Volume 3, Issue 5, May 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Reasoning with Missing Values in Multi Attribute Datasets

Anjana Sharma
Department of Computer Science
Bahra University, Shimla Hills,
H.P., India

Naina Mehta
Department of Computer Science
Bahra University, Shimla Hills,
H.P., India

Iti Sharma
Department of Computer Science
Bahra University, Shimla Hills,
H.P. India

Abstract- The presence of missing data in a datasets can affect the performance of classifier which leads to difficulty of extracting useful information from datasets .Dataset taken for this study is student records of university system that contains some missing values. To compute these missing values three technique are used named as Litwise deletion, Mean/mode imputation and KNN imputation, which result in imputed datasets. On these resulting datasets C4.5 classification algorithm is applied individually. This work analyzes the performance of imputation methods using C4.5 classifier on the basis of accuracy for handling missing data. Weka data mining tool is used for this experimental analysis.

Keywords- Data mining, Missing data, C4.5, KNN (K-nearest neighbor), Mean Imputation, Weka.

I. INTRODUCTION

Data Mining is the process of extracting hidden knowledge from large volumes of raw data. Data mining has been defined as "the nontrivial extraction of previously unknown, implicit and potentially useful information from data. Data Mining is used to discover knowledge out of data and presenting it in a form that is easily understood to humans [1]. Data Mining is the notion of all methods and techniques, which allow analyzing very large data sets to extract and discover previously unknown structures and relations out of such huge heaps of details. This information is filtered, prepared and classified so that it will be a valuable aid for decisions and strategies [3].

A. Missing Data

Missing data or missing values occur when no data value is stored for an instance in the current record. Missing data might occur because value is not relevant to a particular case, could not be recorded when data was collected or ignored by users because of privacy concerns [14]. Most information system usually has some missing values due to unavailability of data. Sometimes data is not presented or get corrupted due to inconsistency of data files. Missing data is a common problem that has a significant effect on the conclusion that can be drawn from the data. Missing data is absence of data items that hide some information that may be important [1].

1) Types of missing data: There are basically three types of missing data, these are:

MCAR- It is probability of missing data on any attribute does not depend on any value of attribute [7]. The term "Missing Completely at Random" refers to data where the missingness mechanism does not depend on the variable of interest, or any other variable, which is observed in the dataset [2]

MAR- The probability of missing data on any attributes does not depends on its own value but value of other attribute [7]. Sometimes data might not be missing at random but may be termed as "Missing at Random". We can consider an entry Xi as missing at random if the data meets the requirement that missingness should not depend on the value of Xi after controlling for another variable [2].

MNAR- Missing data depends on the values that are missing [7]. Sometimes data might not be missing at random but may be termed as "Missing at Random". We can consider an entry Xi as missing at random if the data meets the requirement that missingness should not depend on the value of Xi after controlling for another variable [2].

B. Missing Data Imputation Method

Litwise deletion- It is the simplest way of handling missing data is to delete the subject that have missing values. This method consists of discarding all instances (cases) with missing values for at least one feature. A variation of this method consists of determining the extent of missing data on each instance and attribute, and delete the instances and/or attributes with high levels of missing data [8].Litwise deletion having advantage that it decrease sample size file used for analysis.

Mean/mode imputation- This is one of the most frequently used methods. It consists of replacing the missing data for a given feature (attribute) by the mean of all known values of that attribute in the class where the instance with missing attribute belongs[8].If we assume the existence of true value for each unknown one, we can try to estimate this true value based on the known information. The simple approach for utilizing data-dependent information is to replace unknown value of discrete attributes by most common value(the mode value) and unknown values of continuous attribute by their average values(the mean value)[9].However many such machine learning system use a simple imputer, known as mean

imputation, which replace the missing with the mean value of the attribute over all instances or over all instances of same class or with the most frequently observed value of attribute[10].

K-nearest neighbor(KNN)- In this method the missing values of an instance are imputed considering a given number of instances that are most similar to the instance of interest. The similarity of two instances is determined using a distance Function [8].While the k-nearest neighbor algorithms look for the most similar instances, the whole dataset should be searched. However, the dataset is usually very huge for searching. On the other hand, how to select the value "k" and the measure of similar will impact the result [7]. The choice of an small k produces a deterioration in the performance of the classifier after imputation due to overemphasis of a few dominant instances in the estimation process of the missing values. [8].

