Journal of Theoretical and Applied Information Technology
15th February 2024. Vol.102. No 3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

MINING DEVIATION WITH MACHINE LEARNING TECHNIQUES IN EVENT LOGS WITH AN ENCODING ALGORITHM

DR. V.V. JAYA RAMA KRISHNAIAH[1], DR BANDLA SRINIVASA RAO[2], DUGGINENI VEERAIAH[3], MR. S. SUBBURAJ[4], DR. MOHAMMED SALEH AL ANSARI[5], DR. CHAMANDEEP KAUR[6]

[1]Associate Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India
[2]Professor in CSE, Teegala Krishna Reddy Engineering College, Hyderabad.
[3]Professor, Department of CSE, Lakireddy Bali Reddy College of Engineering, Mylavaram, NTR District, Andhra Pradesh-521230, India
[4]SRMIST Ramapuram Campus, Chennai
[5]Associate Professor, College of Engineering, Department of Chemical Engineering, University of Bahrain, Bahrain.
[6]Lecturer, Computer Science & Information Technology Department, Jazan University, Jizan, Kingdom of Saudi Arabia
[1]jkvemula@gmail.com, [2]sreenibandla@gmail.com, [3]veeraiahdvc@gmail.com, [4]subburajs87@gmail.com, [5]malansari.uob@gmail.com, [6]kaur.chaman83@gmail.com

ABSTRACT

A study field called "commercial operation divergence analysis" tries the identify how a commercial system varies beyond the results which were anticipated. Approaches in this field identify the qualities of a collection of procedure executes may are related to shifts in process efficiency, exposing the characteristics of procedure behaviours which generate undesirable procedure for the process as well as understandings of what process behaviour contributes the greatest efficiency. Success in this scenario can relate to any domain-dependent efficiency measurement as well as the expense, time, and resources factors. The finding of trends from the logs of events using various trend mining approaches is the basis of the present-day company deviation mining methodologies. Such extractor methods are now provide to a small degree of flexibility because they're unable to represent the intricate linkages which could be present in highly variable systems. Within this paper, then offer an innovative decoding strategy for vector-based representations of procedure scenarios, followed by then utilise Machine Learning for a unique approach in the context of Deviance Mining to pinpoint the aspects of a procedure which most significantly affect its efficiency. The outcomes demonstrated how machine learning delivered pertinent as well as emotive conclusions on the event's logs when combined with the suggested Declare-based coding, constituting an effective tool for the analysis of processes.
Keywords: *Machine Learning, Mining Deviation, Encoding, Logs*

1. INTRODUCTION

Effective extraction operations depend on the deep mine's ability to maintain a healthy and secure air atmosphere [1]. A crucial step in the analysis of data is outlier detection. Hawkins states that atypical is "a thing whether diverges sufficiently form other items as to be assumed that it has been produced by an alternate method" [2]. In 2008, the Global Financial Crisis (GFC) and the demise of the coal mining "super cycle" put a stop to a period of production-focused tactics during which operational costs increased faster than output [3]. Because it presents especially challenging compromises the extractive and material extraction sector is a desirable test case for the study of contamination. Individual plants may provide enormous value, up to millions of dollars annually [4]. Studies had lately claimed that the application of Process Mining (PM) might address these drawbacks through enabling auditors to efficiently and primarily automatically analyse all of the databases employing historic &/or present-day information [5]. Nevertheless, a number of issues brought on by the extraction and use of coal assets, including as sinkholes, erosion of soil, landslides, and the demolition of buildings, have had a significant detrimental impact on the daily lives and assets of local populations [6].

Mining processes is a field of study who tries to enhance process enhancements by offering based on reality observations on previous procedure

implementations. The topic sits among system modelling and evaluation and intelligence computing as well as data mine on one's hand. Process variation assessment is described as a collection of methods that allow to contrast more than one event records belonging to various company procedure versions for the purpose to identify the differences between them [7]. A prime instance of contaminated soil include the soils that make up anthracite mine dumps. The sedimentary layers that cover a coal seam are where where the initial soil was formed. The excess soil is typically excavated using various excavators, then delivered into the spoil site via lorries or belt conveyors and deposited form different heights, either with or no choosing the material [8].

Multiple research studies indicate that these last class of computations, machine learning algorithms (MLAs), can be more accurate than statistical methods like discriminant evaluation or logistic regression, particularly if the feature space to be studied is complicated (i.e., once the dimension of the input feature time is believed to be quite large and the connection between the intended contributions along with the feedback transparent include is predicted to be non-linear) and the data sets being used are anticipated to include distinct characteristics [9]. One the contrary, machine learning is a branch of computing which seeks to give machines or different gadgets the capacity to understand sans needing directly controlled. It tries to provide methods and mathematical models for data-driven learning and forecasting. Upon accomplishment, machine learning techniques are used to simulate characteristics of the input in relation to anticipated result, predict production attributes in relation to past information, and characterise the behaviour within the data. A possible approach to predicting wind power using velocity data is machine learning techniques [10]. Machine learning has been immensely successful as information quantities and types have increased because of its ability to examine complex trends in seen information and generate predictive models or choices on fresh data. In the literature, a variety of machine learning methods and algorithms have been published [11].

