Handling Missing Attributes Using Imputation Methods

Bandaru Owmyaswini

Department of Computer Science Sri Venkateswara University, Tirupati

Abstract— *Dealing with missing data is a common challenge in data quality. Many real-world datasets contain missing values. Imputing these missing values facilitates analysis by creating a complete dataset, thereby eliminating the complexity of handling intricate patterns of missingness. While conventional imputation methods are easy to implement, they introduce bias into the data. This approach combines the K-Nearest Neighbors predictive model with PART and Ripper algorithms adapted for missing value imputation. The objective of this evaluation is to assess the impact of missing data on the data mining task of learning discovery processes. The initial step in handling the dataset itself may pose challenges, as it requires addressing missing attributes.*

I. INTRODUCTION

Missing data, also referred to as missing attributes, pertains to data values that are absent for a variable of interest. Addressing the missing data challenge is arguably one of the most common issues encountered by AI practitioners when analyzing real-world data. In various applications spanning from gene expression in computational biology to sentiment analysis in social networks, missing data exists to varying extents. Given that many statistical models and AI algorithms require complete datasets, it is crucial to handle missing data appropriately. Missing data imputation poses a significant and complex challenge in AI and data mining. From sample collection to field experiments and clinical trials to data analysis, numerous challenges arise at each stage of the mining process. Addressing missing values is imperative as incomplete data collected since the inception of data collection may introduce biases affecting the quality of subsequent analyses. Therefore, missing values should be addressed and replaced before extracting meaningful insights.

Several missing data imputation methods have been proposed, yet there is no universally superior imputation technique. The objective of these methods is to estimate the missing values of the data using the available information. It is essential to address the issue of missing attributes before applying any data mining technique; otherwise, data extracted from incomplete records containing missing values may lead to erroneous conclusions. To enhance the accuracy of predictions using the available data, missing values in datasets should be removed or imputed in the preprocessing stage before utilizing the data for analysis.

In general, handling missing data in model building involves two distinct challenges: addressing missing values and model construction. This work introduces the application of the KNN algorithm for imputing and assembling incomplete data. In this approach, missing attributes are imputed using KNN in the initial stage, followed by classification accuracy assessment using an SVM classifier on the imputed dataset.[1,6,7]

II. 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 common of surveyed respects for various neighbors [4]. It's anything but a fundamental and dazzling system. The inspiration driving the KNN calculation is that models with comparable highlights have relative yield respects. The assessment deals with 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 [9].

KNN is the clearest assessment in crediting missing qualities. In this strategy the missing evaluations of an occasion are attributed a huge load of closest neighbor for a model and substitutes the missing information by computing the customary of non-missing qualities to its neighbors [2][3]. The closeness of two models is settled utilizing a parcel work. Segment breaking point can be Euclidean and Manhattan. In this work we have considered the Euclidean parcel work. Precisely when the k-closest neighbors' strategy is related with the test information, the presumption execution yields result nearest to those for the principal information with no missing attributes, and the figure model's show is steady regardless of when the missing information rate increments.

III. RIPPER CALCULATION

The Repeated Incremental Pruning to Produce Error the Repeated Incremental Pruning to Produce Error Reduction (Ripper) is a characterization calculation intended to create rules set straightforwardly from the preparation dataset. The name is drawn from the way that the guidelines are adapted steadily. Another standard related with a class worth will cover different properties of that class. The calculation was intended to be quick and viable when managing huge and boisterous datasets contrasted with choice trees. During the developing period of the calculation, an avaricious methodology of learning is applied, for example each standard is learned each in turn. In datasets with exceptionally huge measurements, this causes over-fitting of the information. This thus expands the order mistake rate essentially if the calculation is tried with information with missing qualities [10].

3.1 Part

PART is a different and-vanquish rule student. The calculation delivering sets of rules called choice records which are arranged arrangement of rules. Another information is contrasted with each standard in the rundown thusly, and the thing is relegated the class of the primary coordinating with rule. PART constructs a fractional C4.5 choice tree in every cycle and makes the best leaf into a standard [5].

