IMPLEMENTATION OF MISCLASSIFICATION DATA USING PREDICTIVE CLASSIFICATION TECHNIQUES

Dr.S.Kanchana

Assistant Professor, Department of Computer Science
Faculty of Science & Humanities, SRM Institute of Science & Technology, Kattankulathur 603202, Chennai, India

Abstract: The misclassification data leads to an extensive and persistent problem in many fields, which affect the accurate results that can be drawn from the analysis. The main action for dealing with missing values are to avoid, all incomplete data are implement with the rest of the information for designing process and change the incomplete values to perform the approximate calculation. This article proposes with the assistance of data mining approach, which can automatically predict whether the customer will subscribe a long term deposit. Research community proposed various suitable algorithms during the prediction of incomplete data analysis in terms of data mining approach. Various classifications of learning assignments are available in machine learning techniques including supervised and unsupervised techniques. This article describes the categories of various machine techniques in order to stand for the predictive analysis of misclassification data. The main advantage of these techniques is to generate more accurate data analysis without human expertise. This article organizes the process with few significant methods like Naïve Bayesian (NB), Decision Tree (DT), and Adaptive Boosting (ADAB) for the prediction of the misclassification data and also performs the comparison studies using predictive methods to find the best accuracy rate among them.

Keywords – Adaptive Boosting, Decision Tree, Machine Learning, Naïve Bayesian, Supervised, Unsupervised.

I. INTRODUCTION

The enormous development in the performance of accumulating, freezing and dispatching massive quantity of knowledge also expanded the quantity of data for knowledge discovery or data analysis applications. Data mining algorithm consist of two sections. Each section defines a number of data mining algorithms at a high level of techniques like classical and next generation techniques. Classical techniques [1] consist of statistics, neighbourhoods and clustering, and in next generation techniques reside trees, networks, and rules. These two sections have been split up based on when the data mining technique was established and when it became technically sophisticated enough to be used for trade, especially for helping the development of customer relationship management systems. Also, it helps to understand the uneven differences in the techniques and well equipped so as not to be confused by the vendors of different data mining tools. Statistical techniques are not data mining, which can be consumed by the data and used to design patterns and build predictive models. Clustering is the process of classification of physical or abstract objects into classes of similar objects and dissimilar objects into another cluster. The advantages of clustering techniques [2] are the flexibility to make changes and assist once in singling out useful features so that they identify different groups. Broadly used in many operations like consumer research, pattern recognition, data classification and digital image processing, Clustering help traders to determine specific groups in the customer base and also distinguish customer groups based on the purchasing patterns.

In the branches of life science, clustering helps in plant and animal classification, classify genes with related features and obtain foresight into frames which are inherent in populations. It assists in categorizing information on the network for knowledge discovery. Finally, missing value problem can be handled by various missing values attribution methods. Three main strategies are available to deal with missing data. Missing values attribution methods are suitable only for missing values introduced by Rubin (1976) who proposed the concept called Missing Completely At random (MCAR), which explained the complete option for a random case of an original set of data identified. Inferences identified only on the original data sets and also identified the complete cases of random are eligible to find the mechanism for missing data imputation. Complete case analysis proceeds with initial strategy applied by researchers to provide accurate inference for the final estimation. The syndrome of missing values is directly related to other reasonable clarification. The techniques are enhancing the probability of ignorable mechanism which is used for various operations in terms of capturing important information. In terms of learning, inference, and prediction theory, the incomplete analysis processes start with the study of missing data. This knowledge provides the difference between two basic classifications to handle the data which is missing at random and not missing at random.