Predicting how a business operation will behave in the years to come is an essential corporate competence. Procedure prediction, a form of statistical analysis used in management of business processes, uses information from previous process occurrences to forecast future ones [12]. Customer service representatives adjusting to requests about the amount of time left until an issue has been settled are a few examples of use instances. Other use cases include production managers forecasting the length of a manufacturing procedure for improved scheduling and higher utilisation or case supervisors determining probably violations of regulations to reduce business risk [13]. One kind of procedure mining work, called procedure learning, looks for a model that describes the behaviour of an organization's process using information about how it has previously been executed. The log of events is mapped onto a procedure model using a method known as a process identification procedure, which guarantees the model in question is a good representation of the behaviour shown in the event log [14].

This research pioneers the exploration of mining deviations in event logs through the innovative application of machine learning techniques, coupled with an advanced encoding algorithm. By addressing the nuanced challenges associated with identifying and understanding deviations within complex event data, this study seeks to contribute novel insights to the field. The proposed approach not only advances the capability to detect deviations effectively but also introduces an encoding algorithm that enhances the interpretability of the mined patterns. The work's scope extends to providing a more comprehensive understanding of deviations in event logs, offering valuable implications for anomaly detection, process optimization, and the broader domain of data-driven decision-making. Through this research, we aim to create a foundation for improving the efficiency and accuracy of deviation detection in diverse domains, fostering advancements in both theoretical frameworks and practical applications of machine learning in event log analysis. The key contribution are as follows:

- Introduction of a novel decoding strategy for vector-based representations of procedure scenarios in commercial operation divergence analysis, addressing limitations in capturing intricate linkages within highly variable systems.
- Pioneering a unique approach by integrating Machine Learning with Declare-based coding, enhancing the flexibility of extraction methods and enabling the identification of aspects significantly impacting process efficiency.
- Demonstration of the effectiveness of the proposed methodology in deviance mining, showcasing how Machine Learning, combined with Declare-based coding, provides pertinent and emotive conclusions

from event logs, serving as a powerful tool for process analysis.

- Significantly advancing the field by offering a more nuanced and effective means of understanding and addressing deviations in commercial operations, contributing to improved efficiency assessments and resource allocation in diverse domains.

The introduction establishes the importance of commercial operation divergence analysis, introducing a novel decoding strategy with Machine Learning. Related works highlight gaps in current methodologies, motivating the need for an innovative approach. The paper addresses the problem statement, presents a novel methodology integrating Machine Learning with Declare-based coding, and discusses empirical results and implications in the result and discussion section, concluding with advancements in process deviation analysis.

2. RELATED WORKS

Richetti et al. [15] proposed to determine the aspects of a procedure which most affect its efficiency, they first use Treatment Learning as an original method in the realm of Deviation Mining. This is a novel encoding method enabling vector-based representations of process occurrences. The suggested encoding method may find more expressive solutions since it is built on Declaring restriction framework fulfilment. Using publicly accessible logs of events from actual procedures, they do a number of tests that contrast our suggestion to the state-of-the-art activity decoding methods. The findings demonstrated that behavioural learning offered actionable and more descriptive insight from events logs when combined with our suggested Declare-based encoding, making it a useful tool for the analysis of processes.

Al-Shehari et al. [16] proposed the use of feature resizing and quick encoding strategies are used in the framework to alleviate the potential skew of identification outcomes that might emerge from an ineffective decoding procedure. The artificial minority sampling too much method (SMOTE) is additionally employed to alleviate the data set's balance problem. In order to discover a highly precise classification which can identify data leakage events carried out by malevolent outsiders throughout the crucial time when they depart an organisation, renowned machine learning methods are used. By applying our mathematical framework on the CMU-CERT Insider Threat Dataset and contrasting its results with the real world, we demonstrate the notion behind it. The results of the experiment demonstrate that our framework outperforms other methods which have been evaluated on the identical data in terms of detecting internal leakage of information events, with an AUC-ROC value of 0.99. The suggested framework offers practical approaches to deal with potential bias and class imbalance concerns in order to design a system that effectively detects insider data leaking.

Roldán et al. [17] proposed an approach that uses technologies like augmented reality and data mapping to teach workers in assembly operations. Firstly, skilled employees do assembly in accordance with their knowledge using a fully immersive environment. The next step is to use process mining methods to extract assemble model in the logs of events. Lastly, to understand the groups what the expert employees incorporated into the framework, learner employees utilise an improved immersion display with suggestions. Construction block experiments were designed as a toy example, and studies on a group of participants have been conducted. The outcomes demonstrate the suggested education system's competitiveness against more traditional options. It bases itself on procedure mining and mixed reality. In terms of mental effort, vision, learning, outcomes, and how they perform, user ratings are also superior.

Helm et al. [18] proposed 38 procedure mining instances related to health care reported from 2016 to 2018 that discussed the instruments, methods, and methodologies used as well as specifics on how the log data were found to have been medically significant. Utilising the common clinical coding schemes SNOMED CT and ICD-10, researchers then connected the diagnostic characteristics of the patient encounter setting, clinical speciality, and diagnosis of illness. The possible results of utilising a standardised method for categorising medical terms and events log data using common clinical codes are also highlighted.

