A Recruitment Big Data Approach to interplay of the Target Drugs

Wael Alzyadat, Mohammad Muhairat, Aysh Alhroob, and Thamer Rawashdeh

Faculty of Science and Information Technology, Al-Zaytoonah University
Amman, Jordan
e-mail: Wael.alzyadat@zuj.edu.jo
Faculty of Science and Information Technology, Al-Zaytoonah University
Amman, Jordan
e-mail: drmohairat@zuj.edu.jo
Faculty of Information Technology, Isra University Amman, Jordan
e-mail: aysh@iu.edu.jo
Faculty of Science and Information Technology, Al-Zaytoonah University
e-mail: Thamer.r@zuj.edu.jo

Abstract

The various model that has been used to predict, datamining, and information retrieval are useful to use through the traditional database, due to big data the prediction should derive in a different role that conduct the hidden structure data based on a stability scale to allow discovering accrue unsupervised drug data. Especially, the drug data must be understandable to analysts. Following this approach, conduct the stability drug data through computation methods are quality measurements, preprocess data, k-mean cluster, and decision tree. This approach seeks to identify the data by two dimensions (vertically and horizontally), which extrapolations, compilation, and interpretation values of the dataset while considering individual attributes. A comparison with clusters defines the set for features using balance value by K-mean algorithm to determine the k clusters that consider the set of features based on two values 0 and 1, which given the discernible between dependent and independent class target, and pinpoint the relationship among them.

Keywords: Big Data, Discretize, k-mean cluster Stability, Target drug.

1   Introduction

The data life cycle is inception from the capture data or collects data from specific domain or from several different domains to transfer in actionable information consist of the process steps for acquiring the information, the primary process predefined the context of raw data to be able to determine the type of data structure. [1].

The elaboration type of data structure need converge in advance addressing by data typing and navigation, which able to trapped in three characteristics are structured data, semi-structured, and unstructured data; the type of structured data is organized resource navigation, and predefined data type through paramount processes such resource manager, allocated processing context, data movement, and tiers of storage, which the performing structured data compatibility for vertical scale and horizontal scale like SQL; the umbrella of globalization is a major impact factor to accentuated the semi-structured data when integrating dataset from different resources (navigations) or overhang format such as XML, Yet, unstructured data is absences the data type and navigation. Typically, the contentions aspects of techniques data architectures. The extensive filed of data architectures, which uses techniques or/and algorithms such area as Artificial Intelligent (AI), statistics, and genetic algorithm. Not long ago the data mining technology can be applied to a large number of databases and files, through the running data mining process considering to the input dataset is structured data that why lead on context aspect using data pre-processing to improve the raw data. Furthermore, the navigation matter resolves by data management, model, and pattern; it relies on a data warehouse to incentivize the knowledge discovery process.

The data warehouse intended for data mining applications utmost large volume of data from multiple sources from different platforms and systems, it dealt with data must be reformatted. However, the prominent issue is up to data from various sources must be cleaned and installed in the data model of the data warehouse. For instance, Intermittently Synchronized Databases (ISDBE) when the client request update from server, the sending deal with a specific type of communication such as unicast or multicast.

One of the keys drives of expansion quantities to data era, depending on communication include all devices on networks with a multitude of environments which the propagation of digitalization perspective, whether it is collected or create from sophisticated end-points consider the right techniques for transfer the data by storage, management, and data locality such as parallel, distributed data processing, and distributed file system [2].

The outstanding challenge during processes of transformation, access, sharing and data locking were overcome through different repositories or gathered into a central compute across perturbed from data structures and contention flow rate

such as streaming data which the behind of the view what the control techniques ensure the non-interference during executing transaction and serializability schedules. Typically, concurrency control techniques based on timestamp is one possibility to handle the issues in DBMS level. The aforementioned issue challenge is to realize in medicine data from different technical aspects.

contemplate finding an answer for what is medicine Big Data? conspicuously must dive individually to big data and medicine keywords to acquire the interaction between both concepts, then repurposing as medicine big data. The medicine term used key characters, the primary is medication names which each have at least generic name and brand name, the route character explain how the medication get in blood such as oral (by mouth) through digestive system, nasal, buccal, sublingual, eye drops and ear drops, transdermal, topical medications, subcutaneous, enteral, rectal and vaginal, and inhaled; the medications affect character classified by desired effect (therapeutic effect), side effect, tolerance and dependence, interactions, no apparent effect, and paradoxical effect, the essential medicines is dosage forms [3] [4], The drugs consist of chemical and biological properties per each molecule.

