Developing An Accurate Model for Comprehensive Analysis Of Mammographic Masses In Clinical Settings

Jalam Likhitha

Department. of Computer Science Sri Venkateswara University, Tirupati

Abstract—*Mammography, a key tool in early threat detection, drives chest screening programs aimed at spotting infections early. These programs, managed by BI-RADS, yield vast data analyzed by radiologists. This study focuses on AI models predicting mammography outcomes, aiming to reduce unnecessary biopsies. Naïve Bayes and K-Nearest Neighbor algorithms were tested, with Naïve Bayes showing the highest accuracy at 85.43%.*

I. INTRODUCTION

Breast cancer stands out as one of the most prevalent diseases affecting women. In 2016 alone, approximately 246 thousand new cases of invasive breast cancer were reported, alongside 61 thousand cases of non-invasive forms. It's an arduous journey for any cancer patient, serving as both a challenge and a constant vigil. Early detection becomes imperative due to the high mortality rates associated with advanced stages of the disease. Mammography emerges as the cornerstone in diagnosing breast cancer, recognized for its reliability and widespread use. The Breast Imaging Reporting and Data System (BI-RADS), established by the American College of Radiology, initially categorized mammogram results into four classifications, later expanded to six. Mammography is lauded for its cost-effectiveness and efficiency in identifying risks during the preclinical stage, although chest screening programs have not been fully optimized for early disease detection.

Clinical evaluation utilizing the BI-RADS scale may necessitate further biopsy before a conclusive diagnosis can be made by the expert. Biopsy results can vary, ranging from benign to malignant growths. While some biopsies may confirm benign conditions, the necessity for biopsy arises when the expert's confidence in the patient's BI-RADS assessment from the mammogram is uncertain. Alarmingly, nearly 70% of biopsies yield benign outcomes, a considerable number that could potentially have been avoided. Radiologists exhibit considerable variability in interpreting mammograms, prompting the utilization of Fine Needle Aspiration Cytology (FNAC) in such cases. However, FNAC's typical accuracy rate stands at 90%, leaving room for potential misdiagnosis.

The primary objective of the BI-RADS system is to categorize patients into those with no evidence of breast cancer (benign) and those with strong indications of malignancy, aiding in treatment decisions based on mammographic findings and patient demographics. This study aims to assess the proficiency of experts in distinguishing the severity of mammographic abnormalities based on BI-RADS classifications, biopsy outcomes, and patient age. [1]. [3]. [5]. [7].

II. CLASSIFICATION

Approach refers to the systematic process of developing a model or framework that characterizes and interprets data categories and concepts, with the aim of using this model to predict the classes of items whose class labels are unknown. Data analysis within this approach can be delineated into two primary stages: the learning phase, where a classifier is constructed to describe a predetermined structure of classes or phenomena by analyzing the dataset comprising descriptive feature tuples and their associated labels. In the subsequent phase, the model is deployed for prediction by initially assessing the predictive accuracy of the classifier established during the learning phase, utilizing test data. Classifier accuracy on a given test set of tuples is the proportion of tuples correctly classified by the classifier. If the accuracy exceeds a certain acceptable threshold, the classifier can be employed to predict future tuples with unknown class labels.

Description serves as a form of data analysis utilized to construct models that describe large data categories. Framework, on the other hand, is a data mining technique employed to predict group membership for data instances. It stands as one of the fundamental methodologies in data mining and finds applications in various domains such as pattern recognition, anomaly detection, customer relationship management, and supervised learning. The primary objective of description analysis is to construct a model from a vast repository of training data, where the target class labels are known, and subsequently utilize this model to categorize unseen instances.

Classification stands as one of the most prevalent and widely used data mining techniques. It entails mapping data into predefined groups or classes and is typically referred to as supervised learning since the classes are predetermined prior to data analysis. Approach, once again, underscores the process of developing a model that discerns data classes, aiming to predict the class of items whose class labels are unknown. The selection of an appropriate model is contingent upon the evaluation of a substantial volume of training data. Descriptive datasets abound with latent information that can be leveraged for informed decision-making. [4][5] [6] [8].

