An Experimental approach on Missing Value Imputation for Classification

Daggupati Venugopal

Department of Computer Science, Sri Venkateswara University, Tirupati

Abstract— *In information mining measure the best task of data preprocessing is missing worth credit. Attribution is a quantifiable cooperation of replacing missing data with subbed values. Various clinical scientific datasets are typically lacking. Banning lacking dataset from the first dataset can bring more issue than adjustments. The customary techniques for attributions are quite easy to do yet introduce biasness in the data. In this paper, a data attribution system with a K-Nearest Neighbors (KNN) is proposed to settle the issue of missing data. This framework combines K-Nearest Neighbors insightful model for Support Vector Machine (SVM) be adapted to various attribution. The objective of this assessment is to address the impact of missing data on the data mining undertaking of learning disclosure measure. The secret stage in dealing with the dataset may itself challenge since this advancement requires overseeing missing characteristics. In our paper the AI methodologies for missing worth credit have been researched using Mammogram mass data from UCI file. The examinations have shown that the last classifier execution increases when Support Vector Machine (SVM) is used*.

I. INTRODUCTION

Missing data (or missing characteristics) is portrayed as the data regard that isn't taken care of for a variable in the view of interest. The missing data issue is apparently the most broadly perceived issue experienced by AI specialists while stalling genuine data. In various applications going from quality verbalization in computational science to outline responses in social sciences, missing data is accessible to various degrees. As various verifiable models and AI estimations rely upon complete enlightening assortments, it is basic to reasonably manage the missing data. Missing information credit is a veritable and testing issue in AI and information mining. Beginning from the social gathering of tests through field tests and clinical preliminaries to performing depiction, there are various difficulties at each stage in the mining strategy. It is has been an inescapable issue in information appraisal since the beginning of information assortment can have inclination that effects the possibility of the canny social occasion introductions. So missing attributes ought not out of the ordinary and supplanted preceding examining remedial information [1].

Two or three missing quality credit procedures were proposed recorded as a printed copy and there exists no generally best attribution strategy. The objective of missing worth credit strategies is to fill the missing appraisals of the article involving the open data in the thing. It is crucial for manage the maze of missing attributes before applying any procedure of information mining; overall, the data disengaged from instructive record containing missing qualities will instigate the technique for wrong principal drive. To chip away at the accuracy of suspicion with the accommodating information, missing a catalyst from dataset ought to be removed or credited in the pre-arranging stage preceding involving the information for figure.

By and large, depiction with missing information concerns two obvious issues, managing missing qualities and model social event. This work isolates the acquaintance of the KNN calculation with credit and assembling inadequate information. Utilizing this technique, in a first stage, the missing qualities are credited with KNN, and beginning there ahead, the game-plan exactness is performed by a SVM classifier utilizing the changed set.

II. MISSING DATA HANDLING MECHANISMS

A couple of procedures have been applied in data mining to manage missing characteristics in informational collection. Data with missing characteristics could be dismissed, or an overall steady could be used to fill missing characteristics (dark, not material, perpetuation, for instance, trademark mean, attribute mean of a comparable class, or a computation could be applied to find missing characteristics [5]. Missing data attribution technique suggests a philosophy to fill missing potential gains of an enlightening list to apply standard procedures which require completed instructive assortment for assessment. These techniques hold data in insufficient cases, similarly as attribute potential gains of related factors.

Missing data attribution procedures are designated irrelevant missing data credit methods, which consolidate single attribution strategies and different attribution strategies, and non-immaterial missing data credit procedures which integrate likelihood based techniques and the non-likelihood based methodologies. A single attribution procedure could fill one motivating force for each missing worth and it is more generally used at present than various credits which displace each missing worth with a couple of possible characteristics and better reflects testing variability about certified worth.

2.1 Strategies for Handling Missing Data

- **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.

