AI Approaches for Imputing Missing Values in Mammogram Mass Data

Chappidi Sravanthi

Department of Computer Science, S V University, Tirupati

Abstract— *In data mining, one of the main challenges in data preprocessing is handling missing values. Imputation, the process of replacing missing data with substituted values, is crucial for ensuring accurate analysis. Many clinical diagnostic datasets often contain missing values, and excluding these incomplete datasets can introduce more problems than solutions. Traditional imputation methods are easy to implement but may introduce bias in the data. This paper proposes a data imputation technique using K-Nearest Neighbors (KNN) to address the issue of missing data. The method combines KNN predictive modeling with Support Vector Machine (SVM) for improved attribution. The aim of this study is to assess the impact of missing data on the data mining process of learning discovery. Handling missing values in the dataset is a challenging task. Our study explores AI techniques for missing value imputation using Mammogram mass data from the UCI repository. The findings indicate that classifier performance improves when Support Vector Machine (SVM) is employed.*

I. INTRODUCTION

Missing data, a common issue in data analysis, refers to the absence of data values for a variable of interest. It poses challenges for AI experts across various fields, from computational science to social sciences. Managing missing data is crucial as it can impact the quality of analysis and modeling. Various techniques for handling missing values have been proposed, but there is no universally best method. The goal of these techniques is to impute missing values using available data. Handling missing values is essential before applying data mining methods to ensure accurate analysis. This study explores the performance of the KNN algorithm for imputing missing data and its impact on model accuracy, using SVM for classification based on the imputed dataset [1].

II. MISSING DATA HANDLING MECHANISMS

A few techniques have been applied in information mining to deal with missing qualities in data set. Information with missing qualities could be disregarded, or a worldwide consistent could be utilized to fill missing qualities (obscure, not material, endlessness, for example, characteristic mean, trait mean of a similar class, or a calculation could be applied to discover missing qualities [5]. Missing information ascription procedure implies a methodology to fill missing upsides of an informational index to apply standard strategies which require finished informational collection for examination. These procedures hold information in deficient cases, just as ascribe upsides of associated factors.

Missing information attribution strategies are delegated unimportant missing information ascription techniques, which incorporate single ascription techniques and different attribution techniques, and non-insignificant missing information ascription techniques which incorporate probability-based strategies and the non-probability-based strategies. A solitary attribution technique could fill one incentive for each missing worth and it is more usually utilized at present than numerous ascriptions which supplant each missing worth with a few conceivable qualities and better reflects testing changeability about genuine worth.

2.1 Missing characteristics can be organized into three sorts:

1. **Missing Completely at Random (MCAR)** values in a dataset are said to miss absolutely aimlessly if events that lead to a particular data thing being missing are self-governing of both the perceptible components and imperceptible boundaries of interest and subsequently happens through and through at unpredictable. It is portrayed as when the probability that the data is missing isn't related to either the specific regard which ought to be gotten or the game plan of watched responses. The datasets only occasionally have MCAR data.
2. **Missing at Random (MAR)** when the missing assessments of some quality are not aimlessly appropriated over the discernments but instead are scattered inside something like one model, they are said to miss unpredictably. Missing absolutely unpredictably is described as when the probability that the data is missing isn't related to either the specific regard which ought to be gotten or the game plan of watched responses.

3. **Not missing at Random (NMAR)** is generally called non immaterial missing quality. If the characters of the data don't meet those of MCAR or MAR, by then they fall into the grouping of missing not at subjective. For the present circumstance the missing data is dependent on the assessments of the quality. It is the trickiest design as it incorporates missing characteristics that are not heedlessly passed on over the discernments.

2.2 Strategies for Handling Missing Data

2.2.1 Mean Imputation

A champion among the practically every so often used techniques. This is the most easy ways to deal with property missing data is to replace each missing a motivator with the mean of non-missing assessments of the variable [6]. This strategy moreover its obstructions the scattering of the attributed variable can get extraordinarily deformed, considering the way that each missing worth is apportioned a comparable credit.

2.2.2 Lit wise erasure:

In this strategy, cases with any missing characteristics are deleted from an examination. It is moreover called absolute case assessment, considering the way that simply cases with complete data are held.

III. METHODOLOGY

3.1 Missing characteristics using K-Nearest Neighbor (KNN):

The K-Nearest Neighbor (KNN) is one of the attribution methods used to treat missing worth. KNN credit approaches are neighbor based methodologies where the attributed regard is either a regard that was assessed for the neighbor or the typical of assessed regards for different neighbors [2] [3]. It is an essential and stunning strategy. The motivation driving the KNN computation is that models with similar features have relative yield regards. The estimation works on the explanation that the attribution of the dark models ought to be conceivable by relating the dark to the known by some partition or closeness work [4].