The Big Data concept considering to five main characters which each covering the dimensions of data such as behavior, content, and producing; the popular character focus on quantity of data sets that identify the capacity of storage in repositories as data centers, Moreover, present by huge massive amount of data termed volume [5, 6]; the velocity conduct to generate data through different sources such as streaming data that generate from digital devices to collect from environment which obvious when use the GPS or sensors [7], consider the big data source form categorize by structure , semi-structure, and unstructured data that character responsible via veracity character [8].

The growing drug information rapidly increasing which inescapable to use Big Data, one of example to inferred the volume in CAS Registry contained 30 million compounds, even in year 2008 the number of compounds were 17 508 756 million after 10 years the amount increases 12 902 970 million compounds by this speed generation data crucial to consider the velocity. Meanwhile, the CAS Registry repository store the compounds by three levels are graph, node, and bound the Variety of data [9].

2   Related Work

In this section will explain and discuss the drug discovery by data bases explain the key terms data storage, data capacity, extract data, and retrieve data.

According to [10] Big Data is a hot topic in health care and biomedical research. In contrast, still many unknowns around the practical use of Big Data in medicine,

also [11] reinforces it hasten to improve both quality and efficiency of care. Undoubtedly the medicine disciplines demand of the Big Data era [12].

The production of drug discovery by data collection, and integration are ongoing steepest into data base that request store, analysis and, management to able to find out right resources as Big Data successfully achieved the target from different field such as social media and ecommerce. Herein, we make investigation to literature including drug discovery and Big Data strategy.

The ZINC database contain over 230 million purchasable compounds presents as 3D as well as over 750 million analogue search [13] the substance resources from Zinc 15 presents the structure data by groups were attributes, hybrids, aliases to related resource attributes, computed properties, and related resources; the attributes constraints by three dimensional the first is a name of attributes (present the meaning of content), type of attributes (integer, string, char, Mol, and date) and the searchable attributes (true or false), the number of attributes are 10.Moreover, the hybrid group same dimensional the number in hybrid group is 30; the aliases to related resource attributes involve 21 items; the computed properties group cover 10 driven as interlink among other groups [14].

These groups are major to indicate the related resources group to achieve inescapable outcome. In this paper we selected four related resources are activities, observation, prediction, and patterns. The life cycle of substance data from ZINC describes under level of structure database and Database Management (DBMs), in figure 1 illustrate the deploy data repository that disseminate to website zinc.org.
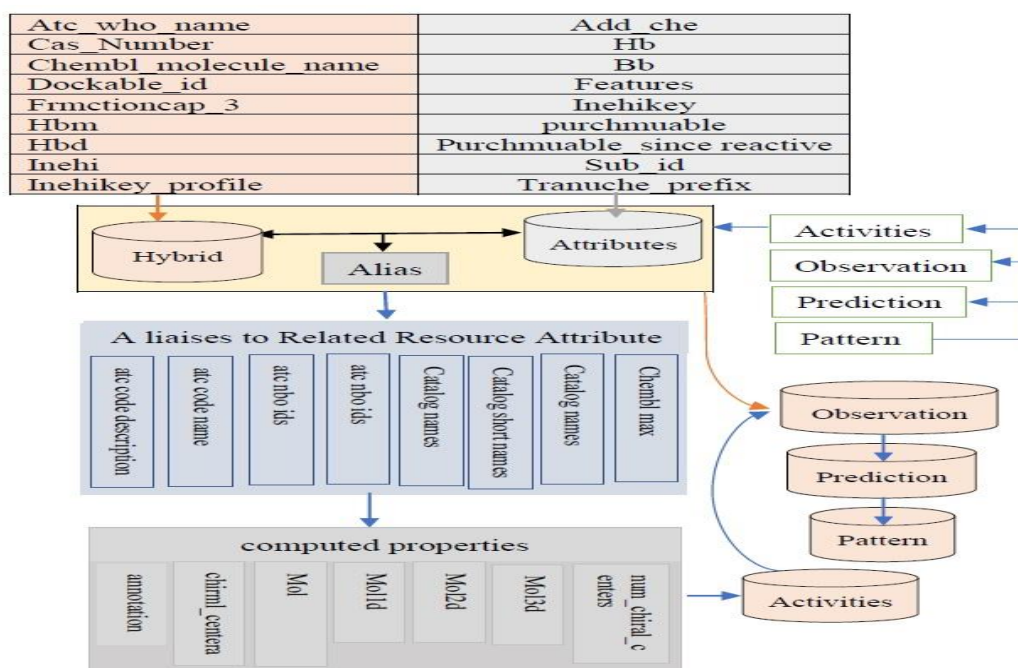

