International Journal of Engineering Research and Modern Education (IJERME)
Impact Factor: 7.018, ISSN (Online): 2455 - 4200
(www.rdmodernresearch.com) Volume 8, Issue 2, 2023

BIG DATA PREPROCESSING USING ENHANCED DATA QUALITY RULES DISCOVERY MODEL (EDQRM)

K. Dharani* & Dr. G. Abel Thangaraja**
* Assistant Professor, Department of Computer Science, Sri Nehru Maha Vidyalaya College of Arts and Science, Coimbatore, Tamilnadu
** Assistant Professor, Department of Computer Science, Sri Krishna Adithya College of Arts and Science, Coimbatore, Tamilnadu

Cite This Article: K. Dharani & Dr. G. Abel Thangaraja, "Big Data Preprocessing Using Enhanced Data Quality Rules Discovery Model (EDQRM)", International Journal of Engineering Research and Modern Education, Volume 8, Issue 2, July - December, Page Number 33-41, 2023.

Abstract:

In the Big Data Era, data is the center for any governmental, institutional, and private organization. Endeavors were equipped towards extricating profoundly important bits of knowledge that can't occur assuming data is of low quality. Hence, data quality (DQ) is considered as a vital component in big data processing. In this stage, bad quality data isn't entered to the Big Data value chain. This paper, proposed the Enhanced data quality Rules discovery model (EDQRM) for assessment of quality and Big Data pre-processing. EDQRM discovery model to improve and precisely focus on the pre-processing exercises in view of quality requirements. Characterized, a bunch of pre-processing exercises related with data quality dimensions (DQD's) to automatize the EDQRM process. Rules improvement is applied on approved rules to stay away from multi-passes pre-processing exercises and disposes of copy rules. Directed tests showed an expanded quality scores in the wake of applying the found and optimized EDQRM's on data.

Key Words: Big Data; Data Quality Evaluation; Data Quality Rules Discovery; Big Data Pre-Processing

1. Introduction:

These days, a large portion of organizations consider data as a resource in a time where practically all business vital choices depend on insights collected from the data. Initially, data is incomplete and could contain a great deal of discrepancies, and irregularities like poor, absent and incomplete data. These data anomalies are brought about by many variables including specialized and human component. Big Data goes through all periods of its lifecycle; such stages incorporate data processing, investigation, and visualization. Nonetheless, without clean, reliable, and complete data these stages won't prevail. However, any data processing remains exceptionally delicate when data isn't reasonable and fit to be processed. This outcome in an unusable data and analysis brought about by variables like terrible data preparation, nature of data, its arrangement, its starting point, and its sort. A ton of data is collected during the business lifecycle making data that could be caught as unstructured, lacking quality of wanted boundaries.

Figure 1: Data Preprocessing Steps

These organizations could incorporate the power, energy, social, retail, web based business, thus numerous other existing or impending modern areas. Every area has data qualities are intended for its domain or nature of business. Thus, the data must be helpful for its clients to be assessed. A domain's Data Quality should

be suitably melded to its particular domain making it a seriously perplexing science to address. Thus, this paper centers around working on Quality of the Data in beginning phases before it is consumed for additional processing or analysis. It expects to determine quality issues for huge data sets involving processes as manual endeavors generally fall flat for enormous volumes of data. Raw, real-world data as text, images, video, and so on, is muddled. Not exclusively may it contain blunders and irregularities; however it is frequently incomplete, and doesn't have a customary, uniform plan. Although data preprocessing is an incredible asset that can empower the client to treat and handle complex data, it might consume a lot of processing time. It incorporates many disciplines, as data preparation and data reduction procedures as should be visible in figure 1.

Data preprocessing isn't simply restricted to traditional data mining undertakings, like arrangement or relapse. An ever increasing number of specialists in original data mining fields are giving increasing consideration to data preprocessing as a device to work on their models. This more extensive reception of data preprocessing methods is resulting in variations of known models for related frameworks, or totally clever recommendations.

Data Quality Dimensions:

Data Quality dimensions are a way to survey the quality of data. These might be Intrinsic or/and Contextual. Despite the fact that there is no norm or widespread administrative definition on these quality dimensions, the endeavor here is to settle on the generally acknowledged ones comprehensively. As such the Contextual might change per business domain, application or relevance. The most well known intrinsic dimensions incorporate Accuracy, Consistency, Uniqueness, Timeliness, Validity lastly and Completeness.

