Application and Comparison of Majority Weighted Minority Oversampling Techniques and Random OverSampling Examples Data Balancing Methods on the Vertebral Column Dataset

Mrs. M.PUSHPALATHA[1], N.INDIRA[2]
[1]Assistant Professor, Department of Computer Science, Padmavani Arts & Science College for Women,
[2]M.Phil scholar, Department of Computer Science, Padmavani Arts & Science College for Women.

In recent years, Data science has emerged as most conspicuous multidisciplinary field. The raw data that are gathered from multiple sources usually have issues like missing data, imbalance data, scaling, normalization and so on. Hence, practically almost all raw dataset need to be converted to more application suitable form. From these facts, it is apparent that data preprocessing is extremely necessary phase of any data analysis task. Despite the fact that not all preprocessing method need to be applied on a solitary dataset, but one or more could be required to form usable formatted dataset. Like any dataset, the Vertebral column dataset also isn't always similarly disbursed to the class label. So right here to solve the problem, in this work, two data balancing algorithms namely Majority Weighted Minority Oversampling Techniques (MWMOTE) and Random Over Sampling Examples (ROSE) applied. Furthermore, simulation results are generated using R Software and assessed estimating processing time. As a result, ROSE algorithm takes less processing time than MWMOTE to balance the data.

Keywords: Data Balancing, ROSE, MWMOTE.

I. INTRODUCTION TO DATA IMBALANCED

In the present time of AI and information mining, numerous genuine applications work on datasets mostly for performing investigation and creating suggestions and expectations. For playing out these counts the dataset ought to be appropriately balance however now and again it is seen that these datasets are unevenness in nature. Prompting the issue of imbalanced information [1]. The data which has an inconsistent appropriation of tests among classes is known as imbalanced information [2]. Class distribution, i.e., the proportion of instances belonging to each class in a data set, plays a key role in any kind of machine-learning and data-mining research. However, the real world data often suffer from class imbalance. The class imbalance case has been reported to exist in a wide variety of real-world domains, such as face recognition, text mining, software defect prediction and remote sensing [3]. The Imbalanced information ordinarily alludes to a characterization issue where the quantity of perceptions per class isn't similarly dispersed. The Figure 1 derives as a lot of information/perceptions for one class (alluded to as the dominant part class), and many less perceptions for at least one different classes (alluded to the minority class) is known as imbalanced data. The balanced data is defined to the class labels distribution is equal [4].

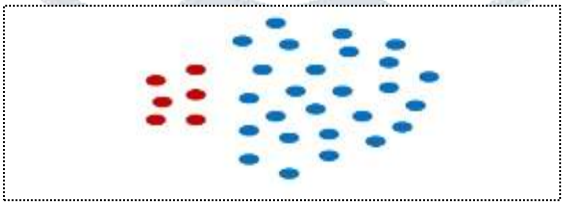

Figure 1: The Data Imbalance Distribution

There are three main approaches are used imbalanced data,

• Data-level methods that modify the collection of examples to balance distributions and/or remove difficult samples.

•Algorithm-level methods that directly modify existing learning algorithms to alleviate the bias towards majority objects and adapt them to mining data with skewed distributions.

• Hybrid methods that combine the advantages of two previous groups [5], [6].

There are many domains dealing with imbalanced data sets, such as, medical diagnosis, network monitoring, fraud detection, risk management, helicopter gear-box fault monitoring, earthquakes and nuclear and explosions, text classification, oil spills detection and so on[7].

II. LITERATURE REVIEW

Table 1 provides an idea about the related work to it their merits and demerits.

