PROLOGUE

Since antiquity the intuitive notions of continuous change, growth and motion, have challenged scientific minds. Yet, the way to the understanding of continuous variations in systems and perceiving various dimensions of data was opened only in late 1980s when modern computer science emerged and rapidly developed in close conjunction with electronics and allied sciences.

Data mining is one of the fastest growing fields in the computer industry. Once a small interest area within computer science and statistics, it has quickly expanded into a field of its own. The major reason that data mining has attracted a great deal of attention in information industry is due to the wide availability of techniques turning data into useful information and knowledge. The spectrum of techniques in data mining opens to predictive and descriptive approaches.

Decision making and data mining has a set of positive differences that encourage researchers of database technologies. Decision making systems are a derivative of traditional database systems with their variants; knowledge databases, deductive databases and analytical databases. Data mining is encapsulated in decision making process to steer the data driven enterprise applications.

"No Quality Data, No quality mining results". Data quality is a key issue with data mining and other correlated area such as machine learning, in order to increase the accuracy of the mining process and results as well. As real-world data tend to be incomplete, noisy and inconsistent, data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

The methods from statistics and probability derive the probable values as plausible values to replace, however the plausibility may not be correct. The
hypothetical frameworks and algorithms as carried out in this research uses corpus of digest values of tuples that are rich with clues to replace the plausible values.

Missing data is available everywhere causing a variety of problems in data analysis, leading to loss of statistical power and bias. Little and Rubin's Statistical Analysis with Missing Data, Second Edition (2002, Wiley) is a classic which worked out much of the theoretical foundation for analyses of incomplete data and contributed several practical methods. The expectation–maximization (EM) algorithm and variants for the computation of maximum likelihood estimates of statistics and missing values are discussed extensively, including discussions of properties (e.g., convergence rates) that are important in applications. Indeed, many real-world data sets contain missing values and it is often regarded as a difficult problem to cope with. "Sometimes, values are missing due to unknown reasons or errors and omissions when data are recorded and transferred".

The single and multiple imputation methods are based on the probabilistic methods. The single imputation method derives a set of plausible values for the replacement of a missing value in the database. The multiple imputation method derives a set of plausible values for the replacement of a set of missing values in the database, which is visualized as a stochastic process. Therefore, missing data can seriously affect the results of KDD processes. Should we ignore missing data or assume that excluding missing data causes sufficient risk reaching invalid and insignificant results? The Mean and Regression methods are coded in C/C++ and their results also evaluated.

The capabilities of both generating and collecting data have been increasing tremendously. The advances in data collection tools range from scanned text and image platforms to satellite remote sensing systems. In addition, popular use of World Wide Web as a global information system has flooded us with a tremendous amount of data and information. Due to this explosive growth of data, there is an urgent need for new technique and automated tools that can intelligently assist us in transforming the vast amounts of data into useful
information and knowledge. The short commons of the existing downstream applications are to maintain an error free data, at the very operational level, with which at a greater risk, stringent validation techniques and security policies are enforced.

The inset of the research has objectives, for designing a model or mechanism that is fool-proof of unpredicted values, being the values for imputation in the missing places and the mechanism is cost effective and can be able to implement practically by the Efficient Imputation Method. The process of imputation proposed in the research hosts a framework containing components that support data storage, sampling mechanism, attribute selection and search and impute mechanism. From the keen observations made on the references and experiments the Data Set availability for the data mining experiments is of three categories namely Real Data Set, Benchmark Data Set and Synthetic Data Set. The Real Data Set is the transaction data that is available in the field of interest like business applications. The Benchmark Data Set, the data mining algorithms are designed to solve a particular viewpoint of the problem of the “Market-Basket Analysis”. The Synthetic Data Set, the candidate databases are freely available, as for the study purpose, the properties of the data in the database can be assumed and using Random Sampling methods of Statistics, the data can be generated with the required properties to test and algorithm and its implementation.

Although the algorithms complexity is measured asymptotically, mathematical analysis and empirical analysis of every module or function has been carried out to derive practical time complexity. Python Orange tool is used to describe the visualization and sampling of data. Python Orange tool samples data based on selected list of attributes. The probability distribution of tuples for that sample is adjudged using the Sampling mechanism in the tool. Visualization of data using parallel coordinates and radar chart is apt for sampled data and gives clear picture distributed samples.
Famous data sets for experiment have been selected such as 22 attributes - *agaricus-lepiota* data, 12 attributes – *cyclo-hexane* chemical data, 9 attributes – *tic-tac-toe* data, 6 attributes – *mammographic* data, 20 attributes – *hepatitis* data. Comparative analysis of the algorithms has been studied and their results are tabulated. Substantial difference in performance evaluation is found, when varied samples are chosen for experimentation. "The rate of increase of attribute selection samples influences the increase of time". Reliability analyses have also been performed on the algorithm when working on various types of data (as cited above).