



جامعة عمان العربية  
AMMAN ARAB UNIVERSITY

**ESTIMATION OF NULL VALUE  
IN RELATIONAL DATA BASE SYSTEM  
USING K-NEAREST NEIGHBOR AND DECISION TREE**

**PREPARED BY  
SARAH AL - SAMADI**

**Supervisor: Prof. Dr. Alaa Al-Hamami**




**This thesis is submitted for the Degree of Master in Computer  
Science**

**Department of Computer Science  
College of Computer Sciences and Informatics  
Amman Arab University**

**2013**

## RESOLUTION OF THE EXAMINING COMMITTEE

This dissertation , titled "Estimation of Null Value in Relational Data Base System Using K-Nearest Neighbor and Decision Tree " , has been defended and approved on July 7<sup>th</sup>. 2013.

<u>Examining committee</u>	<u>Title</u>	<u>Signature</u>
Prof. Dr. Gassan Kanaan	Chair	
Prof. Dr. Alaa Al-Hamami	Member and Supervisor	
Dr. May Haikel Riadh	Member	

## **Dedication**

**I dedicate this work to:**

**My parents,**

**My brother and sister,**

**My supervisor Prof.Dr.Alaa Al-Hamami**

**Mr.Hassan Abdel Razek**

**My dear Friends**

**For their love and support.**

## Acknowledgment

**“All Praises and Thanks to ALLAH”**

I would like to thank my supervisor Prof. Dr. Alaa Al-Hamami who provided me with his full support, encouragement and guidance to get this dissertation in its present form .Without his help and support, this work would not have been achievable. He was available at all times I needed his help.

My sincere thanks go to the Dean of Graduate College of Computer Studies, all of the lecturers, administration and staff of Amman Arab University for Graduate Studies.

I also extend my thanks to my father for his help in refining this dissertation and to Mr.Hassan Abdel Razek For the training course on “Oracle language”

## Table of Content

Dedication .....	III
Acknowledgment .....	IV
Table of Content .....	V
List of Figures .....	VII
Abstract .....	XI
Arabic summary.....	XII
<b>CHAPTER ONE INTRODUCTION .....</b>	<b>1</b>
1.1 INTRODUCTION .....	2
1.2. INTRODUCTION TO DATABASES .....	3
1.3 INTRODUCTION TO MISSING VALUES .....	5
1.4 INTRODUCTION TO DATA MINING .....	10
1.4.1 ESTIMATION .....	10
1.4.2 PREDICTION .....	11
1.4.3 CLASSIFICATION .....	11
1.4.4 ASSOCIATION.....	12
1.5 INTRODUCTION TO K- NEAREST NEIGHBOR ALGORITHM (K-NN) .....	15
1.5.1 INTRODUCTION TO DECISION TREE ALGORITHM.....	19
1.5.2 Organization of the Contents: .....	21
<b>CHAPTER TWO LITERATURE SURVEY .....</b>	<b>22</b>
2.1.INTRODUCTION .....	22
2.2.The Research PROBLEM .....	23
2.3 PREVIOUS WORK.....	24
<b>CHAPTER THREE THEORETICAL DESIGN .....</b>	<b>35</b>
3.1.INTRODUCTION .....	35
3.2.Description of the proposed flow diagram for the frame work to treat the missing values.....	36
3.3Components of the new Method .....	40
3.3.1 Decision Tree .....	40
3.3.2 K – nearest neighbor algorithm.....	40

3.3.2.1K Mean – Nearest Neighbor Algorithm Flow Chart.....	42
3.3.2.2.K Frequency – Nearest Neighbor Algorithm Flow Chart.....	44
3.3.3 Checking the Database .....	46
3.4 New Full Algorithm for a Frame Work to process Missing Values.....	47
<b>CHAPTER FOUR THE EXPERIMENTAL WORKS .....</b>	<b>51</b>
4.1 Introduction.....	52
4.2 Execution of the Frame Work (Estimation of Null Value in Relational Database System ).....	52
4.3 Examples of the Use of the Frame Work for Estimation of Null .....	53
4.3.1 1alue in Relational Database System .....	53
4.3.2 Example(1) – An Error Value Input .....	53
<b>CHAPTER FIVE CONCLUSIONS AND .....</b>	<b>75</b>
<b>RECOMMENDATIONS FOR FUTURE WORK .....</b>	<b>75</b>
5.1 INTRODUCTION .....	76
5.2 CONCLUSIONS .....	76
5.3 RECOMMENDATIONS FOR FUTURE WORK.....	78
<b>REFERENCES .....</b>	<b>80</b>

## List of Figures

<b>Figure number</b>	<b>Figure name</b>	<b>Page</b>
Figure (1)	Concepts of Database System	4
Figure (2)	Mathematical Operation	7
Figure (3)	The Truth Table for the 3-Valued Logic	8
Figure (4)	Diagram Depicts Generic 3-Tier Architecture for Data Mining	12
Figure (5)	The Euclidean Distance Between Two Vectors $X_r$ and $X_s$	16
Figure (6)	The Scoring of a Validation Input Vector $x_0$ Using Some of Its "Neighbors" in the Training Data Set	18
Figure (7)	An Example of a Decision Tree Using Flow Chart Symbols	20
Figure (8)	Diagram of the Proposed System	25
Figure (9)	Flow Diagram for Frame Work (ESTIMATION OF NULL VALUE IN RELATIONAL DATA BASE SYSTEM)	40
Figure (10)	Flow Chart for the K Mean – Nearest Neighbor Algorithm	43
Figure (11)	Flow Chart for the the K Frequency – Nearest Neighbor Algorithm	45

<b>Figure (12)</b>	<b>Flow Chart for Frame Work (ESTIMATION OF NULL VALUE IN RELATIONAL DATA BASE SYSTEM)</b>	<b>48</b>
<b>Figure (13)</b>	<b>Example (1) Main Interface</b>	<b>54</b>
<b>Figure (14)</b>	<b>Example (1) Input and Updating Records Interface</b>	<b>54</b>
<b>Figure (15)</b>	<b>Example (1) Rules Interface</b>	<b>55</b>
<b>Figure (16)</b>	<b>Example (1) Error Message</b>	<b>55</b>
<b>Figure (17)</b>	<b>Figure (15) : Example (1) K-Nearest Neighbor Records</b>	<b>56</b>
<b>Figure (18)</b>	<b>Example (1) Records Account</b>	<b>57</b>
<b>Figure (19)</b>	<b>Example (1) Frequency Records View</b>	<b>57</b>
<b>Figure (20)</b>	<b>Example (1) Rules Interface</b>	<b>57</b>
<b>Figure (21)</b>	<b>Example (1) Display of the Estimated Value</b>	<b>58</b>
<b>Figure (22)</b>	<b>Example (2) Input and Updating Records Interface</b>	<b>59</b>
<b>Figure (23)</b>	<b>Example (2) K-Nearest Neighbor Records</b>	<b>60</b>
<b>Figure (24)</b>	<b>Example (2) Records Account</b>	<b>60</b>



Figure (25)	Example (2) Frequency Records View	61
Figure (26)	Example (2) Rules Interface	61
Figure (27)	Example (2) Display of the Estimated Value	62
Figure (28)	Example (3) Input and Updating Records Interface	63
Figure (29)	Example (4) Input and Updating Records Interface	64
Figure (30)	Figure (28) : Example (4) Rules interface	65
Figure (31)	Figure (29) : Example (4) One – Nearest Neighbor Record	65
Figure (32)	Figure (30) : Example (4) Records Account	66
Figure (33)	Example (4) Frequency Records View	66
Figure (34)	Example (4) Rules Interface	67
Figure (35)	Example (4) Display of the Estimated Value	67
Figure (36)	Example (5) Input and Updating Records Interface	68
Figure (37)	Example (5) No - Nearest neighbor Record	69
Figure (38)	Example (5) Rules Interface	69
Figure (39)	Example (5) Display of Estimated Value	70

X

<b>Figure (40)</b>	<b>Example (6) Main Interface</b>	<b>70</b>
<b>Figure (41)</b>	<b>Example (6) Rechecking Report - Page (1) of (941)</b>	<b>71</b>
<b>Figure (42)</b>	<b>Example (6) Rechecking Report - Page (941) of (941)</b>	<b>72</b>

**ESTIMATION OF NULL VALUE  
IN RELATIONAL DATA BASE SYSTEM  
USING K-NEAREST NEIGHBOR AND DECISION TREE**

**Abstract**

In real life, there are lots of relational database systems which the user, when using them, faces problems related to missing values, whether these values are unknown to the user or are non-existent which lead to statistics and reports retrieved from these systems not being representative of the original stored data. For these reasons many researchers tried to do the relational database estimation of missing values with a high estimated accuracy rate.

The present treatment is through estimating and replacing the missing values with approximate values obtained from the development of an algorithm that combines the two algorithms of data mining for estimations and predictions. The treatment also checks and validates the stored and estimated values through comparing them with the business rules.

The two algorithms used in the treatment are:

- Decision Trees algorithm.
- K-nearest neighbor algorithm after amendment through this research work.

The implementation of the present treatment shows treatment validity with a success rate of up to 80%.

## Arabic summary

تقدير القيم الفارغة

في نظام قواعد البيانات العلائقية

باستخدام طريقة ك-ا قرب جار و شجرة القرار

### الخلاصة

في الحياة العملية ، هناك العديد من أنظمة قواعد البيانات العلائقية التي عند استخدامها يواجه المستخدم مشاكل تتعلق بالقيم المفقودة سواء كانت هذه القيم غير معروفة للمستخدم أو غير موجودة في واقع الحال والتي ينتج عنها احصائيات و تقارير مستخلصة من قواعد البيانات غير ممثلة للبيانات الأصلية المخزونة. و للأسباب هذه حاول العديد من الباحثين استخدام بعض الطرق لتخمين القيم المفقودة و بمعدل دقة عالية.

تتضمن المعالجة الحالية تخمين واستبدال القيم المفقودة بقيم تقريبية مستحصلة من تطوير خوارزمية تدمج اثنتان من خوارزميات تنقيب البيانات للتخمين والتنبؤ . كما وتتضمن المعالجة أيضاً التدقيق والتحقق من صحة البيانات المخزونة والمخمنة من خلال مقارنتها مع القواعد الموضوعية.

إن الخوارزميات المستخدمة في المعالجة الحالية هي:

• شجرة القرار.

• ك-ا قرب جار بعد التعديل من خلال البحث الحالي.

إن تطبيقات المعالجة الحالية تظهر صحة تحقق المعالجة وبنسبة نجاح تصل الى

**80%**

# CHAPTER ONE

# INTRODUCTION



## 1.1 INTRODUCTION

The need for database systems arose in view of the existence of huge amount of data required to be stored and handled in order to eventually provide useful information to beneficiaries and decision makers. The database systems can be defined as integrated software programs used to facilitate dealing with stored data like retrieving, organizing and converting data from mere stored facts to useful information from which new facts can be derived from the stored data.

Advances in technology have led to exciting new applications of database systems, some of which are represented by new media technology which made it possible to store images, audio clips, and video streams digitally whereby these types of files are becoming important components of multimedia databases, and geographic information systems which can store and analyze maps, weather data, and satellite images. In addition data warehouses and online analytical processing systems are used by many companies to extract and analyze useful business information from very large databases to support decision making, and also real-time and active database technology are used to control industrial and manufacturing processes .Database search techniques are being applied to the World Wide Web to improve the search for information that is needed by users browsing the internet [1].

As it is known, the database is an image of the real world; therefore the user, when using these systems and because the real world is incomplete, faces problems related to missing values in the entry, in the retrieval and also in

dealing with these missing values in the statistics and reports required by the user, whether the values are unknown, existent but not available or attributes that do not apply to this tuple.

In this research, we will handle the treatment of these missing values through their estimation using the following data mining algorithms:

- Decision tree algorithm for prediction and classification.
- K-nearest neighbor algorithm for estimation, prediction and classification.

## 1.2. INTRODUCTION TO DATABASES

The Data Base (DB) has become a key element in the information systems and their applications and forms the foundation of the information technology. It is possible to define the database as "a collection of interrelated data that are stored together to serve multiple applications without a harmful or unnecessary redundancy; the data are stored in a way to be independent from the program using that data" [2].

Dealing with the database can be done through Data Base Management System (DBMS) which can be defined as the mean to help users in creating, updating, executing and implementation of the database as a system, and containing all the programs which satisfy this purpose [2].

In fact, the database and the DBMS software can be called the database system. Figure (1) explains some of the concepts of a simplified database system environment [1]:

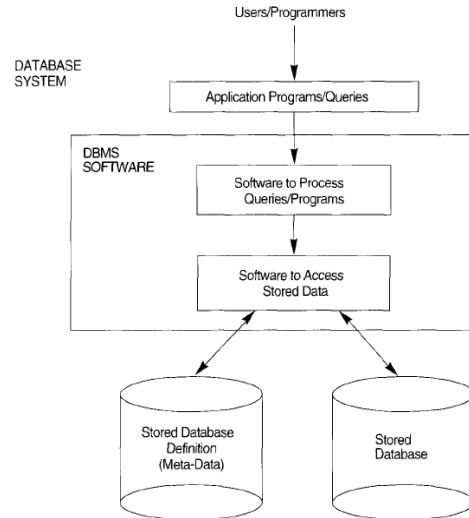


Figure (1) Concepts of Database system [1]

The main objective for setting up a system for a database is to facilitate dealing with the data stored in the database, and for easy access and dealing with it by users and according to assigned authorities for each user, and to provide an integrated view of the available data which reflects the reality as well as to get quick answers to unexpected queries and these are few of the many benefits that can be obtained from such a system.

In 1970, the relational system by E.F.Codd appeared which replaced the hierarchical system and the network system, and by the beginning of 1980 E.F.Codd presented the Relational Data Base Management System (RDBMS) which formed a major revolution for both users and designers and represented a great development in database [1].

"RDBMS stands for Relational Data Base Management System, and RDBMS data is structured in database tables, fields and records. Each RDBMS table consists of database table rows and each database table row consists of one or more database table fields "[3].



"RDBMS stores the data into collection of tables, which might be related to common fields (data base table columns). RDBMS also provide relational operators to manipulate the data stored into the database tables. Most RDBMS use SQL as database query language "[3].

The Structured Query Language (SQL) is the language of the databases which can be classified into the followings:

1. Data Definition Language (DDL): This language supports the definition or declaration of database objects [2].
2. Data Manipulation Language (DML): This language supports the manipulation or processing of such objects [2].

### **1.3 INTRODUCTION TO MISSING VALUES**

Any system designed for a database aims at providing integrity of the data and some of the problems that we often face when retrieving data or predicting values from the database appear when the databases contain missing values called "Null Values".

An important concept for Null Values is that it is used to represent the values of attributes that may be unknown or may not apply to a tuple. These special values are called null and are also known as undefined values [1].

The reasons behind the Null Values can be classified according to the type of the Null Value which can be of two types according to E.F.Codd way of representing the missing data in the relational model.

The followings are the types of Null Values and their causes [4]:

1. A-MARKS: data applicable but not known:

Which means that the data in fact exists but now it is missing such as the date of birth which is inevitable for everyone, but it is unknown to the user entering the data in the database or that the database available to us contains missing values.

2. I-MARKS: data is inapplicable:

Which means that the data does not exist in reality; for example the number of children of an unmarried man/woman is not zero, because he /she is not married and is Null Value and cannot be considered zero or a value , and there is no value that we can possibly enter in this field. As well the phone number where there is originally no phone number in which case it is inapplicable.

To define the value of the “Null Value” one can say that it cannot be a zero in a field numeric in kind and cannot be a space (" ") in the case of a text field. It can be said that the Null Value is an indicator of the absence of a value, whether applicable or inapplicable but not zero or a space.

When a DBMS deals with a database containing missing values, problems may appear of which few are shown in the following:

1. When joining more than one table (LEFT /RIGHT JOIN, FULL JOIN) a problem may appear with the Null Value, even if the tables are originally complete, as it is possible that the problem of the Null Value appears when joining these tables.