C. Classification Algorithm

Classification is a supervised learning method. It means that learning of classifier is supervised in that it is told to which class each training tuples belongs. Data classification is a two step process. In the first step, a classifier is build describing a predetermined set of data classes or concepts [15]. This is the learning step, where classification algorithm builds classifier by analyzing or learning from a training set made up of database tuples and their associated class labels [15]. The data classification process has two phases, these are:-

Learning- Classification algorithm analyzed the training data. Classifier is represented in the form of classification rules. This phase is also viewed as learning of a mapping or function, Y=f(X) which predict the associated class label y of a given tuple X. Mapping is represented in the form of classification rules and used to categorize future data tuples and also provide deeper insight into database contents[15].

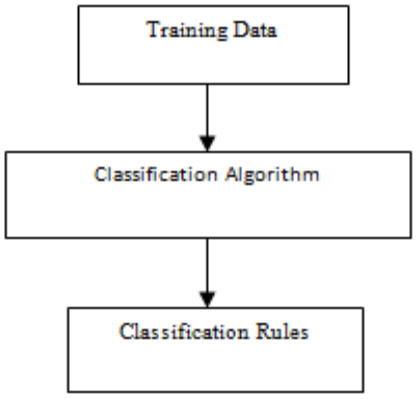

Fig.1 Learning Stage

Classification- To estimates the accuracy of classification algorithm test data is used. If the accuracy is considered acceptable, the rules can be applied to classification of new data tuples. Accuracy of a classifier on a given test set is percentage of test set that are correctly classified by classifier. The associated class labels of each test tuples is compared with learning classifier class prediction for that tuple [15].

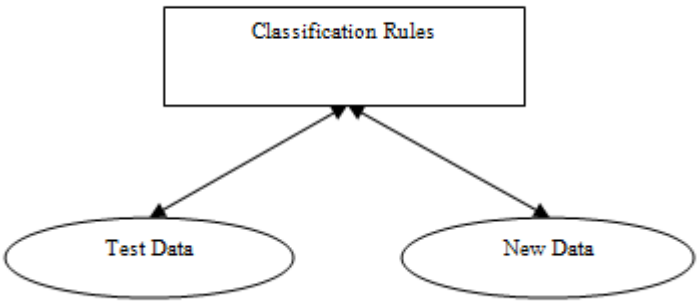

Fig. 2 Classification Stage

1) ***Decision tree induction:*** A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision [17]. Decision tree is a flow chart like tree structure, where each internal node is denoted by rectangles and the leaf nodes are denoted by ovals. The principle idea of a decision tree is to split data recursively into subsets so that each subset contains more or less homogeneous states of target variable (predictable attribute). At each split in the tree, all input attributes are evaluated for their impact on the

predictable attribute. When this recursive process is completed, a decision tree is formed [5]. It is the most commonly used algorithms because of its ease of implementation and easier to understand compared to other classification algorithm [6] [17]. In machine learning it becomes a foretelling model with remarkable properties, able to manage a great deal of data. Usually, a decision tree is built with learning techniques from a set of data previously arranged into two groups: the training and the test set. The former is employed in order to build the structure of the model, the latter to test its accuracy. In our trees, nodes describe a particular classification and ramifications constitute the set of characteristics bringing to the classification. Consequently, every internal node represents a class composed of the union of the classes associated with his children nodes, and the predicate associated with every node is called split condition. [4].

II. PROBLEM DEFINITION

In this work, we are taking examination cell database of a university system that contains student's records. Some of the attributes of this database are student personal details, State of domicile and Qualification details etc. In these records, some of data values are missing. To handle these missing values, missing data techniques are used. Techniques that are used for handling missing values in this work are-

- Ignoring the tuples containing missing data.
- Imputing the missing values using attribute mean value.
- Imputing the missing values using KNN (K Nearest Neighbor).

After applying all of these techniques we get a three complete / imputed dataset. These datasets are input to the classification algorithm. Classification algorithm used for this is C4.5 classifier. This algorithm is applied on imputed dataset to analyze which is the best technique for handling the missing values.

A. Objectives

The main objectives of this research work are:-

- It helps in finding an incomplete real world data in the KDD (Knowledge discovery in database) process.
- To explore missing data imputation methods that help in obtaining good quality datasets for better analysis of results.
- Datasets obtained after applying missing data techniques are analyzed using C4.5 decision tree classifier.
- To find out which missing technique is best on the basis of accuracy and efficiency for handling missing values.