Data Profiling:

Profiling data spins around laying out a rule's framework to guarantee effective evaluation of the quality of the data upheld by a particular definition and attributes on the quality of the data.

Big Data Quality:

Data Quality for the as yet developing and developing field of Big Data is in itself a profoundly complicated subject. Huge organizations prior genuinely thought having caught data from different business processes, multiple divisions, sales, profits, geographies, and locations boundaries, and so forth would empower them to decisively amplify their business and spread in additional areas.

2. Literature Survey:

- **Hadoop Pre-Processing:** Dai, H., Zhang (2016) et.al proposed Research and implementation of big data pre-processing system based on Hadoop. Common big data processing platforms adopt the MapReduce programming model to perform application processing. For instance, the organization and estimation technique for Hadoop are as per the following: Hadoop first gathers data and stores them in distributed storage frameworks, which are storage nodes in clusters. Then, at that point, the process nodes read data from the storage nodes and perform map operations. Finally, the process nodes speak with one another and acquire computation results by performing decrease operations.
- **Outlier Detection Algorithm:** Wang, Z., Huang, X., Song, Y., & Xiao, J. (2017) proposed an outlier detection algorithm based on the degree of sharpness and its applications on traffic big data pre-processing. This paper is given to another outlier detection algorithm in view of the degree of sharpness. The proposed algorithm takes a better approach to tackle the outlier detection issue, which utilizes an action in image processing, degree of sharpness, to recognize the outliers. Contrasted with the traditional outlier detection techniques with factual learning, the proposed algorithm has no iterative processes. It creates a smooth curve to depict the general dispersion of the data first and foremost, and afterward processes the sharpness of degree for every data point. Finally, the outliers are perceived as bigger values of the degree of sharpness.
- **Distributed Data Pre-Processing:** Celik, O., Hasanbasoglu, M., Aktas, M. S., Kalipsiz, O., & Kanli, A. N. (2019) proposed Implementation of Data Pre-processing Techniques on Distributed Big Data Platforms. This paper point was to give distributed implementation of certain algorithms for two of the data pre-processing steps: outlier analysis and missing value imputation. The algorithms were executed on Flash and this paper will zero in on the subtleties and execution of these algorithms on various distributed framework arrangements. Flash to analyze this relationship. In any case, there are other big data processing platforms and these platforms can measure up against one another in various perspectives, for example, convenience and installation, execution or scalability.
- **Quality Framework:** Lincy, S. S. B. T., & Kumar, N. S. (2017) proposed an enhanced pre-processing model for big data processing: A quality framework. Processing and examining the heterogeneous, gigantic measure of data to determine valuable insights out of it. With the advancement of big data numerous innovations are being created. The contribution to it should be processed so that the quality data yields quality successful results. A viable pre-processing model is proposed in this paper for the processing of the big data. Utilizing help algorithm and quick mRMR all together methodology can be utilized for the pre-processing of the data. Analysis shows that this hybrid methodology is more viable

and can significantly improve the quality of the data. This approach can yield better execution upon the big data platform utilizing the Spark framework.

- **Semantic Data Pre-Processing:** Yerashenia, N., Bolotov, A., Chan, D., & Pierantoni, G. (2020) proposed Semantic Data Pre-Processing for Machine Learning Based Bankruptcy Prediction Computational Model. The main goal of the research is to develop data pre-processing techniques where Ontologies plays a central role. The Ontology of Bankruptcy Prediction (OBP Ontology) which provides a conceptual framework for a company's financial analysis is built in the widely established Proteg´ e´ environment. An OBP Ontology can be effectively described with a Graph database (DB). A Graph DB expands the capabilities of traditional databases by tackling the interconnected nature of economic data and providing graph-based structures to store information, allowing the effective selection of the most relevant input features for the machine learning algorithm.

3. Proposed Methodology:

This paper proposed a Big Data quality rules generation and discovery model from the quality evaluation results. The quality is assessed before any pre-processing task. This phase gives a very much developed data quality rules to be utilized in the pre-processing. The rule set can be refined by a client master and applied to work on quality aspect.

Data Pre-Processing:

Most techniques in data mining depend on a data set that is evidently finished or noise free. Nonetheless, real-world data is a long way from being clean or complete. In data pre-processing it is normal to utilize techniques to either eliminating the noisy data or to impute (fill in) the missing data. The accompanying two segments are given two missing values imputation and noise sifting.