IV. EXPERIMENTAL RESULTS

This part will give an outline over the accomplished outcomes, the pre-owned information and the examination interaction to order. We have considered the Annealing information from UCI Machine Learning Repository dataset [8]. The appraisals have been driven by utilizing WEKA. It gives numerous information mining calculations and representation instruments for information examination and prescient demonstrating, with graphical UIs that assists client with effectively running these calculations on datasets. WEKA upholds a few standard information mining errands, that are, information preprocessing, characterization, relapse, grouping, highlight choice, and representation.

4.1 Dataset

The Annealing Data set has 798 lines and 38 sections. In this information there are 6 class names, class astute frequencies are displayed in the figure-1 and furthermore the factual synopsis of each property is introduced in figure-2. The standard dataset is isolated into two sets (70% and 30%), one for preparing and another set for testing.

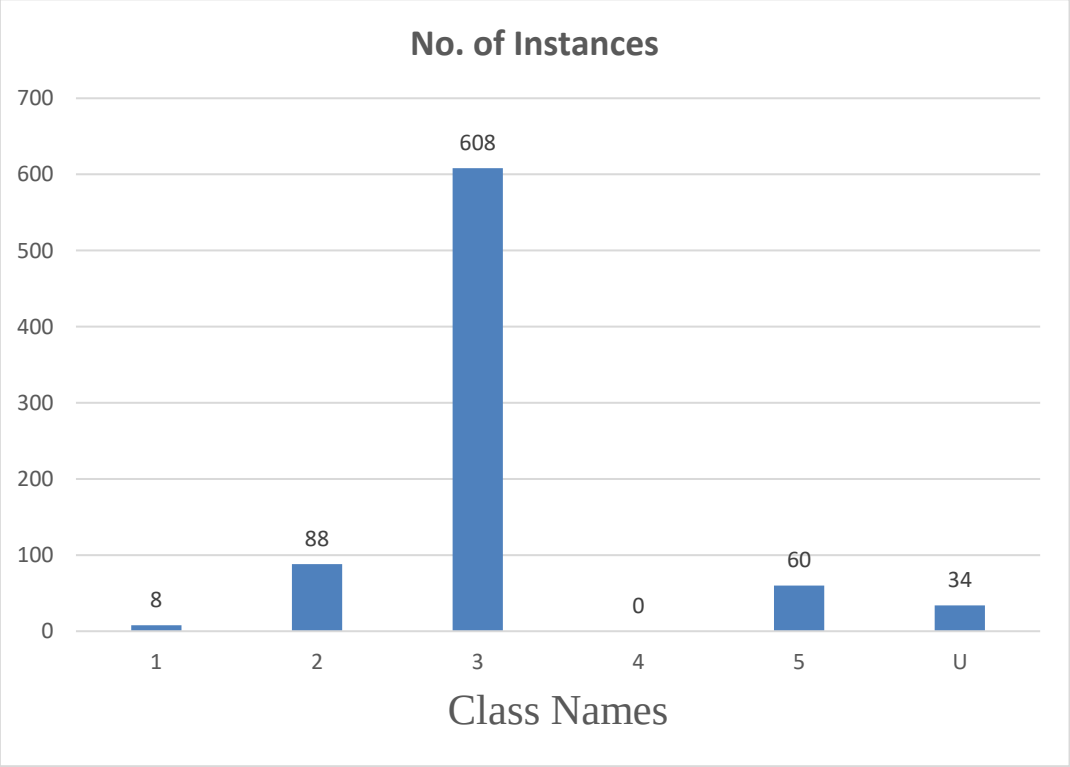

Figure-1: Class wise distribution of six labels

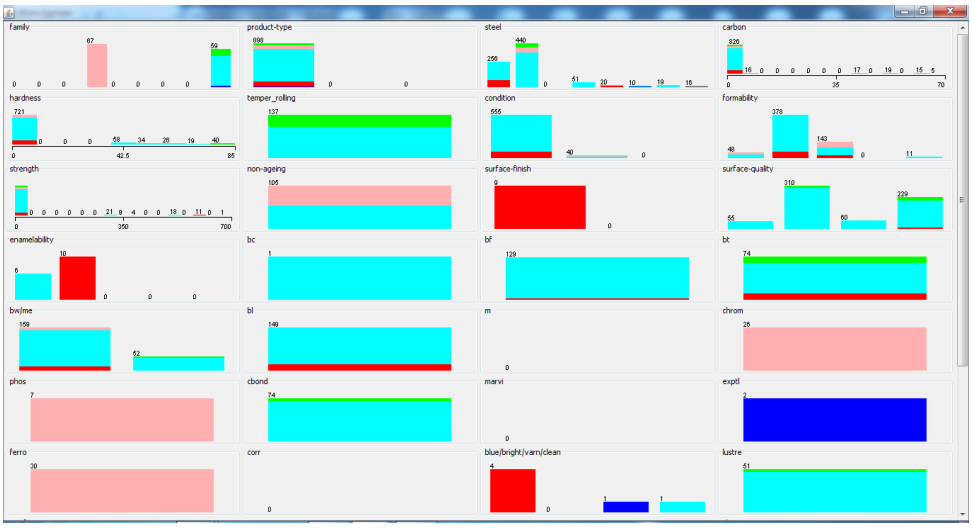

Figure-2: Statistical summary of dataset