III. METHODOLOGY

This fragment gives the compact thought about picked managed models of K-Nearest Neighbor and Naïve Bayes.

3.1 Naive Bayes

The Naive Bayes is a snappy strategy for production of measurable prescient models [66]. NB depends on the Bayesian hypothesis. This characterization strategy investigations the connection between each characteristic and the class for each example to infer a contingent likelihood for the connections between the quality qualities and the class [2] [3]. During preparing, the likelihood of each class is figured by tallying how frequently it happens in the preparation dataset. This is known as the "earlier likelihood" P(C=c). Notwithstanding the earlier likelihood, the calculation additionally registers the likelihood for the occurrence x given c with the suspicion that the qualities are free. This likelihood turns into the result of the probabilities of each single trait. The probabilities would then be able to be evaluated from the frequencies of the occurrences in the preparation set.

3.2 K-Nearest-Neighbors (KNN)

The K-Nearest-Neighbors (KNN) is a non-parametric gathering technique, which is essential anyway incredible all around [1]. The essential thought for k-NN depends after determining the distances between the attempted, and the readiness data tests to recognize its nearest neighbors. The attempted model is then consigned to the class of its nearest neighbor [2].

The K-Nearest-Neighbors (KNN) is a clear anyway convincing procedure for game plan. The KNN estimation is a procedure for gathering objects reliant upon closest planning models in the part space. KNN is a kind of event based learning, or aloof acknowledging where the limit is simply approximated locally and all computation is yielded until gathering [6]

For a data record D to be requested, its K nearest neighbors is recuperated, and these constructions a neighborhood of D. Bigger part projecting a voting form among the data records in the space is by and large used to pick the request for D with or without considered distance-based weighting. Regardless, to apply KNN we need to pick a reasonable motivating force for K, and the accomplishment of collection is a great deal of wards on this value. The critical drawbacks in regards to KNN are (1) its low efficiency - being a slow learning methodology denies it in various applications, for instance, dynamic web burrowing for an enormous vault, and (2) its dependence on the decision of an "incredible worth" for K.

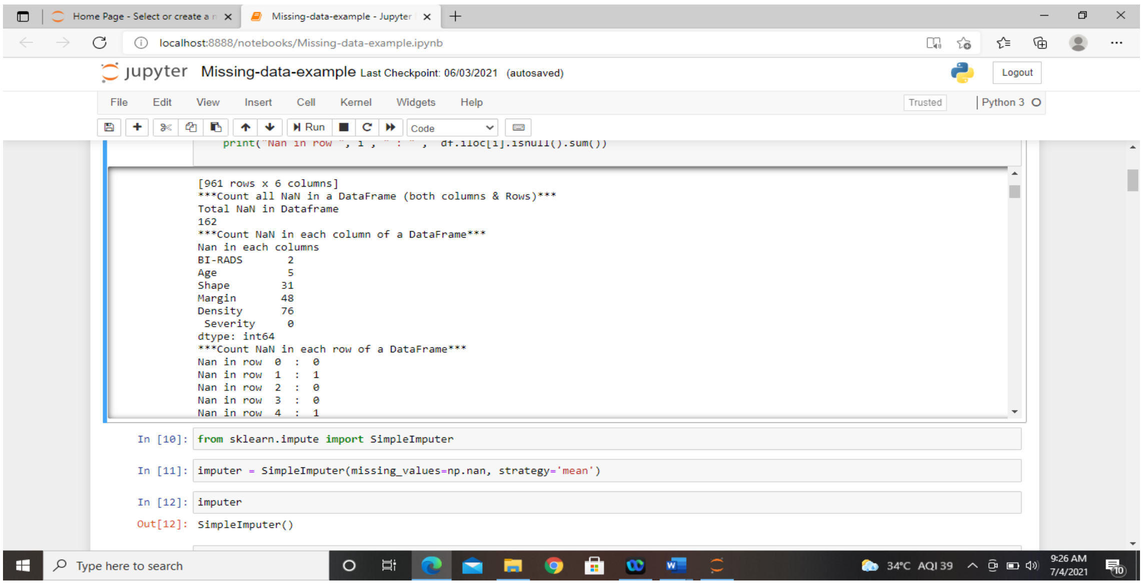
IV. EXPERIMENTAL RESULTS

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.