Fig. 1 Substance of structure data in ZINC

The usability of structure database through zinc is search function. Herein, request the time is comprehensive to achieve the output. The indication of process time was absence because the massive amount of data need monitors through whole resources to capture matching between input and output. Meanwhile, the data access uses the queries for combined few features from different resources. Database rapid increase of size, in 2017 the size of structure data base was 58 million [15], at 2020 the data base increase about 27 million which increase 46.5 % through 275 data sources [16].

One of huge amount chemical data is Chemical European Molecular Biology Laboratory (ChEMBL) [17], which the challenge when using big data is facing difficult to analyze based on a variety of contexts such as target protein, cell line, drug chemical structure, and activity parameter. According to [18] suggested a model named GPCR Signaling Pathways; Combining PTML-ChEMBL Models and GTPγS Binding Assays to carry out predictive

3   Proposed Approach

The proposed approach involves four phases: retrieve data, quality measurement, preprocessing data, and unsupervised, and features selections. The preprocessing data phase includes loops attributes, role attributes, discretize, map, positive class, and remove column, while the unsupervised features selections phase contains the fully connected paths, recall path is dropout the dataset through quality measurement, and the preprocessing data path which the dataset arrangement by attributes and instances. Figure 2 demonstrates a structural representation of the recruitment big data approaches to interplay of the target drugs.
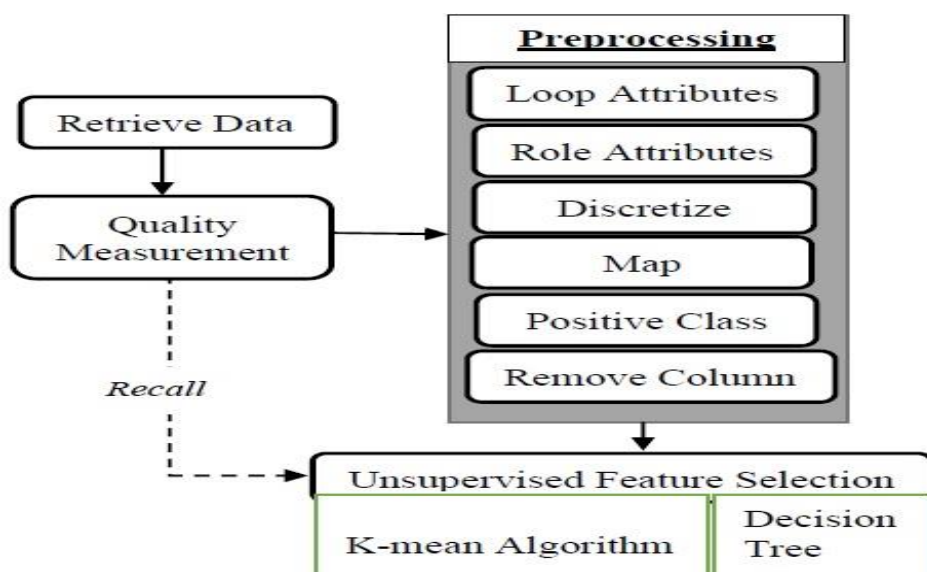

Fig. 2 A recruitment big data approaches to interplay of the target drugs

In our study, the primary data source prepared by Medicare & Medicaid Services (CMS) [19], the datasets consist 1048575 records the constituted twenty-one attributes. the summary for datasets as shown in table.1.