Table 1: Represents to the some methods merits and demerits

S.NO	AUTHOR AND YEAR	METHOD	MERITS	DEMERITS
1.	Mayuri S. Shelke, Dr. Prashant R. Deshmukh, Vijaya K. Shandilya, 2017 [2]	MLP,MWMOTE	After balancing the datasets, the classifiers give more proper and accurate results.	Performances measures of MLP and MWMOTE have different result
2.	Wei Feng , Wenjiang Huang and Jinchang Ren, 2018, [3]	Bagging and margin algorithms	The unsupervised margins achieve better performance.	Bagging achieves poor performances
3	Apurva Sonak, R.A.Patankar, 2015,[4]	Oversampling, under sampling, cost sensitive	Over and under sampling perform well in small dataset. The cost sensitive perform well in large dataset.	Processing time high data loss misclassification
4	Ms. Monica. Ochani, Dr.S.D. Sawarkar, Mrs. Swati Narwane [2019] [6]	Data level approaches, algorithm level approaches	It is very important to balance the imbalance data with effective techniques and at the same time, cost factor should be given attention. The correct classifier techniques and performance evaluation metrics must be applied to achieve good results.	Is not using practical implementation of merits and demerits approaches
5	Adnan Amin, Sajid Anwar, Awais Adnan, Muhammad Nawaz,Newton Howard, Junaid Qadir, Ahmad Hawalah, And Amir Hussain, [2016], [8]	Oversampling algorithms	Cov and LEM2 algorithms achieved higher accuracy	Results to not follow full objects coverage
6	Sotiris Kotsiantis D. Kanellopoulos P. E. Pintelas 2006, [9]	Boosting Algorithms	The minority class is better represented by a larger number of examples.	Poorly performed on small datasets
7	Jinyan Li, Lian-sheng Liu, Simon Fong, Raymond K. Wong, Sabah Mohammed, Jinan Fiaidhi, Yunsick Sung, Kelvin K. L. Wong, 2017, [10]	PSO balancing, Adaptive PSO balancing, BAT balancing, Adaptive balancing	Adaptive Swarm Balancing Algorithms are quick and better solve the imbalance problem of the dataset	The Adaptive Swarm Balancing Algorithms can effectively solve in the large datasets.
8	Chakkrit Tantithamthavorn, Ahmed E. Hassan, and Kenichi Matsumoto, 2018,[11]	ROSE, SMOTE, oversampling, undersmpling	The accuracy of machine learning dataset is improved after apply smote	SMOTE does not perform well with high dimensionality data
9	S.Jayasree, A.AliceGavya 2015,[12]	SSO,MWMOTE	Avoid data loss	The oversampling method causes the over fitting affect the performance
10	Amin Naboureh, Ainong Li, Jinhu Bian, Guangbin Lei and Meisam Amani , 2020, [13]	Hybrid approach	Higher accuracies achieves for both minority and majority class ratio	ROS,RUS,SMOTE,G-SMOTE performance is low

11	Shaheen Layaq, B. Manjula, 2020[14]	Data based and ensemble based approach	Ensemble learning achieves high accuracy	ROU achieves low accuracy
12	Zhongbin Sun, Qinbao Song , Xiaoyan Zhu , Heli Sun , Baowen Xu, Yuming Zhou [2014] [15]	conventional sampling methods, cost-sensitive learning methods, and Bagging and Boosting based ensemble methods,	The proposed method firstly converts the imbalanced binary class data into multiple balanced binary-class data. This is achieved by applying random splitting or clustering to the majority class instances. After that, a specific classification algorithm is applied to the multiple balanced binary-class data to build multiple classifiers.	The conventional sampling methods provide lower performance
13	J. Hoyos-Osorio, A. Alvarez-Meza, G. Daza-Santacoloma , A. Orozco-Gutierrez , G. Castellanos-Dominguez [2021] [16]	RIUS,CRIUS, RUS1,UB4,SBAG4,CUS-AB	RIUS and CRIUS subsample the majority class without losing the underlying structure	RUS1,UB4,SBAG4,CUS-AB achieves poor performance compared to the RIUS AND CRIUS

III. DATA COLLECTION

The dataset that used in this research was extracted from Kaggle website .This dataset has 310 patients' records which classify two classes every patient records is represented as a pattern with 12 attributes, according to the following physical parameters such as pelvic incidence, pelvic tilt, lumbar lordosis angle, sacral slope, pelvic radius, degree spondylolisthesis, pelvic slope, direct tilt, thoracic slope, cervical tilt, sacrum angle and scoliosis. The vertebral column dataset contains the patients classified into one of two categories: Normal (100 patients) or Abnormal (210 patients).

