Original Article

WEATHER DATA ANALYSIS DATA PREPROCESSING

R. Jayakumar[1], Dr. R. Annamalai Saravanan[2]

[1] Research Scholar, Department of Computer Science, Sankara College of Science and Commerce, Tamilnadu, India.
[2] Associate Professor and Head, Dept of Information Technology, Sankara College of Science and Commerce, Tamilnadu, India.

DOI: 10.47750/pnr.2022.13.S08.390

Abstract

In this research, we are working with Big Data for getting, preparing and breaking down data-based data to utilize the data recovered which will help any association. It is an advancing piece of all divisions of industry and business. All organizations in any field, for instance, oil, cash, fabricating hardware, etc produce big data, which can show unquestionably accommodating designs to business chiefs to make and develop their organizations, when the data is accumulated and broke down accurately. It permits us to assemble, store, and unravel colossal proportions of big data to deliver helpful results. Data preprocessing by utilizing SPSS. In the wake of preprocessing, the data is clean, coordinated and decreased. In an end the experiment, SPSS can satisfy fundamentally the majority of the data preprocessing undertakings and give a superior knowledge into the data.

Keywords: Weather Data, preprocessing, SPSS, data cleansing.

Introduction

It's referred to that as of late as the improvement of current culture has advanced, joined by keen frameworks through the Web, and the innovation of the Internet of Things (IoT), this has caused a supported growth in the volume of data around the world, through the various exercises that human beings do in friendly turn of events. This high volume of data has become complicated to manipulate and process. Notwithstanding, and at the same time, there has been a fast growth in computational and stockpiling limit in PCs or in the cloud, as well as a decrease in the expense of this innovation. How much information that is put away everyday in various databases (DB) for instance; from simulations and logical observations, from public bodies, and so on, has permitted gigantic admittance to data in a simple and basic manner, and now and again continuously. These DBs can contain valuable information for various issues and require exceptional techniques to investigate and dissect.

Because of the transversal quality of DM, it's appropriate in any unique circumstance, for example, climatological examination on data from the regular electrical framework or from environmentally friendly power innovation, for example, photovoltaics, the last option corresponding to the contextual analysis of this research work. In the area of meteorology, statistical techniques are generally used to address different weather conditions modeling and prediction issues from observations and numerical model results. Be that as it may, the enormous volume of data accessible today makes these techniques inappropriate.

Big Data

Big Data has become significant for any association that produces an enormous measure of heterogeneous data, which whenever caught, handled, and dissected will uncover designs and give experiences. The Big Data idea

arose when enormous volumes of organized, semi-organized, and unstructured data represented a troublesome undertaking for handling utilizing conventional techniques and databases. The size, speed, and arrangement in which data is produced influence the nature of the information. Data can emerge out of various sources, for example, deal frameworks, client databases, portable applications, sites, machines and constant data sensors created in IoT frameworks.

Big data analytics in IoT requests putting away the data in a few stockpiling advances. Big data implementations will require performing lightning-quick analytics with inquiries to permit associations to acquire fast bits of knowledge and settle on speedy choices. Contingent upon the requirements of the created IoT applications, different logical sorts have been examined in this subsection under constant, disconnected, business intelligence level, and massive level analytics.

Pre-processing of Big Data

In the early stages of any big data program, data preprocessing, or data consistency analysis boost and improve the data values. Usually the lifecycle data pre-processing includes these subsections:

1) The fortification and Integration of data: Data can start from different areas and can be developed/semi-built/unconstructed in various arrangements, spam, and so forth. Data from both of these sources should be homogeneously consolidated with the end goal that data to be remembered for the Big Data network become a typical and last wellspring of data. A component like ETL Extract, Transform, and Load is normal.

2) Improvement and Enrichment of data: Data is arranged from various sources and data are adjusted with other supporting sources with extra data, which is expanded with more information and likely subjectively gotten to the next level. Data from various strong sources are coordinated.