- **Missing Values Imputation:** One big suspicion made by data mining techniques is that the data set is finished. The presence of missing values is, nonetheless, exceptionally normal in the acquisition processes. A missing value is a datum that has not been put away or accumulated because of a faulty examining process, cost restrictions or constraints in the acquisition process. Improperly handling the missing values will effortlessly prompt unfortunate knowledge removed and furthermore off-base ends. Missing values have been accounted for to cause loss of effectiveness in the knowledge extraction process, solid predispositions assuming the missingness introduction system is misused and serious complexities in data handling. Many methodologies are accessible to handle the risky forced by the missing values in data pre-processing. The main choice is typically to dispose of those instances that might contain a missing value. By utilizing most extreme probability strategies, test the estimated probabilistic models to fill the missing values. Since the genuine likelihood model for specific data sets is typically obscure, the use of AI techniques has become extremely famous these days as can be applied keeping away from without giving any prior information.

Big Data Pre-Processing:

This section aims itemizing an intensive rundown of commitments on Big Data pre-processing. Classifies these commitments as per the classification of data pre-processing, number of features, number of instances, most extreme data size oversaw by every algorithm and the model under have been developed. The size has been registered increasing the absolute number features by the number of instances (8 bytes for every datum). For scanty strategies, just the non-meager cells have been thought of. When seen a depiction of the ongoing improvements in Big Data pre-processing will give shorts portrayals of the commitments in the remainder of this part. First portray one of the most well known AI libraries for Big Data: MLlib; which brings many data pre-processing techniques to the Spark people group. Next the remainder of areas will be dedicated to specify those commitments presented sorted and organized. Pre-Processing of data or processing of data's quality at the beginning phase of any Big Data system's lifecycle improve and refine the quality of the data. Data Pre-Processing lifecycle normally cover these sub-processes:

- **Data Consolidation and Integration:** Data might be obtained from multiple locations that be might be in different forms, structured/semi-structured/unstructured, shifting formats, junk, and so forth. Data from this large number of sources should be combined homogeneously to frame a solitary and last wellspring of truth for the data to be utilized in the Big Data system. Advances like ETL Extract, Transform and Load are well known and laid out systems.
- **Data Enhancements and Enrichment:** Data from different sources is united, and during that data details are refreshed utilizing extra information acquired from other supportive sources to make intertwined data that is improved with more information and perhaps at the same time enhanced qualitatively.
- **Data Transformation:** Data transformation includes numerous a stages or maybe sub-processed like catching or pulling data from multiple sources, data ought to be reformatted, standardized, collected, even refreshed utilizing regulatory standards.
- **Data Reduction:** Data reduction is the process of diminishing how much data with the goal that it becomes non-repetitive. This aides in expanding the data storage effectiveness and diminishing

expenses by removing data that isn't significant and holding just the significant parts for that specific work/task.

- **Data Discretization:** This process extracts and isolates data into stretches so it tends to be proficiently used inside accessible mining algorithm and techniques.
- **Data Cleaning:** It is a process which works on the Quality of Data by removing the data which decreases the ease of use of it. The steps engaged with this process are removing the incorrect, incomplete or immaterial data from the data gained so it tends to be processed and broke down upon to extract advantageous value from it.

Enhanced Data Quality Rules Discovery (EDQRM):

The purpose of an Enhanced Data Quality Rule (EDQRM) model is to discover, optimize and generate a set of data quality rules taking into account many parameters as shown in below figure 2:

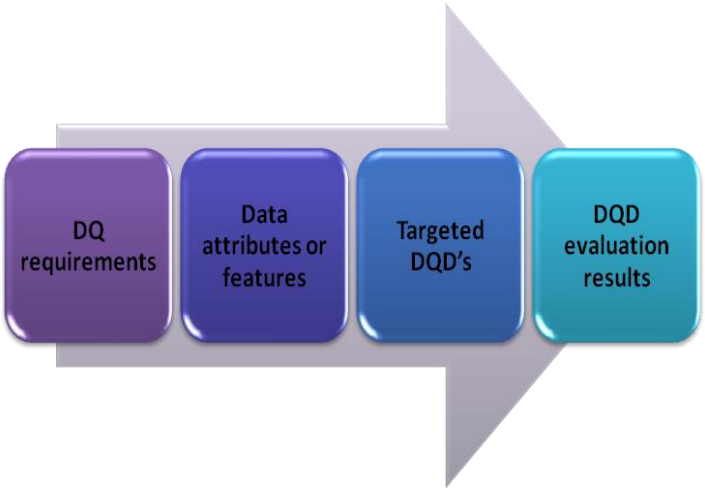

Figure 2: Data Quality Rules Parameters

In this work, are dealing with data quality before the pre-processing phase. These DQR's are essential to address and further develop the data quality while setting the best pre-processing activities.

