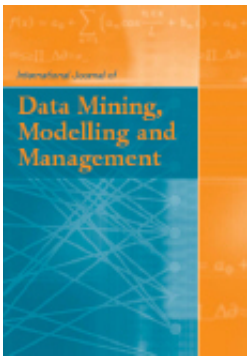


International Journal of Data Mining, Modelling and Management

ISSN online: 1759-1171 - ISSN print: 1759-1163
https://www.inderscience.com/ijdmmm

Optimising data quality of a data warehouse using data purgation process

Neha Gupta

DOI: 10.1504/IJDMMM.2023.10055198

Article History:

Received:	05 August 2020
Last revised:	30 January 2021
Accepted:	07 April 2021
Published online:	04 April 2023

Optimising data quality of a data warehouse using data purgation process

Neha Gupta

Faculty of Computer Applications,
Manav Rachna International Institute of Research and Studies,
Faridabad, 121002, India
Email: nehag2012@gmail.com

Abstract: The rapid growth of data collection and storage services has impacted the quality of the data. Data purgation process helps in maintaining and improving the data quality when the data is subject to extract, transform and load (ETL) methodology. Metadata may contain unnecessary information which can be defined as dummy values, cryptic values or missing values. The present work has improved the EM algorithm with dot product to handle cryptic data, DBSCAN method with Gower metrics has been implemented to ensure dummy values, Wards algorithm with Minkowski distance has been applied to improve the results of contradicting data and K-means algorithm along with Euclidean distance metrics has been applied to handle missing values in a dataset. These distance metrics have improved the data quality and also helped in providing consistent data to be loaded into a data warehouse. The proposed algorithms have helped in maintaining the accuracy, integrity, consistency, non-redundancy of data in a timely manner.

Keywords: data warehouse; DW; data quality; DQ; extract, transform and load; ETL; data purgation; DP.

Reference to this paper should be made as follows: Gupta, N. (2023) 'Optimising data quality of a data warehouse using data purgation process', *Int. J. Data Mining, Modelling and Management*, Vol. 15, No. 1, pp.102–131.

Biographical notes: Neha Gupta earned her PhD from Manav Rachna International University and has total of 14+ year of experience in teaching and research. She is a life member of ACM CSTA, Tech Republic and Professional Member of IEEE. She has authored and co-authored 50 research papers in SCI/SCOPUS/peer reviewed journals (Scopus indexed) and IEEE/IET conference proceedings. She has published books with publishers like Springer, Taylor and Francis, IGI Global and Pacific Book International and has also authored book chapters with Elsevier, Springer, CRC Press and IGI global USA. She is a technical programme committee (TPC) member in various conferences across globe. https://orcid.org/0000-0003-0905-5457.

1 Introduction

1.1 Data quality at ETL in data warehouse

Data warehousing is a network of decision support tools focused on supporting the analyst's knowledge to make quicker and better decisions. Data warehouse (DW) is the

compilation of non-volatile, subject-oriented and organised data. The purposes of DW involve:

1 data extraction (DE) – it is used to collect information from various heterogeneous resources

2 data cleaning – it is used to spot and correct the flaws in the collected data

3 data transformation (DT) – it converts the information from legal presentation to warehouse format

4 data loading (DL) – it includes arranging, associating, evaluating, partitioning, checking integrities, and developing entities

5 refreshing – it is the method of updating the data sources to the warehouse.

In many domains, the DW information can be studied and referred for multiple purposes. It is used by rearranging the products and handling the product portfolios relating to the sales and profit of the year to tune production strategies. The information is used for market research by analysing the customer's interest, their buying time, budget cycle, etc. This analysis plays a vital role in supporting the management of customer relationships in order to make changes according to demand and conditions.

As discussed above, data is the main fuel for any kind of prediction and any kind of operation as well. The initial requirement is to understand the DW components and the DW framework along with the operations which can be done with the data. The information collected from the various resources could be historical information or the accumulated information called as DW. There are various security and quality issues with the data collected from various sources. Data will be collected from various repositories having different formats for the same type of data. To understand the concept, various data formats available in DW s have been explained as follows:

a Structured data.

   This can be historical data that can be compared with database information. The data in the databases can be stored in the form of rows and columns and can be manipulated in the form of rows and columns with simple queries. The queries must target the variables of particular row and column and with simple basic operations on the databases, so that the data can be understood. Most representations of data will be text format and other formats of the database.

b Unstructured data.

   This type of data usually has multimedia content. The multimedia consists of different formats of images, videos and audios. Sometimes we need to consider the concept of live streaming. Live streaming is the data that is captured using Apache spark as the main base. It consists of components that can handle live streaming information. It can be the main source of operations on big data, cloud computing and can also be the source of data management.

c Semi-structured data.

   This format ensures connectivity to the web application. Every web application has some sort of data transfer mechanism and a framework that can hold the data from the user to the database. The information will be carried by those frameworks to the

database. In this scenario, data quality cannot be manipulated as there are different technologies that can be used for the information transfer from the client.

A decision making and the implementation of the predictions can be done with the help of valuable information. This kind of information should be with valid quality metrics and the metrics need to be followed to maintain the great accuracy of the data.

Extract, transform and load (ETL) is the process of handling the data quality with the DW and the process of data quality will be affected when in process of pre-processing.

The quality of data can be slightly compromised based on its functions, such as extraction, transformation, cleaning, and loading. Data is affected by several processes depending upon its environment. Even though, after cleaning and filling, there may be residual dirty data, which should be reported and these remaining dirty data can be are the reason for failure during the process of data cleaning.

2 Literature review

A large amount of data can be stored in a cloud that manages the user data better during the computing process. In the cloud, the most general challenge raised was the quality of the data stored by the user. The quality of data can be improved by utilising certain customary methods with better efficiency. Each unit of data has a certain aspect that has been analysed by various researchers which in turn affects the quality of data. For instance, Dung and Phoung in 2019 handled the missing data which were not in records and estimated the time of measurement thereby improved reliability. Most conventional data mining systems exists and concentrate on the extraction, transformation and loading phases. Data storage is another method which is considered to enhance and further evoke the efficiency. Data access and control is also achieved with several frames. The machine learning concept is also inculcated with a few dummy variables in some approaches.

2.1 Extraction, transformation and loading (ETL)

Idris et al. (2011) analysed the management of data source quality for DW in manufacturing service. It is well known that in DW project, data quality and data source management are the main key factors. Many previous projects have been failed because of the low quality of data and hence, in order to overcome failure, study was conducted on many aspects including total data quality management (TDQM) and quality management system (QMS). These aspects helped to address problems in data quality and detect the best procedure for managing data resource. A high QMS was proposed in managing data source. The proposed approach had the advantage to overcome the failure when compared to traditional approach.

Boufares and Salem (2012) studied the conflicts associated with data quality and heterogeneous data integration. The concepts of non-quality data and the dimensions of data quality are known. The study discussed the importance of textual data in the document for decision making. The data quality in DW is analysed using the textual data. Mostly the documents which are in extensible mark-up language (XML) format describes the several structures involved in it. So, the data quality integration became a greater challenge.

Anand et al. (2013) discussed the data quality issue in the DW at extraction, transformation and loading (ETL) stages. DW are subject oriented, time variant, on-volatile and perform OLAP. The DW architecture has three layers including data source, data staging area and primary DW. During ETL process, data extracted from OLTP database is transformed for determining the data quality. The cause of the problem in data quality is identified during the different stages in the DW.

Bansal and Kagemann (2014) proposed an ETL scheme that utilised semantic techniques to combine and distribute data from numerous sources. The major drawbacks of big data were better querying and development of enhanced applications using data collected from distinct sources. The automatic development of data structure based on semantics algorithms was not generated in the suggested method.

Guo et al. (2015) proposed a new ETL method as TEL that processed on transform, extract and load approach. Virtual tables were applied that illustrated the stage transformation from extraction to the loading stage. This method was performed without the area of data staging or database that stored extracted raw data which was gathered from each system with dissimilar data sources. The outcomes obtained proved that the method proposed has the lower time of response, constant efficiency in execution specifically for large data.

Kholod et al. (2016) described a method that implements altered ETL approach to develop a warehouse for the virtual data. The overall structure and operation metrics were explained along with the present public domain ETL tool. The outcome obtained was utilised in designing of sample warehouse for virtual data. The developed method had the main limitation of reduced speed to execute requests and delays also occurred due to the transfer of data and requests.

Ali (2018a) presented an extended ETL approach that addressed the problems from big data. Various drawbacks and deficiencies were overcome in the present method which was based on the tools of ETL. The component of UDF's was also offered that focused on challenges that degraded the support and its optimisation in earlier frameworks that integrated the part of developed method transformations in the big data environment. The approach also utilised a monitor agent that provides information to the developer of ETL, which further retrieved information. The different aspects of big data were focused and the proposed approach addressed the above-mentioned issues that occurred by big data.

Azeroual et al. (2018) proposed a data collection structure to achieve data quality in former systems by integrating information-based systems. Furthermore, the needs and dismissals among fields of data were examined for better accuracy using research information systems.

Sadiq et al. (2018) utilised the collection, integration and transfer of research information into various research information systems that result in various data errors which may have a variety of negative effects on data quality. To detect and handle errors effectively at an early stage, the clean-up measures and new data cleaning techniques in research institutions need to be determined to improve performance.

Luo et al. (2019) proposed a rotation estimation method that searches for crowd authentication to approve the contribution crowd based on sensor data. The data from sensors used such that data were rearranged with improved reliability. The quality of the data was highly improved by utilising real-time datasets for evaluation.

Tian et al. (2019) proposed a mechanised data authentication framework related to optical character recognition (i.e., OCR technique). Primarily, the images scanned were analysed with the improved OCR machine learning to identify checkbox symbols and script. Recovering data is also ensured to increase efficiency by recognising inaccuracies.