Home Page - Select or create a
Missing-data-example - Jupyter
localhost:8888/notebooks/Missing-data-example.ipynb
jupyter Missing-data-example Last Checkpoint: 06/03/2021 (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3
Run
Code
Type here to search
34°C AQI 39
9:26 AM
7/4/2021

```
    print("Nan in row ", i , " : " ,  df.iloc[i].isnull().sum())
```

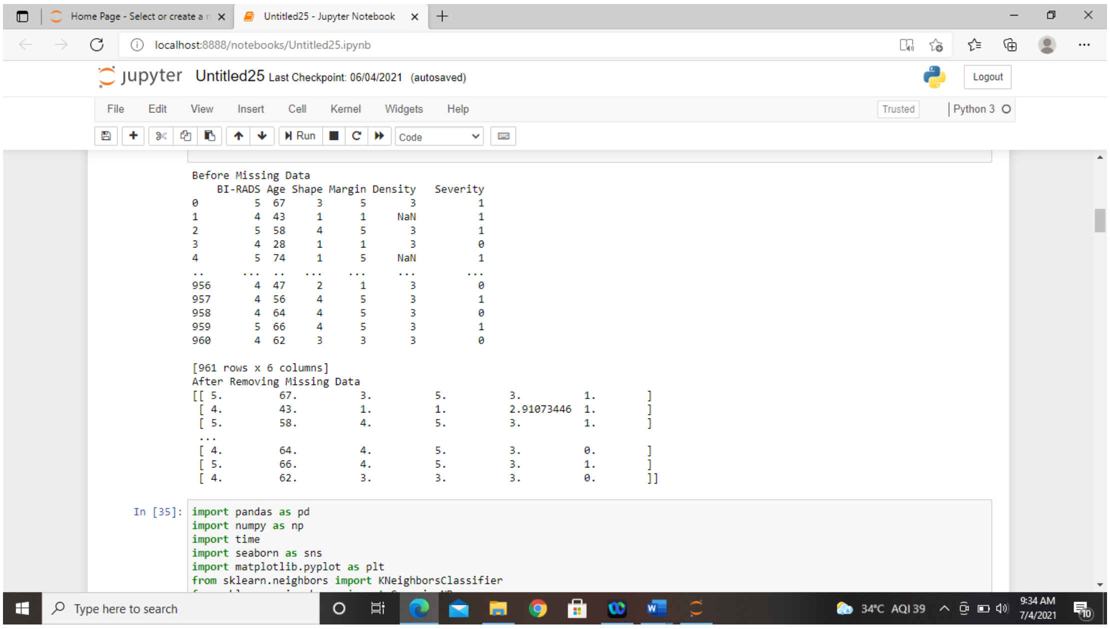
```
[961 rows x 6 columns]
***Count all NaN in a DataFrame (both columns & Rows)***
Total NaN in Dataframe
162
***Count NaN in each column of a DataFrame***
Nan in each columns
BI-RADS       2
Age           5
Shape        31
Margin       48
Density      76
 Severity     0
dtype: int64
***Count NaN in each row of a DataFrame***
Nan in row  0  :  0
Nan in row  1  :  1
Nan in row  2  :  0
Nan in row  3  :  0
Nan in row  4  :  1
```

```
In [10]: from sklearn.impute import SimpleImputer

In [11]: imputer = SimpleImputer(missing_values=np.nan, strategy='mean')

In [12]: imputer
Out[12]: SimpleImputer()
```

