A Multi-purpose Data Pre-processing Framework using Machine Learning for Enterprise Data Models

Venkata Ramana B[1], Dr.Narsimha G[2]
Research Scholar, Department of Computer Science and Engineering, JNTU, Hyderabad[1]
Professor, Department of Computer Science and Engineering, JNTU, Hyderabad[2]

***Abstract*—Growth in the data processing industry has automated decision making for various domains such as engineering, education and also many fields of research. The increased growth has also accelerated higher dependencies on the data driven business decisions on enterprise scale data models. The accuracy of such decisions solely depends on correctness of the data. In the recent past, a good number of data cleaning methods are projected by various research attempts. Nonetheless, most of these outcomes are criticized for higher generalness or higher specificness. Thus, the demand for multi-purpose, however domain specific, framework for enterprise scale data pre-processing is in demand in the recent time. Hence, this work proposes a novel framework for data cleaning method as missing value identification using the standard domain length with significantly reduced time complexity, domain specific outlier identification using customizable rule engine, detailed generic outlier reduction using double differential clustering and finally dimensionality reduction using the change percentage dependency mapping. The outcome from this framework is significantly impressive as the outliers and missing treatment showcases nearly 99% accuracy over benchmarked dataset.**

Keywords*—Standard domain length; domain specific rule engine; double differential clustering; change percentage; dependency map

I. INTRODUCTION

Many enterprises use (probably) use a business data architecture that's an aggregation model, covering all of their details. Most business data models can be conceptual as well as physical. In certain instances, it is self-evident when to create a blueprint. Formal models (often, enterprise data models) seem to be different, And where what was requested has not been completed, or put to use, business use, enterprise data models have been abandoned or remain unfinished. The root cause of these errors is typically is found in a fundamental mathematical error. Initially, it was not obvious what issues the data model wanted to address, and it was not yet clear what was behind these responses. Setting the questions to be asked and the business data model's intent allows things obvious when finished data modeling. There is the option to build business data models unnecessarily, and this causes both cost and time to increase. When problems emerge that need more explanation, go back to the business data model. the use of an enterprise data model is particularly appropriate in the following two cases. The enterprise procedures are being changed due to an extensive reengineering program. Developing an organizational data model in tandem with an enterprise method delivers tremendous benefit to the process reengineering process. The second implementation in business design is derived from a bottom-up method Integration necessitates the use of a logical data model to display the overlaps between different structures.

Pre-processing of the dataset is one of the primary tasks in any data analytics or data dependent researches or projects. The primary component of the pre-processing ensures removal and replacement of the outliers, removal and replacement of the missing values and sometimes the attribute reductions. Also, in some non-trivial situation removal of the critical and sensitive information is also part of the pre-processing method. The work by H. F. Ladd et al. [1] has clearly suggested many case studies where information hiding is highly important without missing any other crucial information. Nonetheless, the generic datasets, unless related to the personalized recommendation systems, come without the personal identification information sets. Thus, the primary task for any data analyst or a strongly data dependent machine learning engineer are to identify and remove or replace the outliers or missing values [14].

The reduction of the outliers and missing values improves the accuracy as proven by many research attempts such as the work by T. Calders et al. [2]. Nonetheless, many of the parallel research works also have suggested that, removing or replacing the outliers or the missing values directly from the dataset without much customization can directly lead to loss of data and result into incorrect classification or clustering. Thus, it is highly important to generate the data pre-processing method suitable to domain from which the data is originally generated. This belief was initially projected by D. Pedreschi et al. [3] in the year 2008. Through many researchers such as S. Hajian et al. [4] have always emphasised on the data security.

Realizing the need for the domain specific data pre-processing and the need for enterprise scale data pre-processing for domain specific outliers and missing value imputation methods, this work formulates a novel multi-purpose framework for data pre-processing [15].

The rest of the paper is formulated such as, in the Section II, the parallel research outcomes are critically analysed; in Section III the used dataset for this research is described; in Section IV the proposed solutions are formulated using the mathematical models; in Section V, the proposed algorithms based on the mathematical models are discussed; in the Section VI the complete framework is elaborated; in the

Section VII the obtained results are discussed; in Section VIII the comparative analysis is furnished; and finally in Section IX, the research conclusion is formulated.

II. Parallel Research Outcomes

The final outcome of any analytical project is to generate the final results in terms of predictions or projections or classifications or clustering. Nonetheless, all these outcomes solely depend on the cleanness of the data. The cleanness of the data primarily refers to the reduction of the missing values, outliers and sometimes the noises present in the spatial datasets. Hence, a good number of research attempts can be seen in order to propose a framework, which is specific in nature to reduce the anomalies from the datasets.

The work by B. Fish et al. [5] has proposed a method to reduce the anomalies from the datasets using the confidence factors and the confidence metric. This method identifies the outliers and missing values from each domain of the dataset and in case any attribute domain has more than half of the values as anomalies, then the confidence matrix decide, whether that specific attribute contributes to the final classification of the data. In case, that attribute showcases less dependencies, then that specific attribute can be completely discarded from the dataset. Regardless to mention, this method is criticized for lesser accuracy due to the information loss, in spite of the better time complexity.

Yet another research attempt by M. B. Zafar et al. [6] have tried showcasing the effect of anomalies in the final prediction from the dataset and up to certain extend, the effects can be ignored, and the pre-processing stages can be completely ignored. Nonetheless, this work is also highly criticised as this method does not suggest any specific boundaries for domain specific dataset treatments.