Table 1 Comparison table of data quality in ETL

S. no	*Methodology and author's name*	*Description*	*Advantages and disadvantages*
1	Rotation estimation method (Luo et al., 2019)	Searched for crowd authentication to approve the contribution crowd based on sensor data	Improved data quality, detected concealed truth
2	TEL (Guo et al., 2015)	Virtual tables were applied that illustrated the stage transformation from extraction to loading stage	Practical, feasible, space efficiency, execution efficiency, lower response time
3	Flexible automatic method (Chand, 2018)	Mechanised process of various user requisites along with the storage in warehouse	Maximum efficiency, some data formats not supported
4	Extended ETL (Ali, 2018b)	Monitor agent that provide information to the developer, which retrieved information	Deficiency were overcome, enhanced efficiency

Source: Author

2.2 Handling missing/mislaid data

Cai and Zhu (2013) formulated a method for hierarchical quality of data through analysis of various big data features. The data with greater quality was one main prerequisite for evaluating and utilising which ensured the data cost. The suggested method includes attributes of quality, dimensions of quality for big data and indexes of quality. Also, this method offered certain advantages of flexibility and better extendable properties by which the requisites of the quality evaluation were acquired. The challenges faced due to the big data usage by organisations were mostly quality assurance through better analysis.

Purwar and Singh (2015) proposed a new hybrid prediction method with missing rate imputation (HPM-MI) that examines several imputation methods using K-means and the best one is applied to a dataset. This proposed procedure is a combination of multilayer perception with K-means. K-means were used to check the class label of particular information before relating classifier. The quality of information was improved drastically by using the best imputation techniques, but these techniques are not applicable to various class imbalanced classification problems. This problem was due to the imbalanced spreading of cases between several classes existent in the dataset.

Pampaka et al. (2016) compare the methods such as stepwise regression and MI models with method from real enhanced example. The value of MI is discussed and the risks involved in discounting missing data are taken. The missing data have a serious effect on imputation and analyse methods.

Enders (2017) presented a practical problem that clinical examiners are likely to meet when applying various imputation, includes a combination of categorical and non-stop variables such as significant testing, item-level lost evidence in surveys, various level missing data and interaction effects.

Ezzine and Benhlima (2018) have reviewed methods and schemes to handle missing data in the context of big data. A design was constructed to predict the missing data based on machine learning. Also, exact attributes were selected that improved the method designed by eliminating data correlation and helped to avoid biased model production. Selecting features were hard in the big data environment as it deals with several numbers of attributes. Imputation of data was not supported when it predicted the imputed values while real values were mean for approach. Also, the imputation of data brought out the uncertainty that was considered for variance estimation.

Gupta and Gupta (2018) reviewed several approaches that were used to handle lost data that included K nearby neighbour (KNN), multiple attributions, case removal, best collective method and so on. The missing data problems brought out partial outcomes, false calculations and some patterns were concealed. In general, the collected data was inadequate and lost values were present that continued the challenge further.

Table 2 Comparison table of handling missing data in ETL

S. no	*Methodology and author's name*	*Description*	*Advantages and disadvantages*
1	SLP-SVR (Dung and Phoung, 2019)	The missing data or not in record were estimated at the time of measurement	Estimated damaged data, improved reliability
2	KNN approach (Ezzine and Benhlima, 2018)	Exact attributes were selected that improved the method designed which eliminated data correlation and helped to avoid biased model production	Outcome were better, increased analysis quality, highly efficient, data were not pre-processed
3	HPM-MI (Purwar and Singh, 2015)	Examine several imputation methods using k-means and best one is applied to dataset	Quality of information was improved, not applicable for various class
4	Stepwise regression and MI models (Pampaka et al., 2016)	Imputing large quantity of missing information points for a binary outcome variable	Real-time data implementation, simple algorithms

Source: Author

Srivastava et al. (2019) analysed that high-quality data is the prerequisite for analysis, manipulation and is also the guarantee of data value. There is currently a lack of systematic analysis, research for quality standards and methods for assessing the value of big data. The paper also puts together reviews of work on data quality and analysed big data systems, data characteristics and discussed quality challenges faced by big data, and formulates a hierarchical data quality model from the data user's perspective. Finally, the paper presents a dynamic data quality assessment process based on proposed framework.

Dung and Phoung (2019) developed a scheme to process missing values by constructing a profile with a standard load (SLP) in line with the previous data load. Further, they combine the scheme with algorithms of machine learning which are comprised of support vector reversion (SVR) compared with random forest and neural network approaches to reconstruct the curve. By this integrated algorithm, the missing data or not in the record were estimated at the time of measurement.

2.3 Crypt information and management

Talib et al. (2016) analysed the issues of increased data sizes in extraction and data evaluation in a timely way to enhance prediction. The calculated process of decision making was better improved by utilising DW. The current and earlier data were stored in the source of a DW. From several sources, the data were collected and later combined in the warehouse which was performed by ETL.

Costa and Santos (2018) explored the role of data structure and modelling in Hive processing times for BDWs, benchmarks multidimensional star schemes and completely denormalised tables with dissimilar scale factors and analysed the effect of appropriate data dividing into these two data modelling strategies. There was a clear benefit in partitioning details, as the data gathered in the questionnaire showed a significant reduction in query compiled code when hive tables were properly partitioned. The findings described highlights the potential benefit of faster manipulation of data dividers as these methods have drastically reduce processing time based on the task being done.

Manogaran et al. (2018) proposed Bayesian hidden Markov model (HMM) with Gaussian mixture (GM) clustering procedure to design the DNA copy count alters across the genome. The projected Bayesian HMM with GM clustering method is matched with a various current approach such as binary division method, pruned strict linear time procedure, sector neighbour procedure.

Sebaa et al. (2018) proposed Hadoop based DW architecture and conceptual information model is used in the field of medicine. It is cost-effective and reports traditional medical DW issues. In Hadoop architecture, information integration is done in the information server, where the replica is automatically generated in more than one data. It stores structured and unstructured data and responsible for answering, writing and reading data queries.

Chandra and Gupta (2018) presented a formal classification of data warehousing and provide a complete view of the fields considered in the research. Later, for potential guidance, current research problems and data warehousing obstacles were outlined. Effective collection, maintenance and review of the massive volume of data, data warehousing has become the most important technology.

Costa and Santos (2018) explored and analysed few architecture patterns and trends in Big Data warehousing systems, including data processing techniques and some streaming issues for BDWs. The paper presented scientific results to help the study and interpretation of several architecture dynamics and developments in Big Data warehousing. The key conclusions of the method included flat tables tend to outperform star patterns, both large and small, but there were situations in which star patterns have certain advantages.

Diop et al. (2019) proposed an architecture built on the projected method for handling time-based quality of data. Implementation architecture defines the situation of prior time-based data managing and how the information stored should be arranged to retain the history deprived of upsetting the software performance. Several techniques had been used in data warehousing and data mining to detect and handle issues in the quality of data. The architecture is structured to advance the DM process by allowing automation of a great deal of information research and anticipation.

Erkayaoglu and Dessureault (2019) presented a framework with mining and warehousing of data that enhanced the process of mine-to-mill. Both the methods were substituted tools that depend on strong data arrangement. These tools were implemented in the advanced mining process that was in line with practical data that predicted the performance of blast in mines. The major algorithm used was random tree classification and flexible boosting over data combined in a warehouse that determined the required active parameters for high efficiency. The limitations in the implementation of this method were eliminated by fuzzy schemes through the corresponding models in the warehouse.

Iam-On (2019) presented a fresh framework called link-built consensus data clustering with the occurrence of missing values which is modelled as information pre-processing tasks. Here the author uses numerous information variations that can endorse the range within a group ensemble. Binary value matrix indicates ensemble data, WCT link depended on matrix used to improve matrix so it gave improved clustering result.

Table 3 Comparison table of crypt information and management

S. no	*Methodology and author's name*	*Description*	*Advantages and disadvantages*
1	Bayesian hidden Markov Model (HMM) with GM (Gaussian mixture) (Manogaran and Vijayakumar, 2018)	Design the DNA copy count alters across genome	Higher efficiency, complex process
2	Hadoop based data warehousing (Sadiq and Dasu, 2018)	Nested partitioning and the solution are applied to DW platform to ensure optimal allocation of health resources	Cost effective
3	MECCA	Enabled new types of applications that provide software components and dependent services across decentralised edge and cloud infrastructures	Complex design, reliability
4	Random tree classification and flexible boosting (Erkayaoglu and Dessureault, 2019)	Data combined in warehouse that determined the required active parameters for high efficiency	Better results, system performance, better security metrics

Source: Author

2.4 Machine learning models and dummy variables

Merino et al. (2014) proposed a method called '3As Data quality in use model'. It is composed of 3 data quality attributes such as operational adequacy, contextual adequacy

and temporal adequacy for retrieving the levels of information quality in projects of big data. The model was integrated into any big data projects as it's free of technologies or pre-conditions. This paper projects the use of the proposed model with working scenarios.

Rehman et al. (2016) proposed an electricity generator forecasting system that forecasts the quantity of power essential at a rate nearer to electricity consumption for the US. It uses big data analytics schema to practice information composed on powder then it applies a machine learning method to train the structure for the calculation stage. The model determines future power generation depending on the collected data and the test output shows that the projected system determines needed power generation is nearest to 99% of real usage. Machine learning with big data can be combined in projecting technique to increase the adeptness and resolve difficult data analytic problem that occurs in a power management system, but load predicting cannot be estimated.

Fatima et al. (2017) offered few approaches based on matching records and restoring of data that eliminated duplication of data with better quality. The performance of the scheme was estimated dependent on the cleaning of data in the warehouse. Pre-processing of data was executed before the phase of data cleaning. The traditional pre-processing comprised of data combination, cleansing and phases of trituration. An approach based on the materialised view increased the overall performance of the DW by query enhancement.

Chen et al. (2017) proposed a convolution neural network-based multimodal disease risk prediction (CNN-MDRP) algorithm using unstructured and structured data to find the precision of risk that was based on mixture feature of hospital. For example, diseases similar to hyperlipidemia, structure information gives a good explanation of the disease. But for a critical disease like cerebral infarction, structure information is not a better way to define the disease. Whereas the CNN-MDRP algorithm gives 94.8% with conjunction speed this was quicker than that of CNN.

Ali (2018b) presented a structure for managing and analysing extremely large and difficult datasets efficiently. For the communication sector, which is required by regulators to handle a significant history of call records of their subscribers, the system can be very efficient, where every single action of a subscriber produces a packet comprising additional 500 attributes. Transactional data analyses permit service providers to fully understand the behaviour of their clients, like deep packet examination and include transactional internet practice of data to clarify consumer's data utilised behaviour. On the conflicts, database engine systems restrict service suppliers to maintain only the lexical level call history compiled at the level of the subscriber.

