Machine Learning Functions in Data Mining and Analytics in the Process Industry

Nausheen Fatima[1] M Abhishek Reddy[2] Mohammad Faiz[3*] Ramandeep Sandhu[4*]

[1,2,3,4] School of Computer Science & Engineering, Lovely Professional University, Phagwara, Punjab, India

nausheen.28838@lpu.co.in itsmeabhi743@gmail.com faiz.techno20@gmail.com ramandeep.28362@lpu.co.in

ABSTRACT: The process sector generates a lot of data from various sources, such as sensors, manufacturing systems, and supply chains. Utilizing analytics and data mining techniques, it is possible to draw useful insights and trends from this data, leading to better decision-making and process optimization. Machine learning, which enables the development of predictive models and automated decision-making processes, is primarily reliant on data mining and analytics. This work investigates how machine learning is applied to analytics and data mining in the process sector. The concept of data mining and analytics is then discussed, it starts off by providing a broad making decision sector and the data management and analytics issues it faces.

1. INTRODUCTION

Analytics and data mining approaches can help businesses in the process industry overcome these challenges by providing a variety of tools and techniques to extract pertinent information from their data. Machine learning is a main technology in analytics and data mining that enables the development of prediction models and automated decision-making procedures. Machine learning algorithms can learn from the data and reveal traditional statistical methods would make it challenging or impossible to identify certain patterns and connections.

The process industry, which creates vast volumes of data from numerous sources, including sensors, industrial systems, and supply chains, is one of the sectors with the greatest rates of data generation[1].

Thanks to these data, organisations in the process industry have a chance to improve their operations, cut costs, and increase efficiency. Nevertheless, because preserving and analysing such enormous amounts of data is a challenging task, conventional analytical tools might not be sufficient to extract significant insights and patterns.

The goal of this essay is to examine how machine learning fits into analytics and data mining in the process industry. The paper will provide an introduction to analytics and data mining as well as a broad review of the process industry and its unique data-related challenges. After that, the study will go into a variety of machine learning methodologies and how they apply to the process industry, such as unsupervised and supervised learning, deep-learning, and reinforcement-learning[2][3].

2. Literature Review

[6] In-depth analyses of decision trees, neural networks, support vector machines, evolutionary this paper provides algorithms and other artificial intelligence techniques for data mining. When it comes to data mining activities, it examines their uses, advantages, and disadvantages.

[7] This review article investigates how machine learning algorithms are used in data mining techniques. It talks about how supervised and unsupervised learning methods, as well as hybrid approaches, can be used for tasks including classification, clustering, association rule mining, and outlier detection. It also talks about the field's difficulties and potential future developments.
[8] This study provides a thorough analysis of machine learning methods utilised in several uses of data mining, including classification, grouping, association rule mining, outlier detection. For various types of data and issue areas, it covers the benefits, drawbacks, and applicability of various algorithms.
[9] In the context of data mining, this survey of the literature focuses on machine learning techniques. It covers a variety of methods and algorithms, combining neural networks, support vector machines, decision trees, and random forests, and offers details on their advantages, drawbacks, and uses in data mining.
[10] This review paper gives a general overview of the machine-learning techniques used in data-mining and analytics for bigdata analysis. It covers the difficulties presented by big data and investigates how machine learning methods may manage enormous datasets. It also discusses upcoming developments and field trends.

3. Similarities and Differences in Data Mining and Data Analysis:

The goal of data analysis and is data mining is the same to derive insights from data but their methods, target audiences, and application domains differ. Data mining is a specific subset that uses cutting-edge algorithms to identify hidden patterns, whereas data analysis comprises a larger range of techniques for viewing and comprehending data[4].

Similarities

i. Both data mining and data analysis include the investigation of data to yield insightful conclusions and knowledge.
ii. The application of statistical and mathematical techniques is required for data analysis and interpretation for both procedures.
iii. Data mining and data analysis both have the objective of identifying patterns, trends, and linkages in the data.
iv. To manage and analyse data, both processes need a number of tools and applications.
v. Both data mining and data analysis support decision- and problem-making across a range of industries[1][12].

Differences

i. Sophisticated machine learning techniques are used to mine huge datasets for hidden links and patterns. In contrast, data analysis involves looking at and evaluating data in order to reach useful conclusions.
ii. Data mining is a subset of data analysis that focuses on the automated discovery of information and patterns in data. Data analysis encompasses a variety of methods, including descriptive statistics, inferential statistics, and exploratory data analysis.
iii. While data analysis may require supporting or reinforcing hypotheses that already exist, data mining typically entails spotting patterns that weren't previously noticed[6].
iv. In contrast to data analysis, data mining usually deals with large datasets and complex patterns, necessitating the use of more complex algorithms and computing power.

