreseeing, determination of the sicknesses, diminishing expenses of medication, settling on continuous choice to save individuals' lives. The Most widely recognized Information mining demonstrating objectives are characterization and expectation which involves a few calculations for the expectation of breast malignant growth.

Breast disease is disintegrated into two sorts of harmless and threatening growths. Harmless cancers are non hazardous growths, they have well-characterizes shapes. They foster gradually in the organ where they showed up without delivering metastatic occurrence. Harmless cancers

are made out of cells that look like to ordinary cells of the breast tissue. Threatening growths are hazardous cancers, as they spread to different parts of the body and bring metastatic cases. Disease cells of threatening cancers have extreme irregularities contrasted and ordinary cells in shape, size and forms, where cells lose their unique attributes. [5]

2 MATERIALS AND METHODS

The University of Wisconsin Emergency clinic's endeavors to analyze and anticipate the visualization of bosom malignant growths absolutely founded on the FNA test are reflected in the WDBC and WPBC data sets. In this methodology, a bosom irregularity is penetrated with a fine measure needle to eliminate liquid, which is then inspected under a magnifying lens for visual qualities.



(a) Benign (b) Malignant

Figure: depicts two images, which were taken from fine needle biopsies of breast as appeared.

In this concentrate on we utilized a machine learning calculation model on the WDBC Datasets to foresee how compelling the model we carry out utilizing for the recognition and analysis of Breast malignant growth. (Wisconsin Breast Cancer Diagnosis) dataset, which was recovered from the UCI bosom disease storehouse, was utilized in this examination. The WDBC dataset contains 10 genuine esteemed ascribes, which are recorded in Table 1. Mean, standard blunder, and most horrendously terrible (mean of the objective worth) are registered for each characteristic, bringing about a sum of 30 credits. What's more, 2 extra ascribes, id and determination class for harmless and threatening cells, are additionally included. The bosom mass was suctioned with a fine needle to make a computerized picture, from which the essential characteristic was extricated. There are 569 occurrences in this dataset, 212 of which are threatening, and 357 of which are harmless [6-8].

Table: Feature Description of cell nuclei [9]

WDBC Dataset			
Feature no	**Features**	**Description**	**Feature number**

1	Radius	Mean of Distance from center to point on the perimeter	3-5
2	Texture	Standard Deviation of grey scale values	6-8
3	Perimeter	Mean size of core tumor	9-11
4	Area	The size of cell area	12-14
5	Smoothness	Mean of local variation in radius lengths	15-17
6	Compactness	Mean of Perimeter2/area-1.0	18-20
7	Concativity	Mean of severity of concave portions of the contour	21-23
8	Concave points	Mean of concave portions of the contour	24-26
9	Symmetry	Symmetricity	27-29
10	Fractal Dimensions	"Coastline approximation"-1	30-32

Other Features are

ID Number

Diagnosis-M for Malignant, B for Benign

The characteristics are portrayed.

The characteristics utilized in the dataset we chose are recorded with depictions in the table underneath. These characteristic depictions are normal ones that are accessible in the dataset that was acquired.

Table: Attributes Description

Attributes	**Respective Description**
id	ID Number
diagnosis	The diagnosis of Breast Tissues (M=Malignant, B= Benign)

radius_mean	Mean of Distance from center to point on the perimeter
texture_ mean	Standard Deviation of grey scale values
perimeter_ mean	Mean size of core tumor
area_ mean	-
smoothness_ mean	Mean of local variation in radius lengths
compactness_ mean	Mean of Perimeter2/area-1.0
concave points_ mean	Mean for the number of concave portions of the contour
symmetry_ mean	-
fractal dimensions_ mean	Mean for "Coastline approximation"-1
radius_se	Standard error for Mean of Distance from center to point on the perimeter
texture_ se	Standard error for Standard Deviation of grey scale values
perimeter_ se	-
area_ se	-
smoothness_ se	Standard error for local variation in radius lengths
compactness_ se	Standard error for Perimeter2/area-1.0
concativity_se	Standard error for severity of concave portions of the contour
concave points_ se	Standard error for number of concave portions of the contour
symmetry_ se	
fractal dimensions_ se	Standard error for "Coastline approximation"-1
radius_worst	"Worst" or largest mean value for Mean of Distance from center to point on the perimeter
texture_ worst	"Worst" or largest mean value for Standard Deviation of grey scale values

perimeter_ worst	-
area_ worst	-
smoothness_ worst	"Worst" or largest mean value for local variation in radius lengths
compactness_ worst	"Worst" or largest mean value for Perimeter2/area-1.0
concativity_worst	"Worst" or largest mean value for severity of concave portions of the contour
concave points_ worst	"Worst" or largest mean value for number of concave portions of the contour
symmetry_ worst	-
fractal dimensions_ worst	"Worst" or largest mean value for "Coastline approximation"-1