III. TECHNIQUES USED FOR DATA RETRIEVAL

Here we use three techniques for data retrieval:-

- Ignoring the tuples containing missing data/lit wise deletion.
- Imputing the missing values using attribute mean value.
- Imputing the missing values using KNN (K- Nearest Neighbor).

Suppose there is dataset that contains missing values. The dataset is shown in table 1 as below:

Table 1: Dataset with Missing Values

Roll No.	Marks in Maths	Marks in Physics	Marks in Chemistry
1	88	78	87
2	87		77
3	90	83	89
4	78	90	82
5	67	85	

Missing data

Ignoring the tuples containing missing data/lit wise deletion- This method consists of discarding all instances (cases) with missing values for at least one feature. This method omits those cases (instances) with missing data and does analysis on the remains. It is the simplest way of handling missing data is to delete the subjects that have missing values. It is available in all statistical packages and is the default method in many programs [8].

Table 2: Dataset with Litwise Deletion

Roll No.	Marks in Maths	Marks in Physics	Marks in Chemistry
1	88	78	87
3	90	83	89
4	78	90	82

After applying litwise deletion all the cases having missing values are deleted from the dataset. The resulted dataset is shown in table 2.

Imputing the missing values using attribute mean value- This is one of the most frequently used methods. It consists of replacing the missing data for a given feature (attribute) by the mean of all known values of that attribute in the class where the instance with missing attribute belongs [8].

Table 3: Dataset with Mean Imputation

Roll No.	Marks in Math's	Marks in Physics	Marks in Chemistry
1	88	78	87
2	87	**67.2**	77
3	90	83	89
4	78	90	82
5	67	85	**67**

Mean 1 Mean 2

In this technique, mean of each attribute that contains missing values is calculated and is replaced in the place of missing values. Mean is calculated as:

Mean = sum of all the values/total number of values.

Mean1= 78+83+90+85/5 = 67.2

Mean2= 87+77+89+82 = 67

Imputing the missing values using KNN (K- Nearest Neighbor) - In this method the missing values of an instance are imputed considering a given number of instances that are most similar to the instance of interest. The similarity of two instances is determined using a distance Function [8]. Distance function can be Euclidean and Manhattan etc. In this work we have considered the Euclidean distance.

The algorithm on how to compute the K-nearest neighbors is as follows-

- Determine the value of K(Nearest neighbors).Value of K will be chosen randomly.
- Calculate the distance between the missing value instance and other training instance ie based upon the value of K. Here Euclidean distance is used for calculating the distance. Euclidean distance is given by the equation as:-

$$D(x,y) = \sum_{i=1}^{n} \sqrt{x_i^2 - y_i^2}$$

- After calculating the Euclidean distances choose the data values those having minimum distance. If the value of K is 5 then we have to choose 5 values that having minimum distance.
- Calculate the mean of these chosen values. The mean is given by the equation as:-

$$M = 1/n \sum_{i=1}^{n} mi$$

- Return M as the output value for missing data.

The advantages of KNN imputation are:

- K-nearest neighbor can predict both qualitative attributes (the most frequent value among the k nearest neighbors) and quantitative attributes (the mean among the k nearest neighbors).
- It can easily treat instances with multiple missing values.
- It takes in consideration the correlation structure of the data.
- It does not require to create a predictive model for each attribute with missing data. Actually, the k-nearest neighbor algorithm does not create explicit models [8].

IV. RESEARCH METHODOLOGY

This research work is two stage procedures. In the first stage, missing data is retrieved using missing data techniques. In the second stage, Analysis is done through C4.5 classifier using weka tool. This analysis has been done on the basis of accuracy and efficiency.

A. Dataset Used

Firstly, we have the database of examination cell of university system which contains student records. There are 100 records of the student in this database which contains various attributes. The attributes are name of the students, Fathers name, address, category, family income, 10^{th} and 12^{th} qualification details etc. Some of the attribute values of student records are missing. The attributes that contain missing values are marks obtained in 10^{th} and 12^{th} class.

B. Data Retrieval

Data retrieval is a method used for imputation of missing values. Imputation is process of replacing missing value with substitute value based on other information [16]. Here we are using three techniques named as litwise deletion, imputation using mean and K-nearest neighbor (KNN) imputation.