Weinzierl et al. [19] proposed several prospective business process monitoring (PBPM) strategies that attempt to forecast potential process behaviours while the procedure is being executed. Methods for predicting subsequent event in particular have considerable promise for enhancing practical company processes. Many of these methods use deep neural networks (DNNs) and take into account data pertaining to the environment where the operation is occurring to provide recommendations that tend to be more reliable. Nevertheless, an in-depth analysis of such methods is lacking in the PBPM literature, making it difficult for academics and industry professionals to decide

which approach is appropriate for a particular event log. To address this issue, they statistically assess the prediction performance among three potential DNN structures using five tried-and-true encoding methods and five context-rich real-world logs of events. They offer four conclusions that might aid researchers and practitioners in developing fresh PBPM methods for anticipating upcoming actions.

3. PROBLEM STATEMENT

The presented literature outlines various approaches in the domain of data-driven analysis, ranging from innovative encoding methods in deviation mining to feature resizing and quick encoding strategies for data leakage detection, augmented reality and data mapping in assembly operations training, standardized categorization of medical terms in healthcare procedure mining, to prospective business process monitoring strategies utilizing deep neural networks. However, a clear problem statement emerges regarding the need for a comprehensive understanding of the strengths, weaknesses, and comparative efficacy of these diverse methodologies. This prompts the formulation of research questions addressing the gaps in existing knowledge, such as the effectiveness of Declare-based encoding in deviation mining, the impact of feature resizing on data leakage detection, the practicality of augmented reality and data mapping in assembly operations training, the benefits of standardized clinical coding in healthcare procedure mining, and the comparative analysis of deep neural network structures in prospective business process monitoring. Addressing these questions will provide a nuanced understanding of the applicability and limitations of these methodologies across different domains, fostering informed decision-making for researchers and industry professionals in their pursuit of effective data-driven solutions.

4. REGARDING DISCOVERY AND DECLARATIVE PROCESS MODELLING

Conventional urgent process diagrams are produced by the majority of mining process methods. These methods work effectively for organised processes since there aren't numerous additional ways an operation may be carried out. Declarative language modelling is suggested as a way to create an improved equilibrium amongst flexibility and guiding support for these types of models, despite the fact that many of these approaches are capable of handling event logs form flexible or unorganised models. Due to expressive modeling's relevance to log files from dynamic or unstructured processes, the potential of mining declarative models has also emerged.

Declaring continues to be the most commonly employed languages for studies regarding declaratory modelling and mineral extraction, although having very little application in business. This is because it's versatile and particularly suited for use in extremely volatile procedures, which are characterised by extreme complexity and variety. The addition enables associations among actions taken upon KiPs to be described using domains limitations as opposed to sequential ordering. Additionally, it enables occurrences in a KiP to signal chronological ties, behavioural consistency restrictions, or choice-of-action relationships in its instances by using these extra notions.

4.1. Deviance mining with machine learning and declare-based encoding of event logs

Machines are the simplest collection of rules that may be used in machine learning to discriminate between circumstances that include numerous highly weighted classes and scenarios with few strongly weighed categories. Machine learning, within contrast to association-rule mineral extraction, specifies a preferred type worth, that serves as a benchmark for weighing various class values and allows it to highlight machines with strong or poor performance as determined by a particular class characteristic in a dataset.

They introduce a unique rules-based technique to analyse company procedure footprints in the next section. By using a machine learner to find those intriguing regulations that have the greatest impact on the results of company procedure cases, our idea builds on previous methodologies centred around rule mining for associations and comparison items sets mine. Usually, indicators of success may be used to track system outputs. As a result, can be seen trace-level indicators of success as trace-level characteristics that may be utilised as class variables in machine learning applications.

Journal of Theoretical and Applied Information Technology
15th February 2024. Vol.102. No 3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

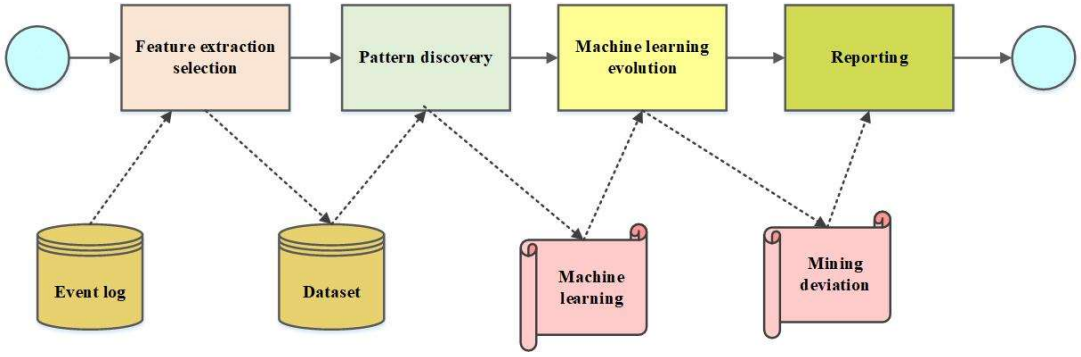

Figure 1: Steps to perform dataset encoding and machine learning analysis

It is crucial to bear in mind which (Process Efficiency Indicator) PPIs can be present at additional levels for abstractness in relation to business procedures, such as at the stage of activity, in which a particular task might be tracked from a PPI without consideration of the outcome of every other task carried throughout the same procedure. At the leadership threshold, it can be more important to keep track of a business' overall efficacy, which is often accomplished by combining the findings of a trace-level PPI. For instance, management is interested in monitoring a service level agreement that requires at least 95% of problems to be resolved within 24 hours, thus they would like to measure the amount of incidents resolved in less than 24 hours. This PPI identifies every procedure trail according to its finish time. It is an incorporation of a tracing-level PPI. Trace-level PPIs are of importance for the purposes of this work.