II. LITERATURE REVIEW

Little and Rubin [3] summarize the mechanism of imputation method. Also introduces mean imputation [4] method to find out missing values. The drawbacks of mean imputation are sample size is overestimated, variance is underestimated, correlation is negatively biased. For median and standard deviation also replacing all missing records with a single value will deflate the variance and artificially inflate the significance of any statistical tests based on it. Different types of machine learning techniques are supervised and unsupervised machine learning techniques [5] summarized in. Classification of multiple imputation and experimental analysis [6] are described in Min Pan et al. [7] summarize the new concept of machine learning techniques like NBI also analysis the experimental results which impute missing values. Comparisons of different unsupervised machine learning techniques are referred from survey paper [8]. To overcome the unsupervised problem Peng Liu, Lei Lei et al. [9] applied the supervised machine learning techniques called Naïve Bayesian Classifier.

III. MACHINE LEARNING TECHNIQUES

Research community proposed various suitable algorithms during the prediction of incomplete data analysis in terms of data mining approach. Machine learning techniques mainly concentrate on the pattern recognition and computational learning theory [10, 11] in artificial intelligence which analyses the term and structure of algorithms in order to handle the predictions of missing data. This specific mechanism is based on framing a structure of algorithms that can learn and build predictions of missing data, which are nearly connected to perform the arithmetic operation through statistics analysis. In order to perform mathematical optimization in terms of machine learning techniques, choose the attributes from the set of available information from the different dataset and achieve the goal of prediction analysis. Learning techniques uses the various applications in terms of predicting data analysis [12] by focussing various basic predictive data analysis techniques in the name of unsupervised learning approach. In parliamentary law, to do predictive analysis by means of machine learning, statistical techniques are linked to each other in terms of methodological structure. All these analysis models provide more innovative information through factual communication to the researchers, developers, and system and data analysts regarding recent movement in data. Various classifications of learning assignments are available in machine learning techniques including supervised, unsupervised and reinforcement learning. This article describes the categories of various machine techniques in order to stand for the predictive analysis of missing data imputation.

Learning techniques depend on the various identical specifications of numerous issues, projected by using an unknown representation of data and are treated as main issues. Such techniques include Analytical Learning, Artificial Neural Network, Boosting, Decision Tree, Naïve Bayes Classifier, and Support Vector Machines from supervised and various clustering techniques from unsupervised machine learning techniques. The concept of machine learning algorithm is to assume the task of unknown data in order to perform the structured unpredicted data, if there are no problems or issues to estimate a skilled set of solution. The following figure 1. Explain the structure of existing system of machine learning techniques.

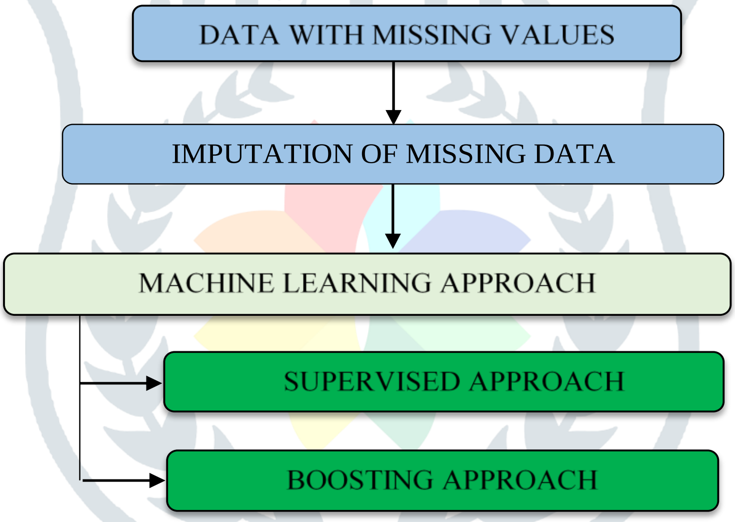

Figure 1. Existing System Structure

Unsupervised learning techniques hold more statistical techniques in terms of different estimation parameters. Self organizing map and adaptive resonance theory are basically implemented in the process of unsupervised machine learning techniques. These models are allow to find the nearest point in mapping the desired inputs and also it permits the number of grouped data which is ideal with many issues in size of the structure and the group of parameter. This technique is proposed to find out the cost estimation of minimized data in terms of priori assumptions. The cost estimation function in the dataset represents the mean value estimation of the concerned data set. Estimation of cost generation process produces more difficulties depending on the application of mutual datasets in terms of posterior probability.