Azgomi and Sohrabi (2018) recommended a structure focused on game theory for the discovery of materialised opinions. A novel static technique, called materialised view selection (GTMV) based on Game Theory, has been proposed according to the framework. Many virtual and real-world databases were used to test the accuracy of the proposed solution. Experimental results indicate that the GTMV approach works better than previous algorithms and significantly outperforms previous approaches. One of the most important issues posed in OLAP was the rapid response to complicated queries. An enormous volume of data used by OLAP systems in data centres was one of the key obstacles to rapidly seeking the response to queries.

Li et al. (2018) developed an effective strategy on dynamic mapping methods based on machine learning that eliminated certain challenges. The responses from the users

were gathered regularly by which the new user responses were also gathered through an online survey.

Iqbal et al. (2018) resented a data modelling method called hierarchical spatial-temporal state machine (HSTSM). This method depends on the function and structure of mammalian brains. It includes various techniques for soft computing and to deal with a huge quantity of information which is categorised by spatial sequential correlations. This method holds great requirement and exploit the potential of allocating with big data and consider as a tool for analytics of big data.

Liono et al. (2019) offered a unique QDaS technique for efficient storage of data and precise organisation of IoT uses. The method included a distinct mechanism for characterised data which utilised an advanced evaluation approach for the quality of data. The data quality is estimated without the need for response from data users or data field awareness. A major advantage of this method was data in the cloud that was stored by the user were reduced through the characterised quality of data approach.

Table 4 Comparison table of machine learning models and dummy variables

S. no.	*Methodology and author's name*	*Description*	*Advantages and disadvantages*
1	GTMV (Azgomi and Sohrabi, 2018)	Game theory for the discovery of materialised opinions	Improved performance, better execution time
2	MLADM (Li et al., 2018)	Effective strategy dynamic mapping and generates automatic queries	Effective user response, customer effective evaluation
3	CNN-MDRP (Chen et al., 2017)	Unstructured and structured data to find the precision of risk that is based on mixture feature of hospital	Fast conjunction speed
4	HSTSM (Iqbal et al., 2018)	Data modelling method to deal with huge quantity of information	Applicable only for specific areas

Source: Author

2.5 Contradictory data

Todoran et al. (2015) assessed the quality of the information based on the dimensions in systems with a distinct approach. The phenomena of the propagated quality supported the queries of sudden quality evaluated in this approach. The method offered to permit the transfer of information to the end-user for both quality of the local and output obtained. Major implementations of this method consisted of an automated system that recognised the target and a diagnosis system that supported codes.

Aubry et al. (2017) states verifiable method and performance which was significantly greater than the unscreened model depended on various test gains, metrics, and estimation of training and statistical tests of predicting each model test localities. Verifiable models were reliable with our facts of fisher's habit dealings and potential supply, whereas the unscreened method represents much larger space of great quality habitat that involves great expanses of great boost habitat of fishers that do not occupy.

Yang et al. (2017) reviewed the significance and advantages of using cloud computing to handling big data in related science domain and Earth. They have introduced upcoming revolutions and agenda of research for cloud computing by helping

in the transformation of velocity, volume, veracity and variety into value of big data for native to universal earth science and apps. Some more aspects of this invention are evolving with expansion and popularisations of Big Data how they work, live, prosper and think.

Münzberg et al. (2018) suggested approach reflects on combining food sources of data into some kind of central database through a collection of extracts, transforms and loads, and the resulting increase in data quality. The data obtained will be forwarded to other health apps for another use by the food data mobile application. In addition, it will be planned to use data profiling methods to define inaccurate, false, duplicate as well as incomplete data so they can be corrected.

Corrales and Ledezma (2018) designed a DC-RM regression framework for cleaning the data. The method developed perform cleaning was estimated the means of real-time datasets taken from machine learning sources. The dataset is cleaned using the proposed method and is later used in improving a similar model of regression.

Qiu and Sun (2019) proposed a unique spatio-temporal data structure that makes data processing and spatio-temporal evaluation easily understood for better RWQ results. In this system, a specific point, representing both position and complex water quality details, was assumed to be the fundamental element of river spaces, and methods of extending a point to the line segment, a flat surface as well as a cube were created to make this concept relevant to various river space generalisations

Table 5 Comparison table of contradictory data

S. no.	*Methodology and author's name*	*Description*	*Advantages and disadvantages*
1	Proposed spatio-temporal data structure (Qiu et al., 2019)	Data processing and spatiotemporal evaluation will be easily understood for better RWQ results	Effective investigation, reduced computational geometry, effective simulation
2	SCDAP (Osman et al., 2017)	Grants new functions to big data analytics structure for smart city	Functions will increase, better capability
3	Proposed mechanised data authentication (Tian et al., 2019)	Images scanned were analysed with the improved OCR machine learning to identify checkbox symbols and script	Increased efficiency, better accuracy, improved efficient
4	Parallel backend (Lavanya et al., 2019)	Distribute work among workforces with the similar tests in active structures in line with a mechanism of load balance	Increased speed, reduced communication, better efficiency

Source: Author

A lot of work has been carried out by various researchers to enhance the data quality in data warehousing. The literature survey discussed above cannot identify metadata during the ETL stage using K-means, EM, DBScan and Ward's algorithm which further affects accuracy. Business models of different applications have different requirements and implementation plan which results in dirty data and mixed values. Various stages of data warehousing such as live data integration, data staging and ETL sometimes generate

sparse data. There is no effective method of handling the scarcity of data. Hence, the data becomes erroneous while migrating from one stage to another. However, there are also some parameters on which researchers are still doing their research. These parameters include machine learning in dummy variables, expectation and maximisation of the datasets and cluster characters of the data.

3 Gaps in the study

The research work discussed points out the various limitations of handling NULL values, cryptic values, contradicting data (dirty data) and dummy values in a dataset. Various research gaps identified during the literature review have been listed as follows:

a The algorithms such as K-mean, DBSCAN and Wards do not give accurate results during the data extracting process using Minskowi distance metrics. Metadata cannot be identified during the ETL stage using these algorithms which further affects accuracy.

b Business models of different applications have different requirements and implementation plan which results in dirty data and mixed values.

c Various stages of data warehousing such as live data integration, data staging and ETL sometimes generate sparse data. There is no effective method of handling the scarcity of data. Hence the data becomes erroneous while migrating from one stage to another.

The above reasons motivated to improve the existing algorithms for handling missing values, identifying contradictory data, adding dummy values, and handling cryptic values in a dataset in order to ensure data quality.

4 Methodology

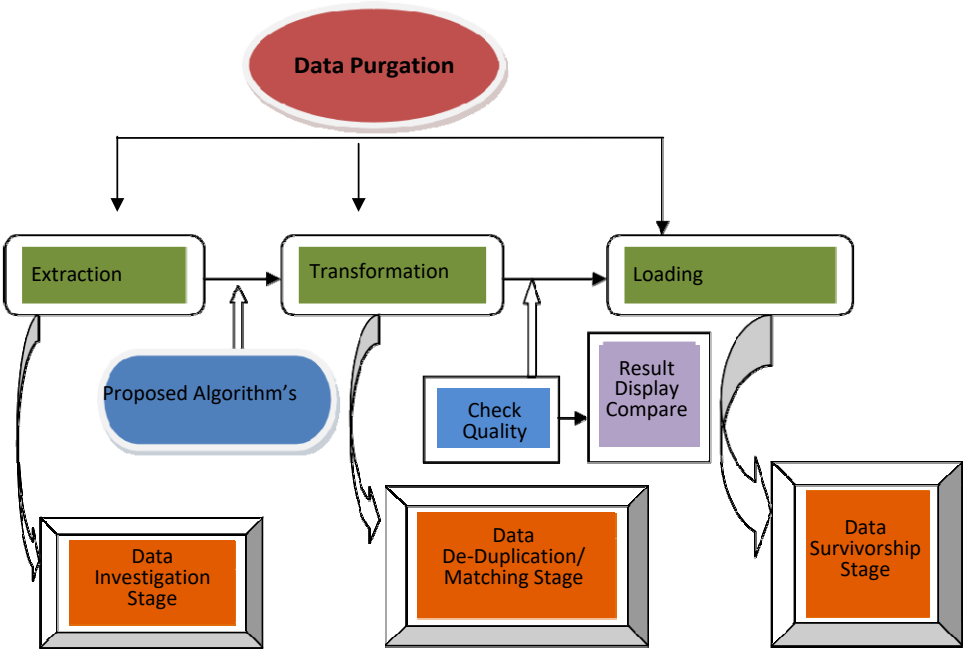
The data quality is dependent on various parameters such as precision, completeness, status update, relevance and consistency across sources of data. The accuracy of data cannot be assured by depending on data entry, storage and manageability.

a The proposed algorithms will help in finding possible solutions to handle missing values, avoid dummy values, cryptic values and contradicting data.

b The algorithms will also help in maintaining the accuracy, integrity, consistency, non-redundancy of data in a timely manner.

This can be achieved by the data purgation process. The Data purgation process consists of three phases. They are DE phase, the DT phase and DL phase. It is represented in Figure 1.

Figure 1 Data purgation process (see online version for colours)



Source: Author

The data purgation process has four stages namely data investigation stage, data standardisation stage, data de-duplication/matching stage, data survivorship stage.

a *Data investigation stage:* in this stage, the data is analysed for determining the errors and patterns. This stage requires all sample datasets to achieve the data quality during the ETL process in DW. Auditing results show frequency reports on various tokens, labels and record patterns. This frequency reports discover the tuning standardisation rule sets for a given heterogeneous datasets.

b *Data standardisation stage:* in this stage, the extracted data is converted into a standard format. This stage helps in the segmentation of data, canonicalisation, correcting spelling errors by using iterative tuning rule sets. Predefined rules are applied to the extracted dataset to obtain a reliable data format.

c *Data de-duplication/matching stage:* in this stage, the duplicate records are identified from the standardised datasets. It also provides the parameters for the blocking and the matching steps. The blocking step is used to minimise the search process and the matching step is used to determine the identity between the records.

d *Data survivorship stage:* this stage is responsible for choosing the datasets that will be loaded into the DWs after the previous stage if the dataset is extracted by merging the data from the different data sources. This stage also explains the overall merging of data during the ETL process.

To implement the data purgation process and to attain data quality before the transformation stage, four algorithms have been implemented. The present work has improved the expectation-maximisation algorithm with dot product to handle cryptic data, DBScan method with Gower metrics has been implemented to ensure dummy values, Ward's algorithm with Minkowski distance has been applied to improve the results of contradicting data and K-means algorithm along with Euclidean distance metrics has been applied to handle missing values in a dataset. These distance metrics have improved the data quality and also helped in providing consistent data to be loaded into a DW. The proposed algorithms are applied after the extraction phase of the data purgation process as represented in the Figure 1. These proposed algorithms will produce a better performance based on accuracy and execution time. The above methodology is implemented on the Net Beans platform using the Java framework.