The idea behind machine learning is this, given a realisable choice, demonstrating the disparities among possibilities could prove more obvious than presenting every single event. As opposed to just listing the details of the present-day scenario, a machine learner quickly determines the critical aspects that most affect that circumstance.

A company's record of events might be transformed into a set of data for the purposes of machine learning. Following that, they describe a compression strategy that mines an archive of processes traces for features using declarative mining of processes. Declare's expressiveness is sufficient to record both basic count of each task and complex related to time interactions between pairs of activity. The idea is to use only one, condensed syntax in this manner to record both basic and complicated themes that could be present in an event log. As far that we comprehend, research has not yet investigated a declarative-language oriented encoding strategy to build vector illustrations of process occurrences.

Table 1: Sequence encoding

h_{id}	σ	boa			bigram		mrs		mra	
		x	y	z	xy	xz	xy	xyz	x,y	X,y,z
h_1	xyzxy	2	1	2	1	1	1	1	1	2
h_2	xyzx	2	1	1	0	1	0	0	1	2
h_3	xyzyzx	3	1	2	1	1	1	1	1	3

Table 1 is a non-exhaustive illustration of characteristics that may be retrieved given the occurrences of events inside the context of multiple activity traces that together make up a log of events P′. The illustration used known coding methods including bag-of-activities (boa), bigram, maximum repeat sequence (mrs), and maximal repeat alphabet (mra). Such encoding techniques track the frequency with which each encoding pattern is present in the process traces. The choice containing directly extracting tracing-level characteristics off the set h_{attr} of an operation trace and adding those into the collection of instance properties j_{attr} is also taken into account by our methodology. It is feasible to convert the incident log P′ to a datasets after extraction properties from a set of activity traces H ′ by transferring each h to H ′ to a dataset instance q, so that each q = $(h_{id}, h_{attr}, c_{name})$, with $q_{1.....n} \in X^{'}$.*Table 2: Event –log using boa encoding.*

h_{id}	x	y	z	et	Pc(c_{name})
h_1	2	1	2	4.50	false
h_2	2	1	1	3.20	true
h_3	3	1	2	7.30	false

The incident log change example's information is shown in the Table 2. Imagine the identical examples traced from before that additionally have additional trace-level characteristics: processing duration (et), in days, as well as effective process conclusion (pc), containing an integer categorization value of "True, False." This provides an illustration of how an event log may be completely transformed into a dataset. A finite a number of occurrences within each trace $h_{1..3}$ may then be encoded using an encoding approach, such as bag-of-activities. Four unique qualities (characteristics) were identified utilising the BOA technique taking into account the peculiarities of the aforementioned activity traces: a, b, and c. It is therefore feasible to create a dataset that includes the gathering of each of the event-driven & trace-level characteristics by taking into account both of the current trace-level characteristics, et & pc. In this manner, the procedure's control-flow and information properties may be examined to one another. In order to connect to the element which serves as the foundation for verifying deviant behaviour, the given name for an attribute of a class (c_{name}) has to be identified in the dataset. The term "c_{name}" must be used to identify a trace-level performance marker that is relevant for examination. False-valued (unsuccessful) footprints are regarded as aberrant instances in our scenario since the effective completion characteristic is specified as a class variable, c_{name} = pc.

4.2. Mining Declare Constraints as Trace-level Attributes

Compared to the currently used series encoding methods, they also suggest a fresh method employing logical process mining in order for extracting traits from periods of happenings. They took into account the Declaration programming syntax and its restriction examples, which offer the primary relationship and presence restrictions forms. They took into account the meaning of Declare restrictions using standard patterns included in both Unrestricted Miner and MINERFul++ declaratory mining algorithms with the goal to execute the discovery of limitations at the track levels. They must stay away from vacuously fulfilled restrictions since pattern fulfilments are the things that we want to engage in. To eliminate simply met restrictions, a different labelling collection of support automaton for vacuity detecting is suggested. In our search process, the comparable routine expressions used by the vacuity detecting support automata have been taken into account. Declarative syntax mining methods now in use seek to identify a collection of restriction patterns to describe the behaviour of a whole event record as one procedure paradigm. To determine if a restriction template is valid and meaningful, these techniques may take into account several threshold characteristics at the event log level, such as support, confidence, and interest factor. Through examining the achievement of a set Announce specifications for every step in the trace, they hope to employ Declare constraints as features at the trace level in this study. Similar to the previously discussed current encoding methodologies, those Declare-based attributes for each process trace may be used to create a collection of examples.

They use declaratory procedure mined approaches to identify whether Declaration requirements was satisfied in every programme tracing $h \in H$, provided an events log P. A number of Predefined limitations have to be established before mine can be done correctly. A Declaration restriction generators collection may represent all of it or a portion of it in this case. It then needs to be paired to a collection of unique occurrences that are recorded on the occurrence log. This occurrence set includes the parameters that Declaring requirement patterns require in order to function, while this mixture produces the collection of characteristics produced by this encode technique. By creating unique ordinary expressions, the list of default requirement templates may be expanded to include additional restrictions as appropriate. The label of the restriction example, that symbolises an abstraction of a restriction (at first used stated in LTL or via an ordinary expressions), plus a group of parameters are combined to form a Declare restriction d, where d = name ({args}). The total amount of parameters differs based on the pattern; for instance, a init restriction theme only requires a single query since it applies to the occurrence who initiates the trace's execution, but the coexisting restraint pattern requires two inputs because it applies whenever two occurrences occur in the same processes trail.