4.2 Results

Our exploratory arrangement went through two stages. Two analyses have been directed for assessing the Ripper and PART calculations with KNN Imputation strategy for missing information. In the primary stage, we thought about the exhibition of the two classifiers with no missing qualities for the dataset. We prepared the classifiers on the preparation dataset utilizing Stratified Cross-Validation of 10-folds. In the subsequent stage, we eliminate the missing qualities utilizing KNN attribution calculation after that we applied the two grouping calculations (Ripper and PART). The consequences of two stage are sums up execution measurements for the prepared models are displayed in the figure-3.

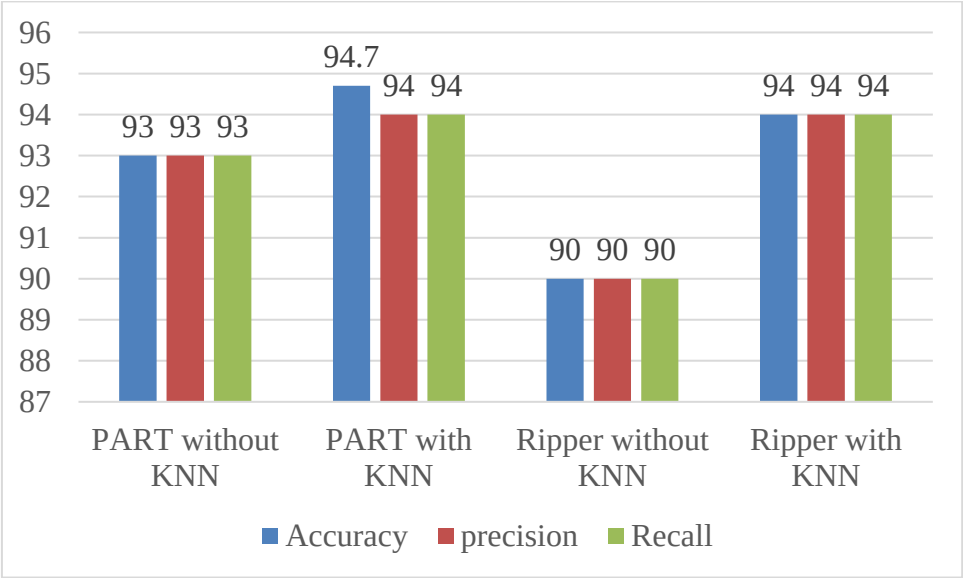

Figure-3: Performance metrics

From the figure-3, we notice the exhibition of PART and Ripper calculations with KNN attribution and without KNN ascription. The PART without KNN dependent on precision has 93%, though the exhibition of PART with KNN dependent on exactness has accomplished 94.7%. Be that as it may, there is an improvement in the precision with KNN missing attribution. The precision rate is expanded 1.4% with KNN attribution.