3) Transformation of data: There are different advances or sub-processes engaged with data handling, for example, assembling or gathering data from various sources, data should be reformatted, compacted, investigated, or changed by administrative requirements.

4) Reduction of data: Reduction of data is the component to minimize data volume to an un-redundant stage. This expects to increase the nature of data stockpiling and save costs by killing data that isn't required and safeguarding just the fitting components for this specific task.

5) Discretization of data: This method collects and divides data into intervals so that the available mining algorithms and techniques can be used effectively.

6) Cleansing of data: A strategy means to upgrade data effectiveness by dispensing with data that decreases data openness. The actions taken in this cycle are to dispense with the mistaken, missing, or immaterial data from gathered data so valuable information can be deciphered and assessed.

These pre-planning steps are critical to change the information to levels sensible or significant for assessment. Data ought to be sifted and transformed for what it's worth from organized, semi-organized, and unstructured sources. Pre-handling steps have prime significance to change the data to levels appropriate or important for investigation. Heterogeneous data ought to be accumulated with joins across source databases is the obligation of the Data conglomeration and capacity stage. The last stage is the Data Analysis stage. It implants sense and pertinence into consolidated data. This cycle is executed by contrasting data qualities with recognize designs.

Existing Methodologies

Big Data Quality Framework

Ashish Juneja (2019) et.al proposed Big Data Quality Framework: Pre-Processing Data in Weather Monitoring Application. Big Data has turned into an approaching piece of all enterprises and business areas today. All associations in any area like energy, banking, retail, equipment, organizing, and so on all create a tremendous quantum of heterogeneous data which whenever mined, handled and broke down precisely can uncover colossally valuable examples for business heads to apply to produce and develop their organizations. Big Data helps in obtaining, handling and analyzing a lot of heterogeneous data to determine important outcomes. The quality of information is impacted by the size, speed and arrangement in which data is produced. Thus, the Quality of Big Data is of incredible significance and significance. We propose addressing different parts of the crude data to further develop its quality in the pre-handling stage, as the crude data may not be utilized with any guarantees. We are investigating processes like Cleansing to fix as much data as practical, Noise channels to eliminate awful data, also sub-processes for Integration and Filtering alongside Data Transformation/Normalization. Propose a Pre-Processing Framework to address the quality of data in a weather conditions observing and determining application that likewise considers an Earth-wide temperature boost boundaries and raises cautions/notifications to caution clients and scientists ahead of time.

Automated data cleaning

Haider, S. N., Zhao (2020) et.al proposed automated data cleaning for data centers. Preprocessing the crude data is a basic stage in AI whose crucial goal is to set up a cleaned and mistake free data set for data scientific calculations. Changing crude data into clean data is an essential necessity in modern and business areas however there are many difficulties which must be tended to separately and physically. Since there is no brought together framework that integrates every one of the expected fields to change crude data into clean data, manual transformation is inadequate and very tedious. We examine a contextual investigation for cleaning data in the data community, contrasting missing qualities filling issues and conjecture and mean worth substitution for missing qualities and propose a mechanized data preprocessing framework for data cleaning. The proposed framework effectively cleans data sets consequently as opposed to managing numerous issues unmistakably and physically.