Considering the occurrence logging instance P ′ from earlier, that includes a collection of three separate occurrences (a, b, and c). Three Declaration requirements init (a), init (b), and init (c) can be produced from a Declaration restriction generator of class init. Every limitations, represented by "1" as a fulfilment or "0" alternatively, makes up as a trace-level attribute-value pairing in the sake of decoding by obtaining an amount matching to the Declaring condition's fulfilment. Common attribute-

value pairings associated with the *init* model, for instance, are as follows: $h'_{attr=((init(a),1),(init(b),0),(init(c),0))}$. The exactly_n model, which counts an exact n of instances of events within the entire track, corresponds to the lone alternative. Activity tracing containing Declare-based attribute-value pairings can then be converted into database objects in a manner similar to that shown in the Table for boa coding. A typical dataset is shown in Table 3 and is made up of objects with characteristics that correspond to an example of Declaration restrictions obtained from the event log P ′. Declare-based characteristics may represent timing connections among actions in a manner that sequence- based set-based encoding methods can't, in contrast with other current encoding methods. For instance, the boa, bigram, mra, and mrs methods do not have an equivalent for the answer (b,c) restriction. Customised constraints for incident sequencing representations of features may nevertheless be defined. Declaring also offers a number of predefined templates that can handle a variety of timing connections between procedure incidents, which is a further advantage. Concerning methodology, each of the four rules may be represented with the current Announce limitation components.

Table 3: event log using declare encoding

h_{id}	Init(x)	Last(x)	Exactly(x)	Response(x, y)	et	pc
h_1	1	0	2	1	4.5	false
h_2	1	1	2	0	3.2	true
h_3	1	1	3	1	7.3	false

4.3. Machine learning Evaluation

4.3.1. Standardized streamflow index (SSI)

Similarly to indicators of severe weather, the majority of investigations used standardised criteria for assessing hydrologic dryness. Two significant standardised indices are flow indices and standardised runoff indices, both which have an analogous theoretical foundation. The sole difference between SSI computations and other computations is that run-off from the surface data are utilised in place of precipitation data. For example, this index displays a correct beta dispersion. As a result, for each month, the total flow values are separately estimated before the SSI is computed.

4.3.2. Gene expression programming (GEP)

Genetics can be made using genetic algorithms in the Gene Expression Programming (GEP) algorithm, which uses communities of people and selects these according to fitness. The GEP method's initial step is to create a main collection of answers. This level can be finished by an unintentional procedure or by using some knowledge about the issue. The chromosomal structures were then visualised as a tree expression and evaluated using a fitting method. In general, processing a number of target issues, also known as fitting problems, allows for the evaluation of the appropriate function. The research process ends and the most effective resolution is determined once the answer has an appropriate standard or if a certain number of iterations have passed. The most suitable response form the latest generation is maintained if the most favourable scenario cannot be discovered, and the remaining options are left to be chosen from. The best people are more likely to have children, based on the decision. For many generations to come, every step has been repeated, and it is anticipated the group in question quality will generally increase as new generations are born. GEP chooses the candidates using the renowned roulette wheel approach. In contrast to genetic algorithms and genetic programming, GEP uses a number of genetic operatives to reproduce modified people. Replica is a procedure designed for preserving a few of the most talented members of this era into the following one. A mutation operator's objective is to insert arbitrary changes into an individual chromosome. To avoid producing people deemed rule-deficient, this operator conducts some of the perfect procedures. GEP employs a one-point and two-point combination, similar to a biological algorithm. The genomic equivalent problem (GEP) employs a single-point and two-point combinations. The kind of two-point combo is a little more intriguing because it can largely switch on and off the chromosomal regions that are not encoded. Additionally, the GEP also performs a different kind of combining known as gene combination, in which genes are entirely combined. To create two new children, this operator randomly chooses genes on both-parent chromosomes that are located in the same location.

4.3.3. Support vector regression

Over the following decades, Support Vector Machine (SVM) evolved into a linear classification algorithm using optimum hyperplane

concept. Utilising statistical learning theory, this approach is used. Additionally, they utilised kernel algorithms to create nonlinear classifications. SVM's classification algorithm serves to categorise problems associated with data into multiple classes, while its regression technique is applied to solve prediction issues. Regression on fit data produces a hyperplane. A given location's deviation from its hyperplane revealed the inaccuracy of that location. The most effective technique for regression analysis is advised is the leastsquares approach. But it can happen that using a least-square estimation for analysis issues in the form of outliers may not be entirely rational, which would lead to the analysis performing poorly. In order to avoid bad performance that is not responsive to minute modifications to the model, a robust estimator should be created. As mentioned, the SVM is built upon the principle of minimising risk, a hierarchy generated by the theory associated with statistical training. a distance from real values termed an error function to employ SVM in regression issues that overlook mistakes in a -insensitive manner. This function's definition translates as follows (1) and (2):

$$P(a, f(d,y) = |a - f(d,y)|_{\varepsilon} \quad (1)$$

$$= \begin{cases} 0 \ \ for \ |a - f(d,y)| \le \varepsilon \\ |a - f(d,y)| - \varepsilon \ if \ |a - f(d,y)| > \varepsilon \end{cases} \quad (2)$$

Below, this mistake function does not take into account errors.