The functional specification of unsupervised learning approach is commonly generalized in mathematical and statistical analysis and also grouping the similar dataset in terms of clustering analysis. The supervised approach defined from the training data is to find the accurate outcome of the testing data. Another part of ensemble approach in the machine learning techniques consists of various boosting approaches and apart from that, we have concentrated only on adaptive boosting techniques. This technique builds to solve the training data set issues by using statistical analysis. Impute the missing values in the training data set by applying three machine learning techniques such as Naïve Bayes Approach (NB), Decision Tree Approach (DT) and Adaptive Boosting Approach (ADAB).

IV. CLASSIFICATION OF PREDICTION TECHNIQUES

Generally machine learning techniques are analysed in different categories in terms of predicting the incomplete values of large datasets, which propose outstanding performance or response from learning system. Basically, the three objective learning techniques in machine learning classifications are supervised, unsupervised and reinforcement learning techniques, which propose to perform the accurate predictions depending on the prior observations based on the classification problem [13, 14]. The main advantage of these techniques is to generate more accurate data analysis without human expertise. Supervised techniques come under the process of machine learning task for providing specific inference from the identified data set. These techniques estimate various training datasets for producing optimal output result which is required for arranging unknown models. With the help of optimal algorithm it can be decided for correcting the class specification for identical attributes, which need to optimise label attributes in the desired data for producing the unexpected result. These techniques perform multiple processes to determine the specific model of practised datasets, such as to collect more information about existing dataset, to examine the value of specific input in terms of data representation, to analyse the complete framework for the existing procedure and respective algorithm, to design the structure of complete data structure, and to estimate the accurate value from the experienced task. In these techniques, cost function represents the mean squared error of mismatch data between the knowledge data in case of the problem domain. This mean squared error is commonly used to minimize the approximate error between the source and the target outcome pairs. Supervised learning techniques are commonly used in pattern recognition in terms of classification and approximation function. It is applicable only for consequence record set to provide sequential feedback about the accuracy of the outcome. The classification of supervised can perform various technical analysis to generate the unknown values in terms of testing data, through which the missing value can be imputed by analysing various learning model.

Bayesian technique was introduced by machine learning algorithm in order to estimate the probability value of given dataset with the classification of probability distribution model to generate unknown value in the dataset. Every attribute is treated as the estimation analysis which consists of training dataset in addition to adding all information in which the class attribute is assessed and then the information is tested in the database in which the attributes are missing. Combination of each class attributes along with non class attribute information is computed using prior probability model. This prior probability of each attribute can perform along with the class attribute of the predictive dataset in the missing values testing the imputation of missing attributes in supervised machine learning techniques. Naïve Bayesian classification model is the new treatment for missing data analysis to generate the unknown data from a large dataset in which to change the missing data. This classifier focuses on every class in training attribute and generates the unknown data in the sample classification techniques. Once the unknown values are imputed using statistical analysis, the accuracy rate of each statistical method in terms of different percentage of missing values is found.