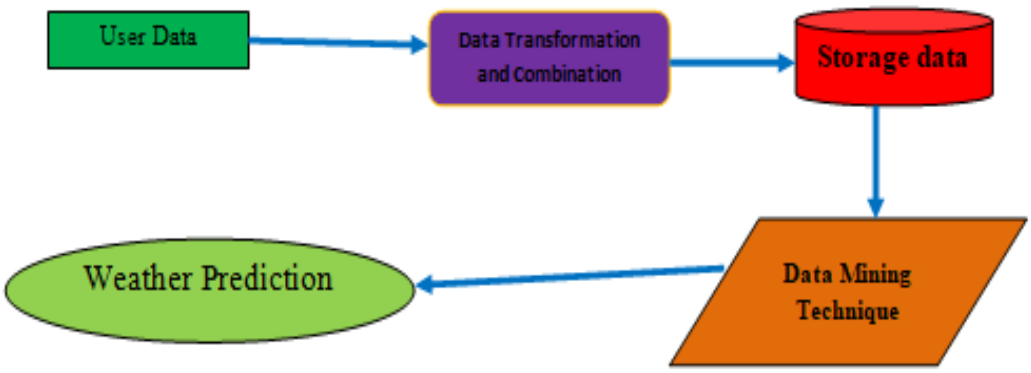
VIM

Arafat, S (2017) et.al proposed VIM: A Big Data Analytics Tool for Data Visualization and Knowledge Mining. With the headway of information innovations and applications, a bountiful measure of data is produced, which draws in both the research local area to use this information for removing information and the business for fostering an information based framework. Representation of data, design mining from datasets and analyzing data float for the various highlights are three exceptionally utilized applications of AI and data science fields. A conventional online instrument coordinated with such highlights will offer huge help for preprocessing the dataset and in this manner separating precise information. In this work, we propose such a data perception device, named VIM, which is an electronic exhaustive device for nonexclusive data representation, data preprocessing and mining reasonable information with float examination of data. In this work, we proposed our data perception, design mining and float examination device, VIM, which is a combination of the three most continuous applications of data mining and Big Data analytics. Our device is equipped for preprocessing any dataset as well as separating information from the dataset.

Proposed Methodology

The system predicts the future weather conditions based on current weather data. The data mining techniques namely Chi square test and Naïve Base statistics are applied on the dataset to extract the useful information from the dataset. The System Methodology shows in figure. 1:

Figure 1. Data preprocessing Flowchart



Data Collection and Preprocessing

The underlying stage in the data mining process is data assortment and preprocessing. The significant stage is data preprocessing on the grounds that main substantial data will yield exact result. The data is utilized in this task gathered from clients. However the data set contained many ascribes, the data preprocessing step thought about just the pertinent information, disregarding the rest. Then, at that point, data transformation is performed, into an organization, which is reasonable for Data mining. Four Attributes are utilized to distinguish the Weather Forecasting.

Naïve Bayes

Naïve Bayes is a wonderful general-learning algorithm for all fields of machine-learning and data analysis. The Naïve Bayes' presumption of attribute freedom is bad for business (no attribute linkage). In most cases, we assume that our attributes are independent of each other.

Meta-learning for Data Pre-processing

Meta-learning is a general interaction utilized for foreseeing the presentation (e.g., prescient exactness) of a calculation on a given dataset. A technique targets tracking down connections between dataset qualities and data mining calculations. This should be possible, since transformations, through the progressions they cause in the dataset qualities, influence the aftereffects of the data mining calculations. Utilizing meta-learning, we can realize this effect and we can rank transformations as per their capacity of working on the end-product of the data mining calculation.

Initial, a meta-learning space is laid out utilizing metadata. The metadata comprises of dataset attributes alongside some presentation measures for data mining calculations on those specific datasets. Then, the meta-learning stage produces a model (i.e., prescient meta-model) which characterizes the area of skill of the data mining calculation. At long last, when a changed dataset (i.e., a transformation was applied on the dataset) shows up, the dataset qualities are removed and taken care of to the prescient meta-model, which predicts the presentation of the calculation on the changed form of the dataset.

Meta-features

Data incorporation is consolidating data from multi sources to a steady distribution center. Various data can be consolidated by matching similar records, for instance, matching the essential key and unfamiliar key or

recognizing in light of sound judgment. In this cycle, there are dependably redundancies on the grounds that an equivalent item can have various qualities and ought to be taken out by connection examination.

Numerical correlation analysis uses a correlation coefficient (also called Pearson's product moment coefficient)

$$r_{A,B} = \frac{\sum(A - \overline{A})\,(B - \bar{B})}{(n-1)\sigma_A\sigma_B} = \frac{\sum(AB) - n\overline{AB}}{(n-1)\sigma_A\sigma_B}$$

Where n is the number of tupls, $\bar{A}\ and\ \bar{B}$ are the respective means of A and B, $\sigma_A and\ \sigma_B$ are the respective standard deviation of A and B, and then $\sum(AB)$ is the sum of the AB cross-product. If result is positive, A and B are positively correlated which means A.