Table 1: The summary for datasets

Object	mean	min	max	std	length of sample	number of NAs
bene_count	2.7	11	21246	4.31	1048575	665007
bene_count_ge65	1.85	0	19078	5.85	1048575	906615
bene_count_ge65_suppress_flag	na		*	na	1048575	0
description_flag	na	S	T	na	1048575	0
drug_name	na	1st tier unifine pentips	zyvox	na	1048575	0
ge65_suppress_flag	na		*	na	1048575	0
generic_name	na	0.9 % sodium chloride	zoster vaccine live/pf	na	1048575	0
npi	1.49	1003000522	1992999122	2.87	1048575	0
nppes_provider_city	na	waldport	zwolle	0.00	1048575	0
nppes_provider_first_name	na		zyra	na	1048575	6
nppes_provider_last_org_name	na	a. omar vento	zziwambazza	na	1048575	8
nppes_provider_state	na	ae	zz	na	1048575	0
specialty_description	na	acupuncturist	vascular surgery	na	1048575	0
total_30_day_fill_count	6.98	11	21290	1.19	1048575	0
total_30_day_fill_count_ge65	6.6	0	19113	1.22	1048575	434273
total_claim_count	5.	11	21290	8.47	1048575	0
total_claim_count_ge65	4.65	0	19113	8.48	1048575	434273
total_day_supply	1.95	9	150499	3.42	1048575	0
total_day_supply_ge65	1.89	0	142159	3.49	1048575	434273
total_drug_cost	3.41	0.11	4069122.91	1.53	1048575	0
total_drug_cost_ge65	2.75	0	3651179.68	1.17	1048575	434273

The ID_ness measure conveys the attributes to indicates the fraction values, using the ID for each attribute divided by numbers of rows, the significant for it determine the behaved for each attribute from datasets which the potential categorize by two outcomes, the first outcome is present many different values, refer to table 2 the highest attributes through npi, nppes_provider_last_org_name, nppes_provider_city and nppes_provider_first_name. Yet, the inspect by ID_ness as high of values are signs the attribute inoperative due to a feature's selection. Second outcome present specific values, occurred at description_flag and bene_count_ge65_suppress_flag.

The contaminations data present various issues. First, absence the index for a variable that reduce to sensitivity analysis. Second, the data mislay an unsatisfactory of measurement along with parameters. Third, it can be atomizing the sampling of data [20] [21]. Refer to missing data term retains the features that scale zero missing [22] such as npi and nppes_provider_last_org_name. In

contrast, the bene_count_ge65 and ge65_suppress_flag both approximate 0.8% from the overall dataset; to handling missing data should be conducted to the sensitivity of dataset by most common approaches are Pairwise deletion, mean substitution, Regression imputation and Maximum likelihood.
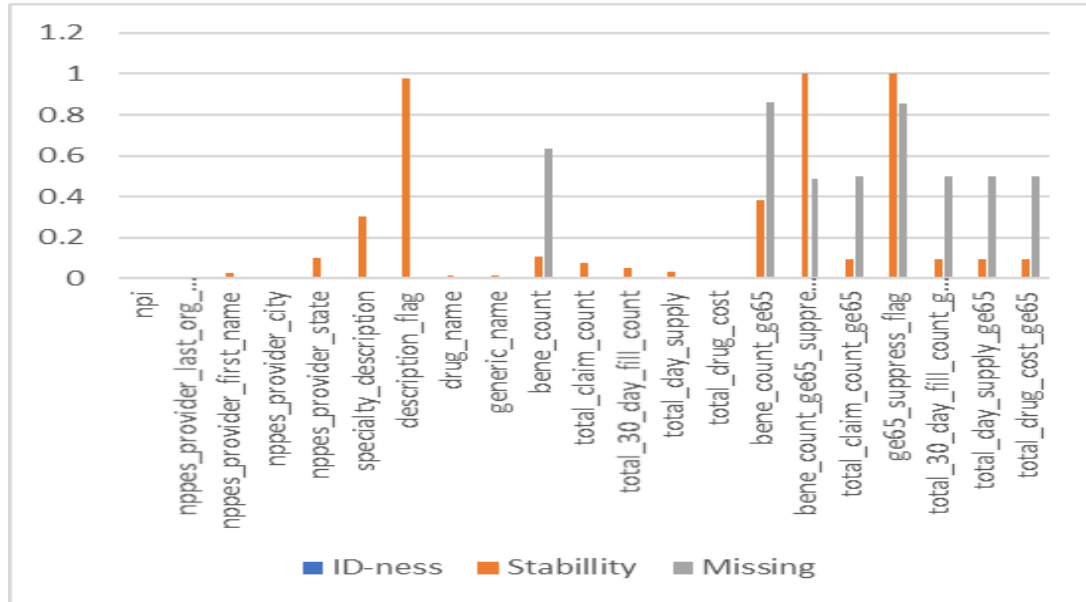

Fig 3 Relationship among ID-ness, stability and missing

The stability is an importance of analysis data, which extrapolations, compilation and interpretation values of dataset while consider to individual attributes. Furthermore, the choice of analytical technique dependent on upon extensive dataset condition. For example, the correlation one of a prime statistical technique were concentrations of quantitative such as linear-free-energy relationships (LFERs) and Pearson [23] both reliant on numerical or binominal to measure by degree -1 to +1 that implies association positive or negative. Unfortunately, the correlation technique for Medicare Provider Utilization and Payment Dataset rectified the values of variables which required task to transform data such as range, normalize, sample, split and/or replace, these are complicated and onerous to determine the feature selection.