2. When functions are used such as AVG, COUNT, SUM with a database containing Null Values, false results may appear; for example, when we have 20 entries and 4 of them are Null Values, when calculating the average, 16 values will be added and divided by 20 and a false result which does not represent the true average shall appear.

3. When performing any mathematical operation and one of the entries is a Null Value, the result will be a Null Value whatever the other values are, as shown in Figure (2):

$$\left( \begin{array}{ccc} X+X=2X & X+A=A & \\ A+X=A & & \\ A+A=A & A+I=I & I+A=I \\ I+I=I & X+I=I & I+X=I \end{array} \right)$$

Figure (2) [5]

4. When performing any logical process or implementing program instructions which are essentially logical operations, e.g. BETWEEN :

$$(X \text{ BETWEEN } A \text{ AND } B) = ((A \leq X) \text{ AND } (X \leq B)) \text{ [6]}$$

We notice that the implementation of this instruction is a logical operation, and as there are three logical values as mentioned by E.F.Codd which are (TRUE, FALSE, UNKNOWN).

The UNKNOWN value represents the logical value of the Null Value. The following (Figure (3)) is the truth table for the 3-valued logic.

3-valued logic				
a	b	a OR b	a AND b	a == b
True	True	True	True	True
True	False	True	False	False
False	True	True	False	False
False	False	False	False	True
True	Unknown	True	Unknown	Unknown
Unknown	True	True	Unknown	Unknown
False	Unknown	Unknown	False	Unknown
Unknown	False	Unknown	False	Unknown
Unknown	Unknown	Unknown	Unknown	Unknown

- Results if condition involves a boolean combination:

Figure (3) [7]

For example, that A=7, B=10, and C is UNK (UNKNOWN) . Then the following expressions have the indicated truth values [8]:

A > B AND B > C : false .....[8]

A > B OR B > C : unk.....[8]

A < B OR B < C : true .....[8]

NOT ( A = C ) : unk.....[8]

In spite of the reasons already mentioned, it is possible to say that the presence of the missing values reflects the reality that the database is an image of the real world.

There are different methods that process and solve the problem of missing values and some of them are shown in the following [9]:

Removal: Ignores or deletes the records that contain missing values and deals only with complete data records. This option becomes non effective when there is a large amount of missing values within a limited amount of data and the deletion of the records that contain missing values leads to the loss of real and important data.

Imputation: Filling in the missing values with default values (max,min,mean or average value) for an attribute which contains Null Values. This option is non-effective because it is possible that, for example one value, the highest value is placed instead of all the missing values and this solution is not accurate.

Special Coding: Filling in the missing values with a certain code indicating the missing value. This option is also not effective as placing indications of the absence of values does not solve the problem.

Some other solutions through filling in the missing values with estimated values using relationships that exist among the complete values in the dataset to estimate the missing data and, filling in the missing values with the estimated values using techniques and algorithms such as decision trees, fuzzy neural network, or other methods which may represent more effective solutions[9].

## 1.4 INTRODUCTION TO DATA MINING

The success of database systems in traditional applications encouraged developers of other types of applications to attempt to use them, like data mining applications which analyze large amounts of data. The data mining is "A field of study that emerges from statistics, machine learning and database systems. As a discipline for data analysis, statistics contribute significantly towards data mining in terms of fundamental theories and methods for data analysis, measures for evaluating significance and relevance of patterns , and so on" [9].

The data mining can be defined by the following: "Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner" [10].

Some of the most important objectives of data mining are [11]:

### 1.4.1. DESCRIPTION

All the processes must be transparent to the user and the results should clearly describe the available data, their classification and the relationship between them.

### 1.4.1 ESTIMATION

Estimation can be described as "similar to classification except that the target variable is numerical rather than categorical. Models are built using "complete" records, which provide the value of the target variable as well as the predictors. Then, for new observations, estimates of the value of the target variable are made, based on the values of the predictors"[11].

### 1.4.2 PREDICTION

It is a process of building a predictive model, well structured, with good predictions using data available to it and applying certain techniques to provide successful futuristic values. In these applications, not only we are interested in obtaining accurate predictions of the future but also in learning the relationship between the characteristics of the observations and the forthcoming events [12].

### 1.4.3 CLASSIFICATION

Can be defined as the “Most commonly used technique for predicting a specific outcome such as response / no-response, high / medium / low value customer, likely to buy / not buy”[13].It is also a process of building a model describing a set of categories of data or principles that have been specified in advance and using this model as a descriptive or predictive model [12].

The scale of accuracy and the error rate of the model can be defined by the following equations [12]:

Precision=number of correct predictions / total predictions.

Error Rate=number of false predictions / total predictions.

These classification techniques are more suitable to descriptive or predictive groups of data.

### 1.4.4 ASSOCIATION

The association task for data mining can be defined as “the job of finding which attributes “go together” and this is most prevalent in the business world, where it is known as affinity analysis or market basket analysis. The association task seeks to uncover rules for quantifying the relationship between two or more attributes [11].

The following diagram depicts generic 3-tier architecture for Data Mining:

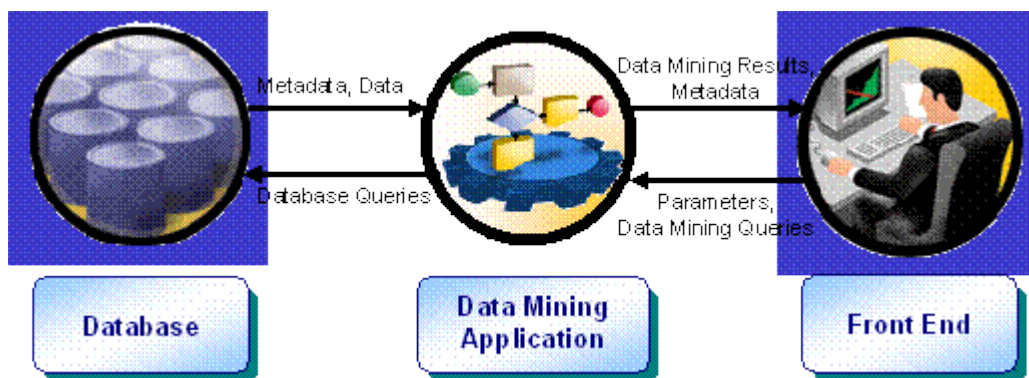


Figure (4) [14]

### PROCESSING MISSING DATA IN DATA MINING

With the emergence of the main frames in the sixties of the last Century, there was an increase in the ability to collect and store large amounts of data in databases specific for each institution or for other purposes. The collected data form a reflection of what exists in reality and is an image of the real world, and hence as the database contains a large amount of data it also contains some missing data . Many treatments were presented to process these missing data and as shown in the following:



In 1966 Afifi and Elashoff presented a review of the publications on missing data and data mining , and in 1972 Orchard and Woodbury presented the expectation – maximization algorithm (EM) to provide unbiased expectations when the data are missing at random (MAR) [ 12 ]. In 1977 A. P. Dempster, N. M. Laird and D. B. Rubin presented another method to get the maximum likelihood using the EM algorithm "Maximum Likelihood from Incomplete Data via the EM Algorithm. This algorithm computes maximum likelihood estimates from incomplete data and presents them at various levels of generality [15]. Little and Rubin in 1987 took statistical analysis with the missing data into consideration. In 1983 Al-Hamami introduced a computational method for dealing with the missing data through the formation of two databases, one complete and another incomplete to provide accurate inquiries and also to provide the correct answers in case there is a shortage of some of the information [12]. In 1993 Agrawal , Imielinski and Swami presented the association rules for the first time in their work, "Mining Association Rules between Sets of Items in Large Databases" [16], and in 1994 Agrawal and Srikant presented two algorithms , "Apriori and AprioriTid , for discovering all significant association rules between items in a large database of transactions"[17]. In 1997 Graham and his group discussed the use of the EM algorithm for the expectations of the means and covariance measurements with the incomplete data. Neural networks were used to build demonstration models by showing groups of data in the search for related variables or groups of variables. Also in 1994 Lewis presented a good look on Genetic algorithms, which are data technologies based on learning for search and optimization problems [12].

Briefly, the research work and publications to date on the various methods of data mining treated the problem of dealing with the missing data through basic theories and algorithms of which some are listed below [11 ] :

- K-Nearest Neighbor Algorithm
- Decision Tree
- Association Rules
- Neural Networks
- Genetic Algorithms

There is no specific theory being used in the selection of data mining technique among the many available and the selection is usually based on experience in this field and on the actual experience with these techniques and their effectiveness. Also the preference between the traditional and the modern techniques could depend on the extent of the availability of the suitable tools, and with increasing experience we can evaluate the options, determine the appropriate option and apply it.

In this research work we will handle the treatment of the missing values through their estimation using the following data mining algorithms:

- Decision tree algorithm for prediction and classification.
- K-nearest neighbor algorithm for estimation, prediction and classification.

## 1.5 INTRODUCTION TO K- NEAREST NEIGHBOR ALGORITHM (K-NN)

Most people, when faced with a problem, tend first to look at similar solutions to the problem they are facing which had been solved earlier. The technique k- nearest neighbor is a classification technique, ie. it classifies similar cases and calculates the number of cases for each class and allocates the new case to the same class to which most of its neighbors belong, as well this technique can be used for estimation and prediction[12].

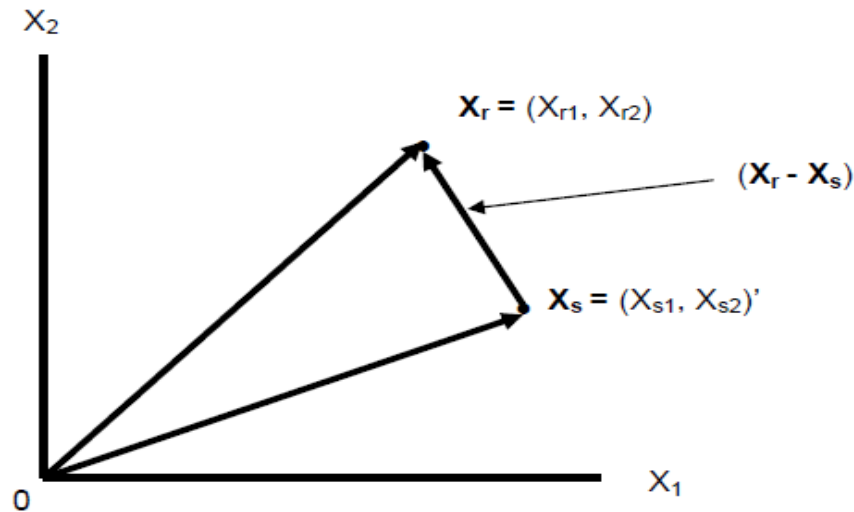
This algorithm determines which records are more similar to the new unclassified record and establishes how these similar records can combine to provide a classification decision for the new record through a combination function. The most basic combination function is simple un-weighted nearest neighbor scores, while the other function is weighted nearest neighbor scores which is used when the neighbors that are closer or more similar to the new record are weighted more heavily than the more distant neighbors [11].

The most common distance function is Euclidean distance, which represents the usual manner in which humans think of distance in the real world[11]:

$$d_{\text{Euclidean}}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_i (x_i - y_i)^2}$$

For the two input variable case, the Euclidean distance between two input vectors  $x_r$  and  $x_s$ , for example, can be easily represented in two-dimensional space. The two vectors  $x_r = (x_{r1}, x_{r2})$  and  $x_s = (x_{s1}, x_{s2})$  are represented in Figure (5) below. The distance between these two vectors is computed as the length of the difference vector  $x_r - x_s$ , denoted by [18]

$$d(\mathbf{x}_r, \mathbf{x}_s) = \|\mathbf{x}_r - \mathbf{x}_s\| = \sqrt{(x_{r1} - x_{s1})^2 + (x_{r2} - x_{s2})^2} .$$



$$d(\mathbf{x}_r, \mathbf{x}_s) = \|\mathbf{x}_r - \mathbf{x}_s\| = \sqrt{(x_{r1} - x_{s1})^2 + (x_{r2} - x_{s2})^2}$$

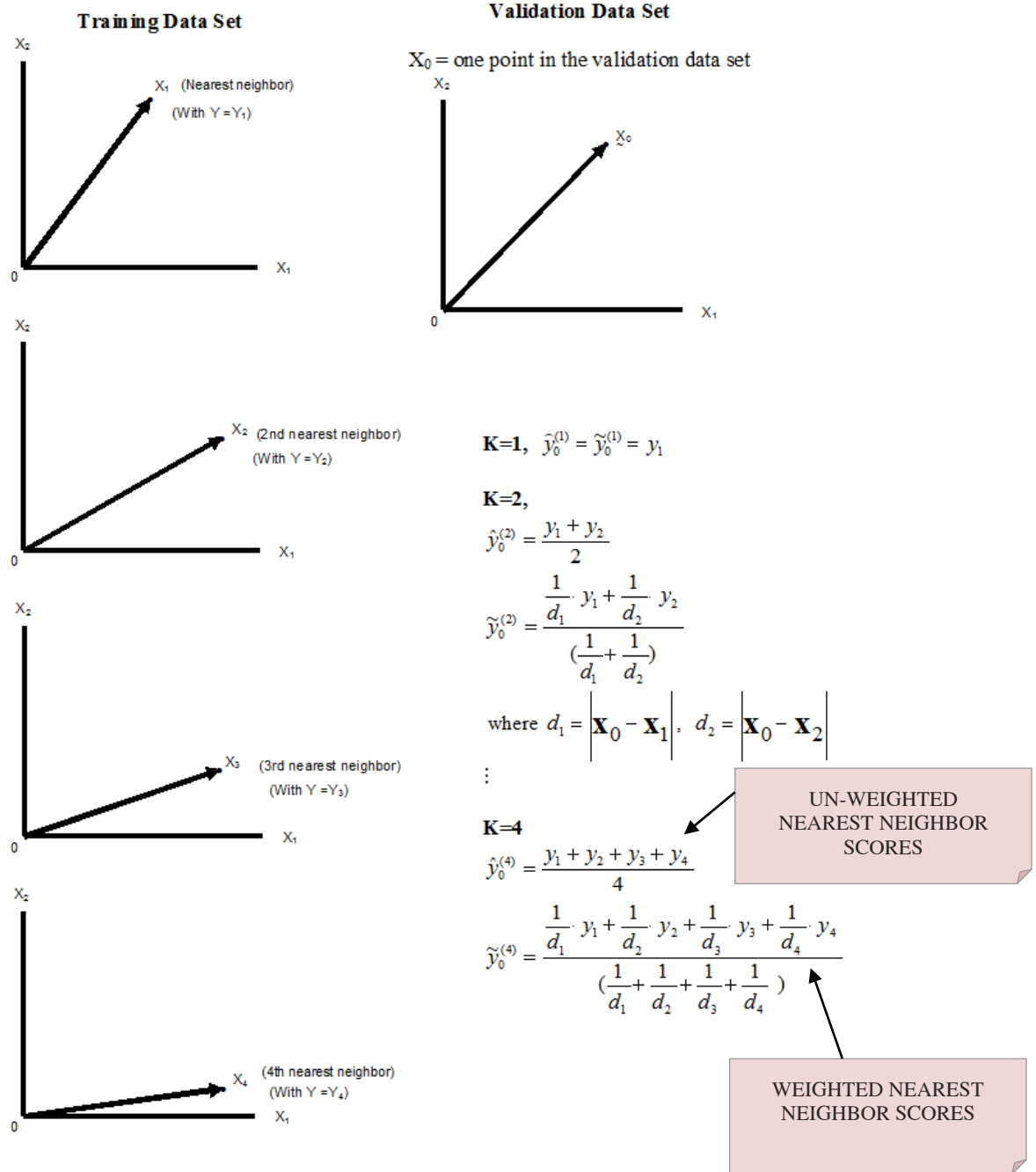
**Figure (5), the Euclidean Distance between Two Vectors  $\mathbf{x}_r$  and  $\mathbf{x}_s$  [18]**

More generally the distance between two  $p$ -dimensional vectors  $\mathbf{u}=(u_1, u_2, \dots, u_p)$  and  $\mathbf{v}=(v_1, v_2, \dots, v_p)$  is calculated as :

$$d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\| = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + \dots + (u_p - v_p)^2} .$$

The K-NN method depends on measurement of distance. To start our intuition on how the K-NN method works, consider the representation in Figure (6) below. On the left-hand-side, four separate points are represented in the training data set, namely,

$x'_1 = (x_{11}, x_{12}), x'_2 = (x_{21}, x_{22}), x'_3 = (x_{31}, x_{32}), x'_4 = (x_{41}, x_{42})$  with their associated output values, respectively,  $y_1, y_2, y_3$  and  $y_4$ . On the right-hand-side, the method represents the first input vector in the validation data set  $x_0 = (x_{01}, x_{02})$  to be "scored." ; ie the method is used to predict the associated output value,  $y_0$ . In this figure,  $x_0$  is positioned so that its closest "neighbor" is  $x_1$  that is distance  $d_1 = d(x_0, x_1)$  from  $x_0$ ; its next closest neighbor is  $x_2$  that is distance  $d_2 > d_1$  and similarly to the other points,  $x_3$  and  $x_4$ , in the training data set such that  $d_4 > d_3 > d_2 > d_1$  [18]. Figure (6) illustrates the scoring of a validation input vector  $x_0$  using some of its "neighbors" in the training data set.



