Assessing the Impact of Missing Data Imputation with K-Nearest Neighbors on the Performance of Decision Tree Classification for Mammographic Data Prediction

P. Chandrika

PG Scholar, Dept. of Computer Science Sri Venkateswara University, Tirupati

Abstract— *Missing data is a common challenge in the field of medical data analysis, particularly when predicting outcomes from mammographic data. The missing data is one of the typical issues of data quality. An enormous part of the certified datasets have missing characteristics. Crediting the missing characteristics simplifies the assessment by making an all out dataset as it kills the issue of managing complex instances of missingness. In this research paper, we investigate the impact of missing data on the performance of supervised learning models, specifically decision tree classification, when applied to mammographic data. The objective of this examination is to address the impact of missing data on the data mining errand of learning disclosure measure. The principal stage in dealing with the dataset may itself challenge since this improvement requires overseeing missing properties.*

I. INTRODUCTION

Missing information (or missing qualities) is depicted as the information respect that isn't dealt with for a variable in the perspective on interest. The missing information issue is evidently the most by and large saw issue experienced by computer based intelligence experts while isolating certifiable information [1][2]. In different applications going from quality articulation in computational science to frame reactions in humanistic frameworks, missing information is open to different degrees. As different valid models and simulated intelligence calculations depend upon complete informational varieties, it is essential to sensibly deal with the missing information. Missing data credit is a true and testing issue in artificial intelligence and data mining. Starting from the party of tests through field tests and clinical preparations to performing portrayal, there are different challenges at each stage in the mining system. It is has been an undeniable issue in data assessment starting from the start of data arrangement can have affinity that impacts the chance of the wise social event presentations [10]. So missing characteristics should be depended upon and replaced prior to researching accommodating data.

A few missing quality credit methods were proposed recorded as a printed rendition and there exists no generally best attribution procedure. The goal of missing worth credit strategies is to fill the missing assessments of the article using the open information in the thing. It is significant for deal with the labyrinth of missing characteristics prior to applying any method of data mining; all over, the information confined from educational record containing missing characteristics will prompt the procedure for wrong significant drive [11]. To work on the precision of assumption with the steady data, missing a motivation from dataset should be removed or credited in the coordinating stage going before using the data for figure. Generally speaking, plan portrayal with missing data concerns two irrefutable issues, overseeing missing characteristics and model social affair.

II. 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 procedures where the credited regard is either a regard that was assessed for the neighbor or the normal of overviewed regards for different neighbors [1]. It's everything except an essential and stunning framework. The motivation driving the KNN estimation is that models with tantamount features have relative yield regards. The evaluation manages the explanation that the attribution of the dull models ought to be conceivable by relating the faint to the known by some fragment or closeness work [9].

KNN is the most clear evaluation in crediting missing characteristics. In this procedure the missing assessments of an event are credited a tremendous heap of nearest neighbor for a model and substitutes the missing data by figuring the standard of non missing characteristics to its neighbors [10][11]. The closeness of two models is settled using a package work. Fragment limit can be Euclidean and Manhattan. In this work we have considered the Euclidean bundle work. Definitively when the k-nearest neighbors methodology is connected with the test data, the assumption execution yields results closest to those for the rule data with no missing credits, and the figure model's show is consistent paying little heed to while the missing data rate increases.

III. METHODOLOGY

The study involves the following steps:

1. Preprocessing the mammographic data, including handling missing values using KNN imputation.
2. Constructing decision tree classification models with varying levels of missing data imputation.
3. Evaluating the classification performance using standard metrics such as accuracy, precision, recall, and F1-score.
4. Comparing the results to understand the impact of KNN imputation on classification performance in the presence of missing data.

3.1 Decision Tree Classifier

Decision tree hypothesis is a regularly utilized information revealing technique for setting depiction structures subject to various covariates or for making presumption calculations for an objective variable [3][4]. This strategy portrays an all inclusive community into branch-like parts that cultivate a bothered tree with a root place point, inner focuses, and leaf focuses. The assessment is non-parametric and can competently regulate monster, tangled datasets without convincing a muddled parametric advancement [5]. Decision trees are classifiers that address their depiction information in tree structure. Every inside place point of a decision tree is a test on a property. Fulfilling that test causes the case being depicted to dispose of one branch from that middle point, attacking the test makes the model take the other branch [6][7].

A Decision tree is utilized to pack a model by beginning at the root community point of the decision tree and following how the property tests direct until a leaf place is fit [4]. Each leaf community point in a decision tree is a choice, i.e., addresses a solicitation. An occasion that breezes up at some specific leaf place point is engineered with the class allocated to that leaf community. A second sort of tree is a class likelihood tree. This has a vector of class probabilities at each leaf instead of a choice. The significant assessment fabricates a tree top down utilizing the standard voracious solicitation rule, taking into account recursive isolating. The allotting wires halting, isolating and pruning rules. Precisely when the model size is sufficiently huge, concentrate on information can be disengaged into arranging and support datasets. Utilizing the arranging dataset to gather a choice tree model and a support dataset to pick the fitting tree size expected to accomplish the best last model.