Figure 3: Workflow of the Proposed Model

The EDQRM Discovery Quality Rules Model is outlined in Figure 3 where the vital components of EDQRMP comprise of: (a) Big Data testing and profiling, (b) Large data quality mapping and evaluation, (c) Huge data quality rules revelation (e) DQR validation and (f) DQR optimization. In the accompanying sections, depict every module, its input(s) and output(s), the primary capabilities, and its roles and communications with different modules.

A. Sampling and Profiling:

Since profiling is an action to find data attributes from at least one data sources. It is viewed as a data assessment process that gives an initial feeling of the data quality detailed in its data profile. This work utilized the (Bag of Little Bootstrap) BLB bootstrap for large data inspecting to effectively test huge data while not

losing precision and reducing evaluation time. Let S a bunch of data tests from the data source: $S = \{s_0, \dots, s_i, \dots, s_n\}$ and P the individual profiles.

B. Quality Mapping / Evaluation Processing:

Mapping should be finished between Data Quality Dimensions (DQD) and the designated data features/attributes. Each DQD is estimated for each characteristic and for each sample. This model is given as quality requirements to comprehend the scores expressed as the quality degree of acknowledgment. The quality requirements can be a bunch of values, a stretch in which values are acknowledged or dismissed, or a single score ratio. Let's note by A, a set of data attributes, D a set of data quality dimensions, and R a set of Quality requirements. The Data Quality Evaluation set DQES $(Q_0(a_j, d_k, r_t), \dots Q_x(a_a, d_b, r_c), \dots)$ each elements is a quality score for a specific attribute, DQD, and a quality requirement. DQES is applied on a set of samples S, which result in a Quality scores represented by QScore containing the DQD quality scores for each attribute.

C. Quality Rules Discovery:

- **Quality Scores Results Analysis:** Each DQD evaluation $Q_x(a_j, d_k, r_t)$, in Data Quality Evaluation Scheme (DQES) generates a quality score $Q_x Score(a_j, d_k, r_t)$. These scores are tested against quality requirements. The quality rules are generated, and attributes fully oppose these rules might be damaged.
- **Pre-Processing Activities Repository and Quality Rules Generation:** The Pre-Processing activities repository storehouse is coordinated as a tuple $Pre - Processing\ activities\ repository(d_k, af_{k,v})$. Every data quality aspect d_k is associated with an initiation capability. For instance, the DQD culmination of a bunch of attributes is assessed to give the ratio of complete data observation inside a bunch of chosen attributes or features of the data. These large numbers of potential outcomes are expressed in the requirements set R in the Quality planning stage. $Q_x Score(a_j, d_k, r_t)$ rule is developed based on the pre-processing activity repository for the failed evaluation score. Each rule is represented by a tuple: the pre-processing activity $Q_x R$ ($Q_x(a_j, d_k, r_t)$ is selected using the Quality mapping element from DQES and the Pre-Processing Activities Repository.
- **Quality Rules Validation:** To validate the found rules a pre-processing is applied on a bunch of Sample S and a revaluation of quality in view of a similar evaluation scheme,

QEP (DQES, S', QScore')

Where QEP means Quality Evaluation Processing, S' means set of samples. An immediate examination of resulting scores from both evaluation results QScore and QScore' is directed to filter the arrangement of valid rules (DQRLV) from the original set (DQRLN). There are two kinds of ineffective rules: rules that didn't further develop the quality score when applied to data, and rules that decline the original quality scores.