**Figure (6), the Scoring of a Validation Input Vector  $x_0$  Using Some of Its "Neighbors" in the Training Data Set [18]**

### 1.5.1 INTRODUCTION TO DECISION TREE ALGORITHM

The aim of the decision tree is either for classification or for prediction and the flexibility of the decision tree which makes it a very attractive option. The study and the use of the decision tree is not only widespread in the fields of probability and statistical pattern recognition but also widely used in various fields such as medicine (diagnosis), computer science (data structures), botany (classification) and psychological (decision theory). It is possible to display the classification trees as images of schemes to facilitate their translation rather than just being digital translations [12].

The decision tree can be defined as “a collection of decision nodes, connected by branches, extending downward from the root node until terminating in leaf nodes. Beginning at the root node, which by convention is placed at the top of the decision tree diagram, attributes are tested at the decision nodes, with each possible outcome resulting in a branch. Then, each branch leads either to another decision node or to a terminating leaf node [11].

Figure (7) provides an example of a Decision tree using flow chart symbols [19].

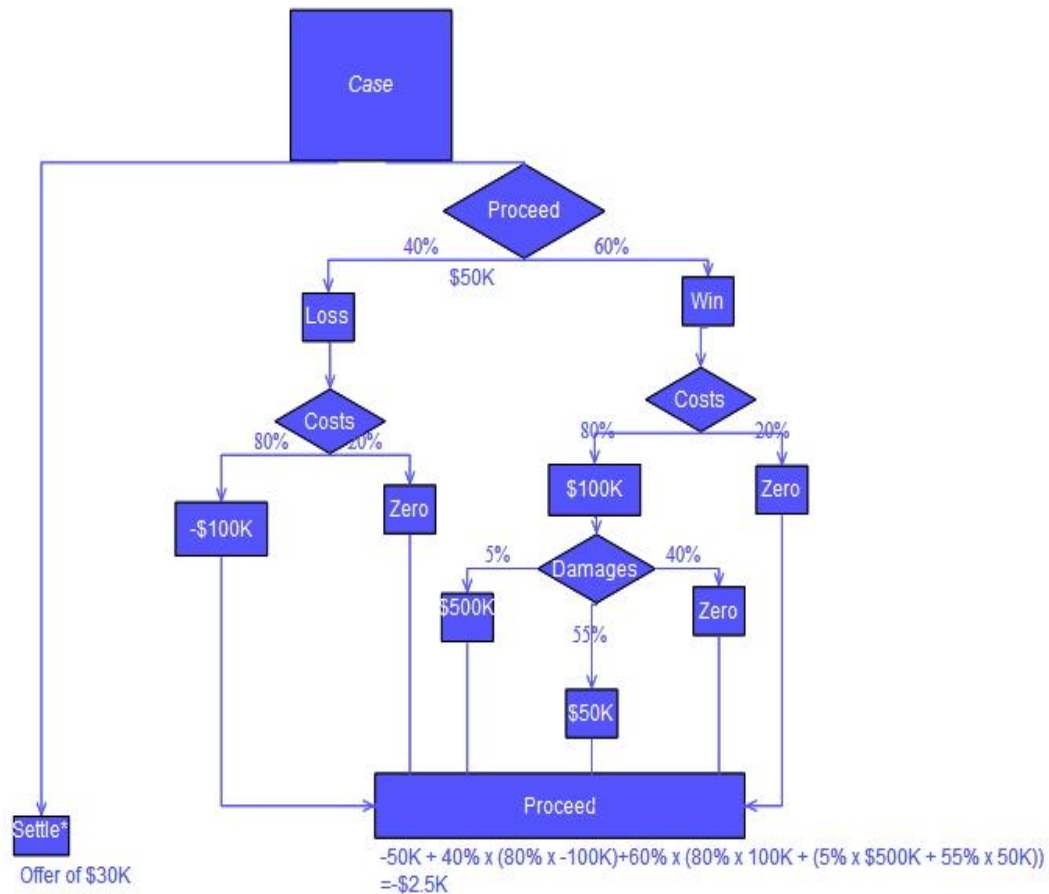


Figure (7) [19]

The decision trees are used to handle missing values . Normally when missing data need to be replaced , the method is to record the number of known elements in the training set that follow each branch to those specific character and then use the most popular branch as a default value for the missing data [12].



### 1.5.2 Organization of the Contents:

- In chapter one, the present research work introduces the concepts of Data Base ,Missing Values, Data Mining , Processing of Missing Data in Data Mining ,K-Nearest Neighbor Algorithm and Decision Tree Algorithm .
- In chapter two, the research problem for the thesis is discussed and some literature surveys are summarized.
- In chapter three, description of the proposed frame work to treat the missing values is introduced showing the components of the new proposed method (Decision tree, K mean – nearest neighbor algorithm, K frequency – nearest neighbor algorithm, checking database) together with description of the full algorithm for the frame work for processing the missing values.
- In chapter four, an example of the execution of the frame work is presented.
- In chapter five, the conclusions and recommendations for future work are presented.

# CHAPTER TWO

## LITERATURE SURVEY

### 2.1.INTRODUCTION

In this chapter, the statement of the problem is illustrated and a summary of some of the previous recent work is given.



## 2.2.The Research PROBLEM

The users of the database systems face the problem of dealing with missing values in data entry, in retrieving results from existing databases, in mathematical treatment of data and in differentiating between the types of missing values whether applicable or inapplicable . There are several reasons behind these missing values; the value is currently not known, it is missed on entry as a result of human error, it does not exist as such or for other reasons. Upon using the databases, the impact of the missing values appear in the statistics and reports that should reflect the real data stored in the databases. For example, when calculating the average of a number of values, in a database containing missing values, the sum of these values is divided by their overall number including the number of the missing values and this is a big mistake. At the same time, we note that it is not easy mathematically to process the Null Values and to differentiate between the reasons for the missing values whether they are applicable or inapplicable; for example, the number of children of a single (unmarried) man/woman entry as Null Value, which is inapplicable.

The treatment of the Null Values in the present research work shall be through the estimation of the applicable values and the marking of the inapplicable values, which will help in the generation of the reports and statistics and in retrieval of data that reflect the reality and the true image of the database.

These estimations and marking are done every time a report or statistics or data retrieval are made so that they reflect real world database by showing the true and estimated values and the final results.

The treatment shall be through estimating and replacing the missing values with approximate values obtained from the development of an algorithm that combines the two algorithms of data mining for estimations and predictions.

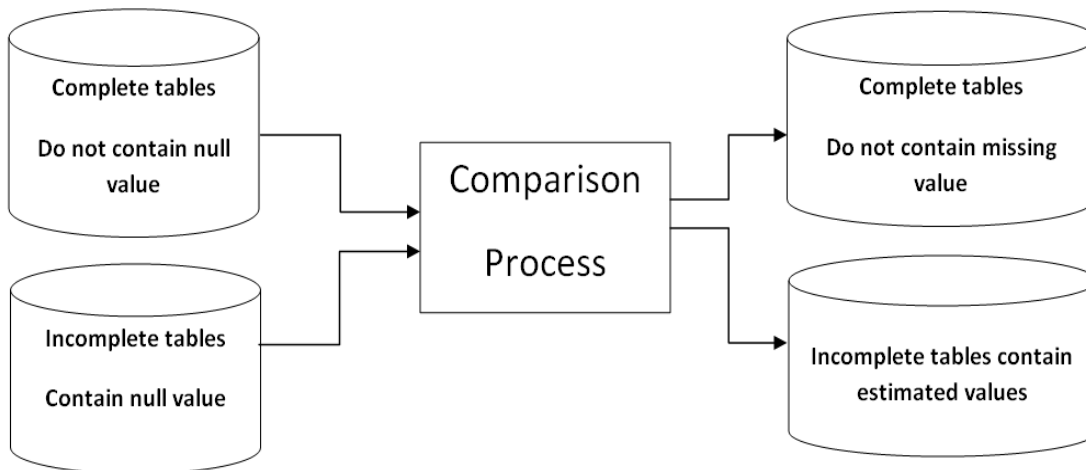
The two algorithms are:

- Decision tree algorithm.
- K-nearest neighbor algorithm after amendment through this research work.

A standard database “ADULT” consisting of (32561) tuples obtained from UCI data repository is used in the present work[20].

Figure (8) shows the flow Diagram that will be used for the development of the proposed algorithm:

Figure (8) Diagram of the Proposed System



## 2.3 PREVIOUS WORK

Below is a review of the most recent work carried out in the field of treatment of Null Values in relational database systems:

**JOHN GRANT, “Null Values in Sql” [21]:** This paper showed that the problem that produces incorrect answers to some queries where a Null Value is included in a table is not that SQL uses three logic values produced by E.F.CODD rather than two logic values produced by Date, but rather in the way that SQL uses the three-valued logic in query evaluation.

**Shin-Jye Lee and Xiaojun Zeng,” A Modular Method for Estimating Null Values in Relational Database Systems” [22] :** This paper presented a modular method to estimate Null Value in relational database systems by constructing automatic fuzzy system, which integrates advantages of fuzzy system and simple linear regression model simultaneously. This proposed method can effectively achieve better performance on relational database estimation.

The presented treatment is lengthy and complex and consists of the following stages:

- A. Partition Determination
- B. Generating Fuzzy Membership Functions based on the  $\alpha$ -cuts of Fuzzy Sets.
- C. Computing Fuzzy Sets with the Method of Least Squares.
- D. Processing the Second Function Approximation by the Method of Simple Linear Regression Model.
- E. Calculating the Differential Rate and Process the Third Function Approximation.

It is possible to find a shorter and less complex method to estimate Null Values.

**Zhang Xia ,Yu Haiyan and Xu Mingzhu ,” Null Value Estimation Method based on Information Granularity for Incomplete Information System”**

**[23]** : This paper presented a Null Value estimation algorithm GRCC(Granular Computing Completer: Null Value estimation algorithm is based on information granularity). The degree of uncertainty of the information system is measured by information granularity; if attributes set A has higher information granularity, it means there are some attribute values in set A occupy master position to classification and the information system has the lower uncertainty degree. GRCC works as follows: choose the attribute set which has the maximum information granularity as the first granular layer; then get the range of the Null Values according to the corresponding attribute value of the tolerance class; finally, choose the minimum attribute value in the range which is mentioned above based on MDL (Minimum Description Length) principle as the Null Value. Thus, the first granular layer is completed. Repeat the above step until the whole incomplete information system is completed. The results show that the rules mined through the method based on GRCC algorithm have high support degree and confidence degree. The uncertainty degree of the incomplete information system can be decreased effectively.

The method they followed for estimation can produce more rules, which means that it can produce more strong association rules, but with the passage of time and with the new data input to the databases , these rules became invalid and gave inaccurate results.

**Muhammad Firoz Mridha and Manoj Banik ,”Performances of Estimating Null Values using Noble Evolutionary Algorithm (NEAs) by Generating Weighted Fuzzy Rules” [24]:** This paper presented a noble technique to estimate Null Values from relational database systems, the technique generated weighted Fuzzy rules from relational database systems for estimating Null Values using Noble Evolutionary algorithms. The parameters (operators) of the Evolutionary algorithms are adapted via Fuzzy systems. They fuzzified the attribute values using membership functions shape. The results of the evolutionary algorithms are the weights of the attributes. The different weights of attribute generate a set of Fuzzy rules. From this they obtained a set of rules. The techniques enable the estimation of the Null Values in relational database systems; this proposed method can get a higher average estimated accuracy rate.

The basic idea of this method which is rule base is close to the tuple having a Null Value. The Null Value can be estimated by the closeness degree of the tuples with respect to the closest rule but as time passes these rules became invalid and give inaccurate results.

**Jing Yang, Ze Jiang, Jianpei Zhang, Lejun Zhang,” A Null Value Estimation Method Based on Similarity Predictions in Rough Sets”[25]:** This paper presented the study of theory and method which is based on how to change the incomplete information system into a complete information system . For the Null Value estimation problem in incomplete system, an improved method is introduced by this paper in view of the weakness of Null

Value estimation based on similar relational algorithm (SIM-EM), proposed in this paper from the perspective of predicting value in collaborative filtering technology. The improved algorithm is good at dealing with sparse rough set, and the accuracy and the mean absolute error is better than the original method, and the quality of estimation is guaranteed.

The method in this paper is feasible, but, the complexity of the algorithm needs to be reduced.

**Shichao Zhang, “Shell-neighbor method and its application in missing data imputation” [26]** :In this paper an approach called SN (Shell Neighbors) imputation, or simply SNI was introduced. The SNI fills in an incomplete instance (with missing values) in a given data set by only using

its left and right nearest neighbors with respect to each factor (attribute), referred to as Shell Neighbors. The left and right nearest neighbors are selected from a set of nearest neighbors of the incomplete instance. The size of the sets of the nearest neighbors is determined with the cross-validation method .Then the SNI is generalized to deal with missing data in datasets with mixed attributes. The results demonstrated that the SNI is more effective than the KNNI (K nearest neighbor imputation) method.

Additional work is needed to apply the SNI approach to real machine learning and data mining application to enable improvement.

**Renu Vashist and M.L Garg , “A Rough Set Approach for Generation and Validation of Rules for Missing Attribute Values of a Data Set “ [27]:** This paper used the most common attribute value approach by replacing all the missing attribute values by most frequently occurring attribute values and



thereby completing the information table . Decision rules are generated using the reduct of the decision table and these rules are also validated using ROSE2 (Rough Sets Data Explorer) software.

This method is based on the generation of rules for estimating the missing values (Null Values), but as time passes and with the new data input to the databases these rules became invalid and give inaccurate results.

**Rohit Raghunathan ,Sushovan De and Subbarao Kambhampati, “Bayes Networks for Supporting Query Processing Over Incomplete Autonomous Databases” [28]:** This paper presented a principled probabilistic alternative that views an incomplete tuple as defining a distribution over the complete tuples that it stands for. The authors’ approach involved mining/"learning" Bayes networks from a sample of the database, and using it to do both imputation (predict a missing value) and query rewriting (retrieve relevant results with incompleteness on the query-constrained attributes, when the data sources are autonomous). This method presented empirical studies to demonstrate that (i) at higher levels of incompleteness, when multiple attribute values are missing, Bayes networks do provide significantly higher classification accuracy and (ii) the relevant possible answers retrieved by the queries reformulated using Bayes networks provide higher precision and recall than AFDs (Approximate Functional Dependencies) while keeping query processing costs manageable.

Learning and inference on Bayes nets can be computationally expensive which might inhibit their applications in handling incompleteness in autonomous data sources.