The following subsections describe the brief explanations of algorithms.

4.1 Identifying cryptic values

Cryptic values are mismatched information in a dataset. EM algorithm has been used to handle cryptic values in a dataset. The EM algorithm is an efficient iterative procedure to estimate the presence of missing or hidden data. Each iteration of the EM algorithm consists of two processes: The E-step, and the M-step. In the expectation, or E-step, the missing data are estimated given the observed data and current estimate of the model parameters. This is achieved using the conditional expectation, explaining the choice of terminology. In the M-step, the likelihood function is maximised under the assumption that the missing data are known. The estimate of the missing data from the E-step is used in lieu of the actual missing data (Jolly and Gupta, 2019c). Convergence is assured since the algorithm is guaranteed to increase the likelihood at each iteration.

In order to identify and correct the mismatched values in a dataset, EM modelling is maximised using the dot product of the GM. The dot product of the GM model has been implemented on the maximisation step of EM modelling using the formula as below equation.

$$p_j(y) = V\left(y; \mu_j, \sigma_j\right) = \frac{1}{\sqrt{2\pi\sigma_j}} \exp\left(-\frac{\left(y-\mu_j\right)^2}{2\sigma_j^2}\right) \tag{1}$$

Datasets used:

a credit: no. of attributes: 10, no. of records: 1,000

b intrusion: no. of attributes: 42, no. of records: 125,974

c air quality: no. of attributes: 15, no. of records: 935

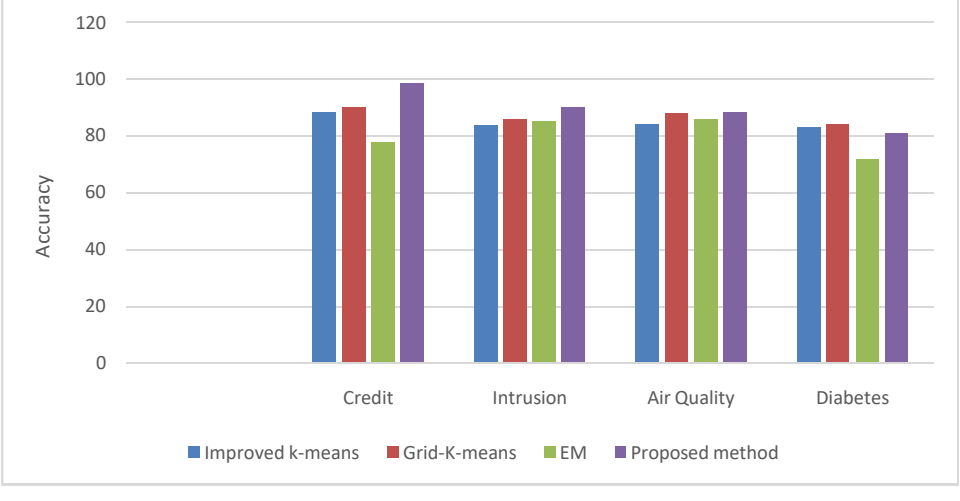
d diabetes: no of attributes: 9, no. of records: 768.

Table 6 Comparison of accuracy in percentage w.r.t to existing methods and proposed method

Dataset	*Improved K-means (existing method)*	*Grid-K-means (existing method)*	*EM (existing method)*	*Proposed method*
Credit	88.43	90.31	78	98.83
Intrusion	83.89	85.96	85.36	90.23
Air quality	84.29	88.23	86.12	88.52
Diabetes	83.24	84.21	72	81.12

Source: Author

Figure 2 Accuracy comparison between existing and proposed methods w.r.t. datasets (see online version for colours)



Source: Author

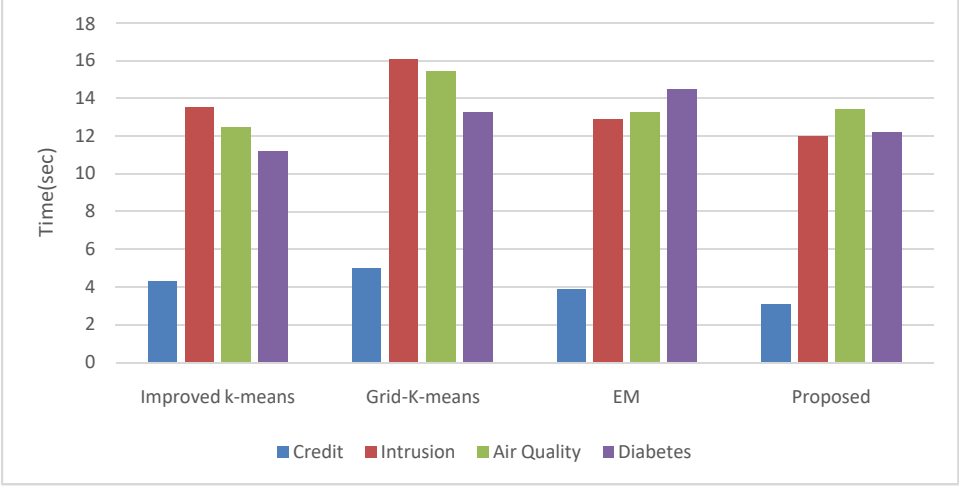
The modified approach changes the maximisation step in order to cluster the exact value of the dataset. As in Figure 2 and Figure 3, the frequent datasets are being tested on the proposed method and the cryptic values are handled in a more efficient way w.r.t accuracy and execution time.

Table 7 Comparison of execution time: time in (sec) w.r.t existing and proposed method

Dataset	*Improved K-means*	*Grid-K-means*	*EM*	*Proposed*
Credit	4.33	4.98	3.89	3.11
Intrusion	13.53	16.09	12.89	12.00
Air quality	12.45	15.42	13.25	13.45
Diabetes	11.20	13.25	14.50	12.20

Source: Author

Figure 3 Comparison of execution time: time (sec) w.r.t existing and proposed algorithm (see online version for colours)



Source: Author

4.2 Adding dummy values to ensure data accuracy

Dummy variable also known as an indicator variable, Boolean indicator, binary variable or qualitative variable is one that takes the value 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome. For the need of high similarity and low variance, DBScan with the value for a position has been implemented. The algorithm uses DBScan clustering with Gower distance metrics to find the maximum and minimum value that is compared with position value. The clustering technique has been implemented to give the result as group 1 and group 2 by adding dummy values as group 2 for 1 and group 1 for 0 but dummy values are not replaceable. The proposed algorithm has helped in separating clusters of high density versus clusters of low density within a given dataset. The overall precision is achieved in the execution of the proposed system and DBScan has shown the improved accuracy with the highest calculating time (Jolly and Gupta, 2020).

Datasets used

a credit card: no. of attributes: 24, no. of records: 30,000

b diabetes: no. of attributes: 9, no. of records: 768

c MNIST: no. of attributes: 13, no. of records: 60,000

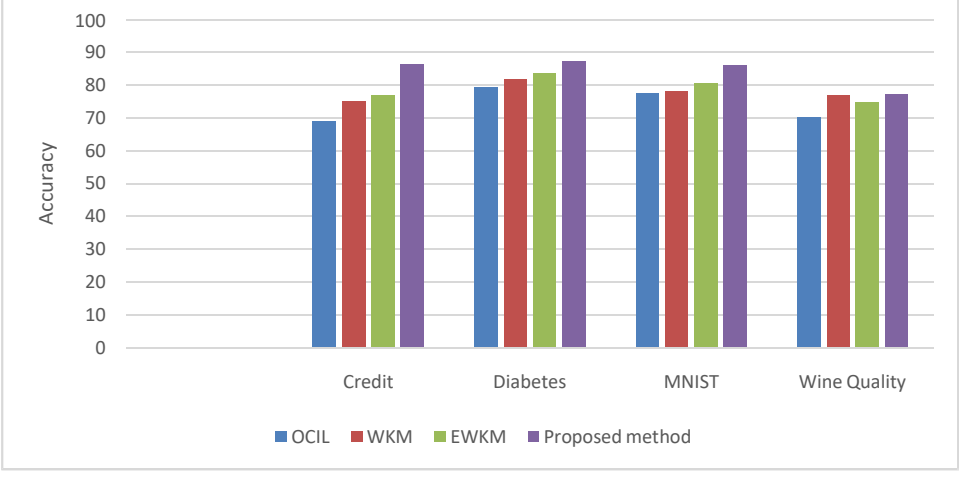
d wine quality: no. of attributes: 12, no. of records: 4,898.

Table 8 Comparison of accuracy in percentage of datasets with existing and proposed method

Dataset	*OCIL (existing method)*	*WKM (existing method)*	*EWKM (existing method)*	*Proposed method*
Credit	69	75	77	86.40
Diabetes	79.30	81.73	83.50	87.20
MNIST	77.45	78.23	80.50	86.21
Wine quality	70.23	76.85	74.89	77.12

Source: Author

Figure 4 Accuracy comparison between existing methods and proposed method w.r.t. datasets (see online version for colours)



Source: Author

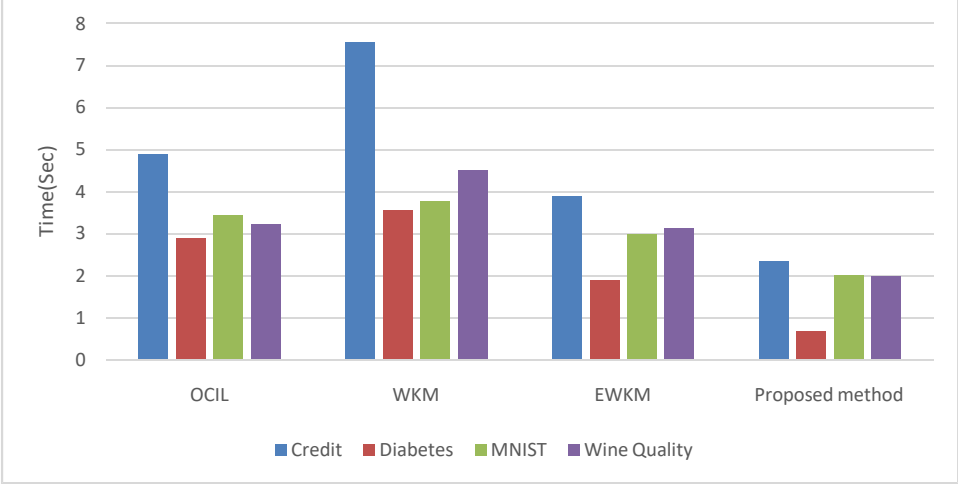
Figure 4 shows the improved accuracy of the proposed method which in turn ensures data quality.