4.3.4. M5 model tree

This technique is an amalgam of machine learning and data mining techniques. Data mining techniques identify several, suitable frameworks before extracting data from a pool of set values. Because data mining techniques differ from statistical approaches because they were established for huge datasets with multiple variables, they were created for smaller datasets with fewer variables. among among the most popular data mining approaches, decision tree-based methods use input data to forecast or categorise target qualities as an output in the shape of an equation having a structure of trees. The M5 modelling trees is a structure of choices that may be utilised for forecasting continous quantitative qualities. Its branches are representations of regression operates, and it has lately sparked a substantial development in classifications and predictions. When contrasted to other theories, the tree algorithm's data has higher precision and is simpler to replicate and comprehend. A tree of choices is composed of four components: the root, the branch, the nodes, and the leaves. The rectangular shape denoted each node, while the connections between them were shown as branches. The tree of choices usually goes from left to right or from top to bottom, with the base (first node) on the very top to make it easier to create. The leaf denotes the conclusion of a series of events. For the reason of minimising the total of the squared variances from the average information for each node, splitting is carried out by one of the predictive variables. Utilising the splitting criterion is the first step in creating a tree model. The M5 algorithm's dividing criteria relies on the accuracy of the usual variation of the numbers acquired in every node that correspond to every class or subcategory. In a consequence of checking every characteristic at that node, dividing criteria determines the amount of erroneous for that component and determines the smallest predicted error type. In most circumstances, the predictive inaccuracy is determined by assessing how well the desired outcomes for hypothetical cases are predicted. SDR, or standard deviation reduction, is (3)

$$SDR = sd(H) - \sum \frac{|H_i|}{|T|} sd(H_i) \quad (3)$$

The total number of specimens approaching all nodes is shown by H , and H_i is the portion of examples which correspond to the nth outcome of a possible test. sd Stands for standard deviation. Up till reaching the final cluster (the leaf), the method of division is repeated multiple times at every node. So when it reaches the leaves, the total of the squared differences above the average information is virtually zero. The consequence is going to be the growth of a huge tree. Using numerous limbs and nodes, it is going to difficult to operate using this large tree; as a result, undesirable branches must be removed to create an ideal and effective tree. There are a total of two ways to prune: (1) while the plant forms its full potential. (2) Trimming following the peak of shrub development. The second strategy begins by forming the largest possible tree before beginning the trimming manipulate, unlike the initial method, which prevents the tree from growing further branching. Choosing the best branch is dependent on reducing errors in prediction.

4.3.5. Evaluation parameters

The root mean square error (RMSE) (4), relative absolute error (RAE) (7), mean absolute

error (MAE) (5), and correlation coefficient (CC) (6) were used to analyse the error values between the anticipated and observed data.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_i - a_i)^2} \quad (4)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|d_i - a_i| \quad (5)$$

$$CC = \frac{(\sum_{i=1}^{n} d_i a_i - \frac{1}{n}\sum_{i=1}^{n} d_i \sum_{i=1}^{n} a_i)}{(\sum_{i=1}^{n} d_i^2 - \frac{1}{n}(\sum_{i=1}^{n} d_i)^2)(\sum_{i=1}^{n} a_i^2 - \frac{1}{n}(\sum_{i=1}^{n} a_i)^2)} \quad (6)$$

$$RAE = \frac{\sum_{i=1}^{n}|a_i - d_i|}{\sum_{i=1}^{n}|d_i - \bar{d}|} \quad (7)$$

when n represents the total amount on assessments, and xi, yi are the anticipated & observed results of the SSI. Complete correlation (CC) among measured and anticipated numbers. Correlation that is direct is shown by values that are positive, and the opposite relationship is indicated by negative values. Additionally, the RMSE and MAE values are errors, therefore smaller values suggest lesser modelling mistakes.

5. RESULTS AND DISCUSSION

The effectiveness of the three models—SVM, GEP, and M5—in projecting the Standardised It Index utilising the SPI and SPEI indices at Navrood station throughout six time delays (a one-month to six-month) is examined in the current work. A 48-month grade was chosen for investigation in this study out of the several scales for predicting SSI since it had a stronger correlation and was predicted by the mathematical models that were provided. The statistical characteristics of the drought indices used in the research region are shown in Table 4.

Table 4: Statistical characteristics of the utilized data

SSI	skewness	coefficient of variation	standerd deviation	maximum	minimum	mean	variable
0.69	0.13	958.7	0.98	1.98	-2.09	0.0011	SPI
0.69	-0.69	19.098	0.99	1.45	-2.023	-0.054	SPEI
1	0.08	530	0.99	1.98	-1.67	0.003	SSI

The Figure 2 shows the Root Mean Square Error (RMSE) values for three machine learning algorithms: GP, M5, and SVR. Each row represents a different evaluation scenario or experiment. The values indicate the accuracy of the algorithms, with lower RMSE values indicating better accuracy. Based on the table, GP consistently has the lowest RMSE values across different scenarios, suggesting it performs better than M5 and SVR in terms of accuracy.