Figure-1: Screen shot of attributes missing records


Home Page - Select or create a
Untitled25 - Jupyter Notebook
localhost:8888/notebooks/Untitled25.ipynb
jupyter Untitled25 Last Checkpoint: 06/04/2021 (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3
Run
Code
Type here to search
34°C AQI 39
9:34 AM
7/4/2021

```
Before Missing Data
    BI-RADS Age Shape Margin Density   Severity
0         5  67     3      5       3          1
1         4  43     1      1     NaN          1
2         5  58     4      5       3          1
3         4  28     1      1       3          0
4         5  74     1      5     NaN          1
..      ...  ..   ...    ...     ...        ...
956       4  47     2      1       3          0
957       4  56     4      5       3          1
958       4  64     4      5       3          0
959       5  66     4      5       3          1
960       4  62     3      3       3          0

[961 rows x 6 columns]
After Removing Missing Data
[[ 5.         67.          3.          5.          3.          1.        ]
 [ 4.         43.          1.          1.          2.91073446  1.        ]
 [ 5.         58.          4.          5.          3.          1.        ]
 ...
 [ 4.         64.          4.          5.          3.          0.        ]
 [ 5.         66.          4.          5.          3.          1.        ]
 [ 4.         62.          3.          3.          3.          0.        ]]
```

```
In [35]: import pandas as pd
         import numpy as np
         import time
         import seaborn as sns
         import matplotlib.pyplot as plt
         from sklearn.neighbors import KNeighborsClassifier
```

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

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

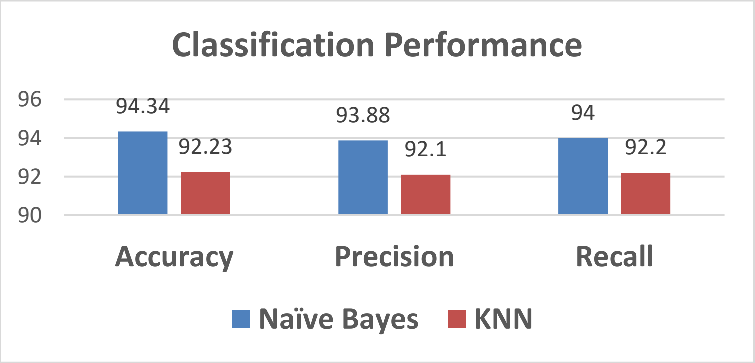


Figure-3: Classification Results

From the figure-3, we observe the performance of Naïve Bayes accuracy has got 94.34%, whereas the performance of KNN accuracy has achieved 93.88%. However, there is an improvement in the accuracy of naïve bayes over KNN model. The naïve bayes accuracy rate is increased 0.46% over the KNN algorithm. In our experimental result the naïve bayes algorithm shows the highest accuracy compared with KNN.

V. CONCLUSION

This paper explores two distinct classification models, artificial neural network and support vector machine, for predicting the severity of breast masses. The proposed method addresses missing values, trains, and optimizes both models. The primary focus is on developing an accurate classification model for clinical analysis of mammographic masses. Experimental results show that the naïve Bayes model outperforms the KNN technique 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] G. Ravi Kumar, K.Nagamani and G.Anjan Babu, “A Framework of Dimensionality Reduction utilizing PCA for Neural Network Prediction”, Lecture Notes on Data Engineering and Communications Technologies, Volume-37, Pages:173 – 180, Springer Nature Singapore Pte Ltd, 2020.

[10] S.Rahamat Basha and Surya Bhupal Rao G.Ravi Kumar, “A Summarization on Text Mining Techniques for Information Extracting from Applications and Issues”, Journal of Mechanics of Continua and Mathematical Sciences, Special Issue, No.-5, PP: 324-332, 2020, Institute of Mechanics of Continua and Mathematical Sciences.