2.1 Feature Scaling (Preprocessing & Splitting of Dataset)

Preceding applying the computation or model, many AI models use Information Preprocessing as the first and most critical stage. Data game plan is a methodology for changing over muddled or unacceptable data for computer based intelligence models. This is the fundamental stage before the computer based intelligence model is developed and executed.

It's possible that we won't get any decontaminated data from time to time. There might be data unmistakable redundancy. We truly need to preprocess our data to dispense with the ambiguities and redundancies in the dataset. We at present utilize the K-nearest neighbor computer based intelligence methodology to research numerical data, however the dataset, we used is in character plan, thusly we ought to change the data over to numerical arrangement. The compromising and innocuous classes are given out the numbers "0" and "1." We apportioned the data gathered (Dataset) into Preparing and testing dataset. The preparation dataset was utilized to set up the structure using the computation that was developed. The testing dataset was answerable for testing. The split of the dataset is fundamental for the model's viability or exactness. From the preparation dataset, the model realizes in isolation.

Dataset	Datatype	Total Records
Breast Cancer Dataset	Training	500
	Testing	69

During the investigation, 500 records were utilized for the Preparation the Dataset and 69 records were utilized for the testing dataset. The above table gives brief depiction of preparing and testing dataset

2.2 Experimental Environment

All the investigation on ML calculation(K-Nearest Neighbor) portrayed during this paper were directed utilizing the scikitlearn library and python programming and Jupyter notebook.

2.3 Data Visualizations

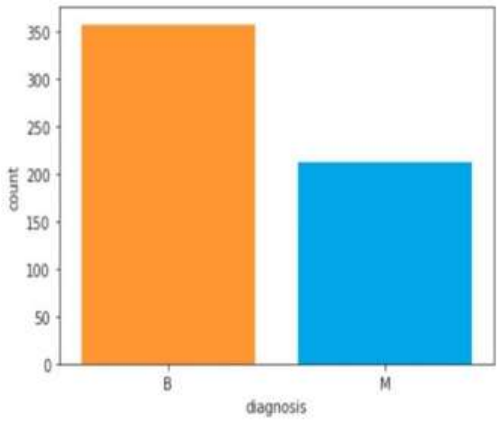


Figure: Count Plot

The above figure shows the plotting of the WDBC dataset, where 357 are the cases of the benign (accounting for 62.74%) and 212 are of malignant (accounting for 32.26%).

2.4 Background

These study will probably decide whether a patient has an inclination toward threatening breast malignant growth or not. ML procedures that gain from models are utilized to accomplish this objective. Rather than providing a PC with a considerable rundown of rules to continue to take care of an issue, AI strategies make models that are prepared to do freely investigating models. A ML model upheld by complex critical thinking philosophies is incredibly precise thanks to these learning standards.

2.5 ML-Based Classifier Model

The ML calculations can make insightful computerization frameworks. Point by point ML classifiers for distinguishing breast malignant growth irregularities are momentarily given in this segment.

2.6 Implementation of Work- K-Nearest Neighbor Classifier

The K-nearest neighbor classifier is one of the most fundamental classifier for design acknowledgment or information characterization. K-Nearest Neighbor is notable as its

interpretation is straightforward and requires less assessment time than some other AI calculations. K-Nearest Neighbor may be used for both characterization and relapse. This is the most straightforward philosophy for portrayal and is called Lethargic Student, considering the way that the progression is done intelligently, and the computations are conceded until course of action. No genuine model or learning is performed during the arrangement stage, regardless of the way that getting ready enlightening record is required; it is used solely to populate an illustration of search space with cases whose class is known. The K-Nearest Neighbor calculation is non-parametric and fit for multi-class request. K-Nearest Neighbor makes no speculations on the fundamental [10-16] data and don't develop the model from the preparation information. The classifier addresses the going with saying " On the off chance that it strolls like a duck, quack like a duck, and seem to be a duck, then, at that point, it is presumably is a duck ": that is, the experiment class settles on the class of its storage room neighbors. The K-Nearest neighbors process begins at testing point and extends the districts until it consolidates K-preparing tests and denotes the testing point x by an enormous vote of those examples.