In the other direction, the work by T. Kamishima et al. [7] have showcased that the missing value imputation can be completely automated using various machine learning methods and the accuracy of this method is also remarkable. Nevertheless, this work does not recommend any specific method to handle the domain specific anomalies as explained in Section IV of this literature. In the same direction, the work by M. Hardt et al. [8] has justified the process of weighted parameters for reduction of anomalies using equality principle. However, during a domain specific pre-processing task, it is nearly impossible to identify the weights as equal in the dataset. Thus, this work also cannot justify the need addressed in this literature.

Yet another approach by M. Feldman et al. [9] recommends that, during a pre-processing task, the knowledge from the previous attempts can be utilized to reduce the time complexity. Using the recommendations from anomaly reduction process from the similar datasets can be utilized on the newer datasets and time complexity can be significantly reduced. Regardless to mention, generating the similarity characteristics from two different datasets are a challenge in itself and the added time complexity shall also be considered. This thought is confirmed by the work of C. Dwork et al. [10].

The two recent research outcomes by Z. Zhang et al. [11] and by J. Kleinberg et al. [12] have recommended using the backtracking methods, which is also adopted in this literature and extended in the Section V.

Further, with the detailed understanding of the parallel research attempts, in the next section of this work, the considered dataset for this research is analysed.

III. Dataset Description

Master and reference data is necessary to ensure continuity across implementations, but it must also be considered scoped to prevent data processing consistency. Since most of the transaction data is almost invariably moved to data centers and monitoring structures, this is predicted to include most organizations' data.

TABLE I. Dataset Description

Attribute Serial #	Dataset Attribute Name	Attribute Alias	Attribute Description	Value Range
1	Employee ID	ID	Unique identification of the employee	Randomized due to identify hiding
2	Job Class	JC	Job category	Retired, Developer, Tester, Student, Etc.
3	Age	AGE	Age of the employee	18 to 70 Years
4	Experience	EXP	Number of years of experiences	0 to 40 years
5	Present Skill Sets	SKILLS_NOW	List of current skill sets	-
6	Upgradation Skill Sets	SKILLS_UP	List of skill sets, which the employee wants to learn	-
7	Job Satisfaction	JS	The level of job satisfaction	0 (Lowest) to 5 (Highest)
8	Job Change Willing ness	JCHA	The desire to change the current job	0 (Lowest) to 5 (Highest)
9	Project ID	PID	Project ID	Randomized due to identify hiding
10	Project Duration	DUR	Duration of the project	In Months
11	Customer Impression	CI	Feedback from the customer	0 (Lowest) to 5 (Highest)
12	Manager Impression	MI	Feedback from the project manager	0 (Lowest) to 5 (Highest)
13	Team Impression	TI	Feedback from the team members (Mean Value)	0 (Lowest) to 5 (Highest)
14	Project Completion Status	CS	Project completion percentage	0 to 100%

In order to carry forward, the research proposed in this work, the 'The Public 2020 Stack Overflow Developer Survey Results' [13] is utilized. The description of this dataset is furnished here [Table I].

Further, based on this domain specific dataset, the formulation of the problems is carried out in the next section of this work.

IV. Proposed Solutions: Mathematical Models

After the critical analysis of the parallel research works and identification of the research problems in the previous section of this work, in this section of the work, the proposed solutions are presented using mathematical models.

This section primarily focuses on four different pre-processing methods as identification of the missing values, conditional outliers, generic outliers and finally reduction of the attributes.

Lemma 1: The detection of the missing values, using the proposed domain count iterative method, reduces the time complexity.

Proof: The domain count of any dataset shall be realized as the maximum number of elements without the missing or null values. Hence, the maximum count will ensure that the maximum number of elements are considered without the missing values and in case of all missing values, the complete tuple is ignored.

Assuming that, the total dataset, DS[], is a collection of multiple domains, D[], and each domain is again collection of multiple data points, D_i. Thus, for a n number of domains or attributes, the initial relation can be formulated as,

$$DS[] = \sum_{i=1}^{n} D[](i) \tag{1}$$

Also, assuming that each domain is consisting of m number of data points, thus, this relation can be formulated as,

$$D[](i) = \sum_{j=1}^{m} D_j \tag{2}$$

Further, assuming that, the method Φ, is responsible for identification of the number of data points without the missing or null values. Then, $\hat{\lambda}$ being the count of data points, this proposed function can be formulated as,

$$\hat{\lambda} = \Phi(D[](i)) \tag{3}$$

Subsequently, the count of data points from each domain can be presented as $\hat{\lambda}[]$ and can be formulated as,

$$\hat{\lambda}[] = \forall \lfloor D[](X) \rfloor \tag{4}$$

Further, assuming the maximum value from the $\hat{\lambda}[]$ collection is δ, then this can be formulated as,

$$\delta = {}_{MAX} \lceil \hat{\lambda}[] \rceil \tag{5}$$

Further for domain the count of the number of data points, Y, must be compared with the maximum data point count, X, using the divide and conquer method as following.

$$\begin{vmatrix} Iff\ \ \delta > \hat{\lambda}[i], \\ Then, Compare\ \delta/2 > \prod_{j=1}^{i/2}(\hat{\lambda}[i]) \end{vmatrix} \\ Else, Compare\ \delta/2 > \prod_{j=i/2+1}^{i}(\hat{\lambda}[i]) \tag{6}$$

Henceforth, if the count of data points is less than the expected count of the data points in first or second half of the domain, then the process must be repeated to identify the missing values only in that half of the domain and the process shall be repeated iteratively to identify all missing values.