4. ALGORITHMS USED

I. Regression: Regression is a supervised learning method that allows the prediction of a continuous output variable from a set of input variables. Regression is a popular method for predicting process variables including temperature, pressure, and flow rate in the process sector[7][8].

II. Decision tree :The term "decision tree" refers to the supervised learning algorithm used for regression and classification. Process optimisation, mistake detection, and quality control are all accomplished in the process sector by using decision trees.

III. Random Forest: Random Forest is an ensemble learning system that combines different decision trees in order to improve prediction accuracy. In the process industry, random forest is used to forecast variables that affect the production process and identify anomalies.

IV. Support-Vector-Machines (SVM): support vector machine (SVM) are a supervised learning method used in categorization. and regression processes. SVM is utilised in the process industry for forecasting process variables and spotting manufacturing process irregularities[9].

V. the K-Nearest Neighbours (KNN) :Frequently employed in classification and regression problems, the K-Nearest Neighbours (KNN) technique is a well-known machine learning algorithm. It is a non-parametric technique that generates predictions based on how closely new data points resemble their k nearest neighbours in the training dataset.

5. Analytics and Data Mining Methodology

In the domains of analytics, data mining, and machine learning, modern computer techniques are utilised to extract knowledge and conclusions from data. These methods can be applied to enhance possibilities, processes, and decision-making in commercial and industrial settings.

In machine learning, algorithms are trained to learn from data and then predictions or judgements are made based on that learning. A subset of artificial intelligence is called machine learning. (AI). Instead of explicitly programming them to accomplish a certain task, it includes creating models that can learn from data. Depending on whether the data has been tagged, supervised or unsupervised or semi-supervised machine learning (that is, have a known outcome)[10].Data mining steps are shown in Fig.1.

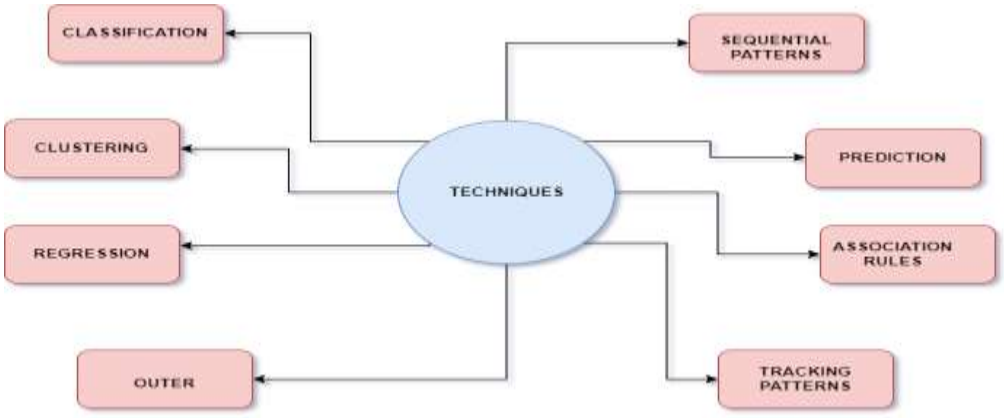

Fig.1: Data Mining Techniques

A. Data Preparation

The first step in data preparation is to gain a basic understanding of the process data, which is followed by picking the suitable data samples for modelling. One of the main objectives of this stage is to remove the dataset from the historical database. The dataset's organisation and techniques for sample and variable-based data selection are additional goals[11][4]. To appropriately extract an operational components of the procedure must be evaluated, and any modifications in operating condition must be noted. dataset with the historical data.