**Kavita Pandole and Niket Bhargava, “Comparison and Evaluation for Grouping of Null Data in Database Based on K-Means and Genetic Algorithm” [29]**

:This paper estimates Null Value in relational database systems by applying K-means method along with genetic algorithm. The aim is to generate right number of clusters and also attain a high accuracy. The experimental results clearly indicated that genetic algorithm produces optimal solution as compared to K-Means method. It is also observed that Genetic algorithm generates right number of clusters, and also achieves higher accuracy rate and thus minimizes error.

This work basically focused on the merits of using evolutionary algorithm for data clustering that contains missing values and clearly compares it to the existing K-Means clustering method.

**A.Azadeh , S.M. Asadzadeh, R. Jafari-Marandi, S. Nazari-Shirkouhi, G. Baharian Khoshkhou ,S. Talebi , A. Naghavi, “Optimum estimation of missing values in randomized complete block design by genetic algorithm” [30]**

:This paper estimated missing values by using Genetic Algorithm (GA) approach in a Randomized Complete Block Design (RCBD) table and compared the computational results with three other methods, namely, Particle Swarm Optimization (PSO), Artificial Neural Network (ANN) and approximate analysis and exact regression method. Furthermore, 30 independent experiments were conducted to estimate missing values in 30 RCBD tables by GA, PSO, ANN, exact regression and approximate analysis methods. Computational results indicated that the best answer (in the last 10-chromosome population) obtained by GA is frequently the same as the missing value, with the mean value being close to the missing observation.

It was concluded that GA provides much better estimation compared to other methods. The superiority of GA is shown through lower error estimations and also Pearson correlation experiment. Therefore, it was suggested to utilize GA approach of this study for estimating missing values for RCBD.

This method to estimate Null Value is complex and it is possible to find a simpler method to estimate Null Values.

**N.C. Vinod and M. Punithavalli, “Performance Evaluation of Mutation / Non-Mutation Based Classification With Missing Data”[31]:** In this paper it was shown that the existence of missing data in databases creates problems in almost all steps of data mining ,and, a method based on KNN combined with imputation and a method which uses a two stage approach without imputation are analyzed. The first method uses a feature-weighted distance metric combined with KNN classifier to handle incomplete data for classification while the second method first divides the dataset into disjoint subsets according to the attributes with missing values. Using these subsets the classification process is performed. Several experiments were conducted and the results proved that both models perform well with missing data.

**Ibrahim Berkan Aydilek, Ahmet Arslan,“A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm” [32]:** This paper presented a study which utilizes a fuzzy c-means clustering hybrid approach that combines support vector regression and a genetic algorithm. In this method, the fuzzy clustering parameters, cluster size and weighting factor are optimized and missing values are estimated. The experimental results demonstrated that the fuzzy c-means genetic algorithm imputation yields a more sufficient, sensible estimation accuracy ratio for suitable clustering data.

The method used to estimate Null Value is complex and it is possible to find a method that is less complex to estimate Null Values.

**Shamsher Singh and Jagdish Prasad, “Estimation of Missing Values in the Data Mining and Comparison of Imputation Methods” [33]:** In this paper it was shown that Missing data is a common problem in data mining. An empty entry in the database sometimes indicates that the value is zero or, in some cases, cannot possibly exist (e.g. an entry for an individual in a field SALARY when the individual is a baby). However, in many cases, a blank field represents an unknown quantity and techniques based on correlation can then be used to complete that entry. Where mapping functions are developed within the techniques used to compute correlations, this should enable the replacement of missing values to be achieved more effectively.

**Pablo H. Ibarguengoytia, Uriel A. Garc´ia, Javier Herrera-Vega, Pablo Hern´andez-Leal, “On the Estimation of Missing Data in Incomplete Databases: Autoregressive Bayesian Networks” [34] :** This paper introduced an approach called an autoregressive Bayesian networks (AR-BN), a variant of Dynamic Bayesian Networks for completing databases which exploits latent variable relations while still benefitting from autoregressive information of the variable being filled. Using AR-BN, new estimated values are calculated using inference in the dynamic model. The results unveiled how the interplay between the variable autoregressive information and the variable relationship to others in the dataset is critical to

selecting the optimal data estimation technique. AR-BN appears as a good candidate ensuring a consistent performance across scenarios, datasets and error metrics.

This method for estimation of missing data is complex and it may be possible to find a less complex method to estimate these data.

**Mohammad Hossein Norouzi Beirami, Mohammad Hossein Nejad Ghavifekr, Rahim Pasha Khajei, “Predicting Missing Attribute Values Using Cooperative Particle Swarm Optimization”[35]:**This paper presented a method using Cooperative Particle Swarm Optimization for predicting attribute values. This method utilizes data records for convergence to missing attribute values without extracting data relations. The method does not need knowledge of a professional person for detecting relation between data. The algorithm has been done on whether casting of Tabriz which has gain for 50 years that shows accuracy 98.45 percent of missing data .The problem with this method is the execution time of the algorithm. Since this method is based on population and needs to be repeated for several generations to get the results, it is thus time consuming in comparison with methods based on rules.

The most popular techniques for estimating the missing values (Null Values) used the generation of rules; but with the passage of time and with the new data input to the databases these rules became invalid and gave inaccurate results. Other techniques to estimate Null Values are more complex and do not rely on the generation of rules and on estimations based on these rules and base their estimates on the data available in the database.

Each of the above mentioned two techniques have a number of drawbacks; in the first and with the passage of time, the use of rules in the estimation became non useful and not valid while in the second technique in which the estimations are based on the data stored in the database, a negative impact is noted and the estimation become inaccurate when the stored data are incompatible and contain anomalous and incorrect values.

It is possible to find methods that are less complex to estimate these Null Values. Our proposed technique for estimating the missing values integrates the use of the rules , selected by the user to provide flexibility to change these rules when needed, together with the use of the data stored in the database, through the use of these rules to check the data stored in the database and then estimate the missing values using K-nearest neighbor algorithm and the Decision tree algorithm and finally checking the results against the available rules to verify the accuracy of estimations.

In this way, we can maintain the accuracy of the available data, as well as the accuracy of estimations so as not to have conflicts with the rules; as well the estimations are based on the stored data and not on the rules and in this way we have achieved the overcoming of the weaknesses embodied in the techniques proposed by previous research work.

# CHAPTER THREE

## THEORETICAL DESIGN



### 3.1.INTRODUCTION

As mentioned earlier, any designed database system aims at providing integrity of the data and some problems appear when users use these systems and deal with missing values which are either unknown to the user or non-existent. The databases, as we know should be images of the real world and thus could contain missing values. In our proposed research work we will treat the presence of missing values in the

database through a frame work to process these missing values, known as the Null Values, by estimating the Null Values through the use of data mining algorithms:

- Decision tree algorithm for prediction and classification.
- K-nearest neighbor algorithm for estimation, prediction and classification.

This chapter shall cover the followings:

- Description of the proposed flow diagram for the frame work to treat the missing values.
- Components of the new method
  - ❖ Decision tree.
  - ❖ K mean – nearest neighbor algorithm.
  - ❖ K frequency – nearest neighbor algorithm.
  - ❖ Checking database.
- New full algorithm for a frame work for processing missing values.

### **3.2. Description of the proposed flow diagram for the frame work to treat the missing values**

The flow diagram for the frame work as shown in Figure (9) uses as input complete and incomplete tables considered as one table and also a business rules table fixing the used rules which are of high flexibility, as determined by the system user. The stored values of the database are checked and tested with the business rules using decision tree to classify the stored values as true or false (entered as error entries contradicting the set rules) or estimated values. The false and estimated values will be replaced by Null Values which in turn will again be estimated. Thus the Decision tree



algorithm is reused to classify the value of the Null Value whether inapplicable as for the case of the number of children of an unmarried man or woman which is considered inapplicable or applicable as the date of birth which is inevitable for everyone, in which case the Null Value is applicable.

In the case the Null Value is classified as inapplicable a value is placed as an indicator for the inapplicable Null Value, while in the case the Null Value is classified as applicable and as known that for each attribute there is a relationship between it and the rest of the relevant attribute as the date of graduation has a close relationship with the date of birth and also an individual income has a close relationship to the educational level and on this basis we find the closest distance between the record containing Null Value and its peers that does not exceed the distance requirements specified by the set rules like comparing the income of an individual who holds a master degree with fellow peers holdings master degree and cannot be compared with those who do not hold any degree, ie, the closest distance within certain conditions.

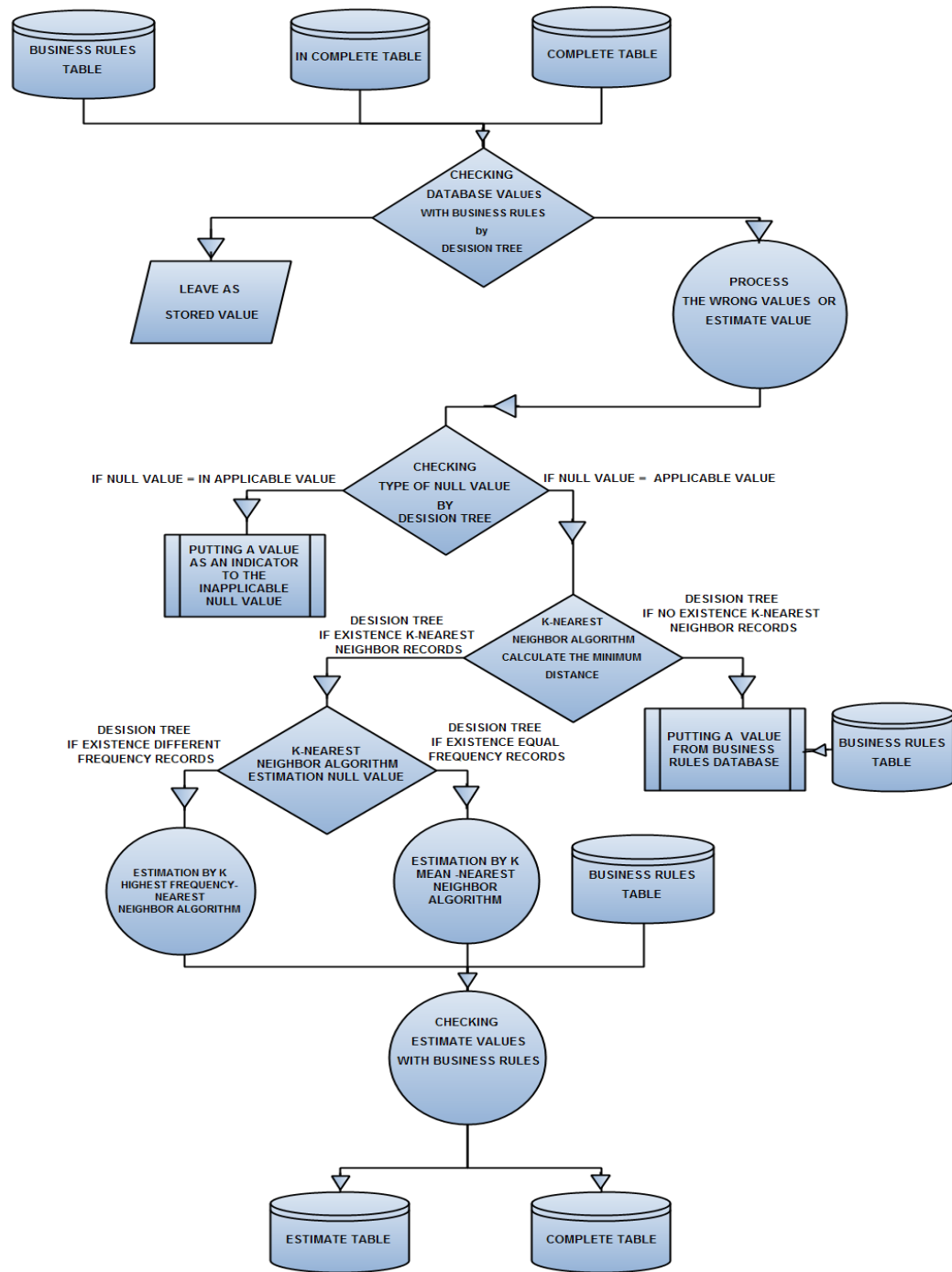
In case no similar records are specified as in the case of a new system containing empty database, and records are entered for the first time and there is no value to compare with; in this case the table of rules is used to estimate the Null Value.

In case close values are found, ie. K-nearest neighbor, the Decision tree algorithm will be used to classify the existing records whether they hold equal or different frequencies.

In case the records hold equal frequencies, then K mean –nearest neighbor algorithm will be used to estimate the Null Value and in case of different frequencies K frequency – nearest neighbor algorithm will be used.

The estimated values will be re-checked by comparing them with the business rules for validation of these values and the output will be a table with actual values and another table with estimated values.

The proposed frame work provides estimates for missing values as well as a precise system to check and validates the stored and estimated values paying attention that the key value cannot be Null Value and cannot be estimated.



**Figure (9)**  
**Flow Diagram for Frame Work**  
**(ESTIMATION OF NULL VALUE**  
**IN RELATIONAL DATA BASE SYSTEM)**

### **3.3 Components of the new Method**

The new method consists from the following:

#### **3.3.1 Decision Tree**

The use of the decision tree in the proposed frame work helps in adding big flexibility to the treatment process of Null Value and in estimating the missing value; for example, the process of classifying the stored value in the database as true, estimated or false, as well as predicting that the value is inapplicable or applicable and also classifying the estimation process based on equal or different frequency of the related records.

To be noted, the decision tree is used in many processes in the frame work, whether for the purpose of classification or prediction, which add flexibility to the treatment.

#### **3.3.2 K – nearest neighbor algorithm**

The basis for the process of estimation in the proposed treatment is the use of K - nearest neighbor algorithm. The main objective behind the use of this algorithm is to estimate the missing value by comparing with similar or near records and estimating on that basis either depending on the value of the higher frequency or on the average of these values; for example in the case of a missing value for an employee's date of birth it is possible to estimate his date of birth from the graduation date of this employee's peers. It is also possible to incorporate more than one attribute related to the estimated value , for example to estimate the annual income of a person it is possible to rely

on the appointment date , academic degree, age and nature the work and combine all these attributes and find the nearest distance either by relying on one attribute or merging more than one attribute or fixing one attribute or a combination of attributes and finding the nearest distance of the other attributes. For example, it is possible to fix the academic degree and the nature of work and find the closest distance to the age and to date of appointment , i.e. to find a number of records near and similar and estimate from them the annual income either based on the average or on the higher repetition of records or merging them with each other such that if the corresponding values for the missing values have different repetitions ,one can rely on repetition in the estimation and the most repeated is the estimated value, and in the case of equal repetitions one can rely on the average of the values corresponding to the missing values and this average becomes the estimated value.

The following figures (10) and (11), as shown in (3.3.2.1.) and (3.3.2.2.) below, respectively show the flow charts for the K Mean – Nearest Neighbor Algorithm and K Frequency – Nearest Neighbor Algorithm:

### 3.3.2.1K Mean – Nearest Neighbor Algorithm Flow Chart

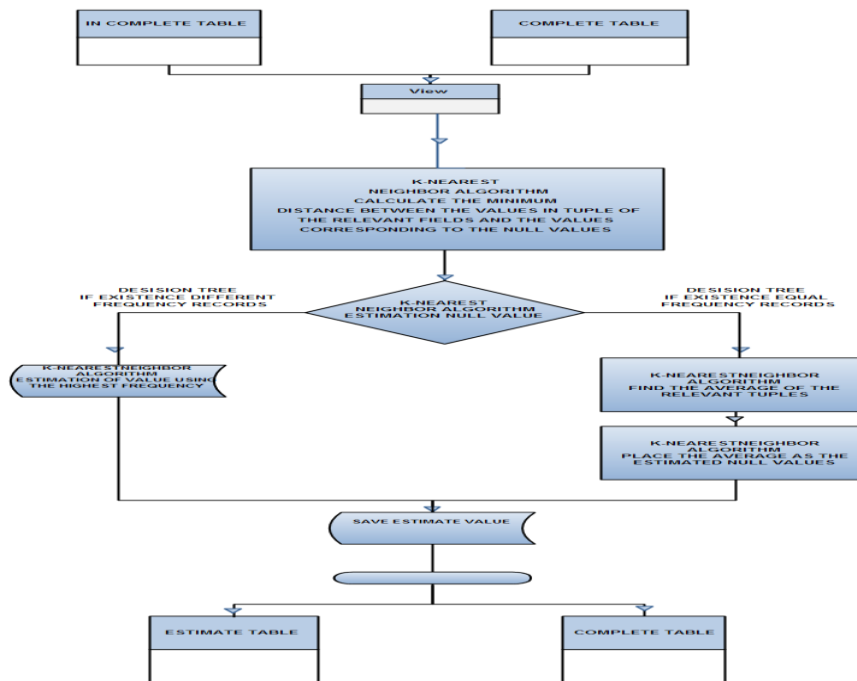


Figure (10) -Flow Chart for the K Mean – Nearest Neighbor Algorithm

**The implementation of the K Mean – Nearest Neighbor Algorithm will be as follows:**

#### **Input: -**

The records stored in two tables (complete table, incomplete table) :

- The value ( $y_*$ ) stored in tuples containing Null Values.
- The value ( $y_j$ ) stored in tuples containing fact values.
- The value ( $x_i$ ) stored in attribute containing actual fact values.