Further, the time complexity of this proposed method is analysed against the generic method.

Assuming that, a total of k number of iterations has to be performed for n number of domains, thus the time complexity, T_1, can be formulated as,

$$T_1 = 1 + n/2 + n/4 + \ldots n/k \tag{7}$$

This can be re-written as,

$$T_1 = O(k \log_2 n) \tag{8}$$

In the other hand, for the similar identification, using the generic methods, thus the time complexity, T_2, can be formulated as,

$$T_2 = k * n \tag{9}$$

It is natural to realize that

$$T_1 << T_2 \tag{10}$$

Hence, the proposed method for outlier detection significantly reduces the time complexity with higher accuracy.

Further, the conditional outliers are addressed and resolved.

Lemma 2: The outliers within the valid range of the data, can be removed using the domain specific rule sets.

Proof: The dataset contains multiple outliers and can be residing in the valid range of data. Thus, the domain specific outliers must be addressed with the valid set domain specific rule engine.

Assuming that, the domain specific rulesets or rule engine, R[], is a collection of individual rules, R_i. Thus, a total number of n rules, this relation can be formulated as,

$$R[] = \sum_{i=1}^{n} R_i \tag{11}$$

Further, from Eq. 1, the dataset is fetched and validated against the ruleset for removal of the outliers as,

$$R[]\left|DS[]\right| \rightarrow \begin{cases} Valid, DS[] \\ Invalid, DS'[] \end{cases} \quad (12)$$

The reduced dataset, $DS'[]$ shall be identified as domain specific outlier reduced dataset.

The following table defines the initial rulesets, specific to this project [Table II]:

TABLE II. DOMAIN SPECIFIC OUTLIER DETECTION RULESETS

Rule #	Ruleset Description		
	Target Rule	**Validation Rule**	**Rule Outcome**
1	Job_Satisfaction >=4	Job_Change>=3	Outlier
2	Job_Satisfaction >=5	Job_Change<=2	Not Outlier
3	Job_Satisfaction >= 3	Job_Satisfaction = Job_Change	Outlier
4	Experience>0	SKILLS_NOW is NULL	Outlier
5	Experience>0	SKILLS_NOW is NOT NULL	Not Outlier
6	Experience<=0	SKILLS_UP is NULL	Outlier
7	Experience=0	SKILLS_UP is NOT NULL	Not Outlier
8	Completion_Status is High	Customer_Rating is low	Outlier
9	Completion_Status is High	Customer_Rating is High	Not Outlier

Further, the generic outliers are addressed and resolved.

Lemma 3: The Double Clustering method, must be utilized to identify the outliers in the dataset.

Proof: Assuming that the complete dataset is denoted as D [] and each attribute in the dataset is assumed to be presented as, A_x for total of n number of attributes. Hence, the following relation can be formed.

$$D[] \rightarrow < A_1, A_2, A_3, ...A_n > \quad (13)$$

Here, each and every attribute is considered to have their own domain with m number of records each and the data elements are denoted as D_i , which can be represented as,

$$A_x[] = \sum_{i=1}^{m} D_i \quad (14)$$

Further, the Euclidian distance between the data points can be considered as the similarity measure and the total distance set is represented as $\lambda[]$, then,

$$\lambda[] = \int_{i=1}^{n} \left|D_i - D_{i+1}\right| \quad (15)$$

Further, the Euclidian distance between the elements of $\lambda[]$ are calculated,

$$\bar{\lambda}[] = \int_{i=1}^{n-1} \left|\lambda_i - \lambda_{i+1}\right| \quad (16)$$

The new $\bar{\lambda}[]$ set defines the relation between the elements based on their similarities.

Furthermore, the repetitive iteration of the Eq. 16 can measure the similarities with deeper and contextual aspect, which can be represented as,

$$\bar{\lambda}_k[] = \int_{i=1}^{n-k} \left|\bar{\lambda}_i - \bar{\lambda}_{i+1}\right| \quad (17)$$

Thus, based on the similarity measures of Euclidian distance of the similarity measures of the elements and the Euclidian distance of the similarity measures of the Euclidian distances, the final cluster centroids can be calculated as,

$$C[] = \bar{\bar{\lambda}}_k[] = \frac{\bar{\lambda}_k[]}{\left|\lambda_i - \lambda_{i+1}\right|_{i=0}^{n}} \quad (18)$$

Further, the attribute reduction process is formulated.

Lemma 4: The domain specific dependency map can build the dimensionality reduced dataset.

Proof: Any two attributes or parameters in the existing dataset shall be compared to identify the change percentage in the complete domain. The parameters with highest amount of change percentages corresponding to the class variable shall define the reduced dataset and the parameters with less change percentage shall not be part of the final dataset.

Assuming that, the domain of the class variable, DC[], is compared with two attributes, D1[] and D2[], from the actual dataset for identification of the change percentages. Assuming, Φ is the function responsible for change detection, thus this can be formulated as,

$$\Phi(DC[] \triangleright D1[]) \rightarrow n_1 \quad (19)$$

And,

$$\Phi(DC[] \triangleright D2[]) \rightarrow n_2 \quad (20)$$

Here, n_1 and n_2 are the change percentages.

Considering, $n_1 < n_2$ and DR[] is the reduced dataset, then as per the proposed lemma, the DC2[] shall be part of the reduced dataset.

$$DR[] \leftarrow DC2[] \quad (21)$$

Similarly, the dependency map can be created such as Table III.