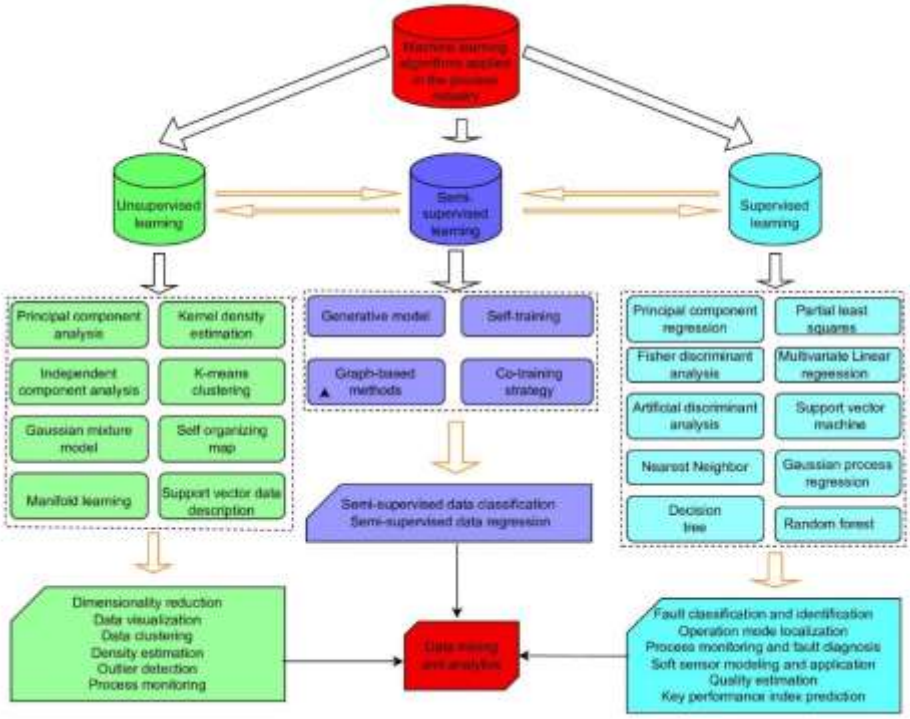

Fig.2: Machine Learning Algorithms Used

II. Data Pre-Processing

Data pre-processing is required’s the enhance the dataset's quality once it has been generated in an earlier stage, and it may be necessary to perform some pertinent data transformations to increase the effectiveness of the data modelling [12] shown in Fig.2.

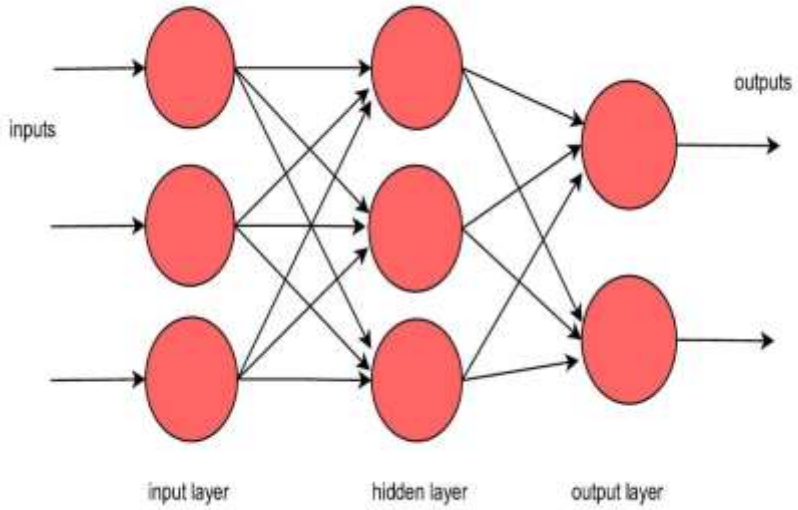

Fig.3: Layers of Data Processor

III. Model Choice, Training, and Performance Assessment

Once the dataset is available, we may decide which machine learning technique is ideal for creating data models. Carefully analysing the data attributes will reveal how complicated the data model is. What sort of machine learning model, for example, ought to we apply to the training dataset? What degree of model complexity ought to be used with the data? Do you need just one model structure[13][14][25]? Or, do you require numerous models [26][27].