1) Litwise deletion: This method consists of discarding all instances (cases) with missing values for at least one feature. A variation of this method consists of determining the extent of missing data on each instance and attribute, and deletes the instances and/or attributes with high levels of missing data[8]. This method is available as a preprocessing option in the weka tool. Firstly the dataset that contain the missing value is loaded into weka tool. Weka support only CSV and ARFF file format. So data base file is converted into CSV format and load into weka. As specified in litwise technique instance that contains missing values are deleted. For performing this technique selects the preprocessor tab and click on the delete instances option. This will result in deletion of all instances that contain missing values.

2) Mean/mode imputation: This is one of the most frequently used methods. It consists of replacing the missing data for a given feature (attribute) by the mean of all known values of that attribute in the class where the instance with missing attribute belongs[8].Again dataset that contains missing values is loaded into weka .On this dataset apply Filtering algorithm. Firstly, select the unsupervised filtering and then use replacemissingvalues option. As a result of this, missing values are replaced through mean of that particular attribute. For numerical attributes, missing value is replaced by mean. And for nominal attributes missing value is replaced by mode.

3) K-nearest neighbor: In this method the missing values of an instance are imputed considering a given number of instances that are most similar to the instance of interest. The similarity of two instances is determined using a distance Function [8].

The algorithm for k-nearest neighbor is as follows-

- Determine the value of K(Nearest neighbors).Value of K=10 will be chosen randomly.
- Calculate the distance between the missing value instance and other training instance. Here Euclidean distance is used for calculating the distance. Euclidean distance is given by the equation as:-

$$D(x,y) = \sum_{i=1}^{n} \sqrt{x_i^2 - y_i^2}$$

- After calculating the Euclidean distances choose the data values those having minimum distance. If the value of K is 5 then we have to choose 5 values that having minimum distance.
- Calculate the mean of these chosen values. The mean is given by the equation as:-

$$M = 1/n \sum_{i=1}^{n} mi$$

- Return M as the output value for missing data.

This algorithm is applied on datasets and missing values are retrieved after performing KNN imputation.

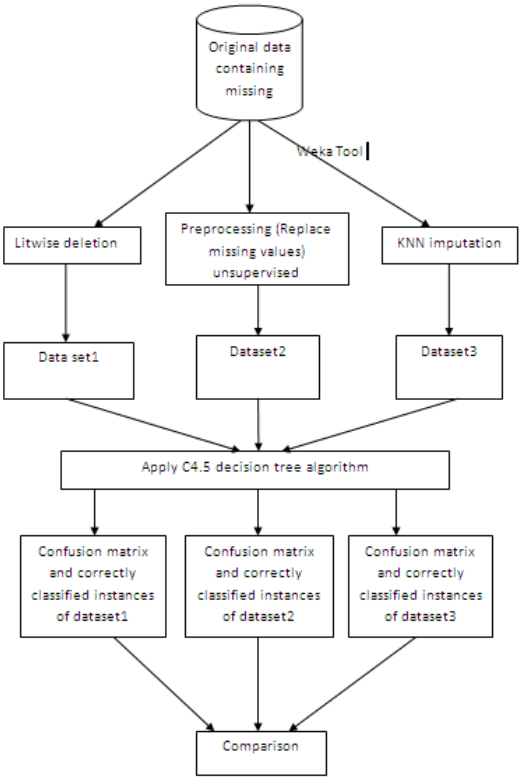

Fig.3 Flow Diagram for Missing Data Handing

V. RESULT AND DISCUSSION

The database that we have taken having records of 100 students. Some of the data values in these records are missing. Database is loaded into weka tool.

A. Experimental Procedure

In our experiment we first load the dataset into weka tool using preprocessing panel.

