An Exploratory Investigation into Missing Value Imputation Techniques for Classification Tasks

Kandikunta Rohith Venkat Sai

Department of Computer Science Sri Venkateswara University, Tirupati

Abstract— *In information mining, the crucial task of data preprocessing is handling missing values. Attribution involves replacing missing data with substituted values, a measurable process. Many clinical datasets commonly suffer from missing values. Excluding datasets with missing values can exacerbate issues. Traditional attribution methods, though simple, introduce bias. This paper proposes a data attribution system using K-Nearest Neighbors (KNN) to address missing data issues. The system integrates KNN with a Support Vector Machine (SVM) model for adaptable attribution. The aim is to mitigate the impact of missing data on learning disclosure measures in data mining tasks. Managing missing features during dataset processing poses a challenge. AI techniques for missing value attribution are explored using mammogram mass data from the UCI repository. Results indicate improved classifier performance with SVM.*

I. INTRODUCTION

Missing data, often referred to as missing characteristics, denotes the absence of data values for a variable within the scope of interest. It stands as one of the most prevalent challenges encountered by AI practitioners when dealing with real-world datasets. Across various applications spanning from gene expression analysis in computational biology to survey responses in social sciences, missing data manifests to varying extents. Given that many statistical models and machine learning algorithms rely on complete information sets, effectively addressing missing data becomes paramount.

Addressing missing data attribution represents a significant and complex problem in both AI and data mining. From the collection of samples through field experiments and clinical trials to data analysis, numerous hurdles present themselves at each stage of the mining process. It has been a persistent issue in data analysis, as missing values from the outset of data collection can introduce biases that impact the quality of subsequent analyses. Therefore, it is expected that missing attributes are identified and addressed before delving into the analysis of empirical data.

Several methods for handling missing data have been proposed over time, yet there is no universally superior attribution technique. The aim of missing value imputation methods is to fill in the missing entries of the dataset using the available information within the dataset. It is imperative to navigate through the complexities of missing attributes before applying any data mining techniques; otherwise, data derived from educational records containing missing values could lead to erroneous conclusions. To enhance the accuracy of predictions based on the available data, missing values in the dataset should either be removed or imputed during the preprocessing stage before further analysis.

In general, dealing with missing data involves addressing two distinct challenges: handling missing values and model building. This study introduces the application of the K-Nearest Neighbors (KNN) algorithm for imputation and classification of incomplete data. In the initial phase, missing values are imputed using KNN, followed by model classification accuracy evaluation using a Support Vector Machine (SVM) classifier on the imputed dataset.

II. MISSING DATA HANDLING MECHANISMS

In data mining, various approaches have been employed to handle missing attributes within datasets. One approach involves disregarding data instances with missing attributes, while another entails utilizing a global constant to fill in missing values (e.g., the mean of the attribute or class). Alternatively, computational methods can be employed to estimate missing values. These techniques aim to impute missing values within a dataset, allowing for the application of standard analysis methods that require complete data.

Missing data imputation techniques encompass both single imputation methods, which fill in one value for each missing entry, and multiple imputation methods, which replace missing values with multiple possible estimates, thereby capturing the sampling variability around the true value more accurately. Additionally, there are nonparametric missing data imputation methods, which include both probability-based and non-probability-based approaches. Single imputation methods are more commonly utilized and involve filling in a single value for each missing entry, while multiple imputation methods offer a more

comprehensive representation of uncertainty by replacing missing values with a range of possible values.**2.1 Strategies for Handling Missing Data**

2.1 Mean Imputation

A hero among the basically now and again utilized procedures. This is the most simple methods for managing property missing information is to supplant each missing an inspiration with the mean of non-missing appraisals of the variable [6]. This procedure additionally its hindrances the dispersing of the ascribed variable can get uncommonly distorted, considering how each missing worth is allocated a tantamount credit.

2.2 Lit wise deletion

In this technique, cases with any missing attributes are erased from an assessment. It is additionally called outright case evaluation, considering the way that just cases with complete information are held.

III. METHODOLOGY

3.1 Missing qualities utilizing K-Nearest Neighbor (KNN)

The K-Nearest Neighbor (KNN) is one of the attribution techniques used to treat missing worth. KNN credit approaches are neighbor based strategies where the ascribed respect is either a respect that was evaluated for the neighbor or the regular of surveyed respects for various neighbors [2] [3]. It is a fundamental and shocking technique. The inspiration driving the KNN calculation is that models with comparable elements have relative yield respects. The assessment chips away at the clarification that the attribution of the dull models should be possible by relating the dim to the known by some segment or closeness work [4].