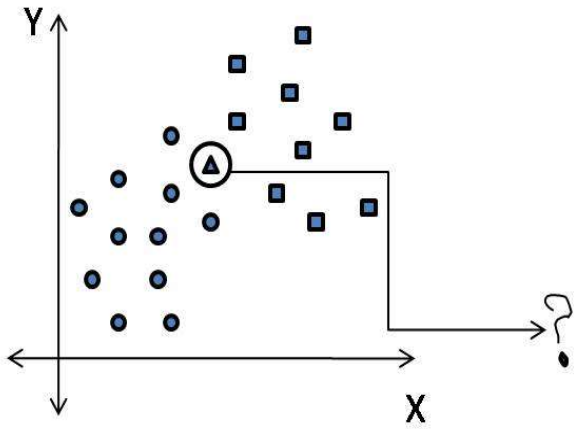


Figure: K-Nearest Neighbor Illustration

The Fig. 2 portrays another significant enlightening components (?) are consigned the three-sided class set apart considering larger part projecting a polling form, among its five nearest neighbors.

A model in the test information is portrayed by discovering the distance to all models in the preparation dataset; the class of the preparation information that gives the most restricted distance chooses the class of the test information. For the two classes, the potential gains of K should be odd to avoid tie. More noteworthy upsides of K will undoubtedly decide ties. [18,19]

There are an extensive variety of distance measurements that can be used for distance assessment in K-Nearest Neighbor classifiers. The decision is to pick one relies upon the multifaceted nature of the issue, the kind of data, and so forth. Most notable included distance procedure in K-Nearest Neighbor computation is Euclidian Distance. The condition to determine distance between two centers can be seen by the going with frame.

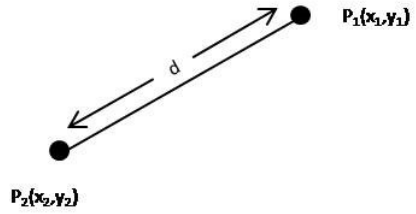


Figure: Euclidian Distance

K-NN gauges class characteristic depending the k Nearest preparation models in the component space. When a dataset is given, it picks the k Nearest examples from the characterized preparing information and decides the class thinking about the most delegate tests. Euclidean distance comparability metric is utilized to choose the areas and determined utilizing Eq. (1) as follows

$$\text{EucledianDistance} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{1}$$

where x_i and y_i are two focuses in Euclidean n-space. After all test tests are characterized by k-NN, the characterization precision is determined with partitioning the quantity of accurately arranged examples by the complete number of tests. Mean absolute error (MAE) is determined by the accompanying Eq. (2) as follows [20]

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - x_i| \tag{2}$$

where y_i is the expectation worth and xi is the genuine worth. Root mean square mistake (RMSE) is determined utilizing Eq. (3) as follows [20]

$$RMSE = \sqrt{\frac{1}{n} \sum_{n=1}^{1} |y_i - x_i|^2} \tag{3}$$

2.7 Performance of a Classifier

Straightforwardly following finishing a computer based intelligence computation, we genuinely need to figure out how successful the model we complete is. The standards for evaluating the sensibility might be laid out on datasets and metric. For assessing our proposed computer based intelligence computation we can utilize different execution assessments. [21] The classifier execution can be assessed by seeing the confusion structures. Confusion Matrix, a two-layered table with each line and each fragment address various classes. Every part in the construction keeps an eye on the absolute test vectors for which the authentic and expected classes are in the line and in the section, solely. The presentation of the depiction calculation is evaluated considering specific assessments like precision, accuracy, recall, and F1-Score. To

decide these, the fundamental step is to convey its confusion matrix and a brief time frame later obtain the True positive, true negative, false positive, false negative variables. [22-24]

Table: Confusion Matrix

		Actual	
		Positive (1)	Negative (0)
Predicted	Positive (1)	True Positive (TP)	False Positives (FP)
	Negative (0)	False Negatives (FN)	True Negatives (TN)

However, if both actual and expected class has the identical label value of one, this would be the outcome in True Positives.

However, if both actual and expected class has the identical label value of zero, this would be the outcome in True Negatives.

When the expected class label is zero, but actual class label is one, this would be the case in false positives.