Table 9 Comparison of execution time: time in (sec) w.r.t existing and proposed method

Dataset	*OCIL*	*WKM*	*EWKM*	*Proposed method*
Credit	4.89	7.55	3.90	2.35
Diabetes	2.89	3.55	1.90	0.68
MNIST	3.45	3.78	2.98	2.01
Wine quality	3.23	4.52	3.14	2.00

Source: Author

Figure 5 Comparison of execution time: time (sec) w.r.t. existing and proposed method (see online version for colours)



Source: Author

4.3 To handle the contradicting data with improved distance metric access

Contradictory data means having mixed values in a particular dataset which in turn affects the data quality (Jolly and Gupta, 2019b). There exist a few metrics that have been designed specifically to handle the contradicting data. These metrics are Gower's and Minkowski distance metric. Contradictory data is incorrect and it is important that such data be investigated and evaluated when analysing a noisy dataset. To handle contradicting data, the Wards algorithm with Minkowski distance has been applied to improve the results. The percentage of retrieving records is increased while using the Wards clustering algorithm which in turn give more accurate results. The following distance metric is applied to increase the accuracy.

$$dist = \left(\sum_{k=1}^{n} |p_k - q_k|^r \right)^{\frac{1}{r}} \tag{2}$$

In Minkowski distance, the range of p has relied on the double points X and Y. The independent variable is X and the dependent variable is Y. The independent variable X is used to prejudice the data from the set, and the predicted value will be Y. If there are a vast number of variations between the actual and predicted variables, then there are high possibilities of error rate in values. Therefore, it is necessary to allocate the dataset as an input value to the machine. For experimental analysis, two clusters have been generated and plotted with resemblances and variations between the clusters and data points present in the cluster. After applying the above-mentioned metrics, the efficiency of cluster evaluation has been improved and has shown greater performance metrics.

Datasets used:

a gene: no. of attributes: 70, no. of records: 7,129

b credit: no. of attributes: 10, no. of records: 1,000

c IMDB: no. of attributes: 7, no. of records: 10,000

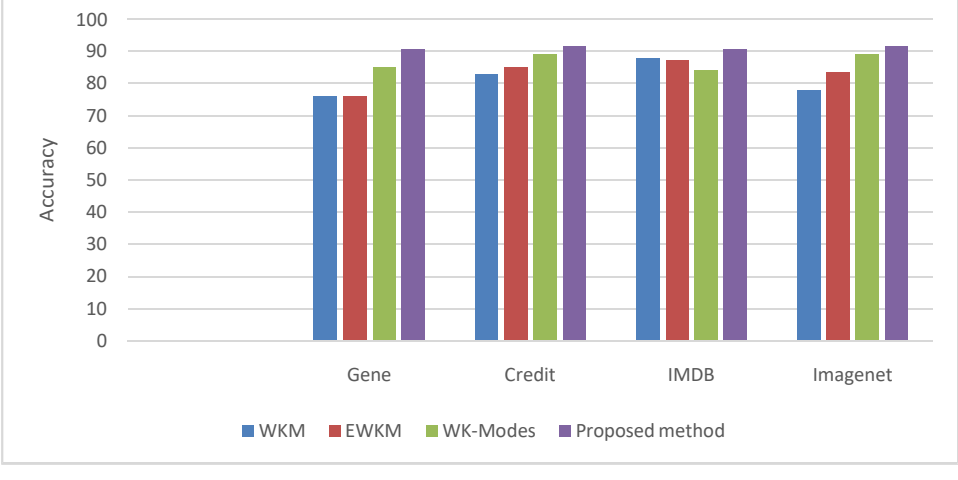
d imagenet: no of attributes: 64, no. of records: 5,620.

Table 10 Comparison of accuracy in percentage w.r.t to datasets with existing methods and proposed method

Dataset	*WKM (existing method)*	*EWKM (existing method)*	*WK-modes (existing method)*	*Proposed method*
Gene	76	76	85	90.70
Credit	83	85	89	91.60
IMDB	88	87	84	90.50
Imagenet	78	83.32	89	91.50

Source: Author

Figure 6 Accuracy comparison between existing methods and proposed method w.r.t. datasets (see online version for colours)



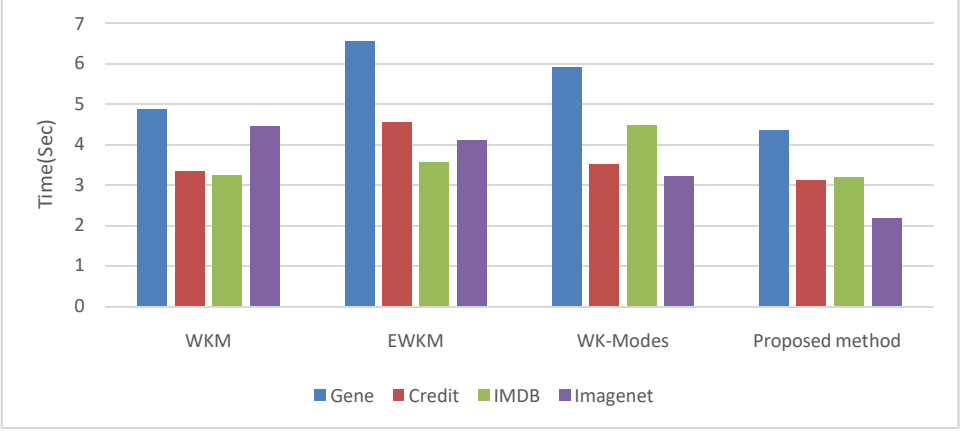
Source: Author

Figure 6 shows the graph plotting for proposed algorithm which ensures the data quality when analysing a contradictory dataset.

Table 11 Comparison of execution time: time (sec) w.r.t existing and proposed method

Dataset	*WKM*	*EWKM*	*WK-modes*	*Proposed method*
Gene	4.89	6.55	5.90	4.35
Credit	3.33	4.56	3.52	3.13
IMDB	3.25	3.56	4.47	3.21
Imagenet	4.45	4.12	3.23	2.20

Figure 7 Comparison of execution time: time (sec) w.r.t. existing and proposed methods (see online version for colours)



Source: Author

4.4 Handling missing data at pre-processing stage

The missing data is an essential factor to select any missing qualities that can be assigned by using some existing methods. The component of the missing information is measured as totally absent, missing partially or missing aimlessly. The instance of the missing data system is not possible without the analysis of the complete dataset. The missing information is generally obscure which is found by using line techniques or segment methods such as deletion techniques, mean value attribution, hot deck attribution methods, K-nearest neighbour algorithm, K-means clustering method, fuzzy K-means clustering methods, multiple and regression attribute methods. The well-known K-means algorithm has bit empirical work for estimation of accuracy and leads to issues of sample value when it is mentioned in close relation to the entire custom of the dataset. Also, it does not give any assurance to the global minimum variable and it is difficult to predict the K value.

These techniques lead to poor performance if the number of samples is lesser than the number of available features. It fails to do complex computations and time-consuming. In the proposed work, the K-means clustering algorithm has been improved using Euclidean distance metrics to handle missing values in a dataset. For numerical missing values in a dataset, the Euclidean distance metric with K means has been implemented to handle numerical data. Euclidean distance metric produces tighter clusters than hierarchical clustering especially if the clusters are globular. This identifies the exact missing position of the relative elements which gives better execution of characterisation. It will provide a unique value for the whole attribute. It leads to the achievement of less prediction average of the classification machine. Clustering of dataset has been done to achieve the better replacement value related to original value, via each cluster with same the weight, age, value entities, for better prediction. The proposed improved K-means clustering technique is more efficient in calibrating the missing values by measuring the distance of nearest entities by locating the positions. This methodology provides a better classification of missing values to the target (Jolly and Gupta, 2019a).

The following distance metric is applied with K-means clustering to increase the accuracy rate:

$$\lim_{1 to cordx.length} Distance = \sqrt{(cordx - cordx[0]^2 + (cordy - cordy[0]^2} \quad (3)$$

The difference between removing missing value in the whole dataset and clustered dataset is time. When clustered data missing value is removed, it will consume less time in comparison to the whole dataset.

The K-mean algorithm along with the Euclidean distance metric is verified on different datasets. The datasets have been obtained from the UCI machine repository.

Datasets used:

a air quality: no. of attributes: 15, no. of records: 935

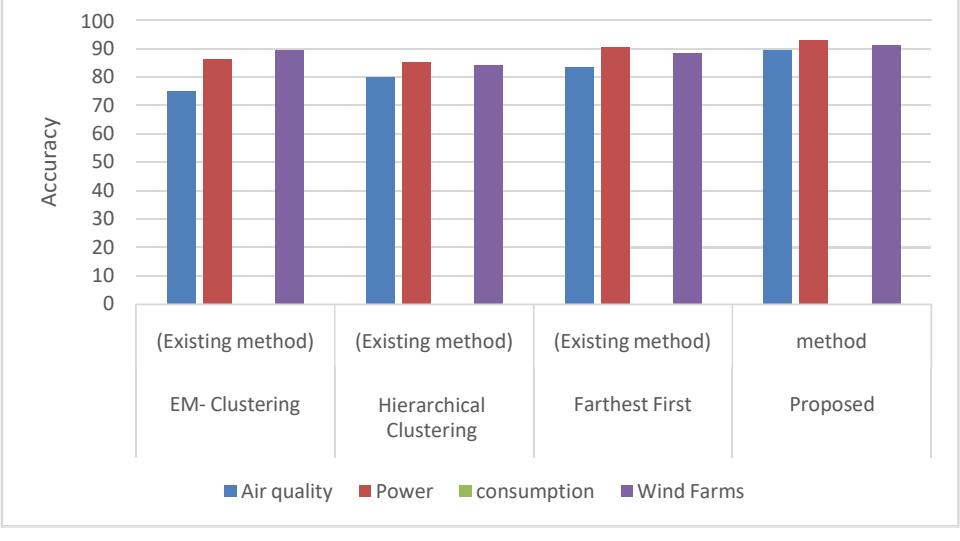
b power consumption: no. of attributes: 9, no. of records: 2,075,259

c wind farms: no. of attributes: 18, no. of records: 25,214.