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 [12] 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

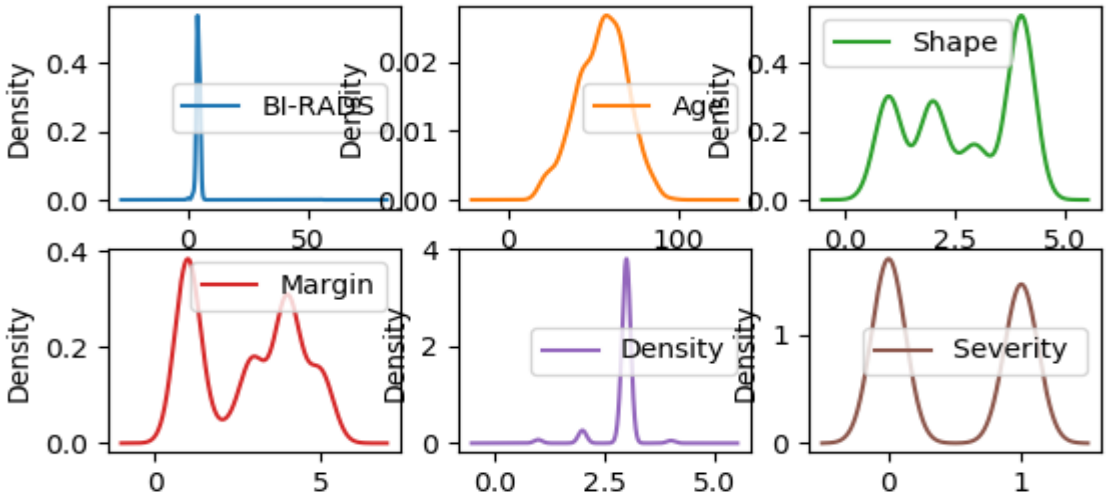

Figure-1: Density plot of Data distribution of each attribute

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 decision tree Classification with KNN Imputation method for missing data. In our Experiment the first step is data pre processing 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.

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.

In the second stage we execute a Decision Tree calculations for forecast of Severity of mammographic dataset. The outcomes that we got for decision tree as displayed in the table-2 with their comparing esteems.

Table-2
Decision Tree Performance

Algorithm	Accuracy	Precision	Recall
Decision Tree	95.87	96	96

From the table-2, we observe the performance of decision tree accuracy has got 95.87%. This research proposes an approach for enhancing the training process of decision tree when dealing with missing data.

The results of our study indicate that the presence of missing data has a notable impact on the performance of decision tree classification models for mammographic data prediction. Without adequate handling of missing data, classification accuracy and other performance metrics suffer, leading to less reliable predictive models.

However, our experiments demonstrate that K-Nearest Neighbors imputation is an effective approach for addressing missing data in mammographic datasets. By imputing missing values using KNN, we observed a significant improvement in classification accuracy and the overall performance of decision tree models. This suggests that KNN imputation is a valuable pre processing step when working with mammographic data, as it helps enhance the robustness of predictive models.

Furthermore, our study highlights the importance of selecting an appropriate value for the KNN imputation parameter, as the choice of K can impact imputation accuracy and, subsequently, classification performance. Careful parameter tuning is crucial to achieving the best results.

V. CONCLUSION

In conclusion, this research underscores the significance of handling missing data in mammographic data analysis and establishes K-Nearest Neighbors imputation as an effective technique to mitigate the negative impact of missing data on decision tree classification performance. These findings contribute to the advancement of predictive modeling in breast cancer detection and diagnosis, ultimately improving healthcare outcomes.

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] D. Hand, H. Mannila, P. Smyth.: Principles of Data Mining. The MIT Press. (2001)

[3] G. Ravi Kumar, K. Tirupathaiah and Prof. B. Krishna Reddy, "Client Churn Prediction of Banking and fund industry utilizing Machine Learning Techniques", International Journal of Computer Sciences and Engineering, Volume-7, Issue-6, e-ISSN: 2347 — 2693, PP: 871-875, June 2019

[4] 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)

[5] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques.2nd ed. San Francisco: Morgan Kaufmann, 2005.

[6] 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.

[7] N.Michael, "Artificial Intelligence – A Guide to Intelligent Systems", 2nd Edition, Addison Wesley 2005

[8] P.N.Tan, M.Steinbach and V.Kumar "Introduction to Data Mining", A: Addision-Wesley, 2005.

[9] Surya Bhupal Rao, S.Rahamat Basha, G. Ravi Kumar, "A Comparative approach of Text Mining: Classification, Clustering and Extraction Techniques", Journal of Mechanics of Continua and Mathematical Sciences, Special Issue, No.-5, January (2020) pp 120-131, ISSN: 0973-8975.

[10] Tahani Aljuaid and SreelaSasi, "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

[11] 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

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