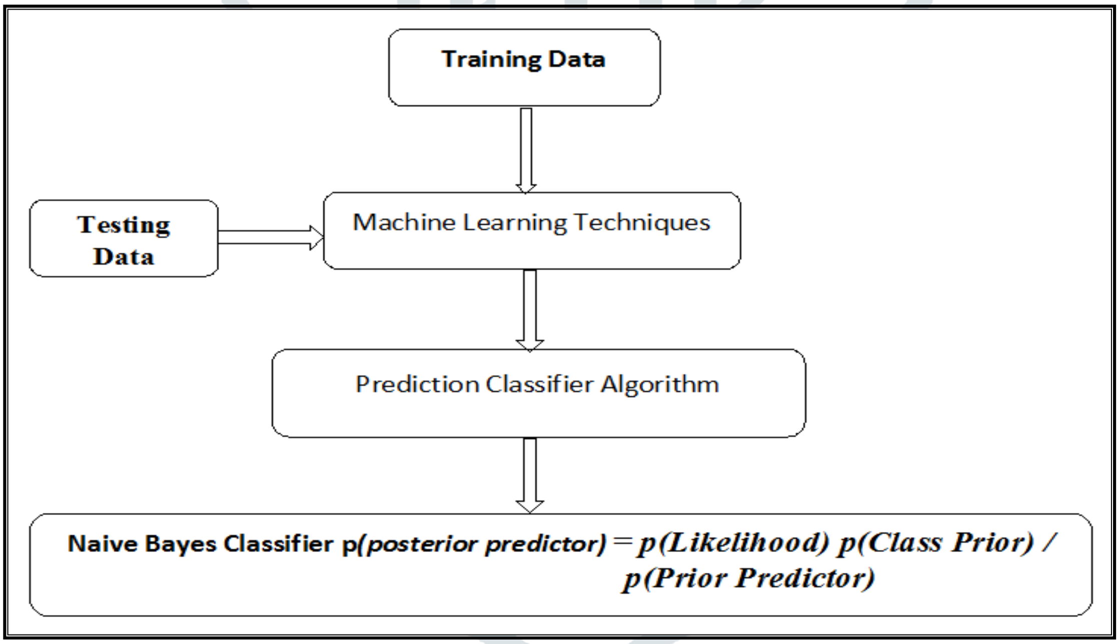
Training Data
Testing Data
Machine Learning Techniques
Prediction Classifier Algorithm
Naive Bayes Classifier p(posterior predictor) = p(Likelihood) p(Class Prior) / p(Prior Predictor)

Figure 2. Process of misclassification data with supervised techniques

The above figure 2 describes the structure using machine learning techniques using prediction classifier algorithm, to calculate the outcome of posterior predictor. Posterior predictor produces the inferences by using three different parameters of probability; they are probability of likelihood, the probability of class prior and the probability of prior predictor. These techniques increase the conditional probability of the target value by means of analysing the imputed known data. This process starts with the large training dataset attribute which comes under known value. Bayes classifier focuses on the independent class of one another and subsequently produces features that do not belong to independent data.

V. EXPERIMENTAL ANALYSIS

This section generates the performance of various predictive techniques such as Mean, Median, Standard Deviation and Naïve Bayesian approach based on the outcome of the evaluation of accuracy rate. The evaluation of this existing system consists of various operations in generating the missing data, applying the predictive technique, imputing the value of missing data and comparing the accuracy rate of imputed data with original data set. All the predictive approaches consist of an equal percentage of missing values ranges in 5%, 10%, 15%, 20% and 25% percentage. The below chart of comparison performance shows the analysis of high accuracy rate, out of which Naïve Bayesian Approach perform high accuracy rate of the imputation of missing data. The following experimental results proved Naïve Bayesian perform high accuracy rate of unknown data. This highly sophisticated approach provides various classification terms to calculate multiclass prediction with the help of Laplace correction in the target variables. Naïve Bayes have limited options for variable selection, which is recommended to focus on pre-processing of information and the variable selection. The following Figures 3 shows the accuracy rate of Naïve Bayesian techniques.