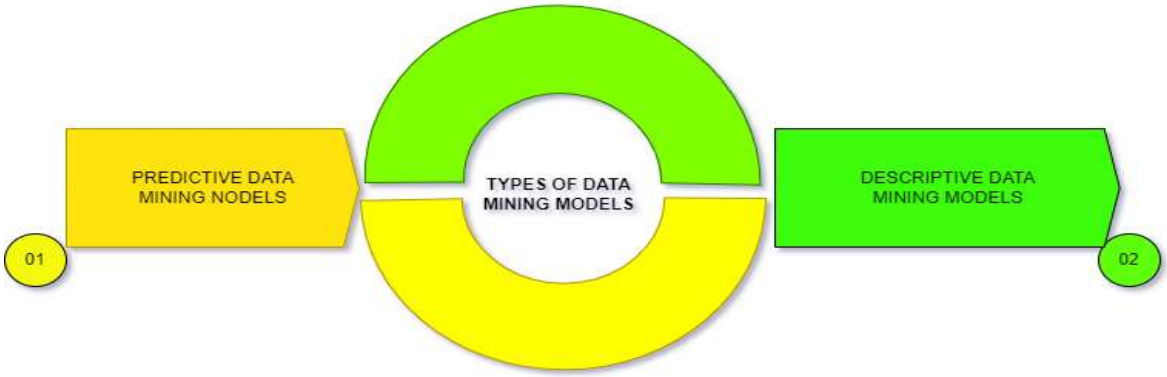

Fig.4: Types of Data Mining Model

IV. Data mining and analytics

Assume that the data model had been trained and validated following the successful completion of the machine learning technique. Currently, the model can perform tasks including data clustering, reducing the dimensionality, visualisation of data, and trend

analysis both offline and online.[1], Process vigilance and fault classification, fault diagnosis, online soft detecting, and quality forecasting [15]. There in Fig (5) [28][29].

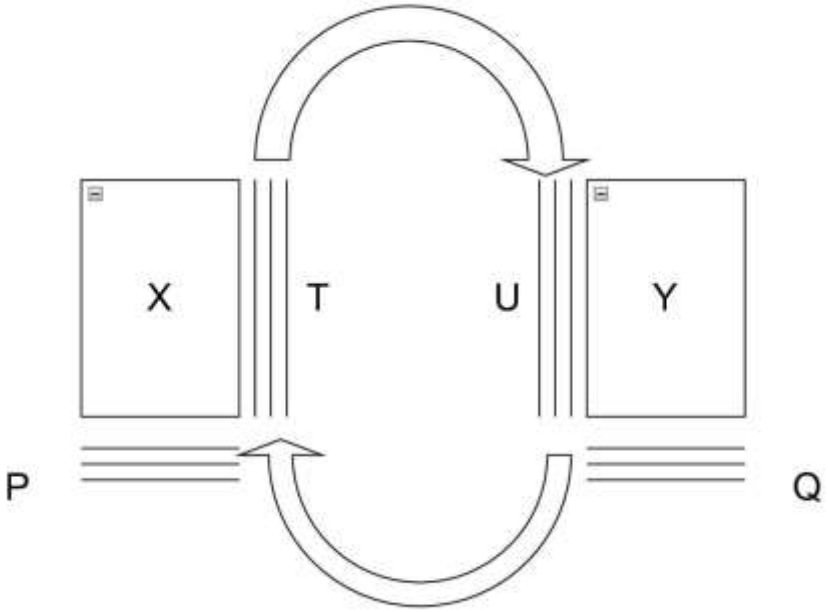

Fig.5: Data Mining Analysis

5. Proposed Model

The machine learning strategy that is suggested for the process industry's use in mining data and analytics looks somewhat like this:

- Data pre-processing: After the raw data has been gathered from multiple sources, it must be cleaned up of noise and inconsistencies. There may be a need for data cleansing, data transformation,[15] and data dimension reduction.
- Feature Selection: The most pertinent in this step, features or variables are selected from the pre-processed data. Numerous feature selection techniques, such as shared knowledge, analysis of principal components, and correlation analysis, are used to achieve this goal [16].
- Model Selection: In this step, a suitable machine learning model is selected based on the features of the problem and the data. It can be necessary to select a classification model, regression model, clustering model, or some other kind of machine learning model for this [30][31][32].

The general formula for regression analysis is:

$$y = a + b(x) + e \ldots \ldots (1)$$

Where:

- The dependent (or response) variable is y.
- The independent variable (also known as predictor variable) is 0.
- The intercept is the value for y when x = 0.

- The slope is defined as the alteration in y for every unit change in x.
- The random variations in y that the model cannot account for is known as the error term.

➢ Model training: The chosen model must next be trained with the pre-processed data and chosen data. A training dataset is used to fit the model's parameters and enhance the model's performance.

➢ Model Evaluation: A validation dataset is used to assess the model's efficacy. The type of model being utilized and the issue being solved decide the appropriate assessment metrics. F1-score, accuracy, precision, recall, and AUC-ROC are examples of common metrics.

➢ Model Deployment: The model is prepared for deployment in the process industry following training and evaluation. To be used for real-time decision-making or process optimization, the model must be integrated into the production process.