KNN is the clearest estimation in attributing missing characteristics. In this technique the missing assessments of an event are ascribed a ton of nearest neighbor for a model and substitutes the missing data by calculating the ordinary of non-missing characteristics to its neighbors. The closeness of two models is settled using a partition work. Partition limit can be Euclidean and Manhattan. In this work we have considered the Euclidean partition work. Exactly when the k-nearest neighbors' methodology is associated with the test data, the assumption execution yields result closest to those for the main data with no missing characteristics, and the figure model's show is consistent despite when the missing data rate increases.

3.2 Support Vector Machine

The SVM is another sort of AI techniques dependent on measurable learning hypothesis. Due to great advancement and a higher exactness, SVM has become the exploration focal point of the AI people group. SVMs are set of related administered learning techniques utilized for grouping and relapse [8]. A few ongoing examinations have revealed that the SVM for the most part are fit for conveying better as far as order exactness than the other information grouping calculations. SVM is based on factual learning hypothesis by Vapnik et al proposed another learning strategy, which is based on a set number of tests in the data contained in the current preparing text to get the best grouping results.

An uncommon property of SVM will be, SVM all the while limits the observational order blunder and expand the mathematical edge. So SVM called Maximum Margin Classifiers. SVM depends on the Structural danger Minimization. SVM map input vector to a higher dimensional space where a maximal isolating hyperplane is developed. Two equal hyperplanes are built on each side of the hyperplane that different the information. The isolating hyperplane is the hyperplane that boost the distance between the two equal hyperplanes. A supposition that is made that the bigger the edge or distance between these equal hyperplanes the better the speculation mistake of the classifier [9].

IV. EXPERIMENTAL RESULTS

The experiments have been conducted by using Python programming language. The Python Scikit-learn is a package for data classification, handling missing data, clustering and visualization. We have considered the Mammographic-Mass UCI Machine Learning Repository dataset [7] for evaluating the efficiency and effectiveness of our proposed algorithm.

4.1 Dataset

The Mammographic-Mass Data set has 961 rows and 6 columns. In this data there are two class labels i.e., The Benign class has 516 instances and Malignant class has 445 instances. Through descriptive statistics we can summaries each attribute of Mammographic-Mass data has shown in the table-1 and also the distribution of each attribute is of density plot is presented in figure-1.

Table-1
Descriptive statistics dataset

	BI-RADS	Age	Shape	Margin	Density	Severity
count	961	961	961	961	961	961
mean	4.35	55.48	2.73	2.79	2.92	0.46
std	1.78	14.44	1.23	1.53	0.37	0.49
min	0.00	18.00	1.00	1.00	1.00	0.00
25%	4.00	45.00	2.00	1.00	3.00	0.00
50%	4.00	57.00	3.00	3.00	3.00	0.00
75%	5.00	66.00	4.00	4.00	3.00	1.00
max	55.00	96.00	4.00	5.00	4.00	1.00

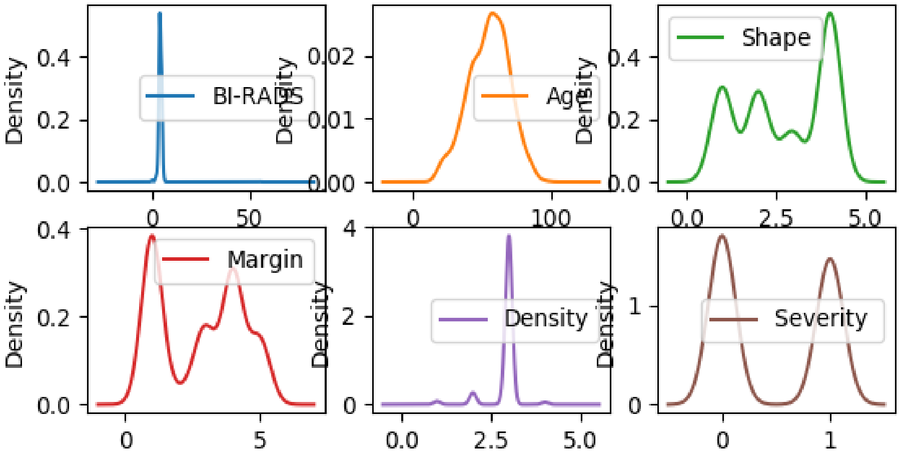


Figure-1: Density plot of Data distribution of each attribute

4.2 Results

The standard dataset is divided into two sets (70% and 30%), one for training and another one set for testing. Two experiments have been conducted for evaluating the SVM Classification with KNN Imputation method for missing data. In our Experiment the first step is data preprocessing for mammography dataset has to go through a cleaning process to remove duplicate records and fill missing data. The performance of a learning model is dependent on the quality features. In this mammography data set 162 instances having missing values, attribute wise missing values are shown in the figure-2.