There are three kinds of normalization methods.

Min-max normalization: $to\left[new_{\min_A} - new\_\min_A\right) + new\_\min_A$

$$v' = \frac{v - \min_A}{\max_A - \min_A}\left(new\_max_A - new_{min_A}\right) + new\_\min_A$$

$$Z - score\ normalization: where\ \mu\ is\ mean\ and\ \sigma\ is\ standard\ deviation$$

$$V' = \frac{v - \mu_A}{\sigma_A}$$

Normalization by decimal scaling: where j is the smallest integer such that $Max(|V^3|) < 1$

$$v' = \frac{v}{10}$$

SPSS Applications in Data Preprocessing (SPSSDP)

Data preprocessing is a significant stage in the KDD project not just on the grounds that it requires an extensive stretch of investment yet in addition on the grounds that the four tasks might be executed a few or no times in the genuine application and there is no sequence between techniques. Starting here of view, data preprocessing is not difficult to be completed by software like SPSS however needs insight and data mining information even another expert information if we have any desire to accomplish good to go data.

SPSS can choose part of the example in Data Editor Window as per the predefined need of a client and dissect just the chose data until the client drops the selection

- Missing Value Analysis estimates the missing value by analysis the internal relations and models of large data sets.
- SPSS analysis generates database Scores
- SPSS Base contains several mining products: Answer Tree, Clementine and Goldminer. Specific technical: Kohonen neural network, regression, factor analysis, decision trees, aggregation, association rules, rule induction, monotonic regression, OLAP environment.
- Clementine can find a model and convert into C language code.

Data calculation

The variable estimation is one of the most significant and broadly applied processes in data examination. It can produce new factors as per SPSS number juggling articulations and capacities are given by the client based on the first qualified cases. As the variable estimation is for all certified cases, each case has its own outcome and the outcome ought to be saved to a predefined variable. The variable data type ought to be reliable with the data kind

of the outcome determined. The variable estimation is one of the most significant and broadly applied processes in data examination. It can produce new factors as per SPSS number-crunching articulations and capacities given by the client based on the first qualified cases. As the variable estimation is for all certified cases, each case has its own outcome and the outcome ought to be saved to a predetermined variable. The variable data type ought to be reliable with the data kind of the outcome determined.