IV. IMPLEMENTATION OF DATA BALANCING ALGORTIHMS WITH RESULT

a) Oversampling techniques

Oversampling is one of the records degree approach. The oversampling methods replicating the minority samples so that the distribution is equal and the statistics is balanced. Random oversampling tries to stability category distribution by using randomly replicating minority classification instances. However, quite a few authors agree that this technique can enlarge the possibility of overfitting occuring, on account that it makes specific copies of present instances. Synthetic Minority Over-sampling Technique (SMOTE), The most famous over-sampling method. It's essential thought is to create new minority classification examples by using interpolating quite a few minority type cases that lie together [7].

In Vertebral Column Dataset class imbalance occurs as the Abnormal label has 68% and Normal label has 32% of data here improve the count of class label Norma is required for better classification performance.

- MWMOTE
- ROSE.

The Figure2 describes to the standard format data have been computed data balancing in oversampling techniques such as random oversampling algorithm (ROSE – Random Over Sampling Examples) and synthetic oversampling algorithm (MWMOTE – Majority Weighted Minority Oversampling Techniques) these two ways the data is balanced and these are generating the new samples data is differed from one to another.

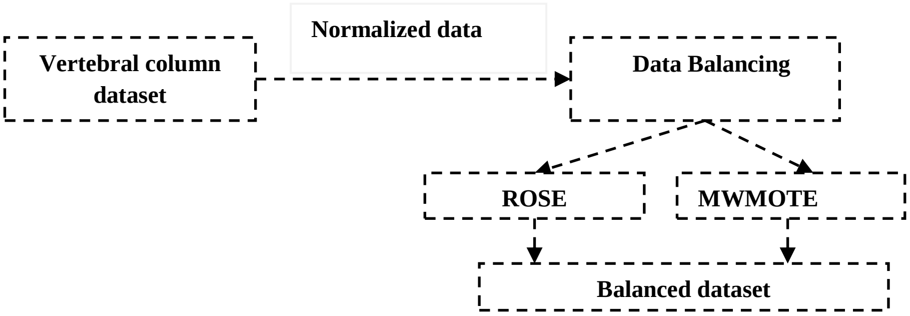

Figure 2: Overview of The Data Balancing

b) Majority Weighted Minority Oversampling Technique (MWMOTE)

SMOTE is a basic algorithm which generates new examples by means of filling empty areas amongst the nice instances. It has a most important setback even though it does no longer notice noisy instances. Therefore it can generate artificial examples out of noisy ones or even between two minority classes, which if no longer cleansed up, can also stop up turning into noise inner a majority classification cluster. Modification for SMOTE method which overcomes some of the troubles of the SMOTE method when there are noisy instances, in which case SMOTE would generate extra noisy situations out of them. MWMOTE (Majority Weighted Minority Oversampling Technique) tries to overcome each problems. It intends to provide greater weight to borderline instances, undersize minority cluster cases and examples close to the borderline of the two classes.

The Figure3 indicates MWMOTE identifies the most vital and difficult to examine minority category samples from the unique minority set, Xmin (Normal) and assemble a set Ximin by way of the recognized samples. Each member of Ximin is given a choice weight in accordance to the significance in the data. MWMOTE generates the artificial samples from Ximin the usage of the weights and produces the output set samples through including the artificial samples to the Xmin.

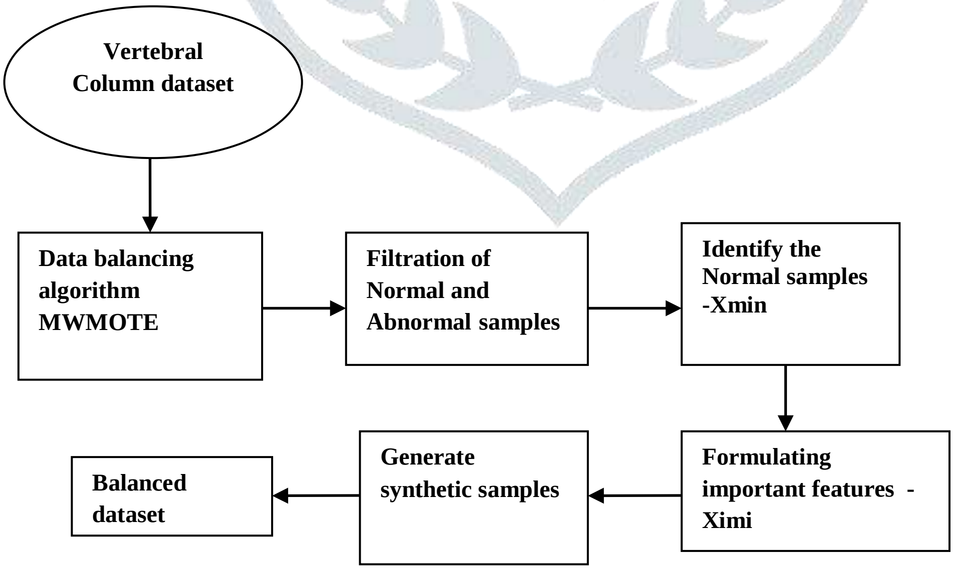
Vertebral Column dataset
Data balancing algorithm MWMOTE
Filtration of Normal and Abnormal samples
Identify the Normal samples -Xmin
Formulating important features - Ximi
Generate synthetic samples
Balanced dataset