Table 12 Comparison of accuracy in percentage of datasets with existing and proposed method

Dataset	*EM-clustering (existing method)*	*Hierarchical clustering (existing method)*	*Farthest first (existing method)*	*Proposed method*
Air quality	74.99	79.85	83.62	89.43
Power consumption	86.32	85.23	90.56	93.12
Wind farms	89.65	84.25	88.23	91.23

Figure 8 Accuracy comparison between existing methods and proposed method w.r.t. datasets (see online version for colours)



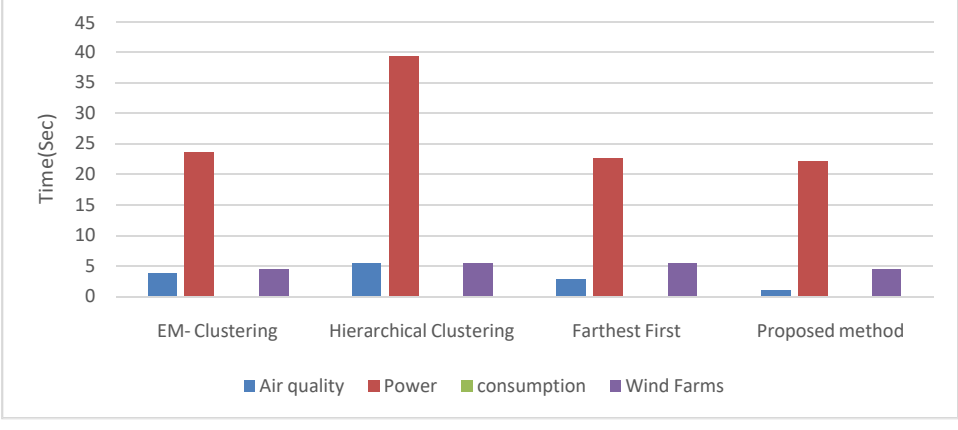
Source: Author

Some more datasets are also tested against NULL values, where the fields were deleted to check the accuracy rate with respect to the actual value.

Table 13 Comparison of execution time: time in (sec) w.r.t existing and proposed method

Dataset	*EM-clustering*	*Hierarchical clustering*	*Farthest first*	*Proposed method*
Air quality	3.89	5.55	2.90	1.10
Power consumption	23.56	39.33	22.63	22.06
Wind Farms	4.56	5.56	5.42	4.50

Figure 9 Comparison of execution time w.r.t. existing and proposed method (see online version for colours)



Source: Author

As shown above in, Figure 8 and Figure 9, the frequent datasets are being tested on the proposed method and the NULL values are handled more efficiently w.r.t accuracy and execution time. As mentioned earlier, after achieving all the above desired results, 'data quality is ensured to attain clean, integrated, historical data in a short time frame for low cost'.

5 Contribution

The present research has been carried out to improve the data quality at the ETL stage of data warehousing by handling missing values, contradicting data, cryptic values and dummy values. Standard data mining algorithms and the distance metric discussed in Section 4 of the long synopsis have been used to improve the data quality of the datasets and also provided consistent data to be loaded into DW. The maximisation step of the EM modelling algorithm has been changed to cluster the exact value of the dataset. Ward's algorithm has given better a percentage of retrieving records using the Minkowski distance metric. The DBSCAN method with Gower metrics has separated the clusters of high density versus a cluster of low density very effectively. K-means clustering algorithm along with Euclidean distance metric has increased the accuracy rate of missing data. The modified algorithms generate more accurate results as compared to the existing algorithms and their variations.

The factors affecting the quality of data are well-explained and it is rectified by using various machine learning algorithm. Each dimension contributing to data quality is well-managed and increased its efficacy at the stage of ETL in data warehousing.

6 Conclusions and future scope

The grave process in DW is purgation and data cleaning, which is performed by extraction, transformation and loading (ETL). It is the process used to fetch data from the OLTP database and converted to the schema of a DW which is fed into DW. Quality of data depends on the schema of database and application programs in ETL. The absence of business rules creates data quality problems and also the reason for the decrease in data quality. The lack of recording the changes made in the source file, inability to track the alterations made in time and date schedule, and absence of periodic refresh at data staging of incorporated data storage systems. Eliminating data from the source at the time of transformation results in the unfortunate consequence that affects the data quality. The proposed methodology will help in accomplishing the data quality during the extraction, transformation and loading process in the DW. The proposed algorithms will help in finding possible solutions to handle missing values, avoid dummy values, cryptic values and contradicting data. The algorithms will also help in maintaining the accuracy, integrity, consistency, non-redundancy of data in a timely manner.

Bunch of cluster characters can additionally be taken in to consideration if the informational indexes are high dimensional. Hard parcelling of clusters to improve information quality can be applied to accomplish non-excess of information. Further, calculations when consolidated together to cause another arrangement of guidelines too can be thought about. In addition to the above-mentioned scope a methodology for building XML DWs using data cleaning and integration can also be area of future research. Structural approach of building an XML document warehouse, specifically to move data from a simple XML database into a given XML DW can also be a promising area. Handling homogeneous and heterogeneous missing data from cloud extraction can also be the area of research. Efficient data initialisation in data stores for mining association laws and by concentrating on calculating aggregate data can help researches in better management of DWs.

References

Ali, A.R. (2018a) 'Real-time Big Data warehousing and analysis framework', *3rd International Conference on Big Data Analysis (ICBDA)*, IEEE, 9–12 March, pp.43–49.

Ali, S.M.F. (2018b) 'Next-generation ETL framework to address the challenges posed by big data', 26–27 March, Vol. 6, No. 2, pp.589–607, IEEE, Vienna.

Anand, N. et al. (2013) 'Modelling and optimization of extraction-transformation- loading (ETL) processes in data warehouse: An overview', *IEEE 4thInternational Conference on Computing, Communications and Networking Technologies (ICCCNT)*.

Aubry, K.B., Raley, C.M. and McKelvey, K.S. (2017) 'The importance of data quality for generating reliable distribution models for rare, elusive, and cryptic species', *PLoS One*, Vol. 12, No. 6, pp.e0179152–e0179165.

Azeroual, O., Saake, G. and Schallehn, E. (2018) 'Analyzing data quality issues in research information systems via data profiling', *International Journal of Information Management*, Vol. 41, No. 2, pp.50–56.

Azgomi, H. and Sohrabi, M.K. (2018) 'A game theory based framework for materialized view selection in data warehouses', *Engineering Applications of Artificial Intelligence*, Vol. 71, No. 3, pp.125–137.

Bansal, S. and Kagemann, S. (2015) 'semantic extract-transform-load framework for big data integration', *Computer (Long. Beach. Calif.)*, Vol. 48, No. 3, pp.42–50.

Boufares, F. and Salem, A.B. (2012) 'Heterogeneous data-integration and data quality: overview of conflicts', *2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, IEEE, March, pp.867–874.

Cai, L. and Zhu, Y. (2015) 'The challenges of data quality and data quality assessment in the big data era', *Journal of Data Science*, Vol. 14, No. 2, pp.536–545.

Chand, K.P. (2018) 'Requirements evocation and analysis using ETL in cloud environments', *International Journal of Information Management*, Vol. 4, No. 2, pp.78–90.

Chandra, P. and Gupta, M.K. (2018) 'Comprehensive survey on data warehousing research', *International Journal of Information Technology*, Vol. 10, No. 1, pp.217–224.

Chen, M., Hao, Y., Hwang, K., Wang, L. and Wang, L. (2017) 'Disease prediction by machine learning over big data from healthcare communities', *IEEE Access*, Vol. 5, No. 2, pp.8869–8879.

Corrales, D. and Ledezma, A. (2018) 'How to address the data quality issues in regression models: a guided process for data cleaning', *Journal of Symmetry*, Vol. 10, No. 4, pp.99–112.

Corrales, D., Corrales, J. and Ledezma, A. (2018) 'How to address the data quality issues in regression models: a guided process for data cleaning', *Journal of Symmetry*, Vol. 10, No. 4, pp.99–112.

Costa, C. and Santos, M.Y. (2018) 'Evaluating several design patterns and trends in Big Data warehousing systems,' *International Conference on Advanced Information Systems Engineering*, Springer, 11–15 June, Estonia, pp.459–473.

Diop, M., Camara, M.S., Bah, A. and Fall, I. (2019) 'Prior management of temporal data quality in a data mining process: an implementation architecture', *Procedia Computer Science*, 3–5 October, Morocco, Vol. 148, No. 1, pp.273–282.

Dung, N.T. and Phuong, N.T. (2019) 'Handling missing data using standardized load profile (SLP) and support vector regression (SVR)', *International Conference on System Science and Engineering (ICSSE)*, IEEE, 20–21 July, pp.414–419.

Enders, C.K. (2017) 'Multiple imputation as a flexible tool for missing data handling in clinical research', *Behaviour Research and Therapy*, Vol. 98, No. 4, pp.4–18.

Erkayaoglu, M. and Dessureault, S. (2019) 'Improving mine-to-mill by data warehousing and data mining', *International Journal of Mining, Reclamation and Environment*, Vol. 33, No. 4, pp.409–424.

Ezzine, I. and Benhlima, L.,2018) 'A study of handling missing data methods for big data', *5th International Conference on Information Science and Technology (CiSt) IEEE*, 4–7 March, Morocco, pp.498–501.

Fatima, A., Nazir, N. and Khan, M.G. (2017) 'Data cleaning in data warehouse: a survey of data pre-processing techniques and tools', *International Journal of Information Technology and Computer Science*, Vol. 3, No. 4, pp.50–61.

Guo, S.S., Yuan, Z.M., Sun, A.B. and Yue, Q. (2015) 'A new ETL approach based on data virtualization', *Journal of Computer Science and Technology*, Vol. 30, No. 2, pp.311–323.

Gupta, S. and Gupta, M.K. (2018) 'A survey on different techniques for handling missing values in dataset', *International journal of Information Technology and Computer Science*, Vol. 5, No. 2, pp.80–91.

Iam-on, N. (2019 'Improving the consensus clustering of data with missing values using the link-based approach', *Data-Enabled Discovery and Applications*, Vol. 3, No. 7, pp.126–134.