**Treatment: -**

- (A) use the K – nearest neighbor algorithm to calculate the distance between the values in tuple of the relevant fields and the values corresponding to the Null Values ( $x_i$ ) and find the closest distance; shortest distance being within the established limits or under the conditions placed and bring the tuples corresponding to the given condition and use these tuples for estimation.
- Use the Decision tree algorithm to classify and decide whether the existing records are different frequency records or equal frequency records
- In the event the records are classified as different frequency records, estimate the Null Value using K - highest frequency -nearest neighbor.
- In the event the records are classified as equal frequency records, find the average ( $y_j$ ) of the relevant tuples.
- Place the average( $y_j$ ) as the estimated Null Value in ( $y_*$ )
- Check the database values; if the database contains other Null Values return to step (A) and continue.

**Output: -**

Table containing the estimated values in place of the missing values.

### 3.3.2.2.K Frequency – Nearest Neighbor Algorithm Flow Chart

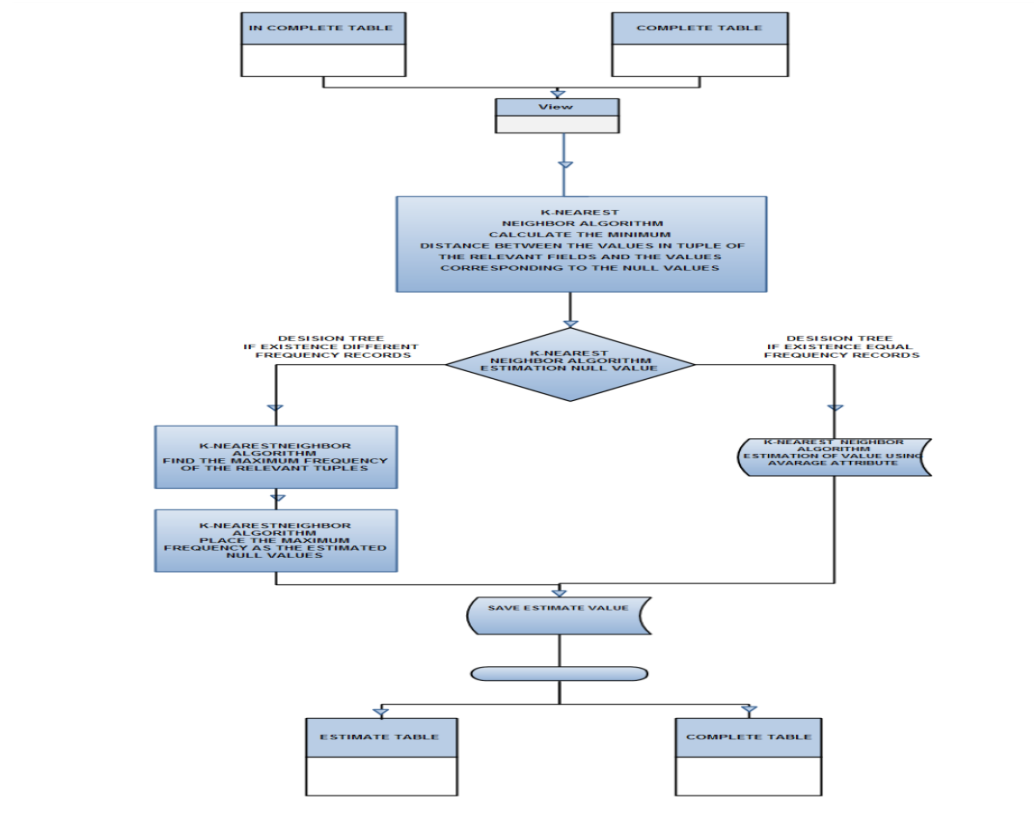


Figure (11) -Flow Chart for the K Frequency – Nearest Neighbor Algorithm

**The implementation of the K Frequency – Nearest Neighbor Algorithm will be as follows :**

#### **Input: -**

The records are stored in two tables (complete table, incomplete table) :

- The value ( $y_*$ ) stored in tuples containing Null Values.
- The value ( $y_j$ ) stored in tuples containing actual fact values.
- The value ( $x_i$ ) stored in attribute containing fact values.



## **Treatment: -**

- (A) Use the K – nearest neighbor algorithm to calculate the distance between the values in tuple of the relevant fields and the values corresponding to the Null Values ( $x_i$ ) and find the closest distance; shortest distance being within the established limits or under the conditions placed, and bring the tuples corresponding to the given condition and use these tuples for estimation.
- Use the Decision tree algorithm to classify and decide whether the existing records are different frequency records or equal frequency records.
- In the event the records are classified as equal frequency records estimate the Null Value using K – mean – nearest neighbor.
- In the event the records are classified as existing different frequency records, find the maximum frequency ( $y_j$ ) of the relevant tuples.
- Place the maximum frequency ( $y_j$ ) as the estimated Null Value in ( $y_*$ ).
- Check the database values; if the database contains other Null Values, return to step (A) and continue.

## **Output:**

Table containing estimated values in place of an incomplete table.

### 3.3.3 Checking the Database

The process of checking and validating the database within the framework is done in two parts of the proposed algorithm.

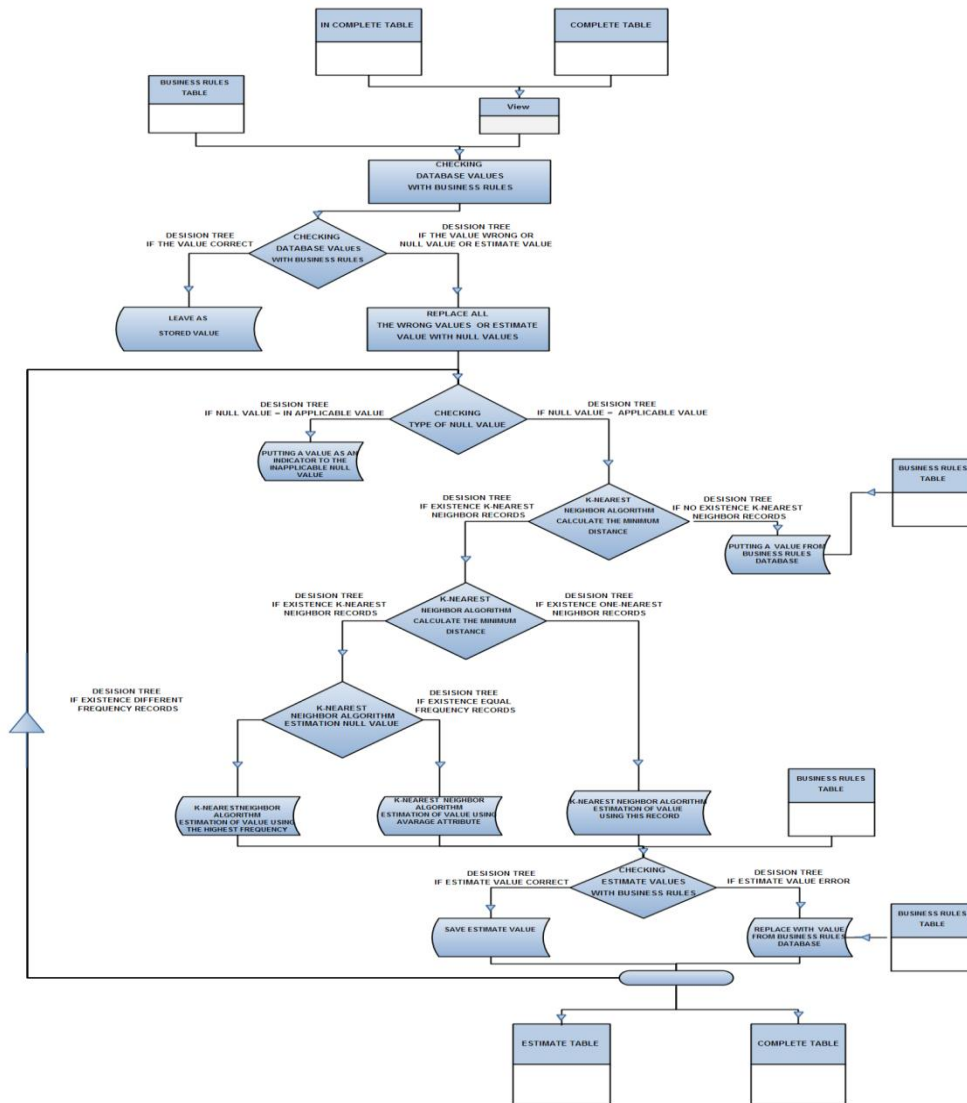
The first part is checking the value stored in the database and in case it is true value, it remains stored as it is, and in case it is false, it is replaced by Null Value and gets re-estimated, while in case the value is estimated it is also replaced by Null Value and also re-estimated for flexibility in estimation, as when we add new records we may get better accuracy of the estimate compared to the previous estimate.

The second part of checking and validating comes at the end of the estimation whereby the true estimated value is checked whether it is within the allowed rules in the business table and in case it is, the estimated value gets stored while if it is false, re-estimation is made from the business table.

Thus, the checking and validation is a process to check the database whether the value is true or false and to correct the error either by estimation or through an approximate value and in this case it helps in maintaining the integrity and correctness of the data.

### 3.4 New Full Algorithm for a Frame Work to process Missing Values

The main aim of the present research work is to process the missing value which also can help in the checking and validation of the database to ensure the correctness of the data contained in the database and as can be seen in the following flow chart detailing the steps of the framework:



The implementation of the frame work will be as follows :

Figure (12) -Flow Chart for the Frame Work (ESTIMATION OF NULL VALUE IN RELATIONAL DATA BASE SYSTEM)

**Input: -**

The main table is divided into two tables:

- The first table contains complete data and no Null Values.
- The second table contains incomplete data ie, any values in the tuples contains at least one Null Value.
- The third table contains business rules.

**Treatment: -**

- Check database values with business rules.
- Use the Decision tree algorithm to classify and decide whether the value is true or false or an estimated value.
- In the event the value is classified as true value, the stored value remains as is.
- In the event the value is classified as false or estimated value, the false or estimated values are replaced by the Null Value.
- (A) Use the Decision tree algorithm to classify and decide whether the Null Value is applicable or inapplicable.
- In the event the Null Value is classified as inapplicable, place a value as an indicator for the value of the inapplicable Null Value.
- In the event the Null Value is classified as containing applicable Null Value use the K – nearest neighbor algorithm to calculate the distance between the values in tuple of the relevant fields and the values corresponding to the Null Values and find the closest distance;

- shortest distance being within the established limits or under the conditions placed and bring the tuples corresponding to the given condition and use these tuples for estimation.
- Use the Decision tree algorithm to classify and decide whether K-nearest neighbor records exist or do not exist.
- In the event no K-nearest neighbor records exist place a value from the business rules table.
- In the event the K-nearest neighbor records exist use the tree algorithm to classify and decide whether one or more k-nearest neighbor records exist.
- In the event there is one-nearest neighbor record, estimate the Null Value using this record.
- In the event more than one K-nearest neighbor records exist, use the tree algorithm to classify and decide whether different frequency records or equal frequency records exist.
- In the event the records are classified as different frequency records, estimate the Null Value using K-nearest neighbor (the highest frequency).
- In the event the records are classified as equal frequency records, estimate the Null Value using K-nearest neighbor (the average attribute of nearest neighbor records).
- Check the estimated values with business rules.
- Use the Decision tree algorithm to classify and decide whether the estimated values are true or false.
- In the event the estimated value is classified as true, store the estimated value.

- In the event the estimated value is classified as false, replace the estimated value with the value from the business rules table.
- Check the database values; if database contains other Null Values return to step (A) and continue.

**Output: -**

A table containing estimated values in place of an incomplete table.

# CHAPTER FOUR

## THE EXPERIMENTAL WORKS



## CHAPTER FOUR

### THE EXPERIMENTAL WORKS

#### 4.1 Introduction

In this chapter, and in order to simplify the way in which the algorithms for the frame work of estimation of Null Value in relational database system are written, we shall present six examples of the execution of the frame work.

#### 4.2 Execution of the Frame Work (Estimation of Null Value in Relational Database System )

A frame work is designed using Oracle language for the estimation of Null Value and this frame work is implemented using the standard database “ADULT” consisting of (32561) tuples obtained from UCI data repository and the “Iraqi Law Number (22) for the year 2008 for Salaries of the state employees” [36] which is applied to the standard database.



## 4.3 Examples of the Use of the Frame Work for Estimation of Null

### 4.3.1 1alue in Relational Database System

### 4.3.2 Example(1) – An Error Value Input

When entering a wrong value as a result of a human error, as shown below, the entered monthly income value is wrong:

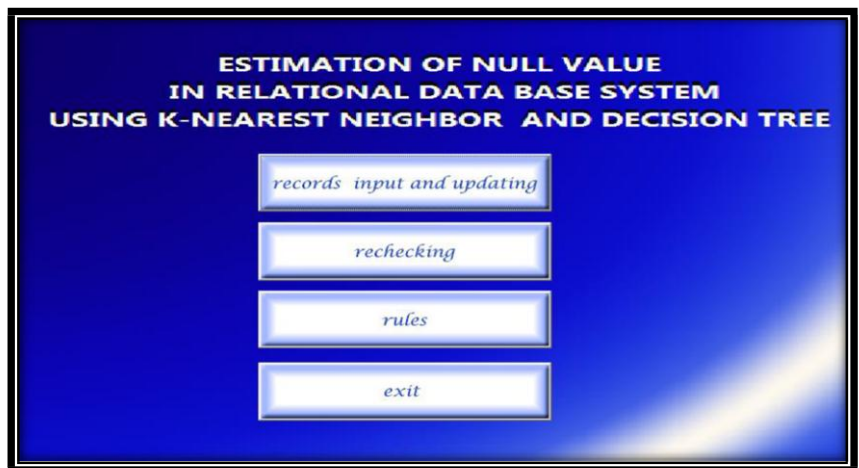


Figure (13) : Example (1)  
Main Interface

ESTIMATION OF NULL VALUE IN RELATIONAL DATA BASE SYSTEM USING K-NEAREST NEIGHBOR AND DECISION TREE			
No	1	Salary	240000
Gender	Male	Scale	D
Education	HS-grad	Degree	8
Education Num	9	Levels	1
Age	42	Marital Reward	50000
Workclass	Self-emp-not-inc	Hours Per Week	40
Marital Status	Married-civ-spouse	Monthly income	10

■ estimated value  
■ inapplicable value

quit gdd delete clear save back  
 << < Search execute Stop Search > >>

Figure (14) : Example (1)

## Input and Updating Records Interface

To treat this situation and before moving to the next record and also before storage, the system will apply the proposed algorithm to cure the problem.