KNN is the most clear assessment in ascribing missing attributes. In this strategy the missing appraisals of an occasion are credited a huge load of closest neighbor for a model and substitutes the missing information by computing the standard of non-missing attributes to its neighbors. The closeness of two models is settled utilizing a parcel work. Parcel breaking point can be Euclidean and Manhattan. In this work we have considered the Euclidean segment work. Precisely when the k-closest neighbors' technique is related with the test information, the suspicion execution yields result nearest to those for the fundamental information with no missing qualities, and the figure model's show is reliable notwithstanding while the missing information rate increments.

3.2 Support Vector Machine

The SVM is one more kind of AI methods reliant upon quantifiable learning theory. Because of incredible headway and a higher precision, SVM has turned into the investigation point of convergence of the AI social class. SVMs are set of related controlled learning procedures used for gathering and backslide [8]. A couple of progressing assessments have uncovered that the SVM generally are good for conveying better to the extent that request precision than the other data gathering estimations. SVM depends on genuine learning speculation by Vapnik et al proposed one more learning system, which depends on a set number of tests in the information contained in the current planning text to get the best gathering results.

An exceptional property of SVM can't avoid being, SVM meanwhile limit the observational request botch and grow the numerical edge. So SVM called Maximum Margin Classifiers. SVM relies upon the Structural peril Minimization. SVM map input vector to a higher layered space where a maximal detaching hyperplane is created. Two equivalent hyperplanes are based on each side of the hyperplane that different the data. The segregating hyperplane is the hyperplane that support the distance between the two equivalent hyperplanes. A notion that is made that the greater the edge or distance between these equivalent hyperplanes the better the speculation.

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

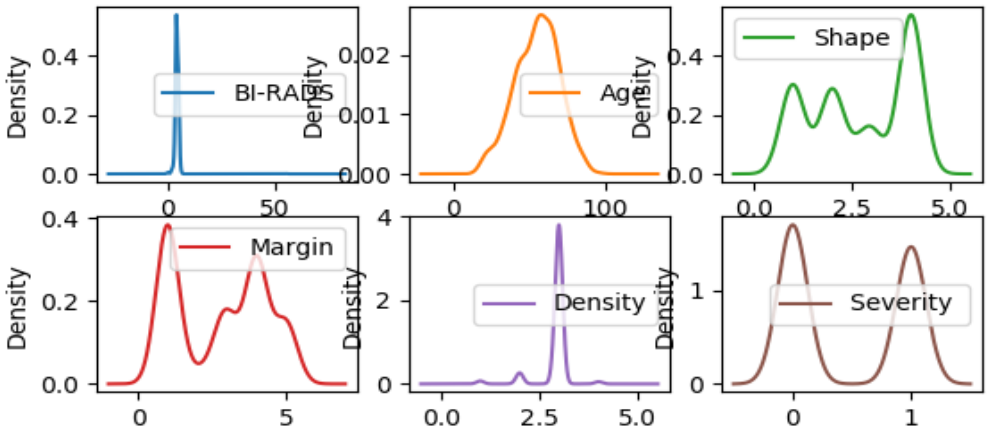

Figure-1: Density plot of Data distribution of each attribute

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.

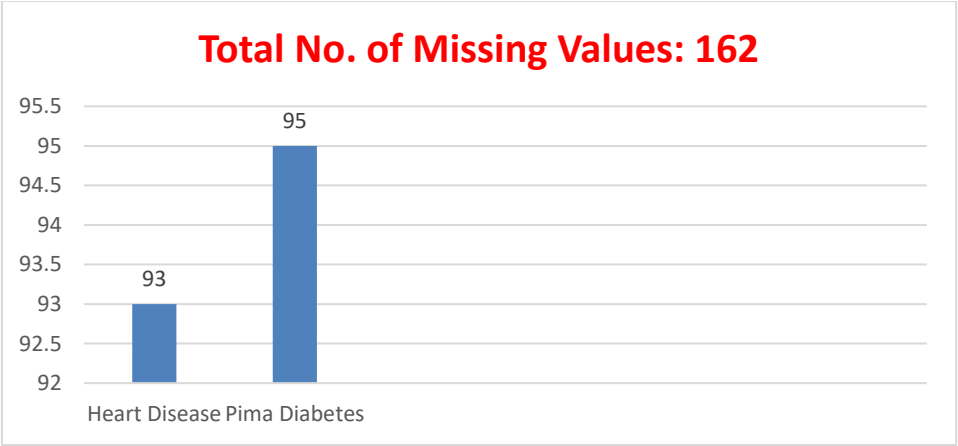

Figure-2: Attribute wise Missing values

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.