Figure 3: MWMOTE Data Balancing Process

An artificial oversampling technique Majority Weighted Minority Oversampling Technique (MWMOTE) is used for oversampling the imbalance data, whose aim is to alleviate the troubles of imbalanced gaining knowledge of and generate the beneficial artificial minority classification samples. In oversampling method, new samples are delivered to the minority type in order to stability the records set. For oversampling ordinarily two strategies are used referred to as random oversampling and artificial oversampling. In random oversampling method, present minority samples are replicated in order to amplify the measurement of a minority class. But right here in this method there is possibilities of essential samples will become uncommon and much less essential attributes may additionally be replicated. That's why, this paper makes use of the approach of MWMOTE based totally on artificial oversampling. In this approach artificial samples are generated for the minority type samples. These new samples add the integral facts to the minority type and prevents its situations from the misclassification.

The MWMOTE algorithm solely will increase the minority pattern classification statistics from the unique dataset does now not exchange the majority type samples it is represented through determine.

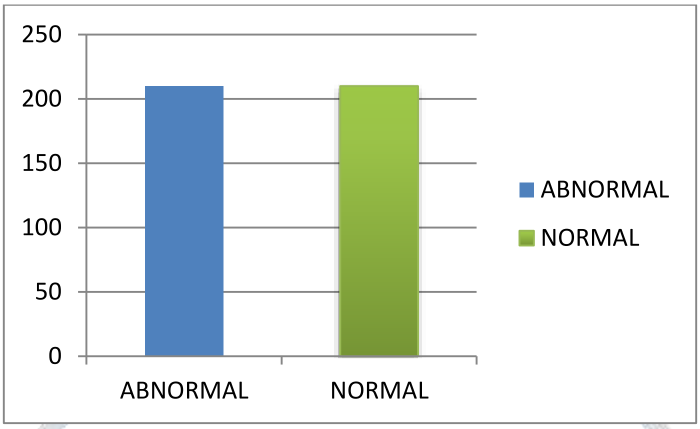

Figure 4: To View MWMOTE Data Balancing

Here, the minority pattern classification (Normal) weights are make bigger a hundred and ten from the authentic Vertebral Column dataset. After the use of this algorithm completely the dataset cases are 420 the category label Normal had 210 situations and the Abnormal type label is the 210 cases these are now not modified it is the authentic statistics value.

c) Boostrap Random Over-Sampling Examples Technique (ROSE)

The ROSE approach makes use of a smoothed-bootstrapping method to draw synthetic samples from the characteristic house neighbourhood round the minority class. ROSE combines oversampling and undersampling via producing an augmented pattern of the facts (especially belonging to the uncommon class). It builds on the era of new synthetic examples from the classes, in accordance to a smoothed bootstrap method Consider a coaching set Tn, of dimension n, whose typical row is the pair (xi , yi), i = 1, . . . , n. The type labels yi belong to the set {Y0, Y1}, and xi are some associated attributes supposed to be realizations of a random vector x defined on R d , with an unknown chance density characteristic f(x). Let the quantity of gadgets in classification Y j , j = 0, 1, be denoted by way of nj & lt; n.

The ROSE system for producing one new synthetic instance consists of the following steps:

- Select y ∗ = Yj with likelihood πj .
- Select (xi , yi) ∈ Tn, such that yi = y ∗ , with likelihood 1/nj .
- Sample x ∗ from KHj (•, xi), with KHj a likelihood distribution situated at xi and covariance matrix Hj.