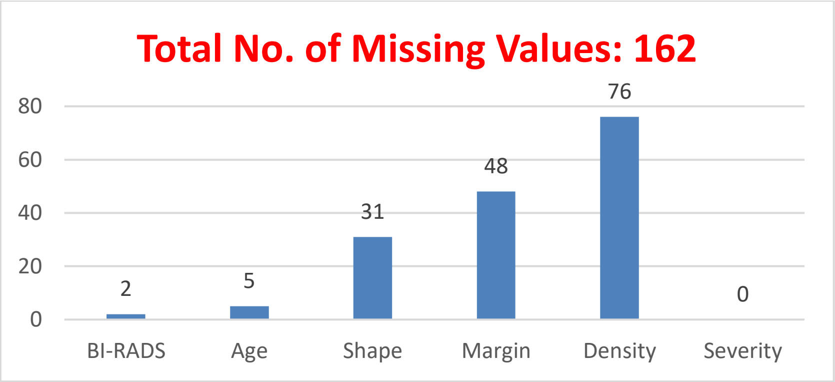
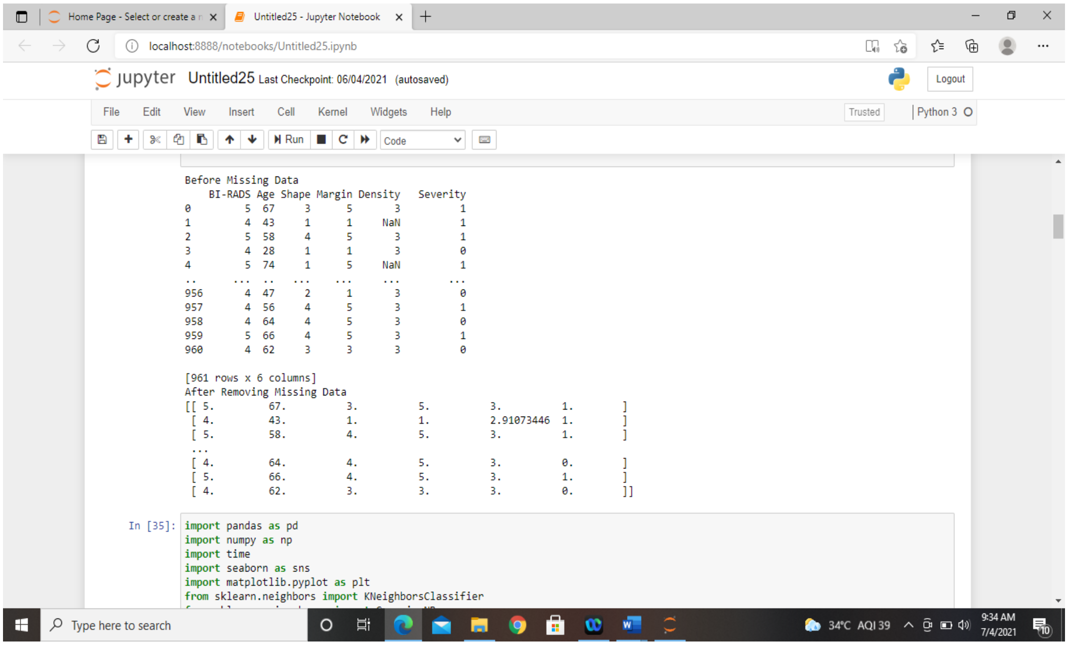


Figure-2: Attribute wise Missing values

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 KNN imputation strategy are used. The missing data results are shown in the screen shots of shown in the figure-3.