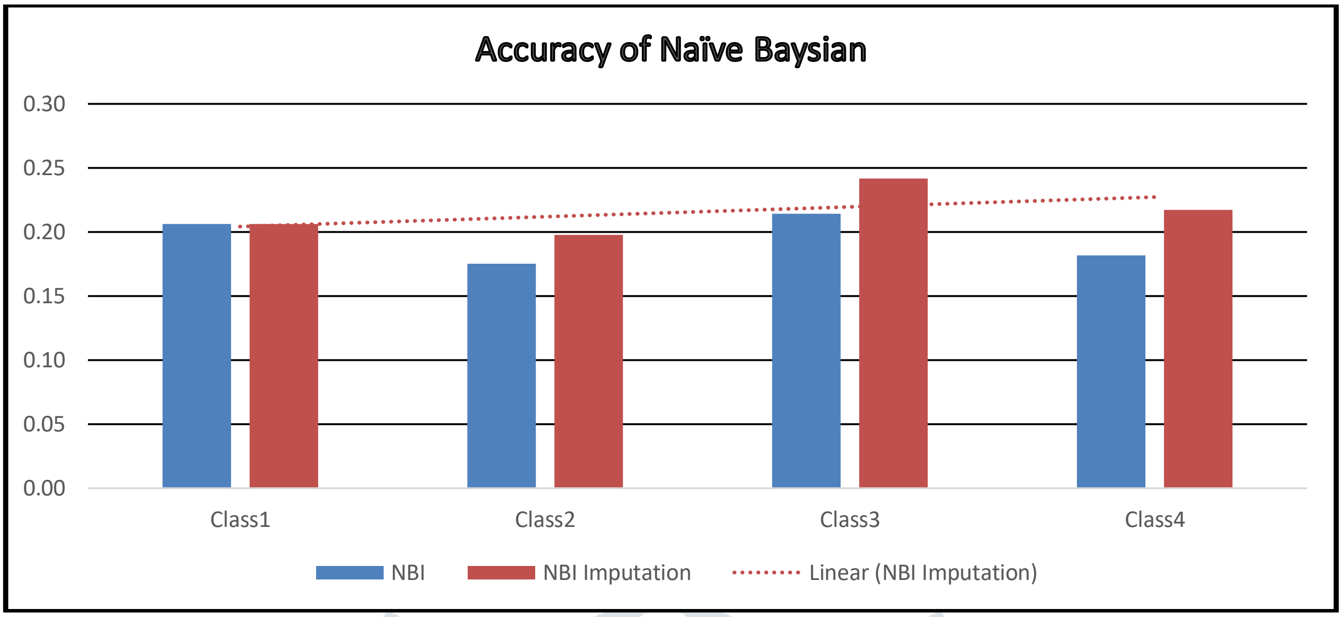

Figure 3. Accuracy of Naïve Bayesian

Naïve Bayes performs fast is very easy to build and is specifically predicts the probability of the unknown data. The accompanying chart compares the data set of missing value starts from 5% to 25% of the missing analysis with the original data set. As per the structure of below figure 4 the percentage rate of missing values is increased, then the prediction of missing values will be diminished.