### Treatment: -

Scale	Education	Num	Degree	Levels	Salary	Min	Max	Description
A	1st-4th	2	10	1	140000	777.78	127777.78	
A	5th-6th	3	10	1	140000	777.78	127777.78	
A	Preschool	1	10	1	140000	777.78	127777.78	
B	7th-8th	4	10	4	152000	844.44	134444.44	
C	10th	6	9	1	185000	1027.78	152777.78	
C	11th	7	9	1	185000	1027.78	152777.78	
C	9th	5	9	1	185000	1027.78	152777.78	
D	12th	8	8	1	240000	1333.33	183333.33	
D	HS-grad	9	8	1	240000	1333.33	183333.33	
E	Assoc-acdm	12	8	5	260000	1444.44	194444.44	
E	Assoc-voc	11	8	5	260000	1444.44	194444.44	
E	Some-college	10	8	5	260000	1444.44	194444.44	
F	Bachelors	13	7	1	296000	1644.44	214444.44	
G	Masters	14	6	3	374000	2077.78	257777.78	
H	Doctorate	16	5	1	429000	2383.33	288333.33	
H	Prof-school	15	5	1	429000	2383.33	288333.33	
Z								Age >= 60

☞ Check the database value (monthly income =10) with the business rules.

**Figure (15) : Example (1)  
Rules Interface**

The education rule for HS-grad is that the monthly income is between [1333.33 and 183333.33]

☞ Use the Decision tree algorithm to classify and decide whether the value is true, false or an estimated value.

☞ In the event the value is classified as false, it is replaced by the Null Value and an error message is displayed.



**Figure (16) : Example (1)**

### Error Message

- ❏ Use the Decision tree algorithm to classify and decide whether the Null Value is applicable or inapplicable.
- ❏ In the event the Null Value is classified as containing an applicable Null Value because the age is less than 60 years, use the nearest neighbor algorithm to calculate the distance between the values in

AGE	WORKCLASS	FNLWGT	EDUCATION	EDUCATION_NUM	MARITAL_STATUS	RELATIONSHIP	SEX	HOURS_PER_WEEK	NATIVE
42	Self-emp-not-inc	170721	HS-grad	9	Married-civ-spouse	Husband	Male	40	United-S
42	Self-emp-not-inc	195897	HS-grad	9	Married-civ-spouse	Husband	Male	40	United-S
42	Self-emp-not-inc	221172	HS-grad	9	Divorced	Not-in-family	Male	40	United-S
42	Self-emp-not-inc	248406	HS-grad	9	Married-civ-spouse	Husband	Male	40	United-S
42	State-gov	83411	HS-grad	9	Married-civ-spouse	Husband	Male	40	United-S
42	State-gov	160369	HS-grad	9	Never-married	Not-in-family	Male	40	United-S
42	?	137390	HS-grad	9	Married-civ-spouse	Husband	Male	40	United-S
42	?	175935	HS-grad	9	Separated	Unmarried	Male	40	United-S
42	?	240027	HS-grad	9	Divorced	Not-in-family	Female	40	United-S
42	Federal-gov	46366	HS-grad	9	Separated	Unmarried	Female	40	United-S
42	Federal-gov	53727	HS-grad	9	Married-civ-spouse	Husband	Male	40	?
42	Federal-gov	65950	HS-grad	9	Divorced	Unmarried	Female	40	United-S
42	Federal-gov	91468	HS-grad	9	Divorced	Not-in-family	Female	40	United-S
42	Federal-gov	141459	HS-grad	9	Separated	Other-relative	Female	40	United-S
42	State-gov	304302	HS-grad	9	Married-civ-spouse	Husband	Male	40	United-S
42	State-gov	455553	HS-grad	9	Never-married	Unmarried	Female	40	United-S
42	Federal-gov	178470	HS-grad	9	Divorced	Not-in-family	Female	40	United-S
42	Federal-gov	284403	HS-grad	9	Divorced	Not-in-family	Male	40	United-S
42	Federal-gov	557644	HS-grad	9	Never-married	Unmarried	Female	40	United-S
42	Local-gov	70655	HS-grad	9	Divorced	Not-in-family	Male	40	United-S
42	Local-gov	104334	HS-grad	9	Married-civ-spouse	Husband	Male	40	El-Save
42	Local-gov	143046	HS-grad	9	Widowed	Unmarried	Female	40	United-S

the tuple of the relevant fields and the value corresponding to the Null Value and find the closest distance; the shortest distance being within the established limits, bring the tuples corresponding to the given condition and use these tuples for estimation.

**Figure (17) : Example (1)**

## K-Nearest Neighbor Records

- Use the Decision tree algorithm to classify and decide whether K-nearest neighbor records exist or do not exist.



- In the event the K-nearest neighbor records exist, use the Decision tree algorithm to classify and decide whether one or more K-nearest neighbor records exist.



**Figure (18) : Example (1)  
Records Account**

- In the event more than one K-nearest neighbor records exist, use the tree algorithm to classify and decide whether different frequency records or equal frequency records exist.
- In the event the records are classified as different frequency records, estimate the Null Value using K-nearest neighbor (the highest frequency).

	CNT	SAL
▶	66	103333.33
	60	53333.33

**Figure (19) : Example (1)  
Frequency Records View**

Monthly income=103,333.33

☛ Check the estimated value with the business rules.

**ESTIMATION OF NULL VALUE  
IN RELATIONAL DATA BASE SYSTEM  
USING K-NEAREST NEIGHBOR AND DECISION TREE**

get    add    delete    save    back    <<    <    >    >>    execute

Scale	Education	Education Num	Degree	Levels	Salary	Min	Max	Description
A	1st-4th	2	10	1	140000	777.78	127777.78	
A	5th-6th	3	10	1	140000	777.78	127777.78	
A	Preschool	1	10	1	140000	777.78	127777.78	
B	7th-8th	4	10	4	152000	844.44	134444.44	
C	10th	6	9	1	185000	1027.78	152777.78	
C	11th	7	9	1	185000	1027.78	152777.78	
C	9th	5	9	1	185000	1027.78	152777.78	
D	12th	8	8	1	240000	1333.33	183333.33	
D	HS-grad	9	8	1	240000	1333.33	183333.33	
E	Assoc-acdm	12	8	5	260000	1444.44	194444.44	
E	Assoc-voc	11	8	5	260000	1444.44	194444.44	
E	Some-college	10	8	5	260000	1444.44	194444.44	
F	Bachelors	13	7	1	296000	1644.44	214444.44	
G	Masters	14	6	3	374000	2077.78	257777.78	
H	Doctorate	16	5	1	429000	2383.33	288333.33	
H	Prof-school	15	5	1	429000	2383.33	288333.33	
Z								Age >= 60

**Figure (20) : Example (1)  
Rules Interface**

Use the Decision tree algorithm to classify and decide whether the estimated value is true or false.

- ☛ In the event the estimated value is classified as true, store the estimated value.
- ☛ Transfer this record from the complete table to the incomplete table and mark the value as an estimated value.



**ESTIMATION OF NULL VALUE  
IN RELATIONAL DATA BASE SYSTEM  
USING K-NEAREST NEIGHBOR AND DECISION TREE**

No	<input type="text" value="1"/>	Salary	<input type="text" value="240000"/>
Gender	<input type="text" value="Male"/>	Scale	<input type="text" value="D"/>
Education	<input type="text" value="HS-grad"/>	Degree	<input type="text" value="8"/>
Education Num	<input type="text" value="9"/>	Levels	<input type="text" value="1"/>
Age	<input type="text" value="42"/>	Marital Reward	<input type="text" value="50000"/>
Workclass	<input type="text" value="Self-emp-not-inc"/>	Hours Per Week	<input type="text" value="40"/>
Marital Status	<input type="text" value="Married-civ-spouse"/>	Monthly income	<input type="text" value="103333.33"/>

estimated value   
  inapplicable value

**Figure (21) : Example (1)  
Display of the Estimated Value**

**58**

#### 4.1.1. Example(2) - The Null Value Being Applicable but Not Known

When the value is unknown to the user entering the data in the database, as shown below, the entered monthly income value is Null Value.

**ESTIMATION OF NULL VALUE  
IN RELATIONAL DATA BASE SYSTEM  
USING K-NEAREST NEIGHBOR AND DECISION TREE**

No	48	Salary	240000
Gender	Male	Scale	D
Education	HS-grad	Degree	8
Education Num	9	Levels	1
Age	31	Marital Reward	50000
Workclass	Private	Hours Per Week	45
Marital Status	Married-civ-spouse	Monthly income	

estimated value      inapplicable value

**Figure (22) : Example (2)  
Input and Updating Records Interface**

To treat this situation and before moving to the next record and also before storage, the system will apply the following proposed algorithm to cure the problem.

### **Treatment: -**

- ١ Use the Decision tree algorithm to classify and decide whether the Null Value is applicable or inapplicable.
- ٢ In the event the Null Value is classified as containing an applicable Null Value because the age is less than 60 years, use the nearest

AGE	WORKCLASS	FNLWGT	EDUCATION	EDUCATION_NUM	MARITAL_STATUS	RELATIONSHIP	SEX	HOURS_PER_WEEK	NATIVE
31	Self-emp-not-inc	281030	HS-grad	9	Married-civ-spouse	Husband	Male	45	United-States
31	Private	48189	HS-grad	9	Married-civ-spouse	Husband	Male	45	United-States
31	Private	99844	HS-grad	9	Never-married	Not-in-family	Male	45	United-States
31	Private	109055	HS-grad	9	Married-civ-spouse	Husband	Male	45	United-States
31	Private	113838	HS-grad	9	Married-civ-spouse	Husband	Male	45	United-States
31	Private	117507	HS-grad	9	Never-married	Not-in-family	Male	45	United-States
31	Private	125457	HS-grad	9	Never-married	Not-in-family	Male	45	United-States
31	Private	130021	HS-grad	9	Married-civ-spouse	Husband	Male	45	United-States
31	Private	132996	HS-grad	9	Married-civ-spouse	Husband	Male	45	United-States
31	Private	142038	HS-grad	9	Married-civ-spouse	Wife	Female	45	United-States
31	Private	152156	HS-grad	9	Married-civ-spouse	Husband	Male	45	United-States
31	Private	163594	HS-grad	9	Never-married	Not-in-family	Female	45	United-States
31	Private	182246	HS-grad	9	Divorced	Not-in-family	Male	45	United-States
31	Private	215297	HS-grad	9	Married-civ-spouse	Husband	Male	45	United-States
31	Private	217460	HS-grad	9	Divorced	Not-in-family	Male	45	United-States
31	Private	225053	HS-grad	9	Divorced	Not-in-family	Male	45	United-States
31	Private	241885	HS-grad	9	Never-married	Unmarried	Male	45	United-States
31	Private	416415	HS-grad	9	Separated	Not-in-family	Male	45	United-States
31	Self-emp-inc	113530	HS-grad	9	Never-married	Not-in-family	Male	45	United-States
31	?	233371	HS-grad	9	Married-civ-spouse	Wife	Female	45	United-States

neighbor algorithm to calculate the distance between the values in the tuple of the relevant fields and the value corresponding to the Null Value and find the closest distance; the shortest distance being within the established limits, bring the tuples corresponding to the given condition and use these tuples for estimation.

**Figure (23) : Example (2)**  
**K-Nearest Neighbor Records**

ت Use the Decision tree algorithm to classify and decide whether K-nearest neighbor records exist or do not exist.



ت In the event the K-nearest neighbor records exist , use the Decision tree algorithm to classify and decide whether one or more K-nearest neighbor records exist.

**Figure (24) : Example (2)**  
**Records Account**  
**60**



- ت In the event more than one K-nearest neighbor records exist, use the tree algorithm to classify and decide whether different frequency records or equal frequency records exist.
- ت In the event the records are classified as equal frequency records, estimate the Null Value using K-nearest neighbor (the average attribute of nearest neighbor records).

	CNT	SAL
▶	10	110000
	10	60000

Figure (25) : Example (2)  
Frequency Records View

ت Monthly income =  $(110000 + 60000) \div 2 = 85000$

ت Check the estimated values with the business rules.

ESTIMATION OF NULL VALUE IN RELATIONAL DATA BASE SYSTEM USING K-NEAREST NEIGHBOR AND DECISION TREE								
Scale	Education	Num	Degree	Levels	Salary	Min	Max	Description
A	1st-4th	2	10	1	140000	777.78	127777.78	
A	5th-6th	3	10	1	140000	777.78	127777.78	
A	Preschool	1	10	1	140000	777.78	127777.78	
B	7th-8th	4	10	4	152000	844.44	134444.44	
C	10th	6	9	1	185000	1027.78	152777.78	
C	11th	7	9	1	185000	1027.78	152777.78	
C	9th	5	9	1	185000	1027.78	152777.78	
D	12th	8	8	1	240000	1333.33	183333.33	
D	HS-grad	9	8	1	240000	1333.33	183333.33	
E	Assoc-acdm	12	8	5	260000	1444.44	194444.44	
E	Assoc-voc	11	8	5	260000	1444.44	194444.44	
E	Some-college	10	8	5	260000	1444.44	194444.44	
F	Bachelors	13	7	1	296000	1644.44	214444.44	
G	Masters	14	6	3	374000	2077.78	257777.78	
H	Doctorate	16	5	1	429000	2383.33	288333.33	
H	Prof-school	15	5	1	429000	2383.33	288333.33	
Z								Age >= 60

Figure (26) : Example (2)  
Rules Interface

The education rule for HS-grad is that the monthly income is between [1333.33 and 183333.33]

- Use the Decision tree algorithm to classify and decide whether the estimated value is true or false.
- In the event the estimated value is classified as true, store the estimated value and mark the value as an estimated value.

**ESTIMATION OF NULL VALUE  
IN RELATIONAL DATA BASE SYSTEM  
USING K-NEAREST NEIGHBOR AND DECISION TREE**

No	48	Salary	240000
Gender	Male	Scale	D
Education	HS-grad	Degree	8
Education Num	9	Levels	1
Age	31	Marital Reward	50000
Workclass	Private	Hours Per Week	45
Marital Status	Married-civ-spouse	Monthly income	85000

estimated value     inapplicable value

**Figure (27) : Example (2)  
Display of the Estimated Value**

62

#### 4.1.2. Example(3) - The Null Value Being Inapplicable

**ESTIMATION OF NULL VALUE  
IN RELATIONAL DATA BASE SYSTEM  
USING K-NEAREST NEIGHBOR AND DECISION TREE**

No	13	Salary	
Gender	Male	Scale	
Education	HS-grad	Degree	
Education Num	9	Levels	
Age	64	Marital Reward	
Workclass	Private	Hours Per Week	40
Marital Status	Married-civ-spouse	Monthly income	

estimated value     inapplicable value

When there is no value that we can possibly enter in this field, as shown below, the entered monthly income value is inapplicable Null Value when the age is more than or equal 60 years.

**Figure (28) : Example (3)  
Input and Updating Records Interface**

To treat this situation and before moving to the next record and also before storage, the system will apply the following proposed algorithm to cure the problem.

### **Treatment: -**

- ❏ Use the Decision tree algorithm to classify and decide whether the Null Value is applicable or inapplicable.
- ❏ In the event the Null Value is classified as inapplicable, place a value as an indicator for the value of the inapplicable Null Value.



- ❏ Transfer this record from the complete table to the incomplete table and mark the value as an inapplicable Null Value .

**63**

### **4.1.3. Example(4) –The Value Being Estimated**

**ESTIMATION OF NULL VALUE  
IN RELATIONAL DATA BASE SYSTEM  
USING K-NEAREST NEIGHBOR AND DECISION TREE**

No	<input type="text" value="11"/>	Salary	<input type="text" value="260000"/>
Gender	<input type="text" value="Male"/>	Scale	<input type="text" value="E"/>
Education	<input type="text" value="Some-college"/>	Degree	<input type="text" value="8"/>
Education Num	<input type="text" value="10"/>	Levels	<input type="text" value="5"/>
Age	<input type="text" value="37"/>	Marital Reward	<input type="text" value="50000"/>
Workclass	<input type="text" value="Self-emp-inc"/>	Hours Per Week	<input type="text" value="75"/>
Marital Status	<input type="text" value="Married-civ-spouse"/>	Monthly income	<input style="background-color: #FFB6C1;" type="text" value="158333.33"/>

■ estimated value

■ inapplicable value    <<   <            >   >>

When an estimated value is stored in the database, as shown below, the monthly income value is an estimated value.