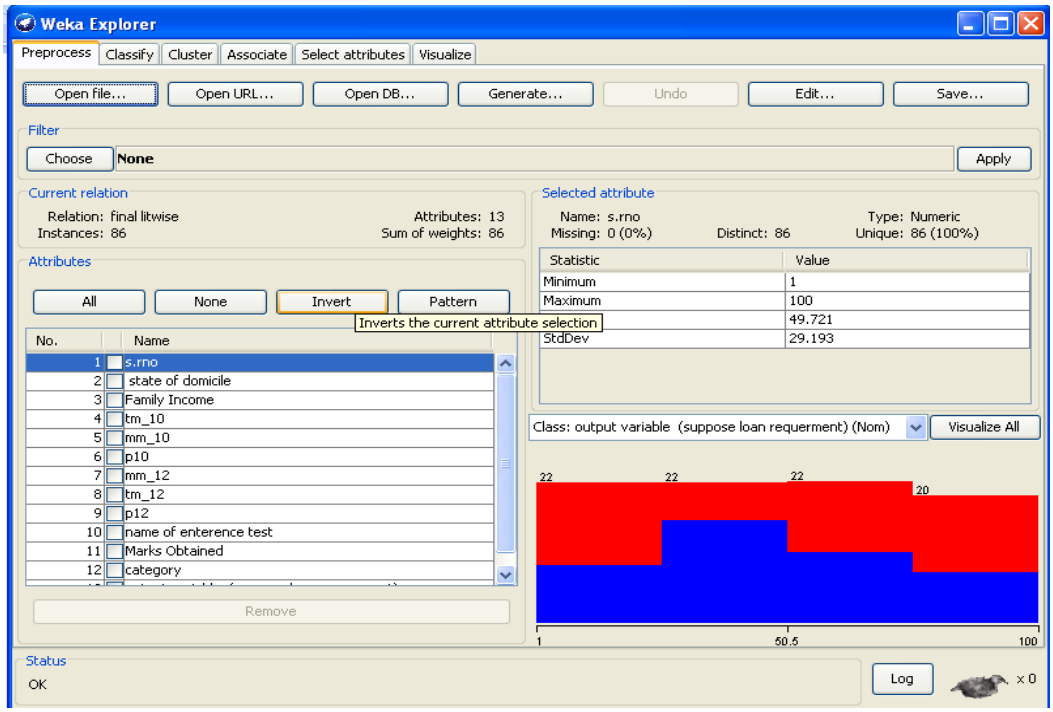

Fig.4 Loaded Dataset into Weka Tool

On this dataset apply filtering algorithm for litwise deletion. Then in the classify panel we choose C4.5 Decision tree classifier and start the analysis using 10 fold cross validation.

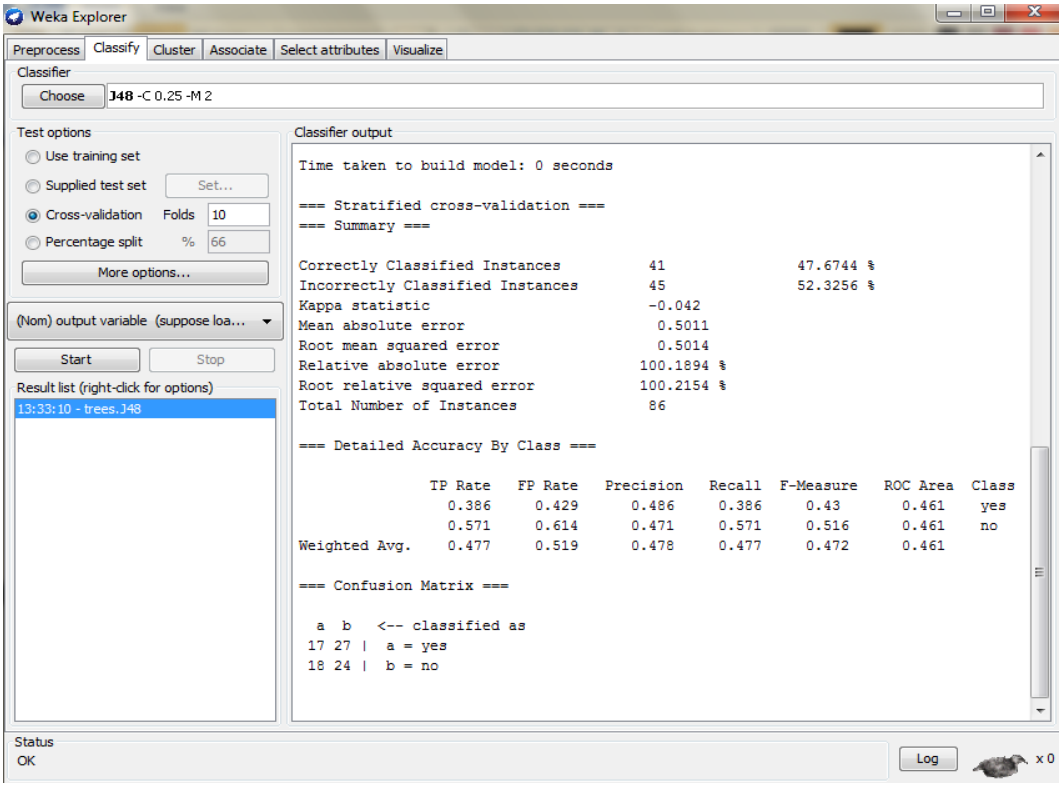

Fig.5 Result of C4.5 Classifier for Litwise Deletion

Apply unsupervised filtering technique by choosing replacemissingvalues option. Then in the classify panel we choose C4.5 Decision tree classifier and start the analysis using 10 fold cross validation.