Here, the dependency map clearly suggests the priority of the attributes to be included in the final reduced dataset, as,

$$DR[] \leftarrow D2 > D3 > D4 > D1 > D5 > D6 : DC \quad (22)$$

TABLE III. DOMAIN DEPENDENCY MAP

	D1	D2	D3	D4	D5	D6	DC
D1	0	8	39	71	75	65	100
D2	72	0	58	71	74	81	100
D3	69	27	0	72	73	82	100
D4	72	55	18	0	74	84	100
D5	71	3	43	70	0	80	100
D6	73	24	65	71	73	0	100
DC	73	24	65	71	73	84	0

Further, accuracy must be verified with time complexity to realize the best possible reduced set.

In the results section of this work, the time complexity and accuracy are analysed for building the final reduced dataset.

Henceforth, in the next section of this work, the proposed algorithms are furnished based on the proposed mathematical models of the solutions.

V. PROPOSED SOLUTIONS: ALGORITHMS

After the detailed analysis of the problems and formulation of the proposed solutions using the mathematical models, in this section of the work, the proposed algorithms are furnished here in this section of the work.

Firstly, the iterative missing value replacement algorithms are furnished here.

Algorithm - I: Detection and Replacements of Missing Values using Standard Domain Length **(DMV-SDL)** Algorithm

Inputs:
Dataset, DS[]

Output:
Final Missing Value Replaced Dataset, DSF[]

Algorithm:
Step - 1. Import the dataset, DS[]
Step - 2. For each attribute in DS[] as DS[i]
  a. Count the non-missing value fields as N[i]
Step - 3. Find Max(N[i]) as CN
Step - 4. For each N[i]
  a. If N[i]/2 < CN/2
  b. Then, Check for missing values in DS[i][0] to DS[i][(N[i]/2)] and Add DS[i] to DST[]
  c. Else If N[i]/2 > CN/2
  d. Then, Check for missing values in DS[i][(N[i]/2)] to DS[i][(N[i])] and Add DS[i] to DST[]
  e. Else,
  f. Mark DS[i] as No Missing Value Fields and Add DS[i] to DSF[]
  g. Repeat Step - 4 for CN/n with n from 4 to CN
Step - 5. For each attribute fields in DST[] as DST[j]
  a. Calculate the domain moving average Avg_DST[j] and replace with missing values
  b. Add DST[j] to DSF[]
Step - 6. Return DSF[] as missing value cleared dataset

The proposed algorithm is primarily based on the divide and conquer method and thus, demonstrates a huge improvement in terms of time complexity.

Also, the proposed algorithm is capable of reducing the total rows if all the fields are missing. In measurements, ascription is the way toward supplanting missing information with subbed values. There are three fundamental issues that missing information causes: missing information can present a generous measure of predisposition, make the taking care of and investigation of the information more challenging, and make decreases in efficiency. In other words, when at least one quality is absent for a case, most factual bundles default to disposing of any case that has a missing worth, which may present inclination or influence the representativeness of the outcomes. Ascription saves all cases by supplanting missing information with an expected worth dependent on other accessible data.

Secondly, the domain specific outlier removal algorithm is furnished here.

Algorithm - II: Outlier Removal using Domain Specific Rule Engine **(OR-DSRE)** Algorithm

Inputs:
Dataset, FDS[]
Rule Engine, RE[]

Output:
Outlier Reduced Dataset, FFDS[]

Algorithm:
Step - 1. Building the rule engine, RS
  a. Rule 1: Job_Satisfaction >= 4 and Job_Change>=4
  b. Rule 2: Job_Satisfaction <= 2 and Job_Change<=2
  c. Rule 3: Job_Satisfaction = 5 and Job_Change = 5
  d. Rule 4: Job_Change >= 3 and Job_Change >= Job_Satisfaction
  e. Rule 5: Experience > 0 and SKILLS_NOW is NULL
  f. Rule 6: Experience <= 0 and SKILLS_UP is NULL
  g. Rule 7: Completion_Status > 50% and Customer_Rating < 3
  h. Rule 8: Completion_Status > 70% and Customer_Rating < 4
  i. Rule 9: Completion_Status > 95% and Customer_Rating < 5
  j. Rule10 to Rule27: Not included in this paper due to page limit constraints
Step - 2. Import the dataset, FDS[]
Step - 3. For each attribute in FDS[] as FDS[i]
  a. If FDS[i][0..n] match (RS)
  b. Then, Mark as outlier and remove
  c. Else, Mark FDS[i] in FFDS[]
Step - 4. Return FFDS[]

Abnormalities, or anomalies, can be a difficult issue when preparing AI calculations or applying factual methods. They are regularly the aftereffect of mistakes in estimations or extraordinary framework conditions and in this way don't depict the normal working of the basic framework. To be sure, the best practice is to actualize an anomaly expulsion stage prior to continuing with additional examination.

Sometimes, exceptions can give us data about confined peculiarities in the entire framework; so, the location of anomalies is a significant cycle due to the extra data they can give about your dataset.

Thirdly, the generic outlier removal algorithm is furnished here.