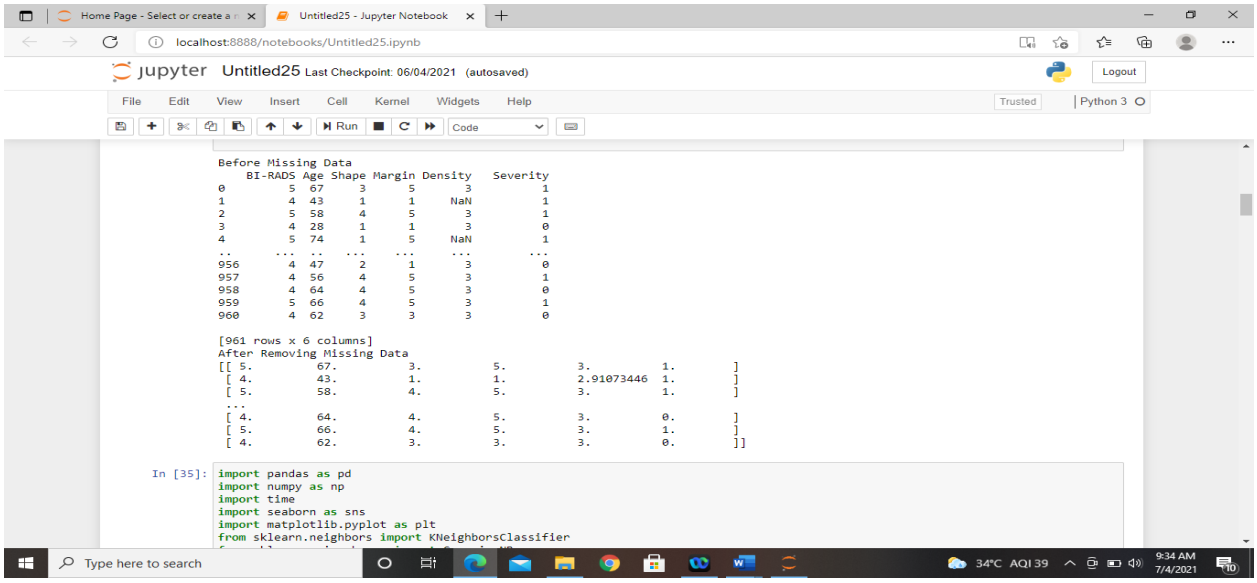

Figure-3: Results of Missing data

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.

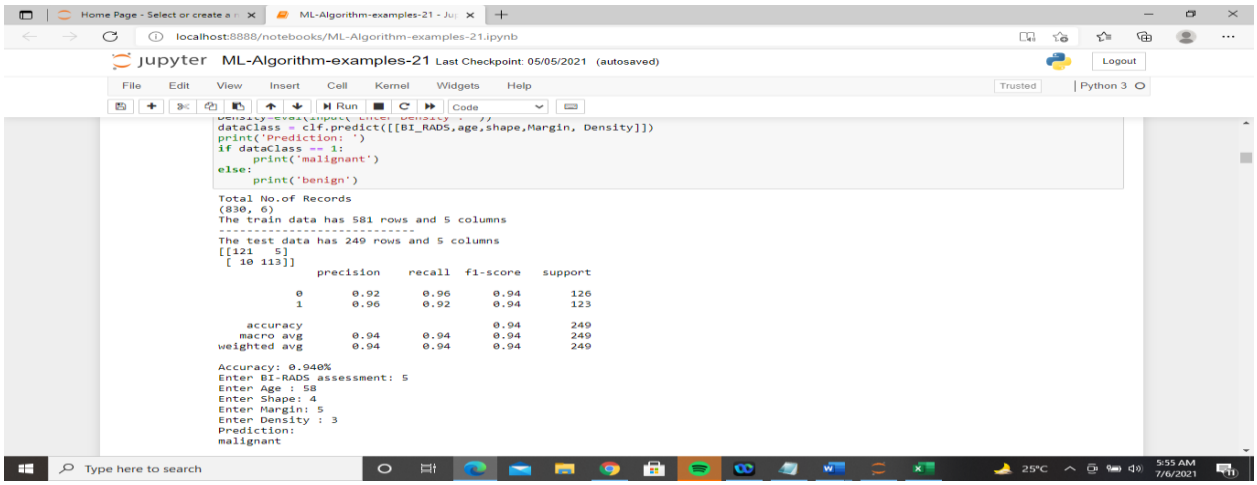

Figure-4: SVM Results after Impute the missing values

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 evaluates existing methods for imputing missing values and introduces a novel approach to address missing data situations, enhancing input accuracy for SVM classifiers and improving prediction, classification, and treatment of mammographic data. The proposed KNN data attribution technique serves as an effective imputation method for SVM classification in the presence of missing data.

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