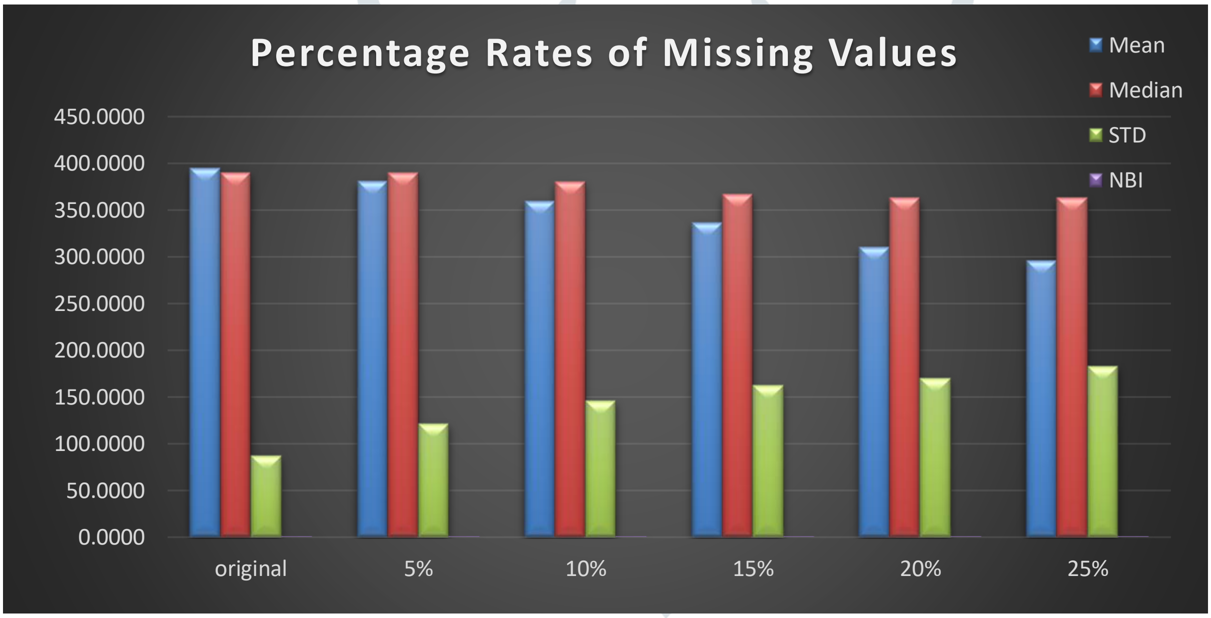

Figure 4. Percentage Rate of Missing Values

Bayes techniques are the simple yet powerful classifier specification to handle the unknown data analysis which is represented in the form of prior and posterior classification of missing data analysis. It makes all the attributes in the dataset for the actual estimation of imputation techniques. Bayesian techniques completely depend on the classification inference between the Bayes model. They are simple and very easy to build without any complication of iterative parameter distribution of machine learning techniques. Inspire of robust and very powerful techniques which produce the estimation of useful information with the flexible model of data analysis. This algorithm generates a simple way of posterior probability in spite of generating the likelihood of probability predictor for an existing class, along with the class prior probability of target analysis and with the help of predictor attribute and finally estimate the prior probability of predictor in terms of conditional independence. During this estimation process, class data can be transferred from prior analysis to posterior analysis by constructing the frequency tables. The numerical variable requires transforming the categorical class analysis of frequency tables.

From the above structure of the Breast Tissue training data set, the calculation of the percentage rates of missing values in the data sets describe the overall construction of unknown value analysis. The above training data set have hundreds of data points with few variables. In such situation, Naïve Bayes performs fast is very easy to build and is specifically predicts the probability of the unknown data. The accompanying chart compares the data set of missing value, starts from 5% to 25% of the missing analysis with the original data set. As per the structure of below figure 5, the percentage rate of missing values is increased, then the prediction of missing values will be diminished.