Algorithm - III: Double Differential Outlier Detection & Replacement (**DDOD-R**) Algorithm

Input:
Dataset, FDS[]

Output:
Outlier Replaced Dataset, FFDS[]

Algorithm:
Step - 1. Import the dataset, FDS[]
Step - 2. For each attribute in FDS[] as FDS[i]
  a. Calculate the element difference as DIFF[j] = Abs|FDS[i][j] - FDS[i][j+1]|
Step - 3. For each element in DIFF[] as DIFF[i]
  a. Calculate the element difference as DIFF_Second[j] = Abs|DIFF[i][k] - DIFF[i][k+1]|
Step - 4. Apply k-Mean Clustering for DIFF[]
Step - 5. Apply k-Mean Clustering for DIFF_Second[]
Step - 6. Identify the outliers for DIFF_Second[]
Step - 7. If DIFF_Second[m] is outlier
Step - 8. Then check,
  a. If DIFF[i][k] is outlier
  b. Then mark FDS[i][j] as outlier
  c. Else if, DIFF[i][k+1] is outlier
  d. Then mark FDS[i][j+1] as outlier
Step - 9. For each outlier in FDS[i][j]
  a. Calculate the moving average and replace the outliers
Step - 10. Repeat from Step – 2 until all outliers are detected
Step - 11. Return the final dataset as FFDS[]

Clustering or grouping is the errand of collection a bunch of items so that objects in a similar gathering are more comparative to one another than to those in different clusters. It is a fundamental undertaking of exploratory information mining, and a typical strategy for factual information investigation.

Fourthly, the attribute reduction algorithm is furnished here.

Algorithm - IV: Change Percentage Oriented Dependency Map based Attribute Reduction **(CPODM-AR)** Algorithm

Input:
Dataset, FFDS[]

Output:
Reduced Dataset, FFFDS[]

Algorithm:
Step - 1. Import the dataset, FFDS[]
Step - 2. For each attribute in FFDS[] as FFDS[i]
  a. Calculate the change percentage, CDP[i] = FFDS[i] with FFDS[0..(i-1)]
Step - 3. For each element in CDP[i]
  a. If CDP[i] < CDP[i+1]
  b. Then, remove FFDS[i] and calculate the Classification accuracy as CA[i]
  c. If CA[i] > CA[i+1]
  d. Then, Assign FFDS[i] to FFFDS[j]
  e. Else, Assign FFDS[i+1] to FFFDS[j]
Step - 4. Return the final dataset as FFFDS[]

VI. Proposed Framework

After the detailed analysis of the proposed algorithms in this section of the work, the proposed framework is furnished and discussed [Fig. 1].

The dataset for this research is adopted from the stack overflow developer survey and identified as one of the prominent datasets for enterprise scale research for pre-processing.

The dataset is distributed in two parts as employee dataset, as described already and project dataset, as described in the previous section of the work.

The proposed framework functions in four phases as in the initial phase the missing values from the employee collection are reduced and generates the missing value reduced dataset for employee collection using the DMV-SDL algorithm.

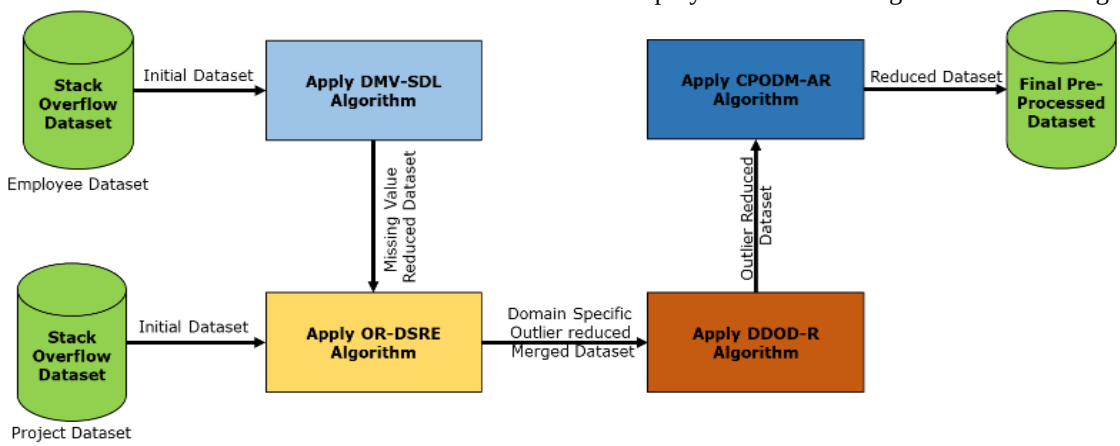

Fig. 1. Multi-Purpose Data Pre-Processing Framework.

The second phase of the proposed framework actually performs two different tasks as reduction of the domain specific outliers from the employee and the project dataset, and further merges the dataset based on the employees' assigned project using the OR-DSRE algorithm.

In the third phase of the proposed framework, the generic outliers are removed using the DDOD-R algorithm from merged dataset with employee and project specific outlines.

In the final phase of the proposed framework, the reduction of the attributes is taken care using the CPODM-AR algorithm where the validation of the reduction process is done using the classification method with the measuring parameters as accuracy and time complexity.

Further, the obtained results from this proposed framework are discussed in the next section of this work.

The dataset for this research is adopted from the stack overflow developer survey and identified as one of the prominent datasets for enterprise scale research for pre-processing.

The dataset is distributed in two parts as employee dataset, as described already and project dataset, as described in the previous section of the work.

The proposed framework functions in four phases as in the initial phase the missing values from the employee collection are reduced and generates the missing value reduced dataset for employee collection using the DMV-SDL algorithm.

The second phase of the proposed framework actually performs two different tasks as reduction of the domain specific outliers from the employee and the project dataset, and further merges the dataset based on the employees' assigned project using the OR-DSRE algorithm.