When the expected class label is one, but actual class label is zero, this would be the case in False Negatives. [25]

For a conjecture model, the introduction of a classifier is critical due to related cost associated with it. An off-base assurance of an infection could have to pay a significant cost for a patient and it might be even of his life. Execution cross not entirely set in stone from these qualities. [26]

2.8 **Performance Metrics** [27]

The most widely recognized measurement for assessing classifier execution is accuracy. Accuracy is determined as the proportion of accurately arranged information focuses to add up to items. Precision is the extent of positive forecasts that are really right. Precision is determined by separating accurately anticipated positive perceptions with complete anticipated positive perceptions. Recall-It is the quantity of right certain outcomes separated by the quantity of every single pertinent sample. F1-Score is the Harmonic method for Precision and recall. So we can observe the F1 Score by utilizing the condition expressed previously.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (5)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (6)$$

$$\text{F1\_Score} = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (7)$$

3 RESULT EVALUATION AND ANALYSIS

We got the accuracy of 95.21 percent after effectively executing an extraordinary K-Nearest Neighbor, displaying that the model is critical for the area and assumption for Bosom compromising turn of events.

Accuracy Obtained - 0.9521276595744

Confusion Matrix

[[116 4]

[5 63]]

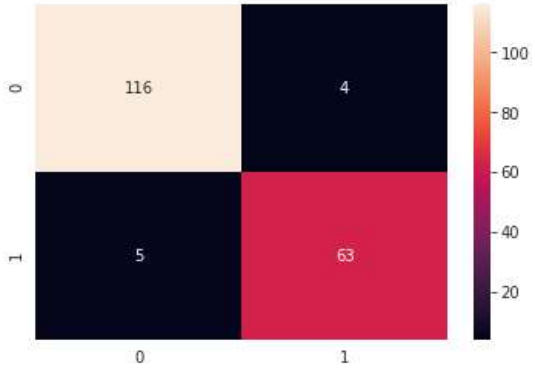


Figure: Confusion Matrix

4 CONCLUSION

Clinical determination, a subset of medical services, is the method involved with distinguishing the infection from the patient's signs and side effects. To do this, the fundamental information is acquired from a few sources, including the actual assessment, the patient's clinical history, and general data. Specialists are exceptionally keen on making brilliant classification calculations for clinical determination. This is for the most part because of the capacity of AI

and information mining calculations to distinguish stowed away patterns in data set highlight matches. In this manner, using clever calculations to classify the clinical datasets makes ready for the production of more compelling clinical demonstrative choice emotionally supportive networks. Our investigation's fundamental objective is to track down the most dependable and solid calculation for distinguishing breast malignant growth. The precision of the calculation displayed in this work is 95.21%. WDBC dataset from the UCI ML repository was used to finish the introduction of the recently demonstrated strategies. Since K-Nearest Neighbor has the most outstanding precision, it was viewed as the best technique. Giving its efficiency to the best super learning strategies is viewed as the most astonishing in malignant growth research; giving its viability to these techniques is viewed as magnificent.

REFERENCES

1 E. Vinoth S.M, P. Valarmathi. Accurate Breast Cancer Prediction using machine Learning Techniques. International Journal of Recent Technology and Engineering, Blue Eyes Intelligence Engineering and Science Publication; March. 2020. pp. 3811-3815.

2 Kaklamanis MM, Flippakis ME, Touloupos M, Christodoulou K. An experimental comparision of machine learning classification algorithms for breast cancer diagnosis. InEuropean, Mediterranean, and Middle Eastern Conference on Information Systems 2019 Dec 9 (pp. 18-30).Springer,Cham.

3 Lahane SR, Chavan N, Madankar M. Classification of Thermographic Images for Breast Cancer Detection Based on Deep Learning. Annals of the Romanian Society for Cell Biology. 2021 May 23; 25 (6): 3459-3466.

4 Lahane SR, Chavan N, Madankar M. Review on Breast Cancer Detection using Deep Learning Methods. Design Engineering. 2021 Jul 7: 2015-22.

5 Saoud H, Ghadi A, Ghailani M, Abdelhakim BA. Using feature selection techniques to improve the accuracy of breast cancer classification. InThe Proceedings of the Third International Conference on Smart City Applications 2018 Oct 10 (pp. 307-315). Springer, Cham.

6 Dangeti P. Statistics for machine learning. Packt Publishing Ltd; 2017 Jul 21.

7 Janghel RR, Singh L, Sahu SP, Rathore CP. Classification and detection of breast cancer using machine learning. InSocial Networking and Computatioanl intelligence 2020 (pp. 269-282). Springer, Singapore.