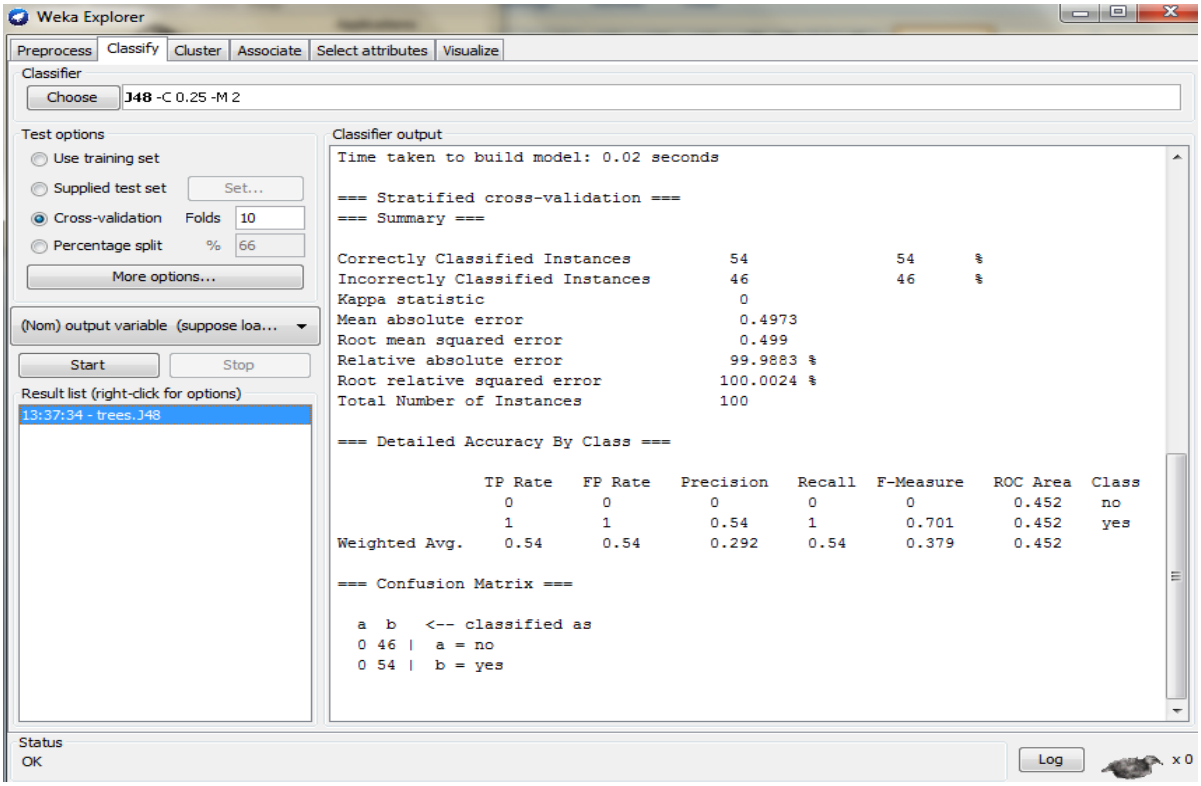

Fig.6 Result of C4.5 Classifier for Mean Imputation

Apply KNN computation on the dataset. Then in the classify panel we choose C4.5 Decision tree classifier and start the analysis using 10 fold cross validation.