In the third phase of the proposed framework, the generic outliers are removed using the DDOD-R algorithm from merged dataset with employee and project specific outlines.

In the final phase of the proposed framework, the reduction of the attributes is taken care using the CPODM-AR algorithm where the validation of the reduction process is done using the classification method with the measuring parameters as accuracy and time complexity.

Further, the obtained results from this proposed framework are discussed in the next section of this work.

VII. Results and Discussions

The obtained results from the proposed framework and the algorithms are highly satisfactory. In this section of the work, the obtained results are furnished and discussed in five segments.

Firstly, the missing value detection and replacement results are observed from the employee dataset [Table IV].

The results are visualized graphically here [Fig. 2].

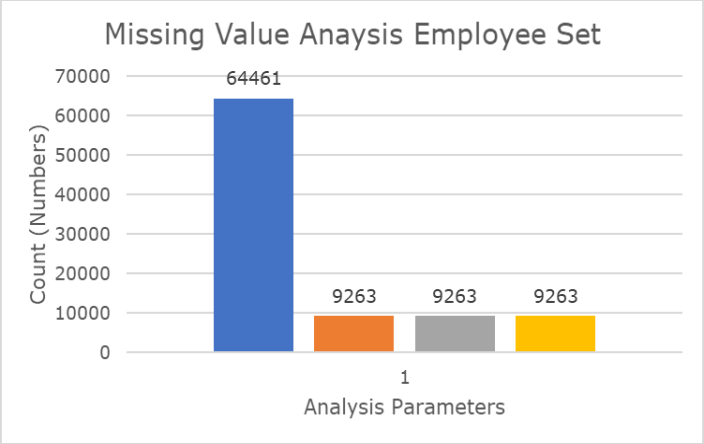

Fig. 2. Employee Dataset Missing Value Analysis

The missing value analysis from the initial employee dataset by the proposed DMV-SDL is highly accurate and demonstrates 100% accuracy.

Secondly, the merged dataset domain specific outlier and missing value analysis, after the merging analysis is here in Table V.

TABLE IV. Employee Dataset Missing Value Detection and Replacement

Total Number of Observation	Initial Number of Missing Values	Missing Values Detected	Missing Values Replaced	Missing Value Detection Accuracy (%)
64461	9263	9263	9263	100

TABLE V. Merged Dataset Missing Value and Domain Specific Outlier Analysis

Total Number of Missing Values Identified	Total Number of Missing Values Replaced	Missing Value Detection Accuracy (%)	Total Number of Outliers Identified	Total Number of Outliers Replaced	Outlier Detection Accuracy (%)
156060	156060	100%	17255	15492	89%

The results are visualized graphically here [Fig. 3].

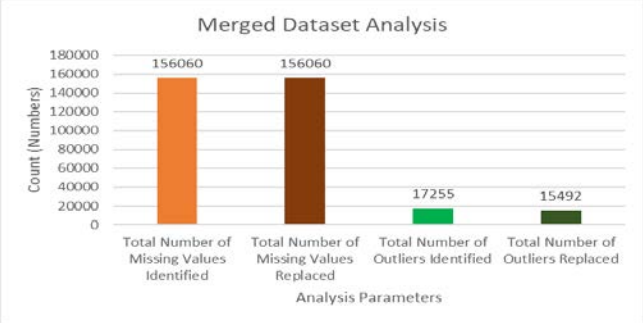

Fig. 3. Merged Dataset Missing Value and Outlier Analysis.

The proposed OR-DSRE algorithm has demonstrated 100% accuracy during the missing value analysis and nearly 90% accuracy during the domain specific outlier detection process.

Thirdly, the generic outlier removal outcomes are furnished here [Table VI].

The results are visualized graphically here [Fig. 4].

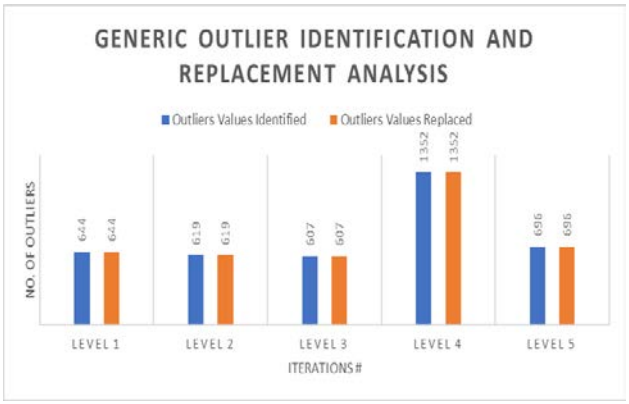

Fig. 4. Generic Outlier Identification and Replacement Analysis.

The iterative outlier identification and removal algorithm have also demonstrated nearly 100% accuracy and the algorithm identifies all the outliers within 5 iterations using the DDOD-R algorithm.

Finally, the attribute reduction results are furnished here [Table VII].