Experiment result

Specificity

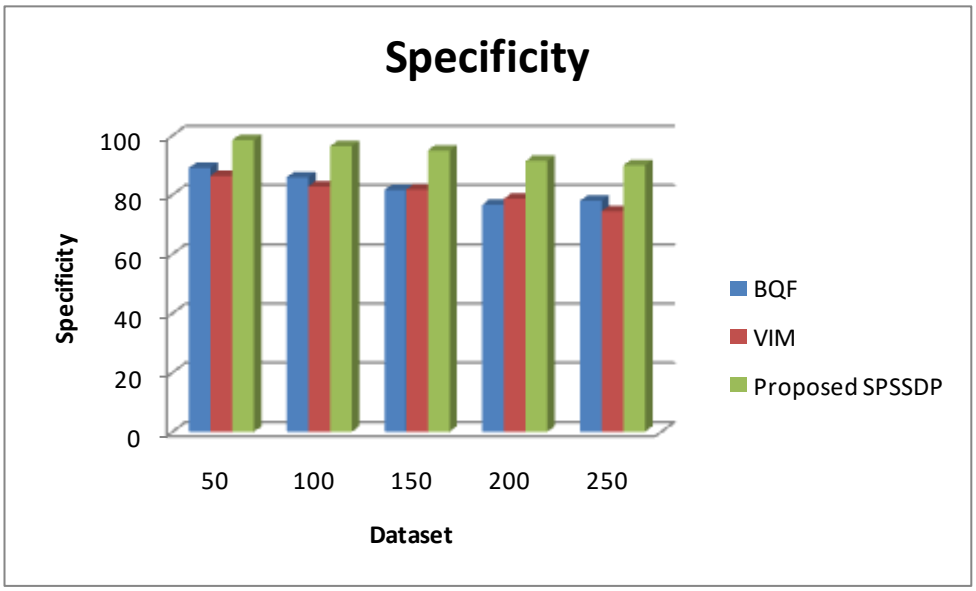
$$Speciicity = \frac{TN}{TN + FP}$$

Table 1. Comparison table for Specificity

Dataset	BQF	VIM	Proposed SPSSDP
50	89.04	86.37	98.45
100	85.89	82.82	96.36
150	81.52	81.74	94.91
200	76.65	78.65	91.38
250	77.91	74.42	89.92

The Comparison table 1 of Specificity Values explains the different values of existing BQF, VIM and proposed SPSSDP. While comparing the Existing algorithm and proposed Improved Relief Algorithm, provides the better results. The existing algorithm values start from 76 to 89, 74 to 87 and proposed SPSSDP values starts from 89 to 99. The proposed method provides the great results.

Figure 2. Comparison table for Specificity



The Comparison Figure 2 of Specificity Values explains the different values of existing BQF, VIM and proposed SPSSDP. While comparing the Existing algorithm and SPSSDP, provides the better results. X axis denote the Dataset and y axis denotes the Specificity ratio. The existing algorithm values start from 76 to 89, 74 to 87 and proposed SPSSDP values starts from 89 to 99. The proposed method provides the great results.

Sensitivity

This is the proportion of patients with diabetes, the positive instances, who are accurately recognized as being diabetic and it is computed as the proportion of true positives (TP) to the amount of TP and false negatives (FN).

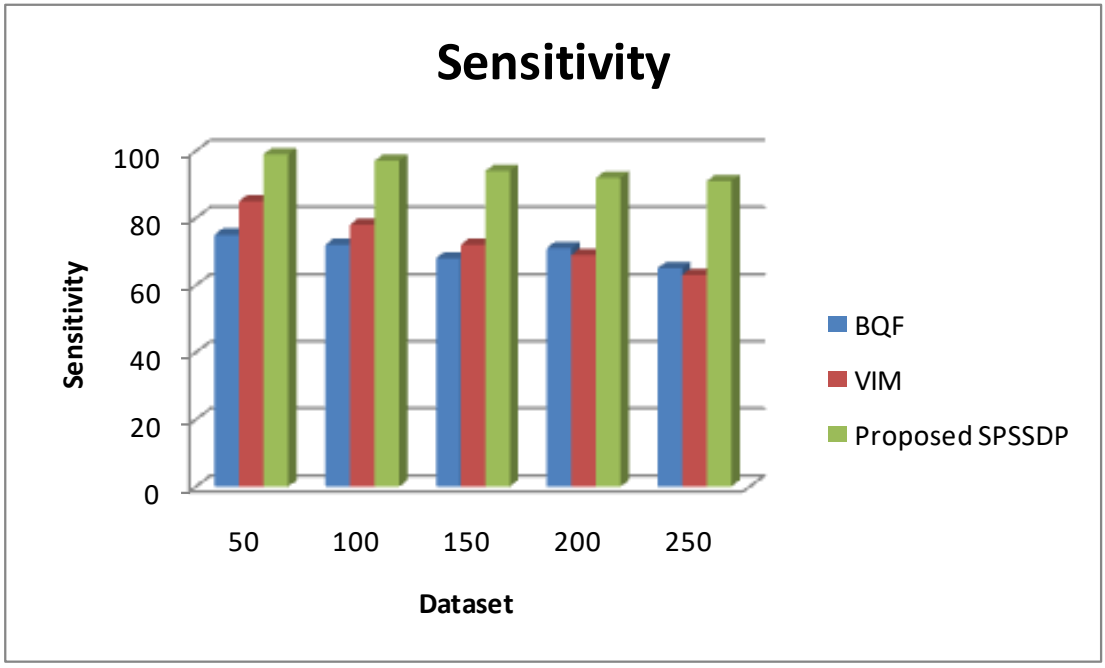
$$Sensitivity = \frac{TP}{TP + FN}$$

Table 2. Comparison table for Sensitivity

Dataset	BQF	VIM	Proposed SPSSDP
50	75	85	99
100	72	78	97
150	68	72	94
200	71	69	92
250	65	63	91

