

***Benha University***

***Faculty of Commerce***

***Department of Statistics,***

***Mathematics and Insurance***

**An Alternative Data Mining Approach for Big Data**

**By**

**Aya Shehata Mahmoud Marzouk**

**Supervised by**

|  |  |
| --- | --- |
| **Prof.**  HEGAZY MOHAMED ZAHER  **Professor of Mathematical Statistics Faculty of graduate studies for statistical research, Cairo university** | **Prof.**  MERVAT MAHDY RAMADAN  **Professor and the Head of Statistics Mathematics, and Insurance department, Faculty of Commerce, Benha University** |

**Dissertation**

***Submitted to the Department of Statistics, Mathematics, and Insurance, Faculty of Commerce, Benha University***

***In a partial fulfillment of the requirements for the degree of Doctor of Philosophy in Applied Statistics***

**2020**

# Introduction

The domains of BD and data mining have meanwhile become significant research fields. While data mining aims to discover patterns, correlations and useful information from raw data which usually be BD and it represents the actual data nowadays.

The main purpose of this study is to help business organizations or businessmen or decision makers in any organization to deal with BD and extract useful information from it by using data mining tools to make the suitable and accurate decision .In past times, the difficulty is when dealing with BD by its huge volume and with its challenging characteristics. This leads to time consuming, more costs and inaccurate decisions for the organization.

To overcome this problem, we propose an alternative data mining approach to overcome BD problems.

The 6 stages of building our technique are: collection and selection, measuring prior BD quality, preprocessing stage, measuring posterior BD quality , processing stage and measuring prior and posterior BD quality. This approach aims to reduce the number of variables by eliminating irrelevant and redundant variables to be easy for the decision maker by utilizing consuming time and computational cost.

Briefly, describe the 6-stages framework of dealing with BD characteristics’ challenges that can be used to deal with BD by reducing its high dimensionality of redundant and noisy features, partitioning it into smaller clusters to be handled in an easier way, extracting the useful information from it and optimize its quality evaluation. We will solve the Volume characteristic of BD by dimension reduction technique and by a popular clustering technique. Moreover, we will solve the veracity characteristic by improving quality levels in a dataset and avoiding inessential data.

# Summary and Conclusion

The proposed technique presented in this dissertation could give good insight into the structure of BD problems and the ability of data mining to suppose solutions of these problems. We can present the following summary of the proposed technique:

1. Browsing through a large data set would be difficult and time consuming, as a result, the accuracy of the discovered patterns can be poor so by using the proposed technique we can overcome these obstacles. While mining on reduced dataset should be more efficient.
2. Applying data cleaning procedures to clean the data by filling in missing values and identifying or removing outliers to achieve the effectiveness which is concerned to the quality and the efficiency which is related to the time required.
3. Examining and evaluating and assessing the quality of BDset and improving the quality of data by outlier detection technique.
4. A variable with values highly correlated to those of another variable will not surely add any new information to the current variables and will add more calculations and parameters to the model so we apply dimension reduction technique based on correlation to overcome the dimensionality problem.
5. Although it is known that the quality is less trustworthy with large amount of BD, we will follow a framework that will retain on BD Quality and we will solve BD characteristics problems.
6. Using a single data mining technique is not sufficient especially when we use it in BD to make suitable and accurate decisions. So, we will use more than one data mining techniques to get rid of the volume veracity and value characteristic of BD for making a decision.
7. We used a real data about electric motor temperature from kaggle.com website.
8. Measuring four quality dimensions consistency, accuracy, completeness, uniqueness of BD before cleaning stage of data.
9. According to the previous quality dimensions, the preprocessing technique will be determined.
10. Improving the quality by the preprocessing procedure through

* Dealing with missing data by a data mining procedures to help in data reduction process
* Outlier detection and removing through a cluster based algorithm

1. Measuring the quality after missing data treatment and outlier detection process.
2. Through the previous procedures, we can combine more than one dimension of data quality, so, we can improve the quality of data and overcome the characteristic of veracity of BD.
3. Constructing a proposed framework for dimensionality reduction based on correlation criterion.
4. By following a proper algorithm the irrelevant variables are eliminated
5. By applying the proposed technique we overcome the volume characteristic of the BD.
6. We can reduce the data size by removing redundant features and clustering, these techniques are not mutually exclusive, they may work together, so, we can get a reduced representation of the dataset that is surely smaller in volume and in processing time.
7. We use the k-means clustering technique to decrease the number of sessions, by dividing the data into three clusters to overcome the volume characteristic of the BD. ,
8. We concluded that there is homogeneity and unbiasedness in the data before and after the reduction process.

# Further Suggestions

Because of ultimate non ending research in BD and DM, we need to increase research in this important area. This study can be extended in different aspects to enclose future points of future research.

The following suggestions are examples of future research:

1. Using a data mining method for imputation missing data.
2. Using different data that contains quantitative and qualitative data (semi structured data) and use the techniques which are suitable for this data.
3. Using different data that contains qualitative data (Unstructured data) and use the techniques which are suitable for this data.
4. Inserting Entropy in solving problems of BD.
5. Using different data and apply 6-stages approach on this data.
6. Using different software in applying 6-stages approach.