Essentially, we draw from the education set an commentary belonging to one of the two classes, and generate a new instance (x*, y*) in its neighborhood, the place the form of the local is decided with the aid of the structure of the contour units of K and its width is ruled through Hj . It can be without problems proven that, given resolution of the classification label Yj , the era of new examples from Yj , in accordance to ROSE, corresponds to the technology of records from the kernel density estimate of f(x|Yj), with kernel K and smoothing matrix Hj. The options of K and Hj may additionally be then addressed with the aid of the massive specialised literature on kernel density estimation. It is rewarding to be aware that, for Hj → 0, ROSE collapses to a preferred mixture of over- and under-sampling. The determine figure5 indicates that the under-sampling approach (down-sampling) randomly samples (reducing) the majority type in order to limit the variety of majority modules to be the equal quantity as the minority type (e.g., faulty class). Oversampling-The over-sampling approach (up-sampling) randomly samples with substitute (i.e., replicating) the minority classification (e.g., faulty class) to be the identical dimension as the majority category (e.g., easy class). Resampling the records of the minority type the usage of a bootstrap resampling approach to repeat modules of the minority category to a defective ratio of 50%.

.

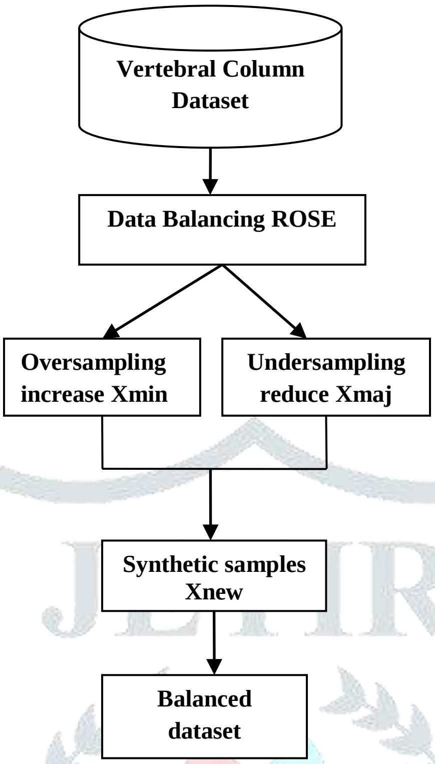

Figure 5: ROSE Data Balancing Process

To generate the new artificial records samples for each minority and majority samples now the dataset is balanced represents the figure6 Rose methods utilized the Vertebral Column dataset the first step resampling the Abnormal type label (majority) the usage of the boostrap resampling strategies to dispose of the Abnormal type to faulty 50% ratio (Undersampling). Resampling the Normal classification the use of boostrap resampling approach to repeat the modules of Normal category to faulty ratio 50% (Oversampling).

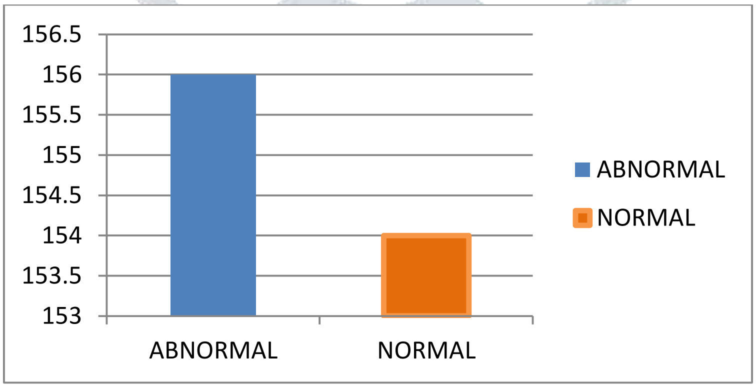

Figure 6: Rose Data Balancing

Generating new artificial pattern of information now the statistics are balanced however these are about equal measurement of the unique dataset the classification Normal 156 situations and the Abnormal 154.

V. CONCLUSION AND FUTURE ENHANCEMENT

The MWMOTE algorithm generates totally 420 samples and time taken is 1.02 sec. The ROSE algorithm generates 310 samples same as original dataset but the class label count changes and the build time is 0.04 sec. Comparison of these two algorithms signifies that the ROSE algorithm takes less processing time than the MWMOTE. These balanced data improves the process of feature selection and classification.