Idris, N. et al. (2011) 'Managing data source quality for data warehouse in manufacturing services', In Electrical Engineering and Informatics (ICEEI)', *IEEE International Conference*, pp.1–6.

Iqbal, R., Doctor, F., More, B., Mahmud, S. and Yousuf, U. (2018) 'Big data analytics: Computational intelligence techniques and application areas', *International Journal of Technological Forecasting and Social Change*, Vol. 153, No. 1, pp.172–195.

Jolly, S. and Gupta, N. (2019a) 'Handing missing / mislaid data to attain data trait', *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, Vol. 8, No. 12, pp.4308–4311.

Jolly, S. and Gupta, N. (2019b) 'Higher dimensional data access and management with improved distance metric access for higher dimensional non-linear data', *International Journal of Recent Technology and Engineering (IJRTE)*, Vol. 8, No. 4, pp.5692–5696.

Jolly, S. and Gupta, N. (2019c) 'AI proposition for crypt information management with maximized EM modelling', *International Journal of Engineering and Advanced Technology*, Vol. 9, No. 2, pp.1287–1291.

Jolly, S. and Gupta, N. (2020) 'Understanding and implementing machine learning models with dummy variables with low variance', *International Conference on Innovative Computing and Communications (ICICC-20)*, 21–23 February, New Delhi, India, Vol. 1165, No. 1, pp.477–487.

Kholod, I.I., Efimova, M., S. and Kulikov, S.Y. (2016) 'Using ETL tools for developing a virtual data warehouse', *International Conference on Soft Computing and Measurements (SCM)*, IEEE, 25 May, Vol. 2, No. 3, pp.351–354.

Lavanya, K., Reddy, L. and Reddy, B. E. (2019) 'Distributed based serial regression multiple imputation for high dimensional multivariate data in multicore environment of cloud', *International Journal of Ambient Computing and Intelligence (IJACI)*, Vol. 10, No. 1, pp.63–79.

Li, Z., Tian, Z., Wang, J., Wang, W. and Huang, G. (2018) 'Dynamic mapping of design elements and affective responses: a machine learning based method for affective design', *Journal of Engineering Design*, Vol. 29, No. 3, pp.358–380.

Liono, J., Jayaraman, P.P., Qin, A.K., Nguyen, T. and Salim, F.D. (2019) 'QDaS: Quality driven data summarisation for effective storage management in internet of things', *Journal of Parallel and Distributed Computing*, Vol. 127, No. 4, pp.196–208.

Luo, T., Huang, J., Kanhere, S.S., Zhang, J. and Das, S.K. (2019) 'Improving IoT data quality in mobile crowd sensing: a cross validation approach', *IEEE Internet of Things Journal*, Vol. 6, No. 3, pp.5651–5664.

Manogaran, G. and Vijayakumar, V. (2018) 'Machine learning based big data processing framework for cancer diagnosis using hidden Markov model and GM clustering', *Wireless Personal Communications*, Vol. 102, No. 3, pp.2099–2116.

Manogaran, G., Vijayakumar, V., Varatharajan, R., Kumar, P.M., Sundarasekar, R. and Hsu, C.H. (2018) 'Machine learning based big data processing framework for cancer diagnosis using hidden Markov model and GM clustering', *Wireless Personal Communications*, Vol. 102, No. 3, pp.2099–2116.

Merino, J., Caballero, I., Rivas, B., Serrano, M. and Piattini, M. (2016) 'A data quality in use model for big data', *Future Generation Computer Systems*, Vol. 63, No. 1, pp.123–130.

Münzberg, A., Sauer, J., Hein, A. and Rösch, N. (2018) 'The use of ETL and data profiling to integrate data and improve quality in food databases', *14th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob) IEEE*, October, pp.231–238.

Osman, A.M.S., Elragal, A. and Bergvall-kåreborn, B. (2017) 'Big data analytics and smart cities: a loose or tight couple', *10th International Conference on Connected Smart Cities 2017 (CSC 2017)*, Lisbon, 20–22 July, pp.157–168.

Pampaka, M., Hutcheson, G. and Williams, J. (2016) 'Handling missing data: analysis of a challenging data set using multiple imputation', *International Journal of Research & Method in Education*, Vol. 39, No. 2, pp.19–37.

Pampaka, M., Hutcheson, G. and Williams, J. (2016) 'Handling missing data: analysis of a challenging data set using multiple imputation', *International Journal of Research & Method in Education*, Vol. 39, No. 2, pp.19–37.

Purwar, A. and Singh, S.K. (2015) 'Hybrid prediction model with missing value imputation for medical data', *Expert Systems with Applications*, Vol. 42, No. 3, pp.5621–5631.

Qiu, Y. and Sun, J. (2019) 'A novel spatiotemporal data model for river water quality visualization and analysis', *IEEE Access*, 23 October, Vol. 7, No. 2, pp.155455–155461.

Qiu, Y., Xie, H., Sun, J. and Duan, H. (2019) 'A novel spatiotemporal data model for river water quality visualization and analysis', *IEEE Access*, 23 October, Vol. 7, No. 2, pp.155455–155461.

Rehman, M.N., Esmailpour, A. and Zhao, J. (2016) 'Machine learning with big data an efficient electricity generation forecasting system', *Big Data Research*, Vol. 5, No. 4, pp.9–15.

Sadiq, S. and Dasu, T. (2018) 'Data quality: the role of empiricism', *Journal ACM SIGMOD Record*, Vol. 46, No. 4, pp.35–43.

Sadiq, S., Dasu, T., Dong, X.L., Freire, J., Ilyas, I.F., Link, S., Miller, M.J., Naumann, F., Zhou, X. and Srivastava, D. (2018) 'Data quality: the role of empiricism', *Journal ACM SIGMOD Record*, Vol. 46, No. 4, pp.35–43.

Sebaa, A., Chikh, F., Nouicer, A. and Tari, A. (2018) 'Medical Big Data warehouse: architecture and system design, a case study: improving healthcare resources distribution', *Journal of Medical Systems*, Vol. 42, No. 1, pp.59–77.

Srivastava, D., Scannapieco, M. and Redman, T.C. (2019) 'ensuring high-quality private data for responsible data science: vision and challenges', *Journal of Data and Information Quality (JDIQ)*, Vol. 11, No. 4, pp.1–8.

Talib, R., Hanif, M.K., Fatima, F. and Ayesha, S. (2016) 'A multi-agent framework for data extraction, transformation and loading in data warehouse', *International Journal Advanced Computer Science Applications*, Vol. 7, No. 3, pp.351–354.

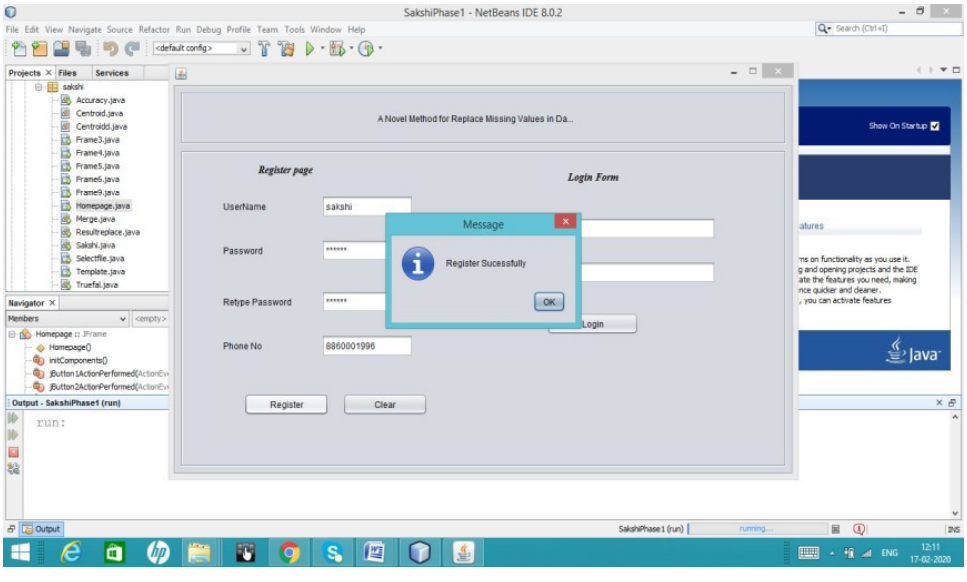
Tian, Q., Liu, M., Min, L., An, J., Lu, X. and Duan, H. (2019) 'An automated data verification approach for improving data quality in a clinical registry', *Computer Methods and Programs in Biomedicine*, Vol. 181, No. 7, pp.104840–104859.

Todoran, I.G., Lecornu, L., Khenchaf, A. and Caillec, J.M.L. (2015) 'A methodology to evaluate important dimensions of information quality in systems', *Journal of Data and Information Quality (JDIQ)*, Vol. 6, No. 2, pp.11–21.

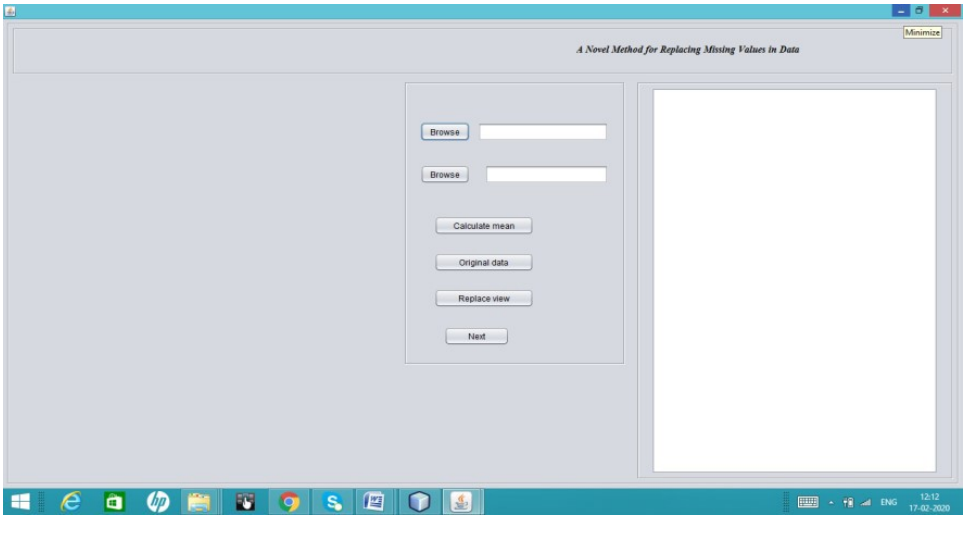
Yang, C., Huang, Q., Li, Z., Liu, K. and Hu, F. (2017) 'Big data and cloud computing: innovation opportunities and challenges', *International Journal of Digital Earth*, Vol. 10, No. 2, pp.13–53.