- **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

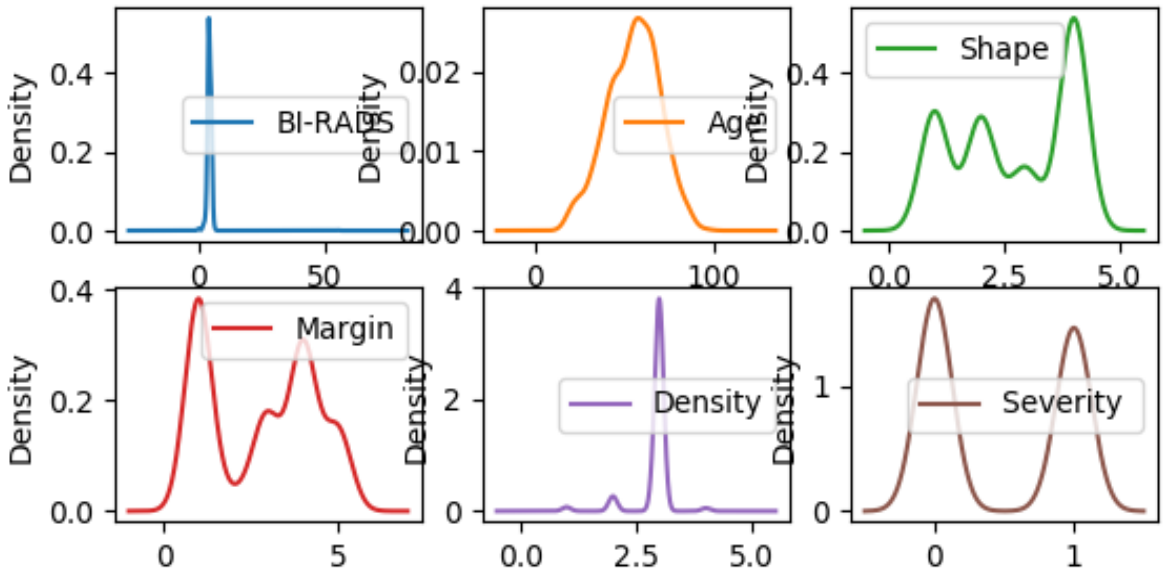

Figure-1: Density plot of Data distribution of each attribute

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.

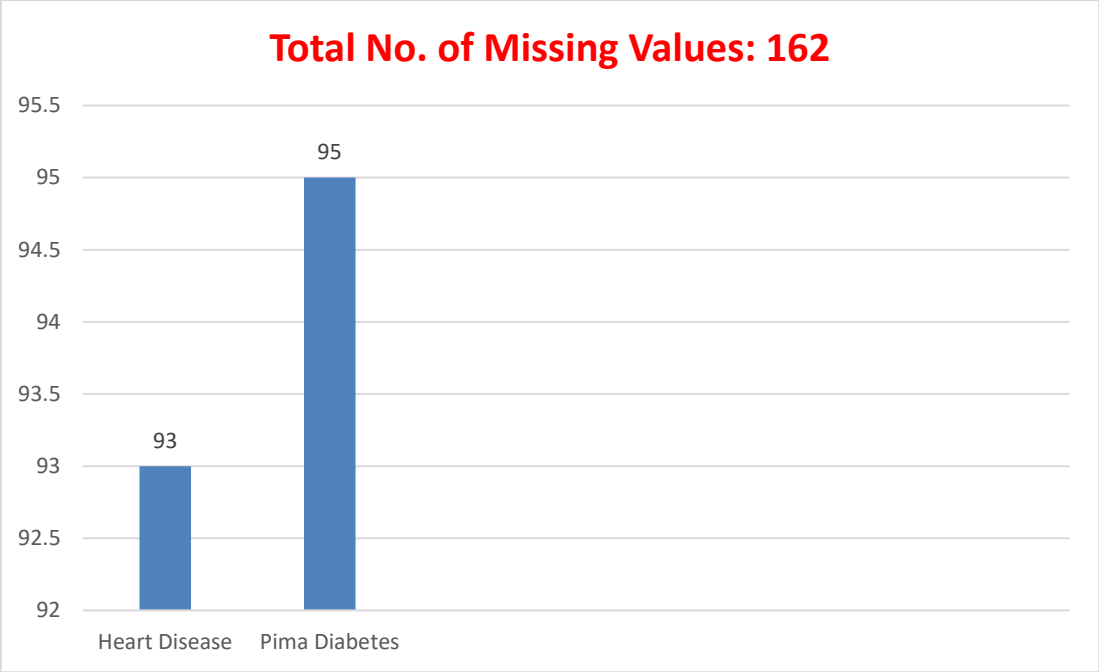

Figure-2: Attribute wise Missing values

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.