In this study, the reliable technique equilibrium with dataset uses ID-ness and missing data to define the stability constant to obtain desirable attributes, shown in figure 3 illustrated the stability for each attributes the equation can be expressed. The count for the most frequent non-missing value for this column divided by the number of rows.

$$\text{Stability\_Attribute} = \sum\nolimits_{(i==1)}^{n} (i(\text{non-missing})/n)) \qquad (1)$$

Where i is the values of attribute to recognize the non-missing data, there was a direct relationship between number of missing data and the validate stability, the n is the number of rows. Refer to figure 3. Indicated the highest attributes stability

are ge65_suppress_flag, bene_count_ge65_suppress_flag and description flag. Since the ge65_suppress_flag, bene_count_ge65_suppress_flag attributes the stability is highest but the involve missing data are 0.85 and 0.48 that required to overlooked and keep the description_ flag significance of feature selection. Meanwhile, to discuss the reason why the first two stability attributes outcome that effect by ID-ness scale.

To rectified the stability constant need condensed the quality measurement using preprocessing concept which is a master key to analysis data process in terms of utilization, reduction and preparation data. It is an involved in set of techniques such as data mining and knowledge discovery mandatory to handle the feature selection, sampling, and compression [24] [25]. First step is loop attributes applies iteration to each attribute. In all iterations essential to perceive the storage data for detection to capture or collect of the data, to determine the source how to generate rate based on their needs. Especially, the real-time data event detects the generate rate scaling by time which the streaming data request monitoring operation rely on buffer such as sensor-captured. On other hand, the multiple resources such as corporate database, its insights about set the conditions and role attributes.
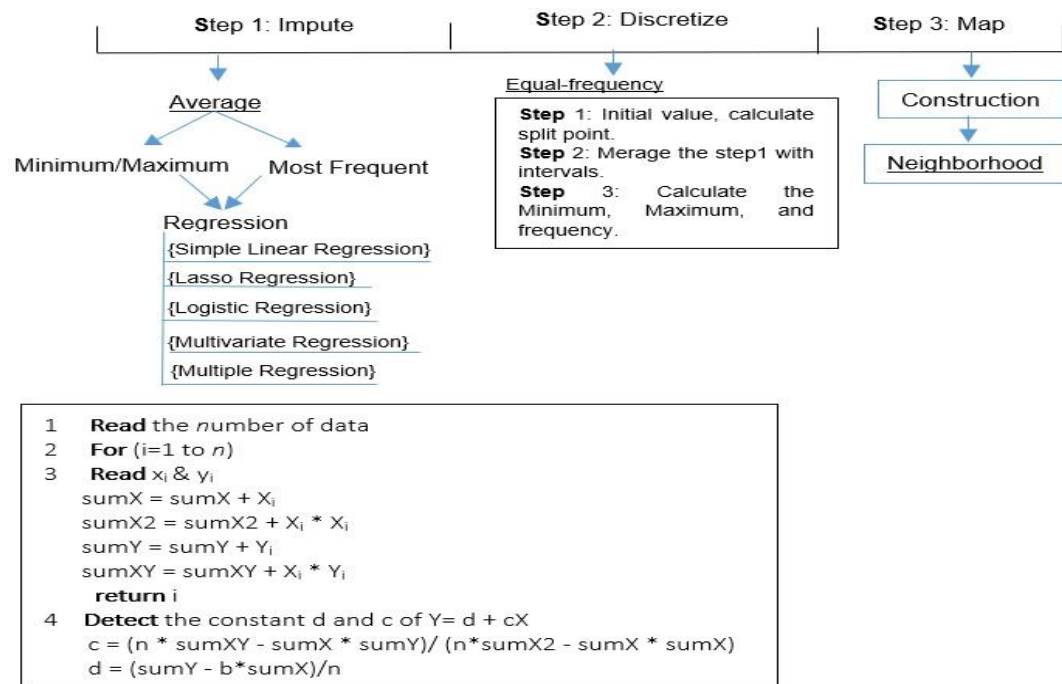

Fig. 4 Algorithm for linear regression

The discretize step effective to spilt point and data distribution with the induction the map dataset and set the boundary of dataset based on three aspects are recursive, rage value, and approximate the instances. The operations through the aspects are similar consideration to switch the map [26] [27], whereas discretization performs the distribution set of feature sets.