TABLE VI. GENERIC OUTLIER IDENTIFICATION AND REPLACEMENT ANALYSIS

Iteration #	**Outliers Values Identified**	**Outliers Values Replaced**
Level 1	644	644
Level 2	619	619
Level 3	607	607
Level 4	1352	1352
Level 5	696	696

TABLE VII. CHANGE PERCENTAGE METRIC

	ID	JC	AGE	EXP	SKILLS_NOW	SKILLS_UP	JS	JCHA	PID	DUR	CI	MI	TI	CS
ID	0	81	81	83	80	84	79	73	8	39	71	75	65	100
JC	21	0	81	83	80	85	76	72	11	58	71	74	81	100
AGE	59	81	0	84	79	84	75	69	27	61	72	73	82	100
EXP	2	81	81	0	79	84	76	72	55	18	73	74	84	100
SKILLS_NOW	6	81	81	82	0	85	79	71	3	43	70	73	80	100
SKILLS_UP	36	80	81	82	79	0	77	73	24	65	71	73	85	100
JS	27	80	81	82	80	84	0	70	61	1	70	73	79	100
JCHA	18	80	80	84	79	84	75	0	24	17	73	73	73	100
PID	8	80	80	83	79	84	76	72	0	40	70	75	67	100
DUR	61	81	81	83	80	85	78	72	61	0	69	73	69	100
CI	52	80	81	84	79	85	77	73	10	8	0	74	73	100
MI	40	80	81	82	80	85	79	70	28	55	70	0	75	100
TI	65	81	82	84	80	85	79	73	67	69	73	75	0	100
CS	100	100	100	100	100	100	100	100	100	100	100	100	100	0

Henceforth, based on the change percentage, the order of the attributes from the highest importance to the lowest is furnished here [Table VIII].

Further, based on the given rank, the attribute reduction process is carried out. The validation of the removal process is based on accuracy of classification and time complexity of processing [Table IX].

It is natural to realize that after the 5th iteration, the time complexity is reduced to a greater scale, but the accuracy has also declined. Thus, the attributes identified till the 5th iteration shall be marked as optimal.

The result is visualized graphically here [Fig. 5].

Thus, based on the final analysis the reduced set attributes are furnished here [Table X].

Further, in the next section of this work, the comparative analysis is carried out.

TABLE VIII. ATTRIBUTE RANKING ANALYSIS

Rank	**Attribute Number**	**Attribute Name**
Class Variable	0	CS
1	13	TI
2	6	SKILLS_UP
3	4	EXP
4	3	AGE
5	2	JC
6	5	SKILLS_NOW
7	7	JS
8	12	MI
9	8	JCHA
10	11	CI
11	10	DUR
12	9	PID
13	1	ID

TABLE IX. FINAL ATTRIBUTE REDUCTION ANALYSIS

Iteration #	**List of Attributes**	**Classification Accuracy**	**Time Complexity (msec)**
1	13,6,4,3,2,5,7,12,8,11,10,9,1	66	188
2	13,6,4,3,2,5,7,12,8,11,10,9	92	152
3	13,6,4,3,2,5,7,12,8,11,10	97	143
4	13,6,4,3,2,5,7,12,8,11	98	101
5	13,6,4,3,2,5,7,12,8	97	99
6	13,6,4,3,2,5,7,12	96	97
7	13,6,4,3,2,5,7	94	96
8	13,6,4,3,2,5	93	95
9	13,6,4,3,2	92	91
10	13,6,4,3	92	87
11	13,6,4	71	76
12	13,6	69	71
13	13	66	70

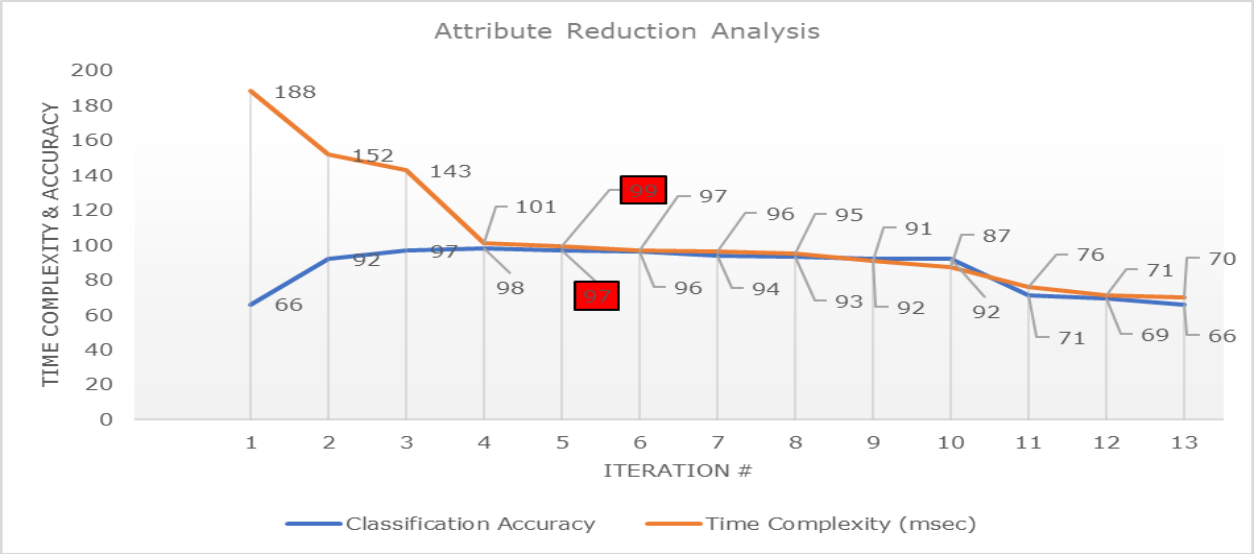

Fig. 5. Attributes Identified Till the 5th Iteration.