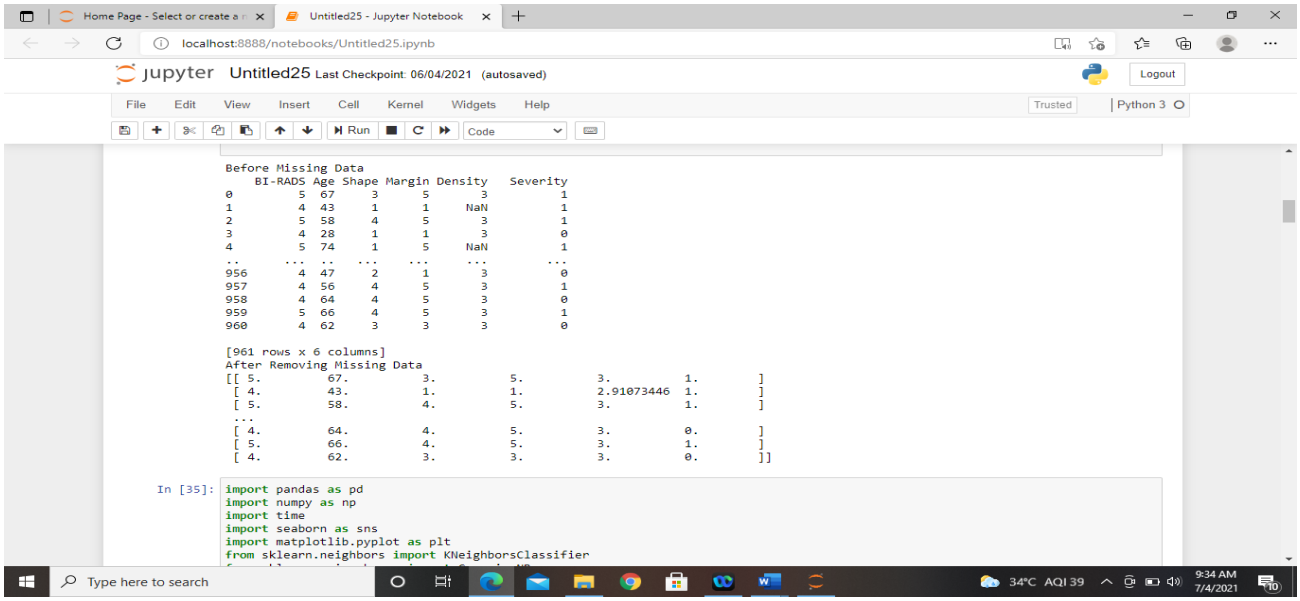

Figure-3: Results of Missing data

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.

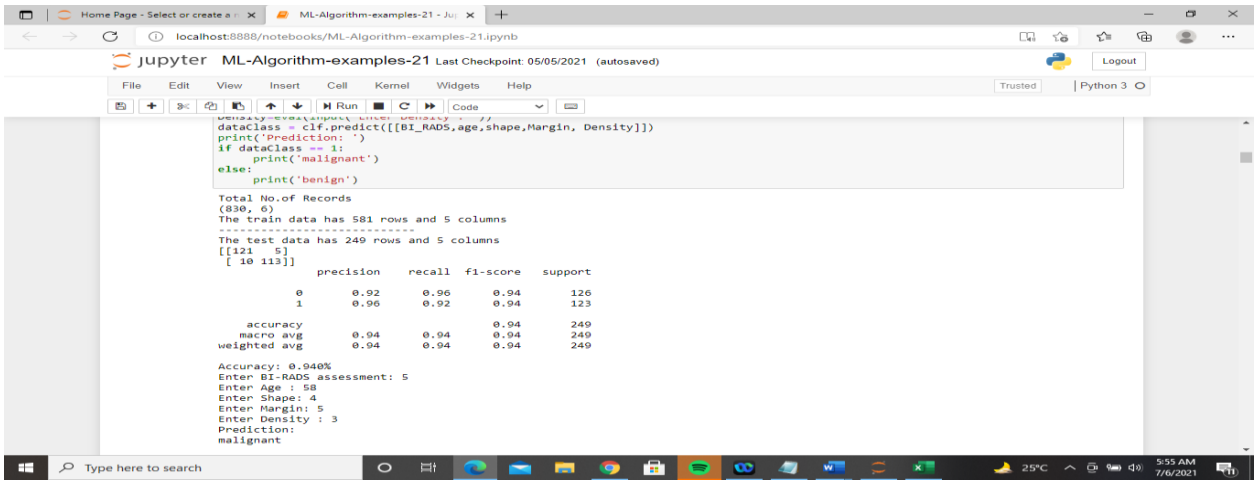

Figure 4: SVM Results after Impute the missing values

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 likewise assesses approaches used to fill missing qualities and proposes a new and better way to deal with handle missing worth circumstance and subsequently empowering to take care of right contribution to the SVM classifier to get better forecast, determination and treatment of the mammographic information. The proposed KNN information attribution technique fills in as a compelling information ascription strategy for SVM characterization on account 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