Algorithm 1: Enhanced Quality Rules Discovery Algorithm

Step 1: Input: (S, A, D, R)
Step 2: $A = \{a_0, \dots, a_j, \dots, a_p\}, R = \{r_0, \dots, r_l, \dots, r_t\}$
Step 3: $D = \{d_0, \dots, d_k, \dots, d_q,\}, S = \{s_0, \dots, s_i, \dots, s_n\}$ *Quality Mapping Selection*
Step 4: Output: $DQES(Q_0(a_j, d_k, r_t), \dots Q_x)$
Step 5: Quality Evaluation Processing: QEP (DQES, S, and QScore)
Step 6: For each Mapped tuple $Q_x(a_j, d_k, r_t)$ *in DQES*
Step 7: For each s_i *in S*
Step 8: $QualityEval\ (Q_x(a_j, d_k, r_t), s_i) \rightarrow Q_x Score(a_j)$ *Evaluation Process*
Step 9: End s_i
Step 10: End Q_x
Step 11: Quality Rules Discovery
Step12:Input:QScores,DQES,$PreProcessingActivity\ (d_k, af_{k,v})af_{k,v}$: $Activity\ function\ for$
Step 13: Output: DQ Rules List: $DQRL\left(Q_x R\left(a_j, d_k, PPA(d_k, af_{k,v})\right)\right)$
Step 14: For each Score tuple $Q_x Score(a_j, d_k, r_t)$
Step 15: Analyze $Q_x Score(a_j, d_k, r_t), PPA(d_k, af_{k,v})$ *Rules Discovery Process*
Step 16: GenerateQRules $() --> Q_x R(Q_x(a_j, d_k, r_t), PA(d_k,$
Step 17: End $Q_x Score$
Step 18: Samples Pre-Processing: Input $(S, DQRL)$ *Output* (S')
Step 19: For each Rule $Q_x R(Q_x(a_j, d_k, r_t), PA(d_k, af_{k,v}))$ *in DQRL*
Step 20: For each s_i *in S*
Step 21: For each a_j , dk
Step 22: Pre-Processing $(Q_x R(a_j, d_k, af_{k,v}), s_i)$ *Data pre-processing*
Step 23: End a_j , d_k

Step 24: End s_i
Step 25: Output: s'_i *Pre-Processed samples*
Step 26: End Q_{xR}
Step 27: Quality Evaluation Processing: $QEP\ (Q_x\ , S', QScore')$
Step 28: Quality Scores Validation: Input (QScores, QScores', and DQRL)
Step 29: Output: DRQLV, DRQLN (V: valid N: not valid Quality rules)
Step 30: For each $Q_x Score(a_j, d_k, r_t\ )$
Step 31: For each s_i *in S*
Step 32: if (ValidScore $(Q_x Score(a_j, d_k, r_t\ ), Q_x\ Score'(\dots)))$
Step 33: Add $Q_x R(a_j, d_k, af_{k,v})$ *to DRQLV* *Rules Validation*
Step 34: else Add $Q_x R(a_j, d_k, af_{k,v}))$ *to DRQLN*
Step 35: End s_i
Step 36: End $Q_x Score$
Step 37: Data Quality rules Optimization (DQRLV)

- **Quality Rules Optimization:** In the final stage, the DRQLV rules are optimized under several situations which may depend on the selected features/attributes. In the following are some optimization schemes that can be applied on the set of rules.
  - The rules are grouped per attributes, dimensions, or pre-processing activity to detect duplicates processing activities repositories.
  - Remove duplicated rules per attributes by grouping all the activities or same activity for multiple attributes.
  - Combining rules targeting same attribute (s) or a set of then ordering the activities per execution priority.
  - Prioritizing the activity function that replaces data attributes in the case of missing values and also for fulfilling data completeness quality dimension.
  - Combining same (DQD, Processing Activities Repository) tuple for many attributes/features in one rule to avoid multi- passes pre-processing.

4. Experiment Results:

Precision:

Dataset	ODA	EPM	Proposed EDQRM
50	66.94	74.91	87.01
100	69.66	71.77	90.87
150	74.12	67.93	92.48
200	79.09	68.05	95.23
250	86.38	65.39	97.52

Table 1: Comparison tale of Precision

The Comparison table 1 of Precision Values explains the different values of existing ODA, EPM and proposed EDQRM algorithm. While comparing the Existing algorithm and proposed EDQRM algorithm, provides the better results. The existing algorithm values start from 66.94 to 86.38, 65.39 to 74.91 and proposed EDQRM values starts from 87.01 to 97.52. The proposed method provides the great results.