**Figure (29) : Example (4)**  
**Input and Updating Records Interface**

To treat this situation and before moving to the next record and also before storage, the system will apply the following proposed algorithm to re – estimate the value.

**Treatment: -**

- ☞ Check the database value (monthly income = 158333.33) with the business rules.

**ESTIMATION OF NULL VALUE  
IN RELATIONAL DATA BASE SYSTEM  
USING K-NEAREST NEIGHBOR AND DECISION TREE**

Scale	Education	Num	Degree	Levels	Salary	Min	Max	Description
A	1st-4th	2	10	1	140000	777.78	127777.78	
A	5th-6th	3	10	1	140000	777.78	127777.78	
A	Preschool	1	10	1	140000	777.78	127777.78	
B	7th-8th	4	10	4	152000	844.44	134444.44	
C	10th	6	9	1	185000	1027.78	152777.78	
C	11th	7	9	1	185000	1027.78	152777.78	
C	9th	5	9	1	185000	1027.78	152777.78	
D	12th	8	8	1	240000	1333.33	183333.33	
D	HS-grad	9	8	1	240000	1333.33	183333.33	
E	Assoc-acdm	12	8	5	260000	1444.44	194444.44	
E	Assoc-voc	11	8	5	260000	1444.44	194444.44	
E	Some-college	10	8	5	260000	1444.44	194444.44	
F	Bachelors	13	7	1	296000	1644.44	214444.44	
G	Masters	14	6	3	374000	2077.78	257777.78	
H	Doctorate	16	5	1	429000	2383.33	288333.33	
H	Prof-school	15	5	1	429000	2383.33	288333.33	
Z								Age >= 60

**Figure (30) : Example (4)**  
**Rules interface**

The education rule for some colleges is that the monthly income is between [1444.44 and 194444.44]

Use the Decision tree algorithm to classify and decide whether the value is true , false or an estimated value.

In the event the value is classified as an estimated value, it is replaced by the Null Value.

Use the Decision tree algorithm to classify and decide whether the Null Value is applicable or inapplicable.

In the event the Null Value is classified as containing an applicable Null Value because the age is less than 60 years, use the K – nearest neighbor algorithm to calculate the distance between the



AGE	WORKCLASS	FNLWGT	EDUCATION	EDUCATION_NUM	MARITAL_STATUS	RELATIONSHIP	SEX	HOURS_PER_WEEK	NATIVE_COUNTRY
37	Private	287031	Some-college	10	Married-civ-spouse	Husband	Male	75	United-States

values in the tuple of the relevant fields and the value corresponding to the Null Value and find the closest distance; the shortest distance being within the established limits, bring the tuples corresponding to the given condition and use these tuples for estimation.

**Figure (31) : Example (4)  
One – Nearest Neighbor Record**

Use the Decision tree algorithm to classify and decide whether K-nearest neighbor records exist or do not exist.



In the event the K-nearest neighbor records exist, use the Decision tree algorithm to classify and decide whether one or more K-nearest neighbor records exist.

**Figure (32) : Example (4)  
Records Account**

In the event one-nearest neighbor record exists, estimate the Null Value using this record.

CNT	SAL
1	158333.33

Figure (33) : Example (4)  
Frequency Records View

Monthly income=158333.33

☞ Check the estimated value with the business rules.

**ESTIMATION OF NULL VALUE  
IN RELATIONAL DATA BASE SYSTEM  
USING K-NEAREST NEIGHBOR AND DECISION TREE**

Scale	Education	Education Num	Degree	Levels	Salary	Min	Max	Description
A	1st-4th	2	10	1	140000	777.78	127777.78	
A	5th-6th	3	10	1	140000	777.78	127777.78	
A	Preschool	1	10	1	140000	777.78	127777.78	
B	7th-8th	4	10	4	152000	844.44	134444.44	
C	10th	6	9	1	185000	1027.78	152777.78	
C	11th	7	9	1	185000	1027.78	152777.78	
C	9th	5	9	1	185000	1027.78	152777.78	
D	12th	8	8	1	240000	1333.33	183333.33	
D	HS-grad	9	8	1	240000	1333.33	183333.33	
E	Assoc-acdm	12	8	5	260000	1444.44	194444.44	
E	Assoc-voc	11	8	5	260000	1444.44	194444.44	
E	Some-college	10	8	5	260000	1444.44	194444.44	
F	Bachelors	13	7	1	296000	1644.44	214444.44	
G	Masters	14	6	3	374000	2077.78	257777.78	
H	Doctorate	16	5	1	429000	2383.33	288333.33	
H	Prof-school	15	5	1	429000	2383.33	288333.33	
Z								Age >= 60

Figure (34) : Example (4)  
Rules Interface

- ☞ Use the Decision tree algorithm to classify and decide whether the estimated value is true or false.
- ☞ In the event the estimated value is classified as true, store the estimated value.

**ESTIMATION OF NULL VALUE  
IN RELATIONAL DATA BASE SYSTEM  
USING K-NEAREST NEIGHBOR AND DECISION TREE**

No	11	Salary	260000
Gender	Male	Scale	E
Education	Some-college	Degree	8
Education Num	10	Levels	5
Age	37	Marital Reward	50000
Workclass	Self-emp-inc	Hours Per Week	75
Marital Status	Married-civ-spouse	Monthly income	158333.33

■ estimated value           
■ inapplicable value    << < Search gexecute Stop Search > >>

**Figure (35) : Example (4)  
Display of the Estimated Value**

**67**

#### 4.1.4. Example (5) - The Null Value Being Applicable but Not Known and with No Nearest Neighbor Records.

When the value is unknown to the user entering the data in the database and no Nearest Neighbor records exist, as shown below, the entered monthly income value is Null Value.

**ESTIMATION OF NULL VALUE  
IN RELATIONAL DATA BASE SYSTEM  
USING K-NEAREST NEIGHBOR AND DECISION TREE**

No	153	Salary	429000
Gender	Female	Scale	H
Education	Doctorate	Degree	5
Education Num	16	Levels	1
Age	43	Marital Reward	
Workclass	Federal-gov	Hours Per Week	50
Marital Status	Never-married	Monthly income	

■ estimated value           
■ inapplicable value    << < Search gexecute Stop Search > >>

**Figure (36) : Example (5)  
Input and Updating Records Interface**

To treat this situation and before moving to the next record and also before storage, the system will apply the following proposed algorithm to cure the problem.

### **Treatment: -**

- Use the Decision tree algorithm to classify and decide whether the Null Value is applicable or inapplicable.



- In the event the Null Value is classified as containing an applicable Null Value because the age is less than 60 years, use the K- nearest

## 68

neighbor algorithm to calculate the distance between the values in the tuple of the relevant fields and the value corresponding to the Null Value and find the closest distance; the shortest distance being within the established limits, bring the tuples corresponding to the given condition and use these tuples for estimation.

AGE	WORKCLASS	FNLWGT	EDUCATION	EDUCATION_NUM	MARITAL_STATUS	O..RELATIONSHIP	SEX	HOURS_PER_WEEK	NATIVE_CC
				[Null Value]					

**Figure (37) : Example (5)  
No - Nearest neighbor Record**

Use the Decision tree algorithm to classify and decide whether K-nearest neighbor records exist or do not exist.



**ESTIMATION OF NULL VALUE  
IN RELATIONAL DATA BASE SYSTEM  
USING K-NEAREST NEIGHBOR AND DECISION TREE**

Scale	Education	Education Num	Degree	Levels	Salary	Min	Max	Description
A	1st-4th	2	10	1	140000	777.78	127777.78	
A	5th-6th	3	10	1	140000	777.78	127777.78	
A	Preschool	1	10	1	140000	777.78	127777.78	
B	7th-8th	4	10	4	152000	844.44	134444.44	
C	10th	6	9	1	185000	1027.78	152777.78	
C	11th	7	9	1	185000	1027.78	152777.78	
C	9th	5	9	1	185000	1027.78	152777.78	
D	12th	8	8	1	240000	1333.33	183333.33	
D	HS-grad	9	8	1	240000	1333.33	183333.33	
E	Assoc-acdm	12	8	5	260000	1444.44	194444.44	
E	Assoc-voc	11	8	5	260000	1444.44	194444.44	
E	Some-college	10	8	5	260000	1444.44	194444.44	
F	Bachelors	13	7	1	296000	1644.44	214444.44	
G	Masters	14	6	3	374000	2077.78	257777.78	
H	Doctorate	16	5	1	429000	2383.33	288333.33	
H	Prof-school	15	5	1	429000	2383.33	288333.33	
Z								Age >= 60

In the event no K-nearest neighbor records exist place a minimum value from the business rules table.

**Figure (38) : Example (5)  
Rules Interface**

The education rule for Doctorate is that the monthly income is between [2383.33 and 288333.33].

Monthly income=2383.33

**ESTIMATION OF NULL VALUE  
IN RELATIONAL DATA BASE SYSTEM  
USING K-NEAREST NEIGHBOR AND DECISION TREE**

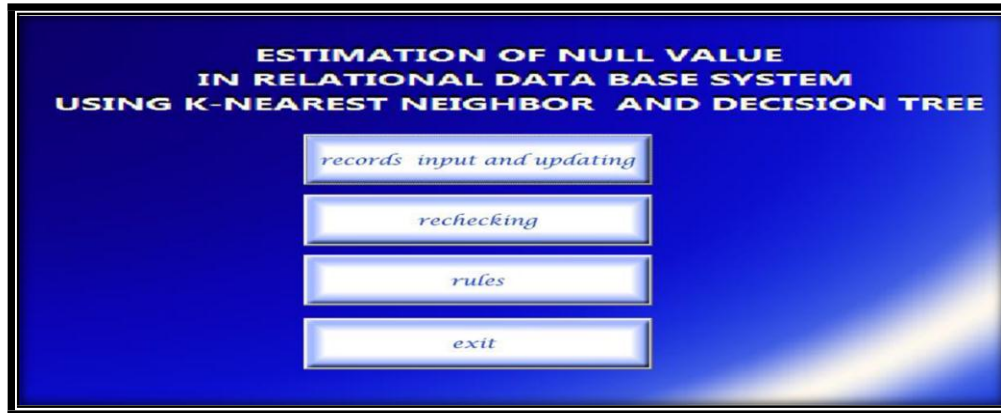
No	<input type="text" value="153"/>	Salary	<input type="text" value="429000"/>
Gender	<input type="text" value="Female"/>	Scale	<input type="text" value="H"/>
Education	<input type="text" value="Doctorate"/>	Degree	<input type="text" value="5"/>
Education Num	<input type="text" value="16"/>	Levels	<input type="text" value="1"/>
Age	<input type="text" value="43"/>	Marital Reward	<input type="text" value=""/>
Workclass	<input type="text" value="Federal-gov"/>	Hours Per Week	<input type="text" value="50"/>
Marital Status	<input type="text" value="Never-married"/>	Monthly income	<input type="text" value="2383.33"/>

estimated value      inapplicable value

**Figure (39) : Example (5)  
Display of Estimated Value**

Store the estimated value and mark the value as an estimated value.

#### 4.1.5. Example(6) - Rechecking the Null Values



The treatment of the Null Values in the present research work helps in re - estimating the Null Values and in the generation of the reports and statistics and in retrieval of data that reflect the reality and the true image of the database through the rechecking as shown in figures (38),(39) and(40).

**Figure (40) : Example (6)  
Main Interface**

Date 07/05/2013 12:45:09

Estimation of Null Value in Relational Data Base System using K-Nearest Neighbor and Decision Tree

Seq	Age	Education	Marital Status	Hours Per Week	Salary	Sal Per Month
1	42	HS-grad	Married-civ-spouse	40	240000	** 103333.33
2	20	HS-grad	Never-married	40	** 240000	** 53333.33
3	32	HS-grad	Separated	30	240000	40000
4	45	Assoc-voc	Widowed	45	260000	** 115000
5	50	7th-8th	Divorced	40	152000	33777.78
6	36	Bachelors	Divorced	40	** 296000	** 65777.78
7	45	HS-grad	Married-civ-spouse	60	240000	130000
8	17	11th	Never-married	12	185000	** 12333.33
9	59	Some-college	Married-civ-spouse	40	260000	107777.78
10	26	11th	Married-civ-spouse	40	185000	91111.11
11	37	Some-college	Married-civ-spouse	75	260000	** 158333.33
12	19	Some-college	Never-married	24	260000	34666.67
13	64	HS-grad	Married-civ-spouse	40		
14	33	Bachelors	Never-married	45	296000	74000
15	33	HS-grad	Married-civ-spouse	40	240000	103333.33
16	61	HS-grad	Married-civ-spouse	40		
17	17	9th	Never-married	24	185000	24666.67
18	50	Masters	Married-civ-spouse	98	374000	253622.22
19	27	Masters	Never-married	35	374000	72722.22
20	30	HS-grad	Divorced	40	240000	53333.33
21	43	HS-grad	Married-civ-spouse	40	240000	103333.33
22	44	Some-college	Married-civ-spouse	40	260000	107777.78
23	35	Some-college	Never-married	40	260000	57777.78
24	25	Some-college	Never-married	40	260000	57777.78
25	24	Some-college	Married-civ-spouse	48	260000	119333.33
26	22	Bachelors	Never-married	15	296000	24666.67
27	42	Some-college	Married-civ-spouse	40	260000	107777.78
28	34	Assoc-acdm	Divorced	45	260000	65000
29	60	Bachelors	Divorced	42		
30	21	HS-grad	Never-married	40	240000	53333.33
31	57	Masters	Married-civ-spouse	40	374000	133111.11
32	41	Prof-school	Married-civ-spouse	60	429000	** 193000
33	50	Some-college	Divorced	45	260000	65000
34	25	Bachelors	Never-married	40	296000	65777.78
35	50	7th-8th	Married-civ-spouse	40	152000	83777.78

Page 1 of 941

Figure (41) : Example (6) Rechecking  
Report - Page (1) of (941)