Let Let's look at the issue of estimating the cost of an apartment based on its size. This problem can be organised as follows:

The graph changes to the following when further information about the house is added as additional variables:

$$Z=\sum wi * xi + b \rightarrow a = \varphi(\text{^z}) \text{ \&^y=a} \ldots\ldots (2)$$

Each neuron is divided into two main blocks:

- computation of z using the inputs $\boldsymbol{xi}$

$$Z=\sum wi * xi + b \ldots\ldots (3)$$

- Computation of a, which is equal to y at the output layer, using z:

$$a = \varphi(\text{^z}) \text{ \&^y=a} \ldots\ldots (4)$$

$\boldsymbol{wi}$ are the weights, bis the bias and is said to be the activation function.

6. Methods of Machine Learning Used in the Processing Industry

This portion offers a thorough analysis of the machine learning techniques lately used in the process industry. Machine learning techniques can be divided into four categories: unsupervised learning, supervision of learning, semi-supervised learning, and reinforcement learning. 80–90% of business applications use both supervised studying and unsupervised learning. Semi-supervised learning techniques have recently been employed for prediction and data analysis, even if reinforced learning hasn't been widely used in the process sector[10][17]**.**

1) Analysing the primary components

A group of correlated variables are transformed via an orthogonal transformation into linearly uncorrelated variables (principal components) by the principal component approach (PCA) statistical technique. Frequently, there are fewer essential components than initial variables.

2) Analyses of Separate Components

To find independent latent components from the observed data, independent component analyse is still in its infancy is first introduced. The objective of ICA is to use optimisation techniques to separate the true components from visible data. Unlike PCA, ICA looks for non-Gaussian, statistically independent latent components. Consequently, ICA might have a greater chance than PCA of obtaining useful information of the non-Gaussian data. Similar to the PCA approach, ICA can be used in the process industry in a variety of ways, include decreasing dimensionality, non-Gaussian information extraction, processes monitoring, visualisation of data, and other things.

3) K-means Cluster

K-means cluster which was primarily developed for signal processing, uses the quantization technique. For clustering on datasets, K-means clustering is still a popular unsupervised machine learning technique. The data samples are separated into k groups using k-means clustering, and each group corresponds to the cluster with the closest mean. The process industry heavily relies on this technique to divide processing data into separate operational modes, fault categories, or product grades. For instance, in batch process monitoring, the segmentation rule of the variable subspace was created[18].

4) Calculating Kernel Density

It is possible to determine the kernel density estimation method, a non-parametric approach to computing the probability density role of a variable that is random (KDE). A key technique for data smoothing is known as kernel density estimation, which uses a limited sample size for training and unidentified, frequently non-Gaussian data distributions to infer information about the population. The distributions of process variables, monitoring data, and other relevant values that are utilised to determine the nature of the process have all been estimated using the kernel density estimate approach in the process industries. Below is a list of some process industry applications using the kernel density estimation method [21][22].

5) Self-Organizing Map

An artificial neural network (ANN) called a self-organizing map (SOM) is trained using a technique called unsupervised learning to produce a low-dimensional, discretized depiction of the training samples' input space. A self-organizing map's building blocks are node or neuron components. Each node has a position within the map space as well as a vector of weights that is the same length as the data that is input vectors. The self-organizing map is a popular mapping from a complex to a smaller data space map[19]. The characteristics of the self-organizing map allow for a wide range of industrial applications, including dimensionality reduction, data visualisation, process monitoring [23][24].