The Comparison table 2 of Sensitivity Values explains the different values of existing BQF, VIM and proposed SPSSDP. While comparing the Existing algorithm and proposed SPSSDP, provides the better results. The existing algorithm values start from 65 to 75, 63 to 85 and proposed SPSSDP values starts from 91 to 99. The proposed method provides the great results.

Figure 3. Comparison table for Sensitivity



The Comparison Figure 3 of Sensitivity Values explains the different values of existing BQF, VIM and proposed SPSSDP. While comparing the Existing algorithm and proposed SPSSDP, provides the better results. X axis denote the Dataset and y axis denotes the Sensitivity ratio. The existing algorithm values start from 65 to 75, 63 to 85 and proposed SPSSDP values starts from 91 to 99. The proposed SPSSDP provides the great results.

Precision

This is the proportion of patients with diabetes, the positive instances, who are correctly identified as being diabetic out of all the diabetic patients and is computed as the ratio of TP to the sum of TP and false positives (FP).

$$Precision = \frac{TP}{TP + FP}$$

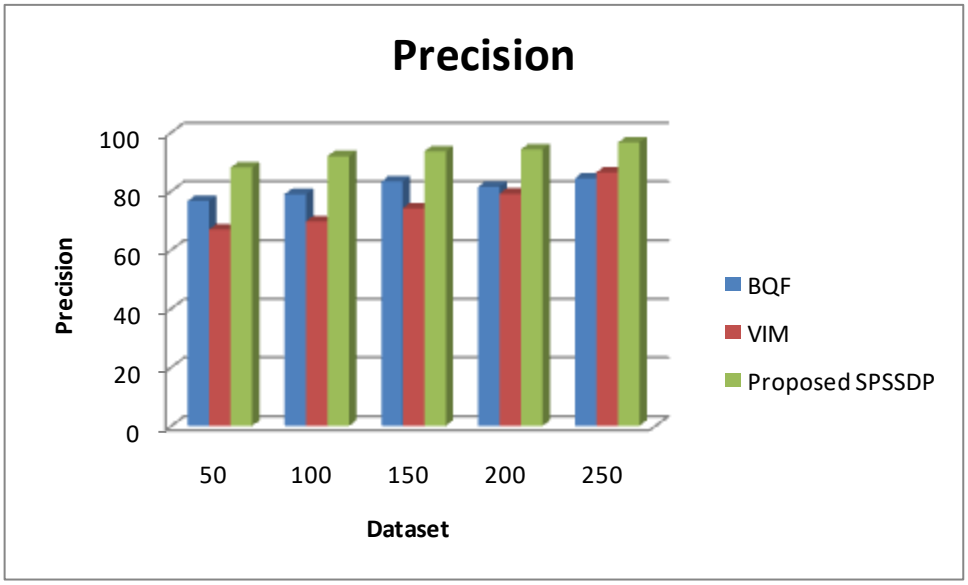
Table 3. Comparison table for Precision

Dataset	BQF	VIM	Proposed SPSSDP
50	76.63	66.94	88.01
100	78.91	69.66	91.87
150	83.26	74.12	93.48
200	81.45	79.09	94.23
250	84.33	86.38	96.52

The Comparison table 3 of Precision Values explains the different values of existing BQF, VIM and proposed SPSSDP. While comparing the Existing algorithm and proposed SPSSDP, provides the better results. The existing

algorithm values start from 66.94 to 86.38 and proposed SPSSDP values starts from 88.01 to 96.52. The proposed method provides the great results.