8 Shang M, M Lakshmi TC. Hands-on Supervised Learning with Python: Learn How to Solve Machine Learning Problems with Supervised Learning Algorithms Using Python. BPB Publications; 2021.

9 Soman KP, Loganathan R, Ajay V. Machine Learning with SVM and other kernel methods. PHI Learning Pvt. Ltd.; 2009 Feb 2.

10 Rajaguru H, Prabhakar SK. K-Nearest Neighbor classifier and K-means clustering for robust classification of epilepsy from EEG signals. A detailed analysis. Diplom. De; 2017 Mar 23.

11 Khanna A, Gupta D, Dey N. Applicaions of big data in healthcare. London, United Kingdom: Academic Press; 2021.

12 Kulkarni P, Joshi P. Artificial intelligence: Building intelligent systems. PHI Learning Pvt. Ltd.; 2015 Feb 26.

13 Raman K. Mastering Python Data Visualization. Packt Publishing Ltd; 2015 Oct 27.

14 Zhu Z, Nandi AK. Automatic modulation classification: principles, algorithms and applications. John Wiely & sons; 2015 Feb 16.

15 Wickham M. Practical Java Machine Learning: Projects with Google Cloud Platform and Amazon Web Services. Apress; 2018 Oct 23.

16 Dougherty G. Pattern recognition and classification: an introduction. Springer Science & Business Media; 2012 Oct 28. p. 100-101.

17 Raschka S. Python machine learning. Packt publishing ltd; 2015 Sep 23.

18 Palaniappan R. Biological signal analysis. BookBoon; 2011.

19 Dougherty G. Pattern recognition and classification: an introduction. Springer Science & Business Media; 2012 Oct 28. p.102.

20 Willmott CJ, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research. 2005 Dec 19;30(1):79-82.

21 Tandon A, Soni S. Introduction to Artificial Intelligence using Python. Book Bazooka Publication; 2020 Jan 24.

22 Ananda Babu T, Rajesh Kumar P. Prediction of Term Labor Using Wavelet Analysis of Uterine Magnetomyography signals. InProceedings of International Conference on Computational Intelligence and Data Engineering 2019 (pp. 29-37). Springer, Singapore.

23 Nicolas B, Jayakumar A, Titus B, Remya Nair T. Comparative Study of Multiple Feature Descriptors for Detecting the Presence of Alzheimer's Disease. InUbiquitous Intelligent Systems 2022 (pp.331-339). Springer, Singapore.

24 Leekha G. Learning AI with Python: Explore Machine Learning and Deep Learning techniques for Building Smart AI Systems Using Scikit-Learn, NLTK, NeuroLab, and Keras(English Edition). BPB Publications; 2021.

25 Haque MA, Shetty S. Leveraging Machine Learning in Financial Fraud Forencics Age of Cybersecurity. InTechnologies to Advance Automation in Forensics Science and Criminal Investigation 2022 (pp. 220-249). IGI Global.

26 Gupta SC, Goel N. Selection of Best K of K-Nearest Neighbors Classifier for Enhancement of Performance for the Prediction of Diabetes. InProgress in Advanced Computing and Intelligent Engineering 2021 (pp. 135-142). Springer, Singapore.

27 Gupta P, Sehgal NK. Introduction to Machine Learning in the Cloud with Python: Concepts and practices. Springer Nature; 2021.

28 Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archieve.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

29 Singhal V, Chaudhary Y, Verma SK, Agarwal U, Sharma MP. Breast Cancer Prediction using KNN, SVM, Logistic Regression and Decision Tree.

30 Mohammad, M., Ghiasi. Implementing decision tree-based algorithms in medical diagnostic decision support systems. (2020).

31 Maglogiannis I, Karpouzis K, editors. Artificial Intelligence Applications and Innovations: 3rd IFIP Conference on Artificial Intelligence Applications and Innovations (AIAI), 2006, June 7-9, 2006, Athens, Greece. Springer; 2006 Aug 29.

32 Soman KP, Loganathan R, Ajay V. Machine Learning with SVM and other Kernel methods. PHI Learning Pvt Ltd.; 2009 Feb 2.

33 Kularathne S. Prediction and data visualization of breast cancer using K-nearest neighbor (knn)classifier... [Internet]. Medium. Analytics Vidhya; 2020 [cited 2022Dec3]. Available from: https://medium.com/analytics-vidhya/prediction-and-data-visualization-of-breast-cancer-using-k-nearest-neighbor-knn-classifier-df7adadc4872