REFERENCES

[1] Miss. Mayuri S. Shelke, Dr. Prashant R. Deshmukh, Prof. Vijaya K. Shandilya, “A Review on Imbalanced Data Handling Using Undersampling and Oversampling Technique” International Journal of Recent Trends in Engineering & Research (IJRTER) Volume 03, Issue 04; April - 2017 [ISSN:2455-1457].

[2] Mayuri S. Shelke, Dr. Prashant R. Deshmukh, Vijaya K. Shandilya, “Efficient Imbalanced Data Handling Techniques through Undersampling and Oversampling Approach” International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue 4, April 2017.

[3] Wei Feng, Wenjiang Huang and Jinchang Ren, “Class Imbalance Ensemble Learning Based on the Margin Theory” www.mdpi.com/journal/applsci , may 2018.

[4] Apurva Sonak, R.A.Patankar, “A Survey on Methods to Handle Imbalance Dataset” International Journal of Computer Science and Mobile Computing, Vol.4 Issue.11, November- 2015.

[5] Bartosz Krawczyk, “Learning from imbalanced data: open challenges and future directions”, This article is published with open access at Springerlink.com, 2016.

[6] Ms. Monica. Ochani, Dr.S.D. Sawarkar, Mrs. Swati Narwane, “A novel approach to handle class imbalance:A Survey”, International Journal of Engineering Development and Research 2019 Volume 7, Issue , ISSN: 2321-9939.

[7] Mohamed Bekkar and Dr. Taklit Akrouf Alitouche,**“** IMBALANCED DATA LEARNING APPROACHES REVIEW**”,** International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.4, July 2013.

[8] Adnan Amin, Sajid Anwar, Awais Adnan, Muhammad Nawaz, Newton Howard, Junaid Qadir, Ahmad Hawalah, And Amir Hussain, “Comparing Oversampling Techniques To Handle The Class Imbalance Problem: A Customer Churn Prediction Case Study”, IEEE Access,2016.

[9] Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas, “Handling imbalanced datasets: A review”,GESTS International Transactions on Computer Science and Engineering, Vol.30, 2006.

[10] Jinyan Li, Lian-sheng Liu, Simon Fong, Raymond K. Wong, Sabah Mohammed, Jinan Fiaidhi, Yunsick Sung, Kelvin K. L. Wong, “Adaptive Swarm Balancing Algorithms for rare-event prediction in imbalanced healthcare data” PLOS ONE https://doi.org/10.1371/journal.pone.0180830 July 28, 2017.

[11] Chakkrit Tantithamthavorn, Ahmed E. Hassan,and Kenichi Matsumoto, “The Impact of Class Rebalancing Techniques on the Performance and Interpretation of Defect Prediction Models”, arXiv:1801.10269v1 [cs.SE] 31 Jan 2018.

[12] S.Jayasree, A.AliceGavya, “Classification of imbalance problem by MWMOTE and SSO” International Journal of Modern Trends in Engineering Science, Volume 02 Issue: 5, 2015, ISSN: 2348-3121.

[13] Amin Naboureh, Ainong Li, Jinhu Bian, Guangbin Lei and Meisam Amani, “A Hybrid Data Balancing Method for Classification of Imbalanced Training Data within Google Earth Engine: Case Studies from Mountainous Regions”, www.mdpi.com/journal/remotesensing , Remote Sens. 2020, 12, 3301; doi:10.3390/rs12203301.

[14] Shaheen Layaq, B. Manjula, “A Recapitulation of Imbalanced Data”, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-3, January 2020.

[15] Zhongbin Sun, Qinbao Song , Xiaoyan Zhu , Heli Sun , Baowen Xu , Yuming Zhou, “Novel ensemble method for classifying imbalanced data” 0031-3203/& 2014 Elsevier Ltd. All rights reserved.

[16] J. Hoyos-Osorio , A. Alvarez-Meza ,G. Daza-Santacoloma , A. Orozco-Gutierrez, G. Castellanos-Dominguez , “Relevent information understanding to support imbalanced data class” 0925-2312/ 2021 Elsevier B.V. All rights reserved.