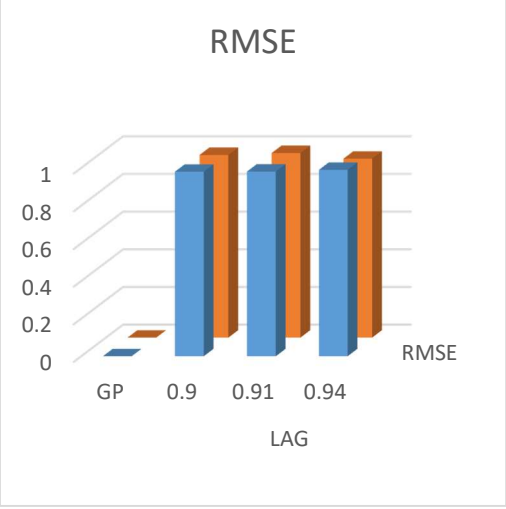

Figure 2: RMSE model

The Figure 3 represents the Mean Absolute Error (MAE) values for three machine learning algorithms: GP, M5, and SVR. Each row corresponds to a different evaluation scenario. MAE is a metric used to measure the average absolute difference between the predicted and actual values, where lower values indicate better accuracy. Based on the table, GP consistently has the lowest MAE values across different scenarios, indicating it performs better in terms of accuracy compared to M5 and SVR.

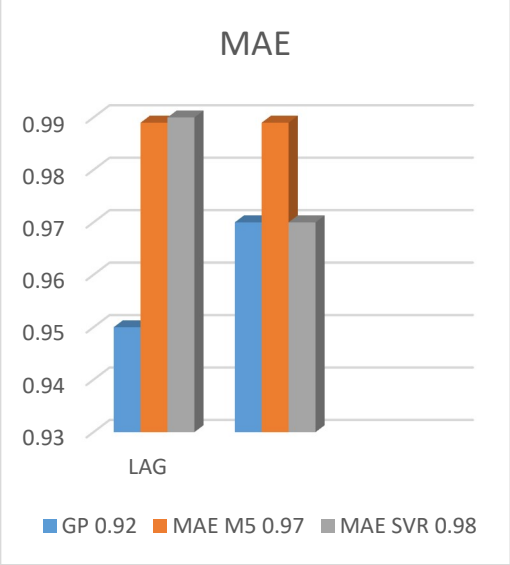

Figure 3: MAE model

The Figure 4 shows the Relative Absolute Error (RAE) values for three machine learning algorithms: GP, M5, and SVR. Each row represents a different evaluation scenario. RAE is a metric used to measure the relative difference between the predicted and actual values, indicating the performance of the algorithms in relation to the magnitude of the target variable. Lower RAE values indicate better accuracy. Based on the table, GP generally has lower RAE values across different scenarios, suggesting it performs better in terms of accuracy compared to M5 and SVR in relation to the magnitude of the target variable.

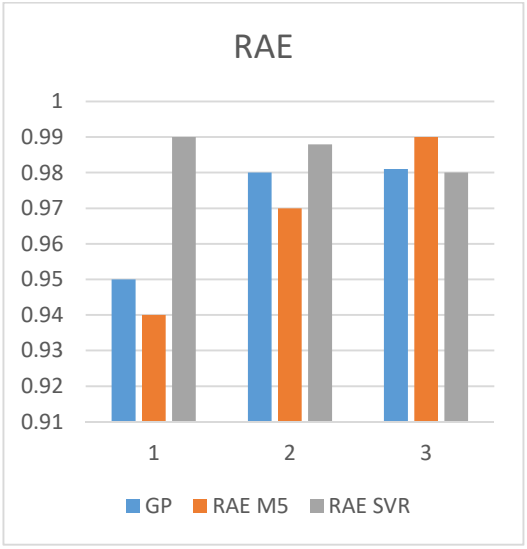

Figure 4: RAE model

Additionally, determined by Pearson's correlation but cross-correlation, it was determined that the USDA hydrological dryness index were better and predicted with a smaller error even though the drought index is better and more dependent on climatic circumstances.

6. CONCLUSION

The research on mining deviations with machine learning techniques in event logs, particularly through the introduction of an innovative encoding algorithm, marks a significant stride in the field. The application of Treatment Learning as a novel approach in Deviation Mining, combined with the original Declare-based encoding method, represents a notable departure from traditional methodologies. The proposed encoding method, founded on the fulfillment of Declaring restrictions, offers a more expressive and nuanced solution for creating vector-based representations of process occurrences. The experimentation, conducted with publicly accessible logs of events from real procedures, underscores the practical utility and superiority of the suggested methodology over state-of-the-art activity decoding methods. The conclusions drawn from this work emphasize the novel contributions and potential impact of the proposed methodology. By showcasing how behavioral learning, when integrated with the Declare-based encoding, provides actionable and more descriptive insights from event logs, the study positions itself as a valuable tool for the analysis of processes. The novelty lies in not only the introduction of Treatment Learning but also in the synergistic application of this method with the innovative encoding algorithm, thereby providing a holistic and advanced

framework for mining deviations in event logs. The implications of this research extend beyond the specific domain investigated, promising to revolutionize anomaly detection, process optimization, and decision-making in diverse sectors. Overall, this work stands as a testament to the innovative possibilities within event log analysis, opening avenues for further exploration and advancements in the broader field of data-driven insights and anomaly detection.

REFERENCE

[1] M. A. Semin and L. Yu. Levin, "Stability of air flows in mine ventilation networks," *Process Saf. Environ. Prot.*, vol. 124, pp. 167–171, Apr. 2019, doi: 10.1016/j.psep.2019.02.006.