4   Result and Discussion

The map expresses the similar data use Euclidean distance criterion to identifies the most important attributes [28]. Afterwards, assigned the status of the attributes through positive or negative; to judge the remove attributes refer to missing value and distance.

Mapping plays an important to declaring attributes and instances, it represents the impact attributes on a clustering to adjust the feature extraction, the key contribution of the map is assigned the group based on a distance measure between features. Furthermore, the construction predefines the schema among to relevant subset without loss information, different techniques consider to map faculty such as map reduce programming (Hadoop), directed acyclic graph parallel processing (Spark), and bulk synchronous parallel processing [6]. Yet, the neighborhood indicating the range of indiscernibility through specified core points within distance use Euclidean metric.
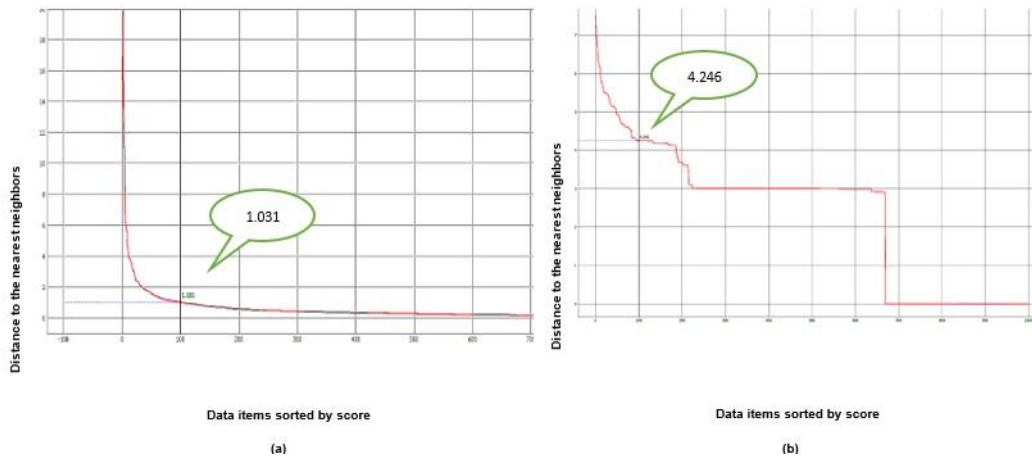

Fig 5 Presents the data sets, (a) the datasets mapping the features without preprocess and discretization performs, (b) use the mapping applied the preprocess and discretization.

The core points from figure 5. (a) is 1.031 describe the corresponding the datasets and detect the clusters based on distance of nearest neighbors which define interior items that determine the ratio of properly feature set. (b) detect the core point by 4.246, while the discretization operation allows to omitted outliers items; it illustrated the clustering dataset by two dimensions are attributes (column) and items (row).

There are three observations of effective the preprocessing and discretization:

1- Dimension reduction: provide the mapping of data to dimensional space, it led to a determine the potential techniques such as the linear transformation is Principal Component Analysis (PCA), which is more informative dimensionality, refer to figure 5. (a) the consistency from a dataset demonstrated for creating decision tree. While proposed approach use recall to

fetch up dataset when quality measurements phase is done that confidently produce valuable set of features.

2- Obtain relevant data: depend on core points and nearest neighbors used for description the variances data. Therefore, to eliminate the features are normally measured on distance score. Therefore, Figure 4 the first step impute process utilizing the features by minimum, maximum, and frequently to filling missing value. Once imputation success the discretization concerned the infer target variables from a specific feature.

3- Informative sets of features: to deal with characteristics by amplitudes that can be detected behavioral of the features which the purpose to find the key features to describes overall datasets.

For discover the closest features belonging to different attributes, to define the set for features using balance value by K-mean algorithm to determine the k clusters that consider the set of features based on two values 0 and 1, based on upon the distance measure assign the similar consist of features. Consequently, the decision tree represents for each attributes a splitting rule to estimate the target (numerical values). The rules belonging to targets intended for generate classes, where a set of classes need way to optimize for each to obtain an interoperability rule. The following processing is reducing the error by regression, which given the discernible between dependent and independent class target, and pinpoint the relationship among them.