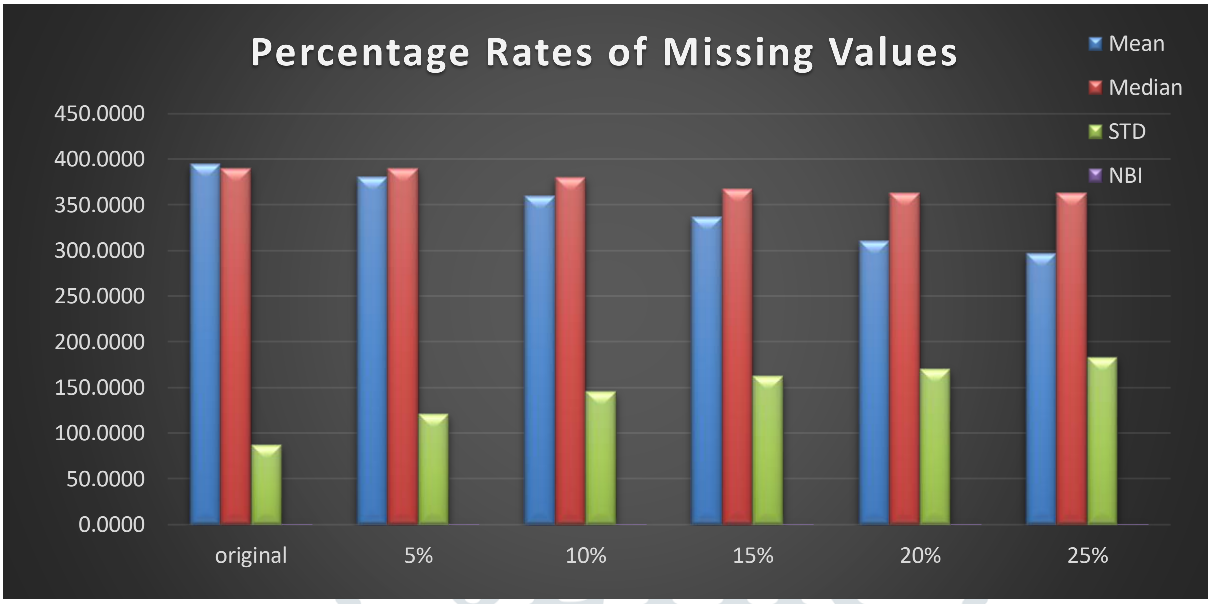

Figure 5. Percentage Rate of Missing Values

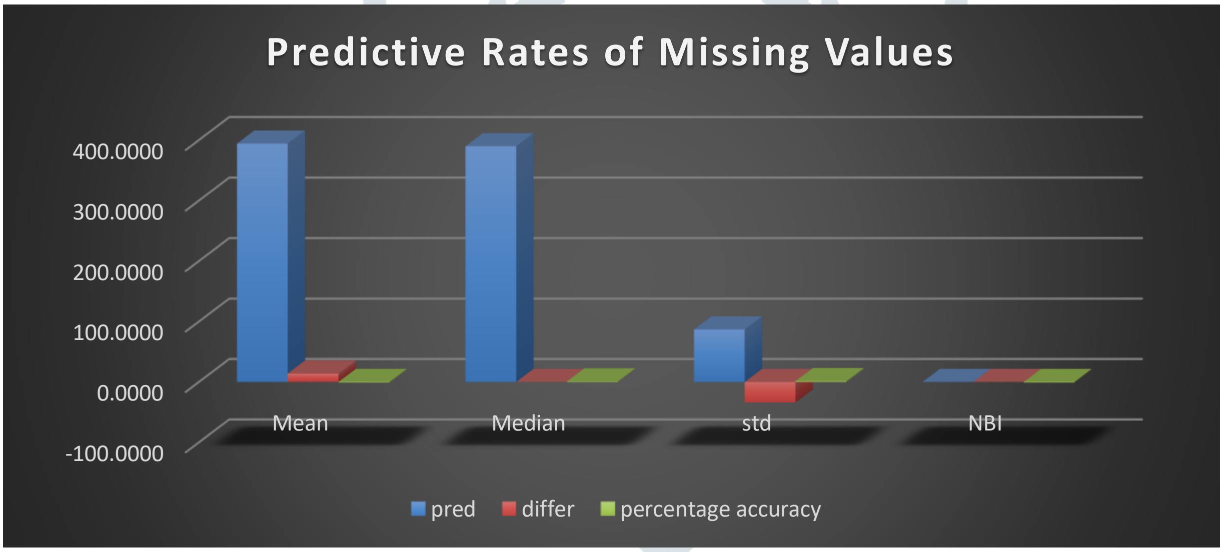

Figure 6. Predictive Rate Using Different Approach

As per the generation of four various operations, the following diagram compares the different construction of missing data as 5%, 15% and 25% respectively. If the performance gap between missing values is increasing, the prediction of missing values perform with less accuracy rate. If the performance of accuracy rate of missing value is high then the generation of prediction of missing values will be low. The above figure 6 is the predictive rates of missing values performed with mean, median, standard deviation and Naïve Bayesian techniques. Three different operations are generated such as prediction of missing values compared with original data sets, the difference between the imputed values using prediction techniques along with original data sets and the estimation of accuracy rate compared with original dataset. All the predictive methods are parallely compared with the performance of original data sets. Mean prediction technique performed imputed data sets values are higher than the original data, the Median process performs the moderate values, Standard Deviation performed imputed values are lesser than the original data and the Naïve Bayesian approach generate accurate values compared with original data set.