[2] Y. Yuan, H. Cao, Y. Zhang, Q. Xie, and R. Yao, "Outlier Mining Based on Neighbor-Density-Deviation with Minimum Hyper-Sphere," *Inf. Technol. Control*, vol. 45, no. 3, pp. 267–277, Sep. 2016, doi: 10.5755/j01.itc.45.3.13164.

[3] J. A. Botín and M. A. Vergara, "A cost management model for economic sustainability and continuos improvement of mining operations," *Resour. Policy*, vol. 46, pp. 212–218, Dec. 2015, doi: 10.1016/j.resourpol.2015.10.004.

[4] J. Von Der Goltz and P. Barnwal, "Mines: The local wealth and health effects of mineral mining in developing countries," *J. Dev. Econ.*, vol. 139, pp. 1–16, Jun. 2019, doi: 10.1016/j.jdeveco.2018.05.005.

[5] P. Zerbino, D. Aloini, R. Dulmin, and V. Mininno, "Process-mining-enabled audit of information systems: Methodology and an application," *Expert Syst. Appl.*, vol. 110, pp. 80–92, Nov. 2018, doi: 10.1016/j.eswa.2018.05.030.

[6] Y. Xu, T. Li, X. Tang, X. Zhang, H. Fan, and Y. Wang, "Research on the Applicability of DInSAR, Stacking-InSAR and SBAS-InSAR for Mining Region Subsidence Detection in the Datong Coalfield," *Remote Sens.*, vol. 14, no. 14, p. 3314, Jul. 2022, doi: 10.3390/rs14143314.

[7] F. Taymouri, M. L. Rosa, M. Dumas, and F. M. Maggi, "Business process variant analysis: Survey and classification," *Knowl.-Based Syst.*, vol. 211, p. 106557, Jan. 2021, doi: 10.1016/j.knosys.2020.106557.

[8] I. Bagińska, M. Kawa, and W. Janecki, "Estimation of spatial variability of lignite mine dumping ground soil properties using CPTu results," *Stud. Geotech. Mech.*, vol. 38, no. 1, pp. 3–13, Mar. 2016, doi: 10.1515/sgem-2016-0001.

[9] V. Rodriguez-Galiano, M. Sanchez-Castillo, M. Chica-Olmo, and M. Chica-Rivas, "Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines," *Ore Geol. Rev.*, vol. 71, pp. 804–818, Dec. 2015, doi: 10.1016/j.oregeorev.2015.01.001.

[10] H. Demolli, A. S. Dokuz, A. Ecemis, and M. Gokcek, "Wind power forecasting based on daily wind speed data using machine learning algorithms," *Energy Convers. Manag.*, vol. 198, p. 111823, Oct. 2019, doi: 10.1016/j.enconman.2019.111823.

[11] Z. Zhu, N. Anwer, Q. Huang, and L. Mathieu, "Machine learning in tolerancing for additive manufacturing," *CIRP Ann.*, vol. 67, no. 1, pp. 157–160, 2018, doi: 10.1016/j.cirp.2018.04.119.

[12] J. Evermann, J.-R. Rehse, and P. Fettke, "Predicting process behaviour using deep learning," *Decis. Support Syst.*, vol. 100, pp. 129–140, Aug. 2017, doi: 10.1016/j.dss.2017.04.003.

[13] J. Evermann, J.-R. Rehse, and P. Fettke, "A Deep Learning Approach for Predicting Process Behaviour at Runtime," in *Business Process Management Workshops*, M. Dumas and M. Fantinato, Eds., in Lecture Notes in Business Information Processing, vol. 281. Cham: Springer International Publishing, 2017, pp. 327–338. doi: 10.1007/978-3-319-58457-7_24.

[14] C. D. S. Garcia *et al.*, "Process mining techniques and applications – A systematic mapping study," *Expert Syst. Appl.*, vol. 133, pp. 260–295, Nov. 2019, doi: 10.1016/j.eswa.2019.05.003.

[15] P. H. P. Richetti, L. S. Jazbik, F. A. Baião, and M. L. M. Campos, "Deviance mining with treatment learning and declare-based encoding of event logs," *Expert Syst. Appl.*, vol. 187, p. 115962, Jan. 2022, doi: 10.1016/j.eswa.2021.115962.

[16] T. Al-Shehari and R. A. Alsowail, "An Insider Data Leakage Detection Using One-Hot Encoding, Synthetic Minority Oversampling and Machine Learning Techniques," *Entropy*, vol. 23, no. 10, p. 1258, Sep. 2021, doi: 10.3390/e23101258.

[17] J. J. Roldán, E. Crespo, A. Martín-Barrio, E. Peña-Tapia, and A. Barrientos, "A training system for Industry 4.0 operators in complex

assemblies based on virtual reality and process mining," *Robot. Comput.-Integr. Manuf.*, vol. 59, pp. 305–316, Oct. 2019, doi: 10.1016/j.rcim.2019.05.004.

[18] E. Helm, A. M. Lin, D. Baumgartner, A. C. Lin, and J. Küng, "Towards the Use of Standardized Terms in Clinical Case Studies for Process Mining in Healthcare," *Int. J. Environ. Res. Public. Health*, vol. 17, no. 4, p. 1348, Feb. 2020, doi: 10.3390/ijerph17041348.

[19] S. Weinzierl *et al.*, "An empirical comparison of deep-neural-network architectures for next activity prediction using context-enriched process event logs," 2020, doi: 10.48550/ARXIV.2005.01194.