5   Conclusion

In this work, we present a theoretical approach to the Medicare & Medicaid Services (CMS) dataset through the big data concept. This task requires verifying context data by two-level horizontal and vertical to identify the quality measures for structure data and their pinpoint sequence events of components. Since the dataset are described by structured data the preprocess data able to mapping data context which is necessary for clustering. The k-mean cluster defines the groups of datasets based on similar features which the outcome based on stability scale, the rules generating from decision trees are high confidence. The limitation of the huge amount of data is hard to visualization inspect, because of that the meaningful data are disappearing. The future work extends the indicate and acquire the meaningful medical data which the raw data adjust to structured data.

References

[1] W. J., Al-Zyadat, R. B., Atan, H., Ibrahim and M. A. A., Murad. (2011), The Directs Impact to Pre-Filtering Process to Weather Dataset, *Journal of Theoretical & Applied Information Technology*, Vol. 26, no. 1.

[2] The digitization of the world from edge to core, [Online]. Available: http://cloudcode.me/media/1014/idc.pdf. [Accessed September 2020].

[3] Standard Treatment Guidelines and Essential Medicines List [Online]. Available: https://www.who.int/selection_medicines/country_lists/Tanzania_STG_05201 3.pdf. [Accessed July 2020].

[4] World Health Organization Model list of Essential Medicines: 21st list 2019, [Online]. Available: https://apps.who.int/iris/bitstream/handle/10665/325771/WHO-MVP-EMP-IAU-2019.06-eng.pdf

[5] A., Al-Hgaish, W. J., Alzyadat, M., Al-Fayoumi, A., Alhroob and A., Thunibat. (2019), Preserve Quality Medical Drug Data toward Meaningful Data Lake by Cluster, *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, no. 3, pp. 270-277.

[6] W. J., Al_Zyadat, F. Y., Alzyoud, A. Alhroob and V. Samawi (2018) "Securitizing Big Data Characteristics Used Tall Array and Mapreduce, *International Journal of Engineering & Technology*, vol. 7, no. 4, pp. 5633-5639.

[7] W. J., Alzyadat, A., AlHroob, I. H., Almukahel and R., Atan, (2019). Fuzzy Map Approach for Accruing Velocity of Big Data, *Compusoft*, vol. 8, no. 4, pp. 3112-3116.

[8] W. j., ALzyadat and A., Alhroob, (2019) Development Planning in the Big Data Era: Design References Architecture, *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, no. 1, pp. 884-887.

[9] A. H., Lipkus, S. P., Watkins, K., Gengras M. J., McBride and T. J., Wills, (2019) "Recent Changes in the Scaffold Diversity of Organic Chemistry as Seen in the CAS Registry, *The Journal of organic chemistry*, vol. *84*, no. 21, pp. 13948-13956.

[10] F., Martin-Sanchez and K., Verspoor, (2014). Big Data in Medicine is Driving Big Changes, *Yearbook of medical informatics*, vol. 23, no. 1, pp. 14-20.

[11] T. B., Murdoch and A. S. Detsky, (2013). The Inevitable Application of Big Data to Health Care, *Jama*, vol. 309, no. 13, pp. 1351-1352.

[12] S. J., Lusher, R., McGuire, R. C., van Schaik, C. D., Nicholson and J., de Vlieg, (2014) "Data-Driven Medicinal Chemistry in The Era of Big Data. *Drug discovery today*, vol. 19, no.7, pp. 859-868.

[13] T., Sterling and J. J., Irwin (2015), "ZINC 15–ligand Discovery for Everyone, *Journal of chemical information and modeling*, vol. 55, no.11, pp. 2324-2337.

[14]   ZINC, "ZINC15," [Online]. Available: http://zinc15.docking.org/. [Accessed October 2019].

[15]   M., Awale, R., Visini, D., Probst, J., Arus-Pous and J. L., Reymond (2017), "Chemical space: big data challenge for molecular diversity, *CHIMIA International Journal for Chemistry*, vol. 71, no. 10, pp. 661-666.

[16]   R. S. O. Chemistry, [Online]. Available: http://www.chemspider.com. [Accessed May 2020].

[17]   ChEMBL, [Online]. Available: https://www.ebi.ac.uk/chembl/, [Accessed 5 February 2021].

[18]   R., Diez-Alarcia, V., Yáñez-Pérez, I., Muneta-Arrate, S., Arrasate, E., Lete, J. J., Meana and H., González-Díaz, (2019). "Big Data Challenges Targeting Proteins in Gpcr Signaling Pathways; Combining PTML-ChEMBL Models and [35S] GTPγS binding assays, *ACS chemical neuroscience*, vol. 10, no. 11, pp. 4476-4491.