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

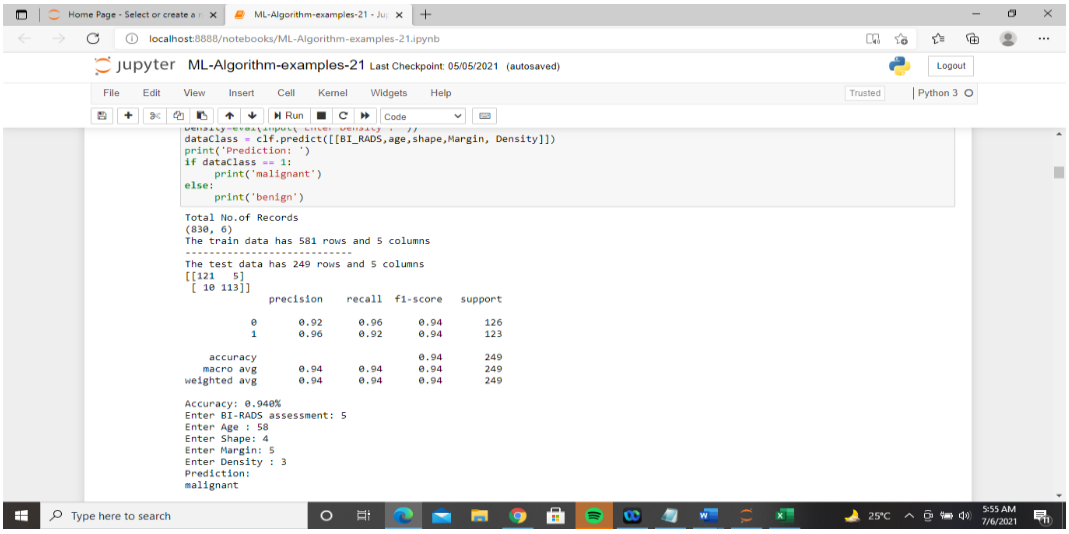
```
Before Missing Data
    BI-RADS Age Shape Margin Density   Severity
0         5  67     3      5       3          1
1         4  43     1      1     NaN          1
2         5  58     4      5       3          1
3         4  28     1      1       3          0
4         5  74     1      5     NaN          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-3: Results of Missing data

In the second stage we execute a SVM calculations for forecast of Severity (kindhearted and dangerous) of mammographic dataset. The outcomes that we got for SVM as displayed in the figure-4 with their comparing esteems.


Home Page - Select or create a
ML-Algorithm-examples-21 - Ju
localhost:8888/notebooks/ML-Algorithm-examples-21.ipynb
jupyter ML-Algorithm-examples-21 Last Checkpoint: 05/05/2021 (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
Type here to search
25°C
5:55 AM
7/6/2021

```
dataClass = clf.predict([[BI_RADS,age,shape,Margin, Density]])
print('Prediction: ')
if dataClass == 1:
    print('malignant')
else:
    print('benign')
```

```
Total No.of Records
(830, 6)
The train data has 581 rows and 5 columns
----------------------------
The test data has 249 rows and 5 columns
[[121   5]
 [ 10 113]]
              precision    recall  f1-score   support

           0       0.92      0.96      0.94       126
           1       0.96      0.92      0.94       123

    accuracy                           0.94       249
   macro avg       0.94      0.94      0.94       249
weighted avg       0.94      0.94      0.94       249

Accuracy: 0.940%
Enter BI-RADS assessment: 5
Enter Age : 58
Enter Shape: 4
Enter Margin: 5
Enter Density : 3
Prediction:
malignant
```

Figure-4: SVM Results after Impute the missing values

From the figure-4, we observe the performance of SVM accuracy has got 94%.

This research proposes an approach for enhancing the training process of SVM when dealing with missing data.

V. CONCLUSION

This paper also evaluates approaches used to fill missing values and proposes a new and better approach to handle missing value situation and thereby enabling to feed correct input to the SVM classifier to get better prediction, diagnosis and treatment of the mammographic data. The proposed KNN data imputation method serves as an effective data imputation method for SVM classification in the case of missing information.

REFERENCES

[1] Alireza Farhangfara, Lukasz Kurganb and Jennifer Dyc, “Impact of imputation of missing values on classification error for discrete data”, 2008 Elsevier, Pattern Recognition 41 (2008) 3692 – 3705

[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] 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.

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

[5] Tahani Aljuaid and Sreela Sasi, “Proper Imputation Techniques for Missing Values in Data sets”, 978-1-5090-1281-7/16, IEEE International Conference on Data Science and Engineering (ICDSE) 2016

[6] Thomas R. Sullivan, Amy B. Salter, Philip Ryan and Katherine J. Lee ,“Bias and Precision of the “Multiple Imputation, Then Deletion” Method for Dealing with Missing Outcome Data”, American Journal of Epidemiology, Volume 182, Issue 6, September 2015, Pages 528–534

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

[8] Vapnik V N, “Statistical Learning Theory”, John Wiley and Sons, New York, USA 1998

[9] Vapnik V N, “The Natural of Statistical Learning Theory”, Springer-Verleg, New York, USA 1995.