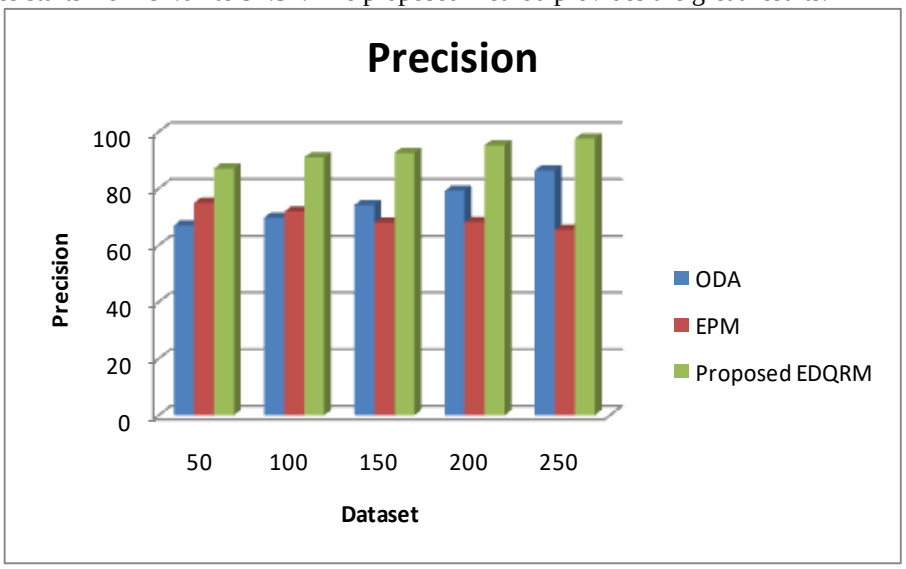

Figure 4: Comparison chart of Precision

The figure 4 Shows the comparison chart of Precision demonstrates the existing ODA, EPM and proposed EDQRM. X axis denote the Dataset and y axis denotes the Precision ratio. The proposed EDQRM values are better than the existing algorithm. The existing algorithm values start from 66.94 to 86.38, 65.39 to 74.91 and proposed EDQRM values starts from 87.01 to 97.52. The proposed method provides the great results.

Recall:

Dataset	ODA	EPM	Proposed EDQRM
20	0.625	0.721	0.836
40	0.663	0.654	0.874
60	0.706	0.598	0.905
80	0.728	0.623	0.941
100	0.752	0.591	0.962

Table 2: Comparison Tale of Recall

The Comparison table 2 of Recall Values explains the different values of existing ODA, EPM and proposed EDQRM algorithm. While comparing the Existing algorithm and proposed EDQRM algorithm, provides the better results. The existing algorithm values start from 0.625 to 0.752, 0.591 to 0.721 and proposed EDQRM values starts from 0.836 to 0.962. The proposed method provides the great results.

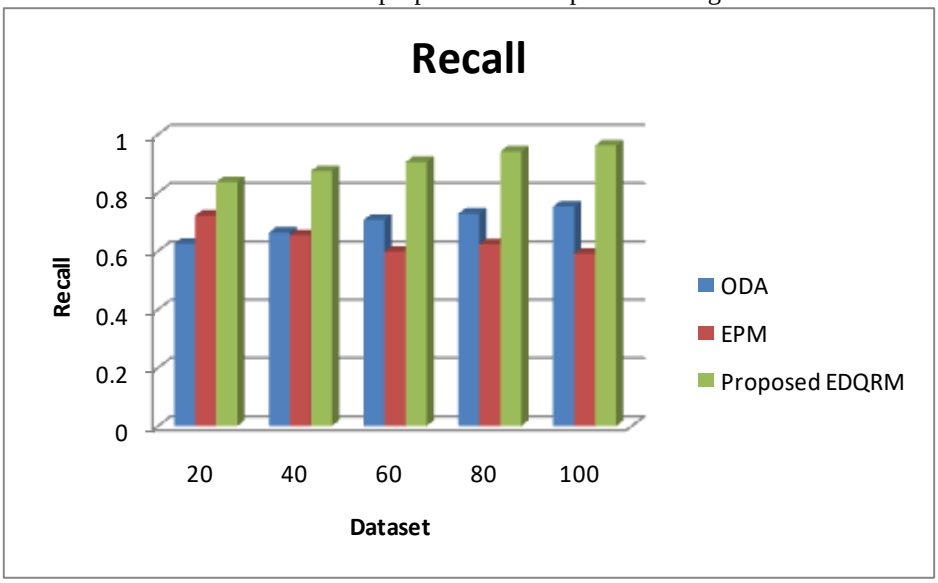

Figure 5: Comparison chart of Recall

The Figure 5 Shows the comparison chart of Recall demonstrates the existing ODA, EPM and proposed EDQRM algorithm. X axis denote the Dataset and y axis denotes the Recall ratio. The proposed EDQRM values are better than the existing algorithm. The existing algorithm values start from 0.625 to 0.752, 0.591 to 0.721 and proposed EDQRM values starts from 0.836 to 0.962. The proposed method provides the great results.