Appendix

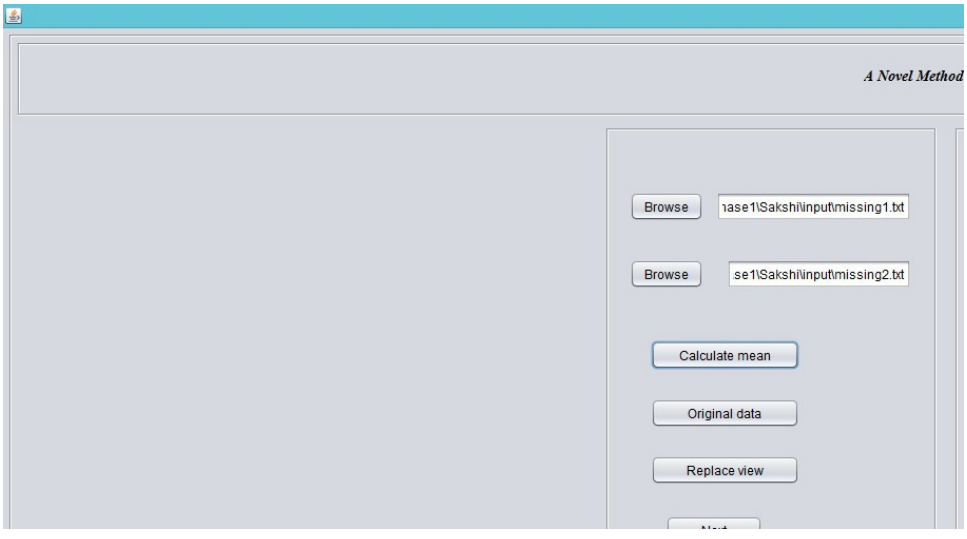
Snapshot 1 Validating the user (see online version for colours)



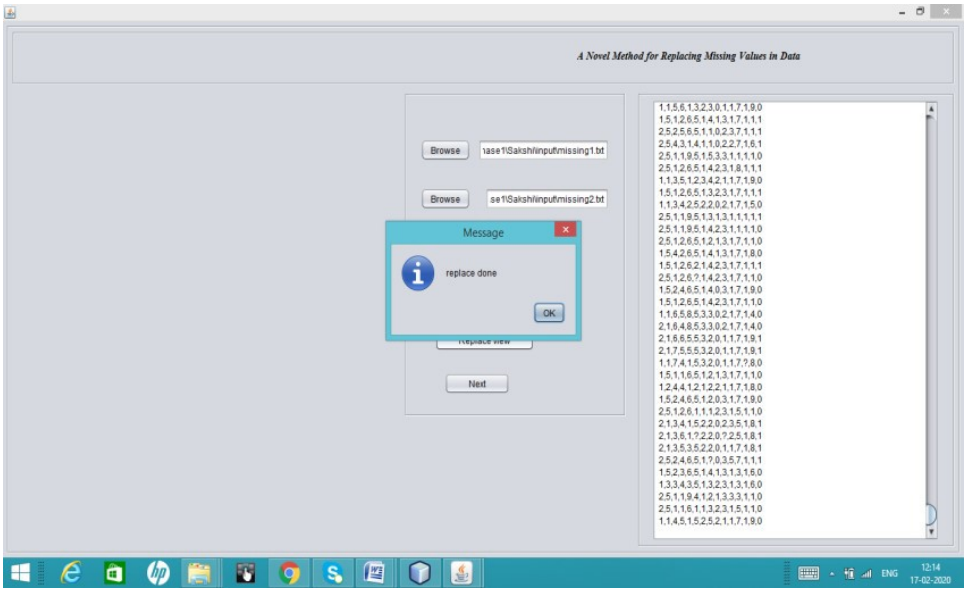
Snapshot 2 Browsing dataset (see online version for colours)



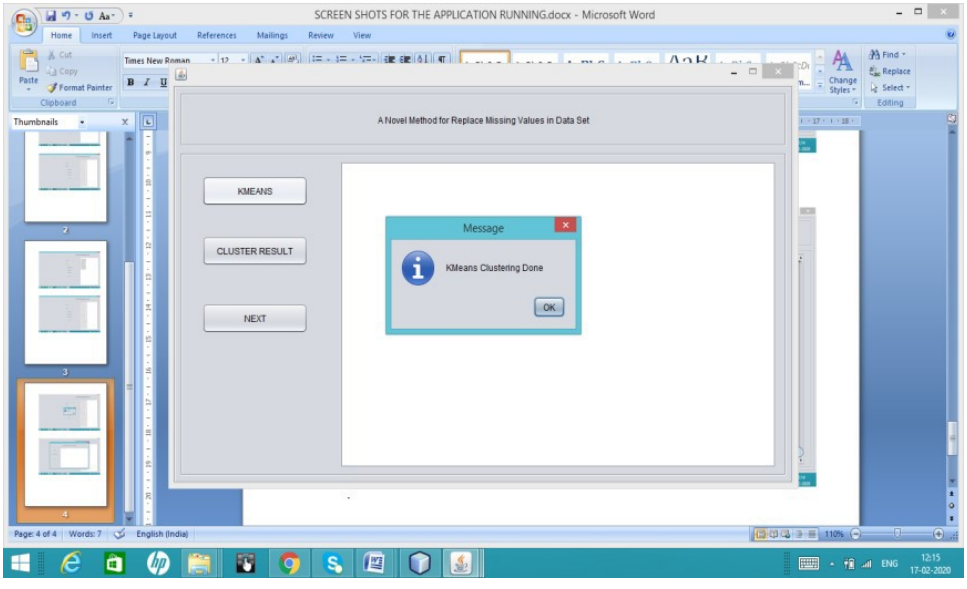
Snapshot 3 Calculating mean (see online version for colours)



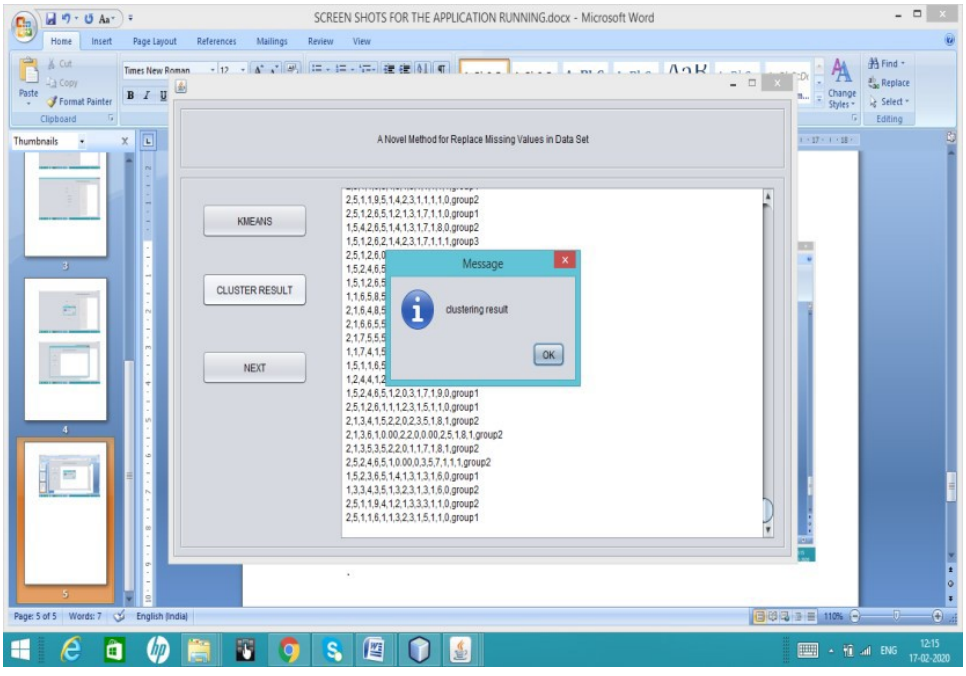
Snapshot 4 Replacement done with respect to mean value (see online version for colours)



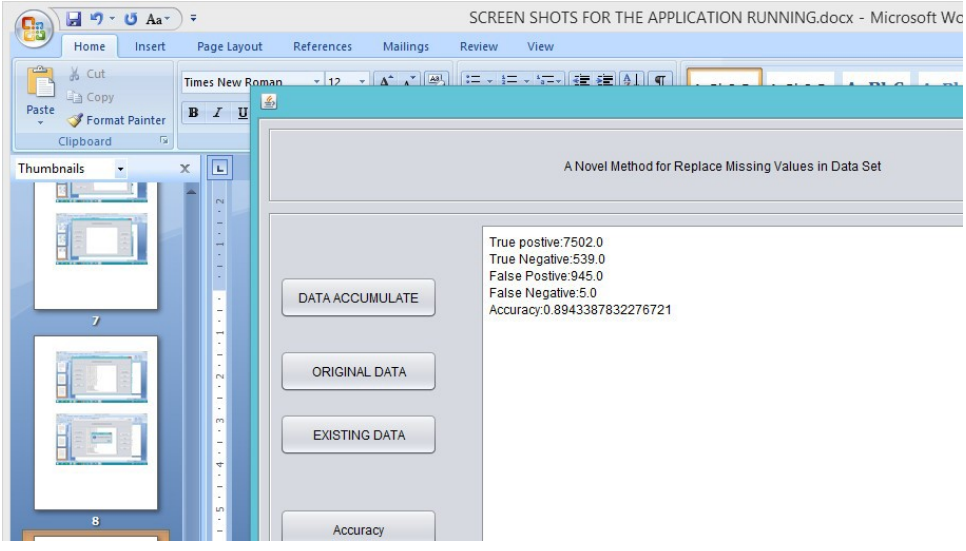
Snapshot 5 Applying K-means (see online version for colours)



Snapshot 6 Evaluating clusters result after applying K-means (see online version for colours)



Snapshot 7 Accuracy checked for improved proposed k-means (see online version for colours)



Snapshot 8 Accuracy for the dataset air quality (see online version for colours)