VI. CONCLUSION

The performance of various predictive techniques such as Mean, Median, Standard Deviation and Naïve Bayesian approach based on the outcome of the evaluation of accuracy rate. The evaluation of this existing system consists of various operations in generating the missing data, applying the predictive technique, imputing the value of missing data and comparing the accuracy rate of imputed data with original data set. The existing research work was performed in order to predict the misclassification of data in a large dataset using effective supervised machine learning approach. The limitations of the existing research works are as follows:

- The existing research works do not carry out the assessment of cognitive approaches to all the parameters in the dataset to predict the missing values using a supervised model.
- Patten knowledge theory was not considered to a satisfactory in the earlier research work to predict a simple path to understand and generate a high accuracy rate.
- The existing work does not analyze the interval of transactions between the demarcation lines of predictive information.

REFERENCES

[1] Gayan Prasad Hettiarachchi, Dhammika Suresh Hettiarachchi, NadeekaNilminiHettiarachchi, and Azusa Ebisuya, “Next Generation Data Classification and Linkage” *IEEE Global Humanitarian Technology Conference* 2014

[2] Pei Zhang, Xiaoyu Wu, Xiaojun Wang, and Sheng Bi, “Short-Term Load Forecasting Based on Big Data Technologies” *CSEE Journal of Power and Energy Systems,* Vol. 1, No. 3, September 2015

[3] R.J. Little and D. B. Rubin. Statistical Analysis with missing Data, John Wiley and Sons, New York, 1997.

[4] R.S. Somasundaram, R. Nedunchezhian, “Evaluation on Three simple Imputation Methods for Enhancing Preprocessing of Data with Missing Values”, International Journal of Computer Applications, Vol21-No. 10, May 2011, pp14-19.

[5] Jeffrey C.Wayman, “Multiple Imputation for Missing Data: What is it and How Can I Use It?” Paper presented at the 2003 Annual Meeting of the American Educational Research Association, Chicago, IL, pp.2-16, 2003.

[6] Mrs.R. Malarvizhi, Dr. Antony Selvadoss Thanamani, “K-Nearest Neighbor in Missing Data Imputation”, International Journal of Engineering Research and Development, Volume 5 Issue 1-November-2012,

[7] dfAlireza Farhangfar, Lukasz Kurgan and Witold Pedrycz, “Experimental Analysis of Methods for Imputation of Missing Values in Databases.

[8] K. Lakshminarayan, S.A. Harp, and T. Samad, “Imputation of Missing Data in Industrial Databases”, Applied Intelligence, vol 11, pp., 259-275, 1999.

[9] Peng Liu, Lei Lei, “Missing Data Treatment Methods and NBI Model”, Sixth International Conference on Intelligent Systems Design and Applications, 0-7695-2528-8/06.

[10] Seema Sharma, Jitendra Agrawal, and Sanjeev Sharma, “Classification through Machine Learning Technique: C4.5 Algorithm Based on Various Entropies” *International Journal of Computer Applications* (0975-8887), Vol. 82, No. 16, November 2013

[11] ShambeelAzim, and Swati Aggarwal, “Hybrid Model for Data Imputation: Using Fuzzy *c* means and Multi-Layer Perceptron” *IEEE* 2014

[12] Yan Li and Manoj A. Thomas, “A Multiple Criteria Decision Analysis (MCDA) Software Selection Framework” *47th Hawaii International Conference on Systems Science* 2014

[13] Samuel H. Hawkins, John N. Korecki, YoganandBalagurunathan, YuhuaGu, Virendra Kumar, SatrajitBasu, Lawrence O. Hall, Dmitry B. Goldgof, Robert A. Gatenby, and Robert J. Gillies, “Predicting Outcomes of Nonsmall Cell Lung Cancer Using CT Image Features” *IEEE* Vol. 2, 2014

[14] Wei Huang, “A Novel Disease Severity Prediction Scheme via Big Pair-Wise Ranking and Learning Techniques Using Image-Based Personal Clinical Data” *Signal Processing* 124 (2016) 233-245 Journal homepage: www.elsevier.com/locate/sigpro