F - Measure:

Dataset	ODA	EPM	Proposed EDQRM
100	0.89	0.72	0.98
200	0.85	0.70	0.96
300	0.86	0.67	0.95
400	0.84	0.64	0.93
500	0.82	0.61	0.92

Table 3: Comparison tale of F - Measure

The Comparison table 3 of F -Measure Values explains the different values of existing ODA, EPM and proposed EDQRM algorithm. While comparing the Existing algorithm and proposed EDQRM, provides the better results. The existing algorithm values start from 0.82 to 0.89, 0.61 to 0.72 and proposed EDQRM values starts from 0.92to 0.98. The proposed method provides the great results.

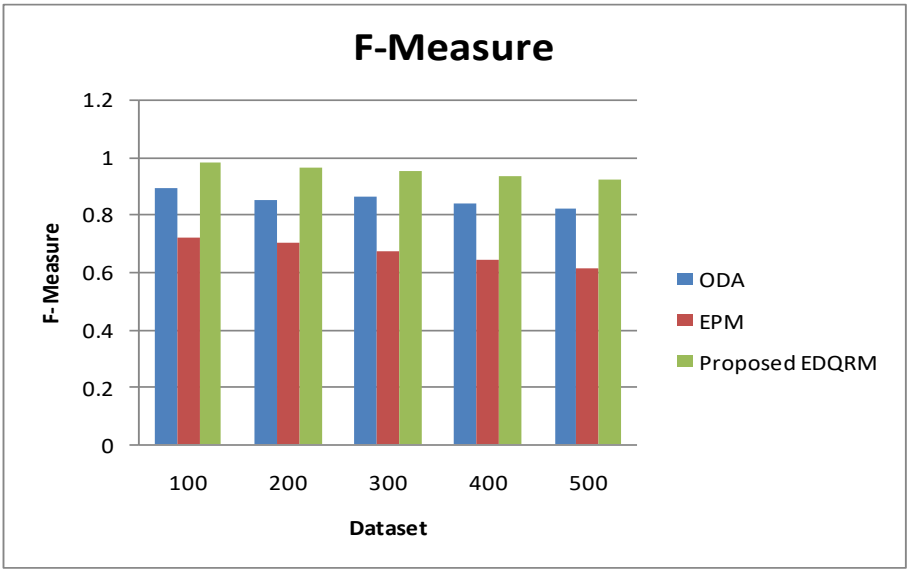

Figure 6: Comparison chart of F -Measure

The Figure 6 Shows the comparison chart of F -Measure demonstrates the existing ODA, EPM and proposed EDQRM. X axis denote the Dataset and y axis denotes the F -Measure ratio. The proposed EDQRM values are better than the existing algorithm. The existing algorithm values start from 0.82 to 0.89, 0.61 to 0.72 and proposed EDQRM values starts from 0.92to 0.98. The proposed method provides the great results.

Throughput:

Dataset	ODA	EPM	Proposed EDQRM
100	0.65	0.36	0.89
200	1.02	0.59	1.18
300	1.45	0.76	1.69
400	1.78	0.98	1.91
500	2.01	1.05	2.35

Table 4: Comparison tale of F -Measure

The Comparison table 4 of Throughput Values explains the different values of existing ODA, EPM and proposed EDQRM algorithm. While comparing the Existing algorithm and proposed EDQRM algorithm, provides the better results. The existing algorithm values start from 0.65 to 2.01, 0.36 to 1.05 and proposed EDQRM values starts from 0.89 to 2.35. The proposed method provides the great results.