Figure 4. Comparison Chart for Precision



The Comparison figure 4 of Precision Values explains the different values of existing BQF, VIM and proposed SPSSDP. While comparing the Existing algorithm and proposed SPSSDP provides the better results. X axis denote the Dataset and y axis denotes the Precision ratio. The existing algorithm values start from 66.94 to 86.38 and proposed SPSSDP values starts from 88.01 to 96.52. The proposed method provides the great results.

Conclusion

Focused on basic concepts and procedures in data mining as well as the importance and tasks of data preprocessing. For the practical part, it is obvious that by using software SPSS, the four major tasks in data preprocessing including data cleaning, integration, transformation and reduction can be easily carried out with no need for programming and can even make some predictions before the actual data mining.

References

1. Agilan, V & Nanduri, U. (2016). Modelling nonlinear trend for developing non-stationary rainfall intensity–duration–frequency curve. International Journal of Climatology. 37. 10.1002/joc.4774.

2. Arafat, S. S. I., Hossain, M. S., Hasan, M. M., Imam, S. M. A.-H., Islam, M. M., Saha, S., Juthi, T. I. (2017). VIM: A Big Data Analytics Tool for Data Visualization and Knowledge Mining. 2017 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE). doi:10.1109/wiecon-ece.2017.8468939

3. Camacho, J., Rodriquez-Gomez, R. A., & Saccenti, E. (2017). Group-wise Principal Component Analysis for Exploratory Data Analysis. Journal of Computational and Graphical Statistics, 26(3), pp. 501-512, doi: 10.1080/10618600.2016.1265527

4. Darji, M. P., Dabhi, V. K., and Prajapati, H. B. (2015). Rainfall forecasting using neural network: A survey, 2015 International Conference on Advances in Computer Engineering and Applications, Ghaziabad, pp. 706-713, doi: 10.1109/ICACEA.2015.7164782.

5. Gantz, J.& Reinsel, D. (2012). The digital universe in 2020: Big data bigger digital shadows and biggest growth in the Far East. International Data Corporation.

6. Haider, S. N., Zhao, Q., & Meran, B. K. (2020). Automated data cleaning for data centers: A case study. 2020 39th Chinese Control Conference (CCC).

7. Hapsari RI, Sugan BAI, Novianto D, Asmara RA, Oishi S. Predictability of Naïve Bayes classifier for lahar hazard mapping by weather radar. In IOP Conference Series: Earth and Environmental Science. 2020;437(1):012049:IOP Publishing.

8. International Data Corporation (2019). The Growth in Connected IoT Devices Is Expected to Generate 79.4ZB of Data in 2025, According to a New IDC Forecast. Retrieved from https://www.idc.com/getdoc.jsp?containerId=prUS45213219

9. Juneja, A., & Das, N. N. (2019). Big Data Quality Framework: Pre-Processing Data in Weather Monitoring Application. 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon).

10. Kalyankar, M.A. & Alaspurkar, S.J. (2013). Data Mining Technique to Analyze the Metrological Data", International Journal of Advanced Research in Computer Science and Software Engineering 3(2), 114-118.

11. Najat N, Abdulazeez AM. Gene clustering with partition around mediods algorithm based on weighted and normalized Mahalanobis distance. In 2017 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS).2017;140-145:IEEE.

12. Prasetya R. Data mining application on weather prediction using classification tree, naïve bayes and K-nearest neighbor algorithm with model testing of supervised learning probabilistic brier score, confusion matrix and ROC. JAICT. 2020;4(2):25-33.

13. Sanjay Chakrabotry, Prof. N.K Nigwani and Lop Dey (2014), "Weather Forecasting using Incremental K-means Clustering", Vol. 8, 2014, pp. 142-147.

14. Saxena, A., Verma, N. & Tripathi K. C. (2013). A Review Study of Weather Forecasting Using Artificial Neural Network Approach. International Journal of Engineering Research & Technology (IJERT), 2(11).

15. Want, R., Schilit, B.N., & Jenson, S. (2015). Enabling the Internet of Things. Computer, 48(1), 28-35, doi: 10.1109/MC.2015.12.