TABLE X. FINAL REDUCED DATASET

Rank	Attribute Number	Attribute Name
Class Variable	0	CS
1	13	TI
2	6	SKILLS_UP
3	4	EXP
4	3	AGE
5	2	JC
6	5	SKILLS_NOW
7	7	JS
8	12	MI
9	8	JCHA

VIII. COMPARATIVE ANALYSIS

After the detailed analysis of the results obtained from the proposed algorithms, in this section of the work, the proposed methods are compared with the parallel research outcomes [Table XI].

It is natural to realize that, the proposed framework deploys more extraction and analysis method for final detection, thus the accuracy in detection of the outliers and missing values are significantly high compared with the existing parallel research attempts. Finally, in the next section of this work, the research conclusion is presented.

TABLE XI. COMPARATIVE ANALYSIS

Research work, Year	Proposed Method	Missing Value Reduction	Outlier Reduction	Domain Information Preservation	Missing Value Detection Accuracy	Outlier Detection Accuracy
M. Hardt et al. [8], 2016	Equality Matrix with Supervised Learning	Yes	No	No	91	-
Z. Zhang et al. [9], 2016	Bias-based Identification	No	Yes	No	-	92
M. B. Zafar et al. [6], 2017	No Reduction	No	No	No	-	-
J. Kleinberg et al. [12], 2017	Risk Score	Yes	Yes	No	95	96
Proposed Framework, 2021	Standard Domain Length, Domain Specific Rule Engine, Double Differential Clustering, Change Percentage Oriented Dependency Map	Yes	Yes	Yes	99	99

IX. Conclusion

This research purposes on the benchmarked dataset by Stack overflow and a synthetic dataset. The proposed DMV-SDL algorithm first processes the employee-related dataset, and due to the nature of the divide and conquer method, the reduction in the time complexity is significant. Further, the stack overflow dataset and the synthetic project-specific dataset are analyzed under the OR-DSRE algorithm for domain-specific outlier imputation and provide a strategic merging of the datasets. Further, DDOD-R algorithm is applied on the merged dataset for generic outlier imputations. The proposed framework demonstrates a nearly 99% accuracy and some cases, up to 100% accuracy. The pre-processed dataset is analyzed under the CPODM-AR algorithm for dimensionality reduction and demonstrates nearly 99% accuracy with reduced time complexity for generic benchmarked classification algorithms. The work finally outcomes into a multi-purpose domain-specific data pre-processing framework for enterprise-scale data to make the data-driven business decisions more reliable.

Future Enhancements: Each pre-processed dataset attribute may be linked to as many timelines as required. In both the dependency properties and dependency forms, this is right (start- and end-attributes). In terms of accuracy, mostly related dataset related libraries are strongly recommended matched with the Original datasets.

References

[1] H. F. Ladd, "Evidence on discrimination in mortgage lending", J. Econ. Perspectives, vol. 12, no. 2, pp. 41-62, 1998.

[2] T. Calders and I. Žliobaitė, "Why unbiased computational processes can lead to discriminative decision procedures" in Discrimination and Privacy in the Information Society, New York, NY, USA:Springer, pp. 43-57, 2013.

[3] D. Pedreschi, S. Ruggieri and F. Turini, "Discrimination-aware data mining", Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 560-568, 2008.

[4] S. Hajian, "Simultaneous discrimination prevention and privacy protection in data publishing and mining", 2013, [online] Available: https://arxiv.org/abs/1306.6805.

[5] B. Fish, J. Kun and Á. D. Lelkes, "A confidence-based approach for balancing fairness and accuracy", Proc. SIAM Int. Conf. Data Mining, pp. 144-152, 2016.

[6] M. B. Zafar, I. Valera, M. G. Rodriguez and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment", Proc. 26th Int. World Wide Web Conf., pp. 1171-1180, 2017.

[7] T. Kamishima, S. Akaho and J. Sakuma, "Fairness-aware learning through regularization approach", Proc. IEEE 11th Int. Conf. Data Mining Workshops, pp. 643-650, 2011.

[8] M. Hardt, E. Price and N. Srebro, "Equality of opportunity in supervised learning", Proc. Adv. Neural Inf. Process. Syst., pp. 3315-3323, 2016.

[9] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger and S. Venkatasubramanian, "Certifying and removing disparate impact", Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 259-268, 2015.

[10] C. Dwork, M. Hardt, T. Pitassi, O. Reingold and R. Zemel, "Fairness through awareness", Proc. 3rd Innov. Theor. Comput. Sci. Conf., pp. 214-226, 2012.

[11] Z. Zhang and D. B. Neill, "Identifying significant predictive bias in classifiers", Proc. NIPS Workshop Interpretable Mach. Learn. Complex Syst., 2016, [online] Available: https://arxiv.org/abs/1611.08292.

[12] J. Kleinberg, S. Mullainathan and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores", Proc. Innov. Theor. Comput. Sci. Conf., 2017.

[13] The Public 2020 Stack Overflow Developer Survey Results, https://insights.stackoverflow.com/survey/2020.

[14] Li, Shaoying, et al. "Inferring the trip purposes and uncovering spatio-temporal activity patterns from dockless shared bike dataset in Shenzhen, China." Journal of Transport Geography 91 (2021): 102974.

[15] Wagner, Joachim. "Exports, R&D and Productivity: A test of the Bustos-model with enterprise data from France, Italy and Spain." MICROECONOMETRIC STUDIES OF FIRMS'IMPORTS AND EXPORTS: Advanced Methods of Analysis and Evidence from German Enterprises. 2021. 217-222.