Likewise, we notice the presentation of Ripper without KNN dependent on precision has 90%, while the exhibition of Ripper with KNN dependent on exactness has accomplished 94%. In any case, there is an improvement in the exactness with KNN missing ascription. The precision rate is expanded 4% with include determination.

4.3 Screenshots

The experimental results are shown in the screen shots from the figures-4 to figures-7

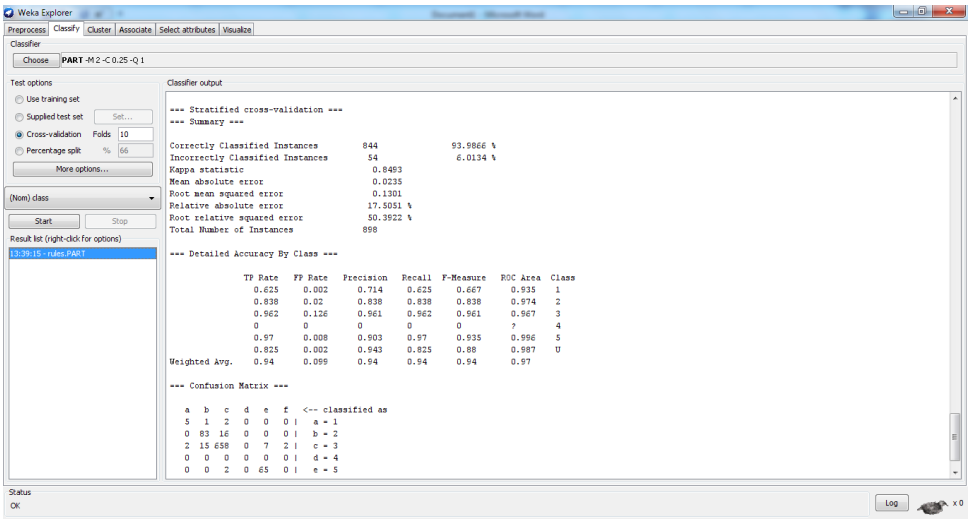

Figure-4: Experimental results of PART without KNN

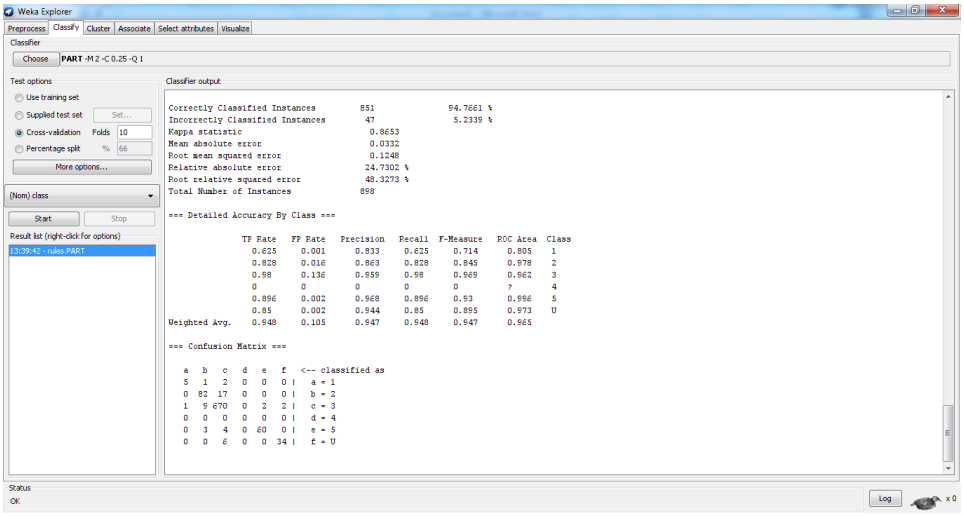

Figure-5: Experimental results of PART with KNN

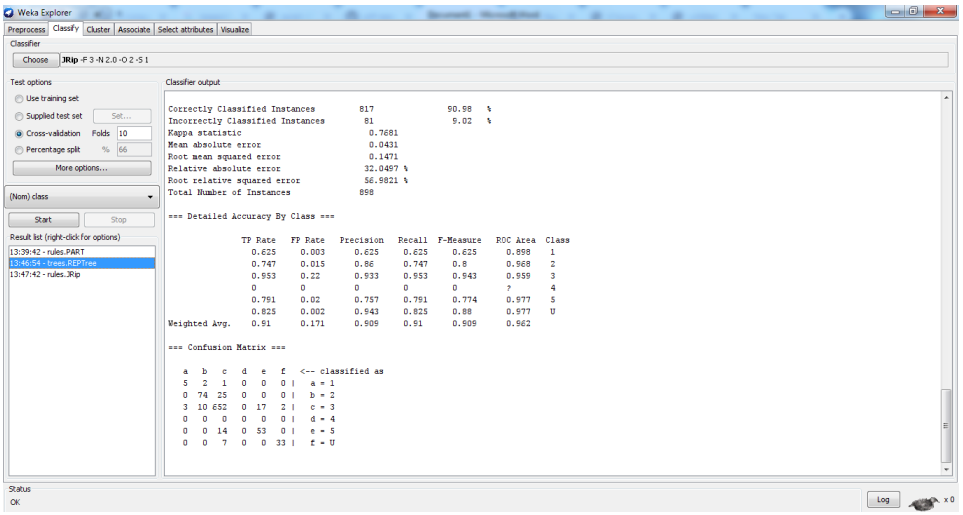

Figure-6: Experimental results of Ripper without KNN

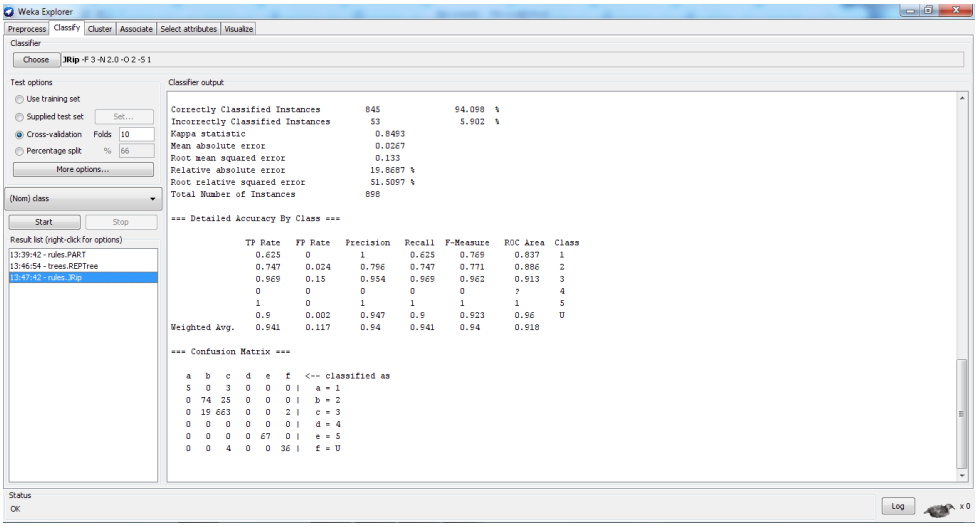

Figure-7: Experimental results of Ripper with KNN

V. CONCLUSION

This paper 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 PART and Ripper classifier to get better prediction. The proposed KNN data imputation method serves as an effective data imputation method for PART and Ripper 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] Keerin P, Kurutach W, Boongoen T (2012) Cluster-based KNN missing value imputation for DNA microarray data. In: 2012 IEEE International conference on systems, man, and cybernetics (SMC). IEEE, pp 445–450

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

[6] 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

[7] Thomas R. Sullivan, Amy B. Salter, Philip Ryan and Katherine J. Lee ,“Bias and recision 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

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

[9] Zhang S (2012) Nearest neighbor selection for iteratively kNN imputation. J Syst Softw 85(11):2541–2552

[10] Zantema, H., and Bodlaender H. L., Finding Small Equivalent Decision Trees is Hard, International Journal of Foundations of Computer Science, 11(2):343-354, 2000. http://dx.doi.org/10.1142/S0129054100000193.