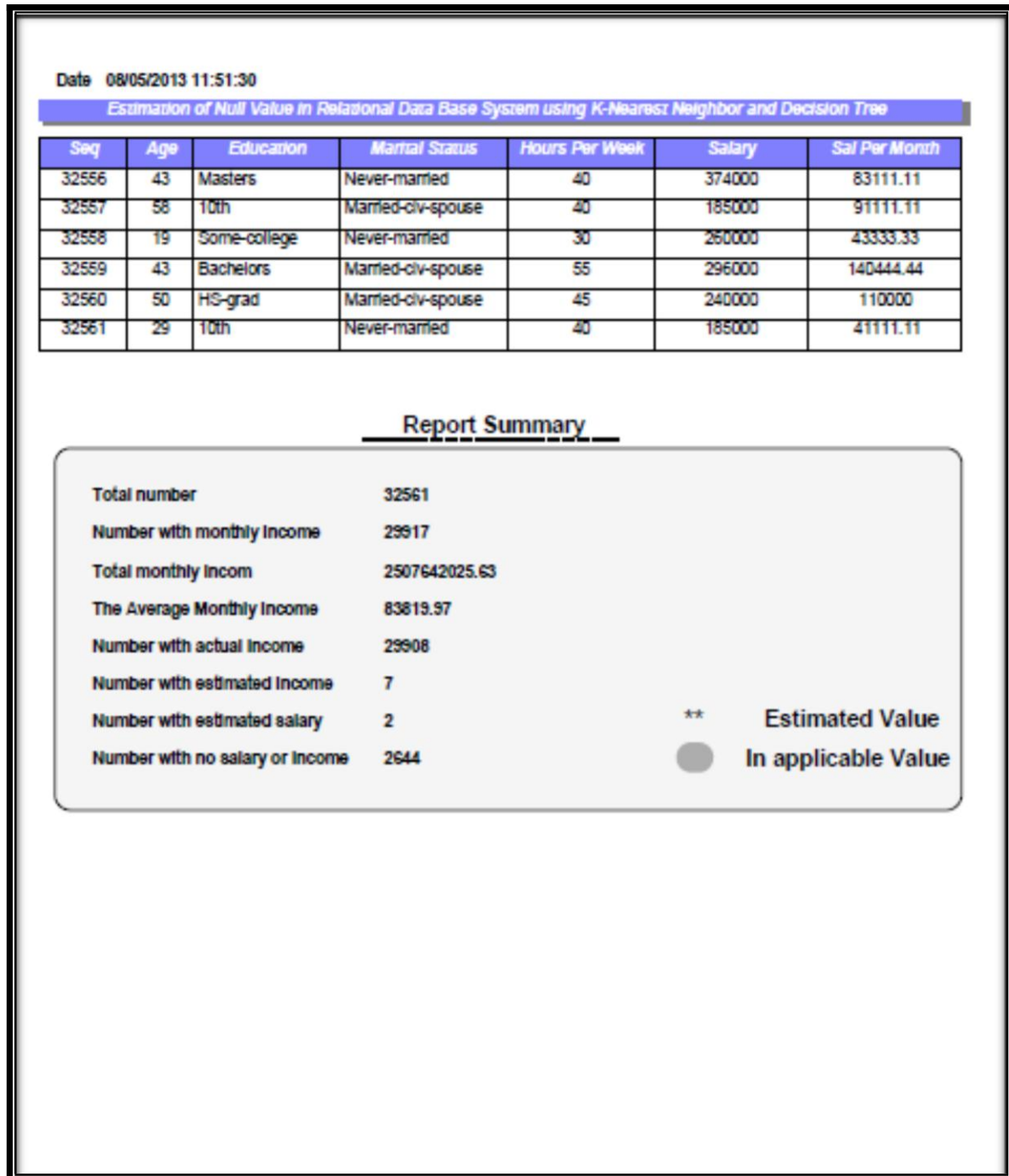
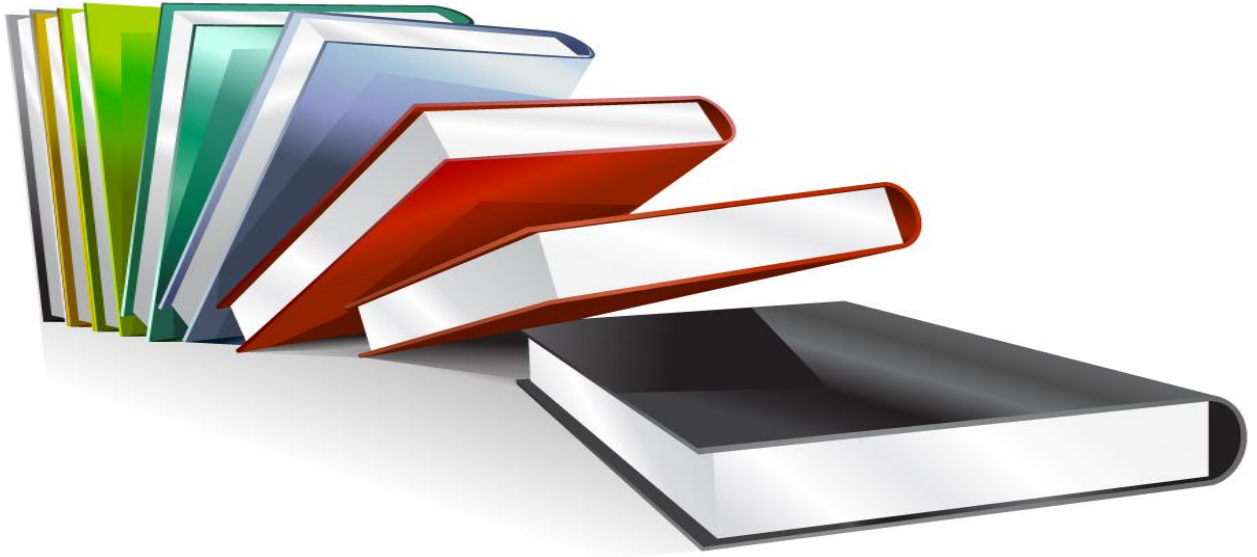


Figure (42) : Example (6) Rechecking  
Report - Page (941) of (941)

# CHAPTER FIVE

## CONCLUSIONS AND

### RECOMMENDATIONS FOR FUTURE WORK



## CHAPTER FIVE

### CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE WORK

#### 5.1 INTRODUCTION

A great number of database systems are designed and used daily in all aspects of life and all users face the problems of dealing with missing values in these systems. In the present work a framework for estimating Null Values in relational database systems is proposed using K-nearest neighbor and Decision tree algorithms. This framework is also used to check the database for integrity and to verify the correctness of the data in these databases. The proposed framework is then implemented on a standard database “ADULT” obtained from UCI data repository and an efficient performance of the estimation is proved.

This chapter includes, in addition to the conclusions of the present work, recommendations for future work and for further studies on the development of the estimation of Null Values in relational database systems.

#### 5.2 CONCLUSIONS

The present research work formulates a technique for estimating the missing values in relational database systems based on the use of certain rules, selected by the user to provide flexibility to change these rules when needed. These rules are used to check the data stored in the database and then estimate the missing values using K-nearest neighbor algorithm and the Decision tree algorithm and the results are finally checked against the set rules to verify the accuracy of the estimations.

The implementation of the proposed framework shows the success of the treatment, whether the type of the stored data, containing the missing values, was a text, numeric or date.

In the case the type of stored data is numeric or date the algorithm is implemented as is while in the case of a text type of data the estimation of the Null Values depends on the K frequency- nearest neighbor algorithm and the Decision tree, and the K mean - nearest neighbor algorithm is not used because of the inability to perform arithmetic operations on texts.

The success rate for the first stage using the estimation algorithms without the adoption of the set rules is up to 76%, while in the second stage, using the estimation algorithms with the adoption of the set rules, this success rate is up to 80%.

Some of the advantages of using the proposed framework are:

- Assists in checking the data stored in the database and ensures the accuracy and validity of the stored information by comparing it with the business rules placed by the user.
- Using the Decision tree helps in the classification of the missing values as applicable Null Values ,and estimating these missing values, or as inapplicable and placing values as indicators for the values of the inapplicable Null Values.
- The framework helps in the estimation of the applicable null values, as well as re-estimating the false and estimated values either through the use of the algorithms K-nearest neighbor (the average attribute of nearest neighbor records) and the Decision tree algorithms for records

- having neighboring records with equal frequency or the use of the K-nearest neighbor (the highest frequency) and the Decision tree algorithms for those records having nearest neighbor records with different frequencies, while for records with no similar nearest neighbor records the business rules are used to estimate the missing values.
- Helps in checking the estimated values through comparing them with the business rules to ensure the accuracy and the validity of the estimation.
- In the process of rechecking the data stored in the database, a summary of the stored data is displayed showing how many of the data are real and how many are estimated with true statistics of the stored data.

The limitations of the proposed framework and the estimation method lies with small size databases as the accuracy of the estimation increases markedly with the increase in the size of the database due to the fact that the estimations are based on the nearest neighbor records.

### 5.3 RECOMMENDATIONS FOR FUTURE WORK

This section offers some suggestions and recommendations for further future work on the development of the estimation of Null Values in relational database systems aiming at achieving higher and higher accuracies:

1. To improve the reliability and efficiency of the designed framework in estimating the missing values using K-nearest neighbor and Decision tree algorithms and in predicting the futuristic data using back propagation algorithm based on the data stored in the database systems.



2. To find ways to repair the databases which contain damaged data by replacing these data by Null Values and then estimating these Null Values through using K-nearest neighbor and Decision tree algorithms based on the remaining undamaged data.
3. Future research work can find many ways to repair the damaged pixels for images to enable the estimation of the pixels values using the two algorithms: K-nearest neighbor and Decision tree based on the presence of undamaged pixels after which fuzzy based image enhancement techniques can be used.

# REFERENCES



## REFERENCES

1. ELMASRI,R., NAVATHE,S.B. (2011). "DATABASE SYSTEMS models ,languages ,design , and application programming (6TH ed.)". Pearson.
2. Al-Hamami, A.H., Al-Aani, S.A. (2008). "CONCEPTS AND APPLICATIONS OF DATA BASES TECHNOLOGY". ITHRAA Publishing and Distribution, Jordan.
3. "DATABASE\_DIRECTORY",[HTTP://WWW.DATABASEDIR.COM/WHAT-IS-RDBMS/](http://www.databasedir.com/WHAT-IS-RDBMS/), access at 15/12/2012.
4. CODD, E.F. (2000). "THE RELATIONAL MODEL FOR DATABASE MANAGEMENT" : VERSION 2. Addison-Wesley.
5. CODD, E.F.(1986). "MISSING INFORMATION (APPLICABLE AND INAPPLICABLE) IN RELATIONAL DATABASE". Addison-Wesley.
6. [HTTP://WWW.SIMPLE-TALK.COM/SQL/LEARN-SQL-SERVER/SQL-AND-THE-SNARE-OF-THREE-VALUE-LOGIC/](http://www.simple-talk.com/sql/learn-sql-server/sql-and-the-snare-of-three-value-logic/) , access at 1/11/2012.
7. "Missing Information" [G51DBS Database System JasonAtkin, [http://www.cs.nott.ac.uk/~jaa/dbs/JAA\\_DB\\_S\\_Lecture12%20-%20final.pdf](http://www.cs.nott.ac.uk/~jaa/dbs/JAA_DB_S_Lecture12%20-%20final.pdf) , access at 29/12/2012.
8. Date ,c.j.,(2000), "An introduction to database systems", Addison Wesley longman , America.
9. Du,H. (2010) . "Data mining techniques and applications an introduction " , nelson education ,ltd ,Canada.
10. Hand,D. ,Mannila,H. ,Smyth,P. (2001)." Principles of Data Mining", MIT Press, Cambridge.

11. LAROSE, D.T. (2005) ."DISCOVERING,KNOWLEDGE IN DATAAn Introduction to Data Mining “, John Wiley & Sons Inc ,New Jersey.
12. Al-Hamami, A.H.,(2008), “DATA MINING” , ITHRAA Publishing and Distribution, Jordan.
13. PETRE,R.,(2012), “Data mining in Cloud Computing” , Database Systems Journal vol. III, no. 3
14. <http://www.executionmih.com/data-mining/technology-architecture-application-frontend.php> , access at 22/2/2013
15. Dempster , A.P., Laird , N. M. and Rubin , D. B. , (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm," Journal of the Royal Statistical Society, B, vol. 39, no. 1, pp. 1–38.
16. Agrawal,R., Imielinski,T.,Swami,A., “Mining Association Rules between Sets of Items in Large Databases”, ACM SIGMOD Conference Washington DC, USA, May 1993
17. Agrawal,R., Srikant ,R., “Fast Algorithms for Mining Association Rules”, Proceedings of the 20th VLDB Conference Santiago, Chile, 1994
18. Fomby,t.b.,(2008), K-Nearest Neighbors Algorithm: Prediction and Classification, , manuscript.
19. [http://en.wikipedia.org/wiki/Decision\\_tree](http://en.wikipedia.org/wiki/Decision_tree) access at 27/2/2013.
20. Repository,U.M.L: <http://archive.ics.uci.edu/ml/datasets.html> access at 1/3/2013.
21. Grant , J. , (2008) . "NULL VALUES IN SQL" , SIGMOD Record , Vol. 37, No. 3.
22. Lee,S. , Zeng,X.,(2008).” A Modular Method for Estimating Null Values in Relational Database Systems” , Eighth International Conference on Intelligent Systems Design and Applications, DOI 10.1109/ISDA.2008.194

23. Haiyan,Y. , Xia , Z. , Mingzhu ,X. ,(2009) . “Null Value Estimation Method based on Information Granularity for Incomplete Information System”,Third International Symposium on Intelligent Information Technology Application.
24. Mridha, M.F. ,Banik, M. ,(2010). “Performances of Estimating Null Values using Noble Evolutionary Algorithm (NEAs) by Generating Weighted Fuzzy Rules”, International Journal of Computer Applications,volume 11-no 9.
25. Yang,J., Jiang,Z., Zhang,J. Zhang,L.,(2010),” A Null Value Estimation Method Based on Similarity Predictions in Rough Sets”, Fifth International Conference on Internet Computing for Science and Engineering, DOI 10.1109/ICICSE.2010.22.
26. Zhang, S. ,(2010).“Shell-neighbor method and its application in missing data imputation” ,appltell journal, volume 35.
27. VashistR. ,Garg M.L ,(2012) . “A Rough Set Approach for Generation and Validation of Rules for Missing Attribute Values of a Data Set “ , International Journal of Computer Applications, volume 42-no 14 .
28. Raghunathan , R. , De , S. ,Kambhampati,S. ,(2012) . “Bayes Networks for Supporting Query Processing Over Incomplete Autonomous Databases”, manuscript.
29. Pandole,K. , Bhargava , N.,(2012),” Comparison and Evaluation for Grouping of Null Data in Database Based on K-Means and Genetic Algorithm”, International Journal of Computer Technology and Electronics Engineering (IJCTEE) , Volume 2, Issue 3.

30. Azadeh, A. ,Asadzadeh, S.M., Jafari-Marandi, R. , Nazari-Shirkouhi, S., BaharianKhoshkhou, G. , Talebi , S., Naghavi , A.,(2012), “Optimum estimation of missing values in randomized complete block design by genetic algorithm”, Elsevier, <http://dx.doi.org/10.1016/j.knosys.2012.06.014>.
31. Vinod, N.C., Punithavalli, M.,(2013),” Performance Evaluation of Mutation / Non-Mutation Based Classification With Missing Data”, International Journal on Computer Science and Engineering (IJCSE), Vol. 5 No. 02 Feb 2013.
32. Aydilek ,I.B., Arslan , A.,(2013),” A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm”, Elsevier, <http://dx.doi.org/10.1016/j.ins.2013.01.021>.
33. Singh,s., Prasad,j.,(2013), “Estimation of Missing Values in the Data Mining and Comparison of Imputation Methods”, Mathematical Journal of Interdisciplinary Sciences, Vol. 1, No. 2,
34. Ibarguengoytia,P.H,García,U.A.,Herrera-Vega,J., Hernández-Leal,P.,(2013), “On the Estimation of Missing Data in Incomplete Databases: Autoregressive Bayesian Networks”, IARIA, ISBN: 978-1-61208-246-2.
35. Beirami,M.H.N,Ghavifekr,M.H.N.,Khajei,R.P.,(2013) “Predicting Missing Attribute Values Using Cooperative Particle Swarm Optimization”, Journal of Basic and Applied Scientific Research, ISSN 2090-4304.

36. <http://www.moi.gov.iq/uploads/Civilian%20staff/%D9%82%D8%A7%D9%86%D9%88%D9%86%20%D8%B1%D9%88%D8%A7%D8%AA%D8%A8%20%D9%85%D9%88%D8%B8%D9%81%D9%8A%20%D8%A7%D9%84%D8%AF%D9%88%D9%84%D8%A9%20%D9%88%D8%A7%D9%84%D9%82%D8%B7%D8%A7%D8%B9%20%D8%A7%D9%84%D8%B9%D8%A7%D9%85%20%D8%B1%D9%82%D9%85%2022%20%D9%84%D8%B3%D9%86%D8%A9%202008.pdf>



جامعة عمان العربية  
AMMAN ARAB UNIVERSITY

تقدير القيم الفارغة  
في نظام قواعد البيانات العلائقية  
باستخدام طريقة ك-اقر ب جار و شجرة القرار

إعداد  
سارة موفق الصمدي

إشراف  
الأستاذ الدكتور علاء الحمامي

قدمت هذه الرسالة لاستكمال متطلبات الحصول على درجة الماجستير في الحاسب  
الآلي

قسم علم الحاسوب  
كلية العلوم الحاسوبية والمعلوماتية  
جامعة عمان العربية  
2013