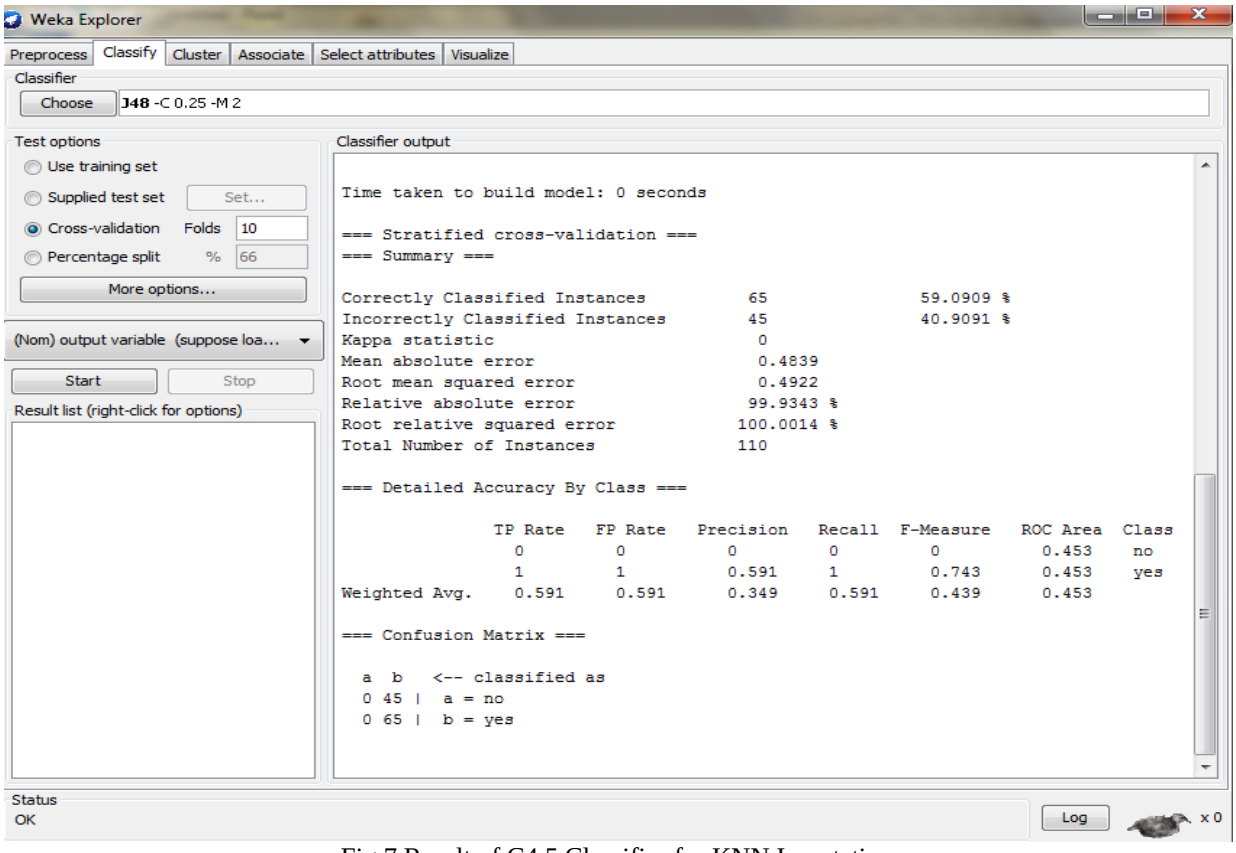

Fig.7 Result of C4.5 Classifier for KNN Imputation

B. Experimental Result

In this study comparison has been done on the basis of Accuracy/correctly classified instance and confusion matrix.

1) Confusion matrix:

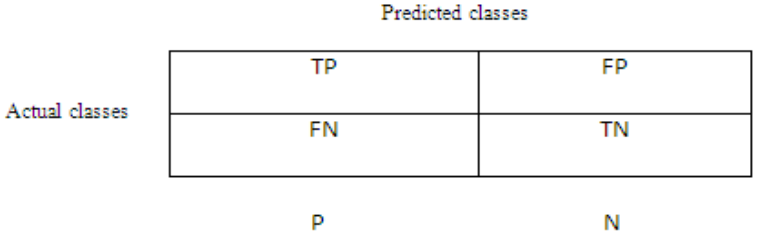

Fig.8 Confusion Matrix

True Positive (TP)-These are the positive tuples that were correctly labeled by the classifier [15].
True Negative (TN)-These are the negative tuples that were correctly labeled by the classifier [15].
False Positive (FP)-These are the negative tuples that were incorrectly labeled as positive [15].
False Negative (FN)-These are the positive tuples that were mislabeled as negative [15].

2) Accuracy: It is the Probability that the test yields correct results.

Accuracy is calculated as = (TP+TN)/(P+N) where, P=TP+FN and N=FP+TN.

Table 4: Comparison of Imputation Techniques using C4.5 Classification Algorithm

Algorithm used C4.5 classifier	Correctly classified instances	Incorrectly classified instances	Accuracy	Mean absolute error
Litwise deletion	41	45	47.6744%	0.5011
Mean/mode imputation	54	46	54%	0.4973
KNN(K-nearest neighbor) imputation	65	45	59.0909%	0.4839

According to experimental results, correctly classified instances for Litwise deletion is 41 and for mean/mode imputation is 54.Correctly classified instances for KNN imputation is 65 which is greater than previous two algorithms. Accuracy of KNN is 59.0909% which is also greater than other two techniques so KNN is a best technique use for missing data imputation as compared to other two techniques.