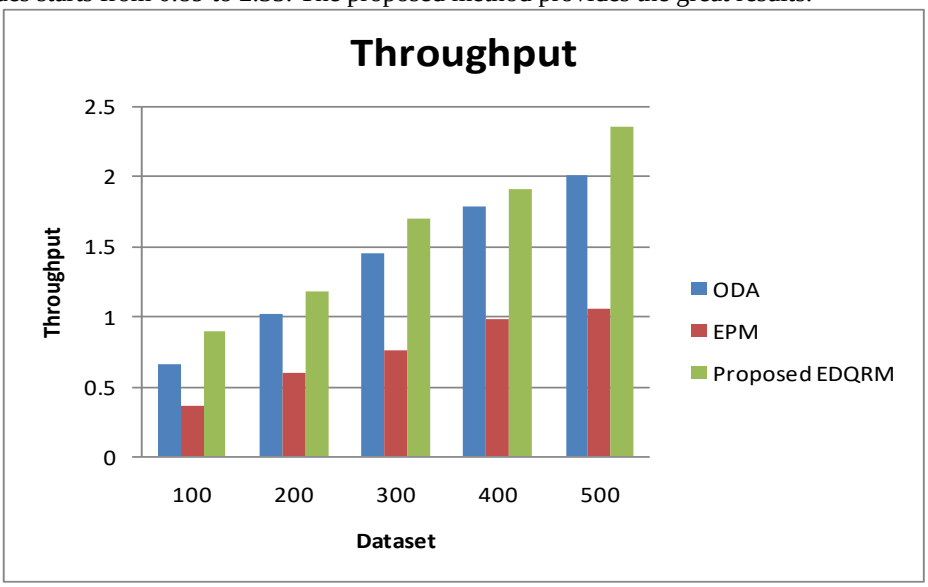

Figure 7: Comparison chart of Throughput

The Figure 7 Shows the comparison chart of Throughput demonstrates the existing ODA, EPM and proposed EDQRM algorithm. X axis denote the Dataset and y axis denotes the Throughput. The proposed EDQRM values are better than the existing algorithm. The existing algorithm values start from 0.65 to 2.01, 0.36 to 1.05 and proposed EDQRM values starts from 0.89 to 2.35. The proposed method provides the great results.

5. Conclusion:

This paper proposed an Enhanced quality-based rule disclosure model to help Huge Data pre-processing. The model depends on separating quality standards from huge data quality evaluation while thinking about a bunch of quality requirements. The examination applied produced rules on huge data tests and afterward re-examined the quality to approve these standards. The worth added element of the EDQRM model is the course of quality rule optimization and the planning between the pre-processing activities and the designated DQD. The trials led on a bunch of enormous data tests demonstrate that quality standards are found, approved, and afterward enhanced to fundamentally further develop the quality in the huge data pre-processing.

6. References:

1. Dai, H., Zhang, S., Wang, L., & Ding, Y. (2016). Research and implementation of big data preprocessing system based on Hadoop. 2016 IEEE International Conference on Big Data Analysis (ICBDA).
2. Wang, Z., Huang, X., Song, Y., & Xiao, J. (2017). An outlier detection algorithm based on the degree of sharpness and its applications on traffic big data preprocessing. 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)
3. Celik, O., Hasanbasoglu, M., Aktas, M. S., Kalipsiz, O., & Kanli, A. N. (2019). Implementation of Data Preprocessing Techniques on Distributed Big Data Platforms. 2019 4th International Conference on Computer Science and Engineering (UBMK).
4. Lincy, S. S. B. T., & Kumar, N. S. (2017). An enhanced pre-processing model for big data processing: A quality framework. 2017 International Conference on Innovations in Green Energy and Healthcare Technologies (IGEHT).
5. P. Glowalla, P. Balazy, D. Basten, and A. Sunyaev, “Process-Driven Data Quality Management - An Application of the Combined Conceptual Life Cycle Model,” in 2014 47th Hawaii International Conference on System Sciences (HICSS), 2014, pp. 4700-4709.
6. F. Sidi, P. H. Shariat Panahy, L. S. Affendey, M. A. Jabar, H. Ibrahim, and A. Mustapha, “Data quality: A survey of data quality dimensions,” in 2012 International Conference on Information Retrieval Knowledge Management (CAMP), 2012, pp. 300-304.
7. Y. W. Lee, “Crafting rules: context-reflective data quality problem solving,” J. Manag. Inf. Syst., vol. 20, no. 3, pp. 93-119, 2003.
8. P. Z. Yeh and C. A. Puri, “An Efficient and Robust Approach for Discovering Data Quality Rules,” 22nd IEEE International Conference on Tools with Artificial Intelligence (ICTAI), 2010, vol. 1, pp. 248-255.
9. F. Chiang and R. J. Miller, “Discovering data quality rules,” Proc. VLDB Endow., vol. 1, no. 1, pp. 1166- 1177, 2008.
10. W. Fan, “Dependencies revisited for improving data quality,” in Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART, 2008, pp. 159-170.