[19]   Medicare Provider Utilization and Payment Data: Part D Prescriber, [Online]. Available: https://www.cms.gov, [Accessed 1 February 2021].

[20]   R. F., Rodriguez, J. M., Castillo, M. P., Castillo, O., Montoya, P., Daza, M. F., Rodríguez and A. M., Angel. (2008), "Hydrocodone/acetaminophen and Tramadol Chlorhydrate Combination Tablets for the Management of Chronic Cancer Pain: A Double-blind Comparative Trial, *The Clinical journal of pain*, vol. 24, no. 1, pp. 1-4.

[21]   H., Kang. (2013), The Prevention and Handling of The Missing Data, *Korean journal of anesthesiology*, vol. 64, no. 5, pp. 402-406, 2013.

[22]   A., Alhroob, W. J., Alzyadat, I., Almukahel, and H., Altarawneh. (2020), "Missing Data Prediction using Correlation Genetic Algorithm and SVM Approach, *population*, *vol. 11*, no.2, pp. 703-709.

[23]   W., Robak, W., Apostoluk, P., Maciejewski, J. A., Pielka and J. N., Kwiotek. (2013), Linear Free Energy Relationship (LFER) Analysis of Dissociation Constants of 8-Hydroxyquinoline and Its Derivatives in Aqueous and Dioxane–Water Solutions, *Journal of Chemical & Engineering Data*, *vol. 58*, no. 6, pp. 1470-1482, 2013.

[24]   N., Shehab, M., Badawy and H. Arafat. (2021), Big Data Analytics and Preprocessing," In *Machine Learning and Big Data Analytics Paradigms: Analysis, Applications and Challenges,* vol. 77, pp. 25-43.

[25]   F., Moraes Gomes, J., Yang, R., Zhangand V. Murdock. (2020), The Role of Attributes in Product Quality Comparisons, in *Proc. of the 2020 Conference on Human Information Interaction and Retrieval*, Vancouver, Canada, pp. 253-262.

[26]   S. Y., Jiang, X., Li, Q., Zheng and L. X., Wang. (2009), Approximate Equal Frequency Discretization Method, *In Proc. of 2009 WRI global congress on intelligent systems*, Xiamen, China, pp. 514-518.

[27]   Kanan, Tarek; Kanaan, Ghassan G.; Al-Shalabi, Riyad; Aldaaja, Amal. (2021), Offensive Language Detection in Social Networks for Arabic

Language Using Clustering Techniques., *International Journal of Advances in Soft Computing & Its Applications*. Vol. 13 Issue 2, p95-111. 17p.

[28]   P. E., Danielsson. (1980), Euclidean Distance Mapping, *Computer Graphics and image processing*, vol. 14, no. 3, pp. 227-248, 1980.

Wael Jumah Alzyadat is currently an Assistant Professor of software engineering. He also works with Al-Zaytoonah University, Jordan. His research area encompasses the area of software analysis, intelligence systems, streaming data, and big data. Moreover, he established more than 25 published articles and achieved two copyrights.

Mohammad I. Muhairat received the Master and Doctor degrees from Kharkov National University of Radio Electronics, Kharkov, Ukraine, in 1997and 2002, respectively. In September 2002, he became an Assistant Professor at the Department of Software Engineering, Al-zaytoonah University, Amman, Jordan, where he became an Associate Professor in July 2011. He is the author or coauthor of more than 35 papers in international refereed journals and more than 10 conference contributions. He held several positions, such as, Software Engineering Chair, Vice Dean of IT Faculty and Deputy Dean of the Deanship of Scientific Research. His research interests cover several aspects across Software Engineering and Big Data.

Aysh M. Alhroob received the Ph.D. degree from the University of Bradford, U.K., in 2010. He is currently an Associate Professor of software engineering with Isra University, Jordan. He joined Isra University as an Assistant Professor with the Software Engineering Department, Faculty of Information Technology. He has a number of published articles in various computer science and software engineering topics in international journals and conferences. In addition, he has published his first book in software testing, in 2010.

Thamer Ahmad AL-Rawashdeh is an Associate Professor at the Al-Zaytoonah University, Jordan. received the Ph.D. degree from Ph.D. in Software Engineering from Universiti Utara Malaysia, Malaysia.