VI. CONCLUSION AND FUTURE SCOPE

Missing values are regarded as serious problem in most of the information system due to unavailability of data and must be impute before the dataset is used. Here we have taken a student records of university system in which some of the student records are missing. To handle these missing values three techniques are used named as Litwise deletion, mean/mode imputation, KNN (k nearest neighbor).Firstly dataset is loaded into weka tool. Apply these missing techniques individually on this dataset, which results as three imputed datasets. Then C4.5/J48 classification algorithm is applied to these replaced datasets and their results are compared in order to evaluate the efficiency and accuracy of missing data techniques on the basis of experimental results accuracy and KNN is greater than other two techniques. So, KNN imputation is a better way of handling missing values.

Future Scope- The proposed work handle missing values only for numerical attributes. Further it can be extended to handle a categorical attribute. Different classification algorithm can be used for comparative analysis of missing data techniques. Missing data technique can also be implemented in matlab.

References

[1] Dinesh J. Prajapati ,Jagruti H. Prajapat, Handling Missing Values: Application to University Data Set .Issue 1, Vol.1 (August-2011), ISSN 2249-6149.
[2] Bhavik Doshi, Handling Missing Values in Data Mining. Data Cleaning and Preparation Term Paper.
[3] Johannes Grabmeier, Andreas Rudolph, Techniques of Cluster Algorithms in Data Mining. Data Mining and Knowledge Discovery, 6, 303–360, 2002.

[4] Gabriele Giarratana, Marco Pizzera, Marco Masseroli, Enzo Medico, Pier Luca Lanzi, Data Mining Techniques for the Identification of Genes with Expression Levels Related to Breast Cancer Prognosis. 2009 Ninth IEEE International Conference on Bioinformatics and Bioengineering.

[5] Anjana Gosain, AmitKumar, Analysis of Health Care Data Using Different Data Mining Techniques. lAMA 2009 978-1-4244-4711-4/09/$25.00 ©2009 IEEE.

[6] Anuja Priyama*, Abhijeeta, Rahul Guptaa, Anju Ratheeb, and Saurabh Srivastavab,Comparative Analysis of Decision Tree Classification Algorithms. International Journal of Current Engineering and Technology ISSN 2277 – 4106.

[7] Liu Peng, Lei Lei , A Review of Missing Data Treatment Methods.

[8] Edgar Acu˜na, Caroline Rodriguez, The treatment of missing values and its effect in the classifier accuracy.

[9] Lars Wohlrab, Johannes Fürnkranz, A Comparison of Strategies for Handling Missing Values in Rule Learning. Technical Report TUD–KE–2009-03. .

[10] Xiaoyuan Su, Taghi M. Khoshgoftaar , Russell Greiner, Using Imputation Techniques to Help Learn Accurate Classifiers .

[11] Gustavo E. A. P. A. Batista and Maria Carolina Monard, A Study of *K*-Nearest Neighbour as an Imputation Method .

[12] V. Srinivasan, Dr. G. Rajenderan, Ms. J. Vandar Kuzhali, M. Aruna, Classify The Student With Missing Value To Calculate Future Semester Result For Placement Record Using Knowledge Acquisition. Journal of Computer Applications, Vol-III, No.3, July - Sept 2010.

[13] Abdullah H. Wahbeh, Qasem A. Al-Radaideh, Mohammed N. Al-Kabi, and Emad M. Al-Shawakfa , A Comparison Study between Data Mining Tools over some Classification Methods. *(IJACSA*) International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence.

[14] Luai Al Shalabi, Mohannad Najjar and Ahmad Al Kayed, A framework to Deal with Missing Data in Data Sets . Journal of Computer Science 2 (9): 740-745, 2006 ISSN 1549-363.

[15] Jiawei Han, Micheline Kamber, Jian Pei,Data Mining Concepts and Techniques Third edition.

[16] Gelman, Andrew, Jennifer Hill,Data Analysis using Regression and Multilevel/ Hierarchical models. Cambridge University Press, 2006.

[17] Sharad Verma, Nikita Jain, Implementation of ID3 – Decision Tree Algorithm.