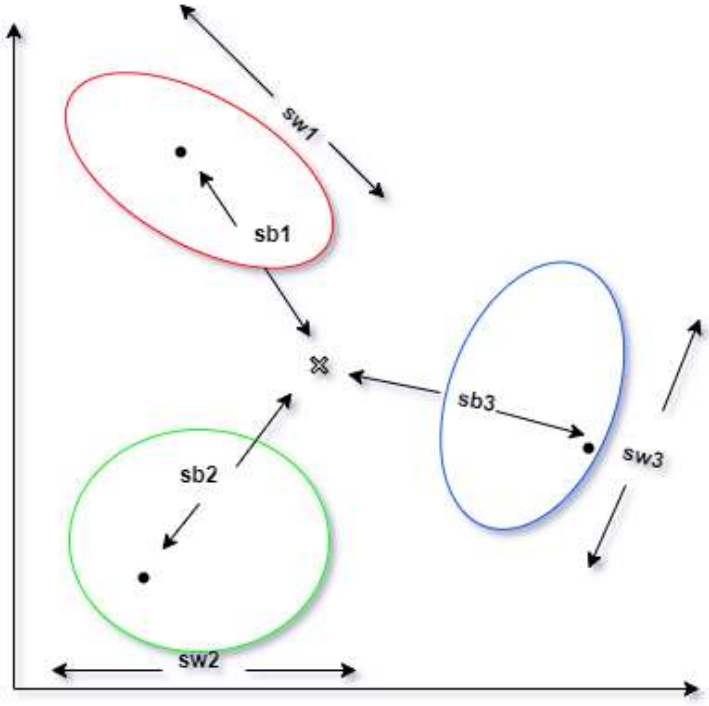

Fig.6: Illustration of the FDA method

7. Perspectives for Future Research

The PCR method has been updated to include more data than the PCA method. Numerous applications in the process sector, including Based on the correlation modelling between the returned main elements using PCA and the variable, condition-related process monitoring, soft sensor modelling, digital quality of the product forecasting, etc., have been created that is dependent[20].

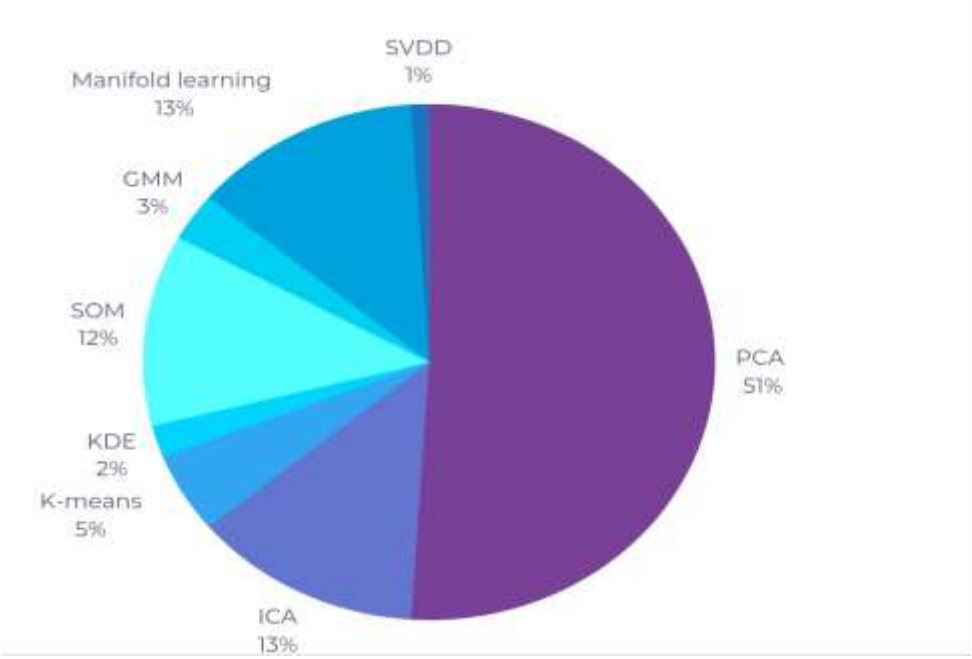

Fig 7: Application status of unsupervised learning methods

Due to the rapid advancement algorithmic learning processes. For instance, data mining and applications for analytics have long made use of neural networks. If complex computations are handled by neural networks, the computational load may become a problem. We can build neural networks that are deep with many layers and increased computing capacity, enabling quick processing and autonomous training dataset learning. Another illustration would be the big data issue, which has recently attracted a lot of attention from various businesses:

A. Machine Learning of Big Data in the Process Industry

Information gathering, data management, and storage have been severely hampered by the rise of the modern process sector. As a result, it is challenging to evaluate the information that is hidden inside those numbers. The process industry has acquired a variety of structured and unstructured data formats, including sensor data, images, videos, audios, log files, and so on, as more measuring equipment has been brought into the sector. In-depth details are offered on the application of data analytics in contemporary industrial processes in the big data era.

B. Sustainability Driven Data Mining and Analytics

Even though the process industry has seen significant with recent developments in energy-saving research as well as worries about contamination of the environment and security, such as greenhouse gas surveillance, sewage treatment, and pollutant analytics, the subject of sustainability has recently attracted a lot of attention. Energy conservation and environmental sustainability in the process industry depend on efficient operations, which include highly advanced control of processes, monitoring, and optimisation. Process data analytics and deep mining could be used to achieve this., which would improve the process's clarity and efficacy for sustainability evaluations.

C. Process Causality Modelling and Analytics

Like additional complex systems like biological ones and social media, the contemporary manufacturing industry is typified by connecting numerous components like operational units, manufacturing machinery, or multiple sensors and devices for measuring. The local activity inside each process element as well as the interactions between various process parts influence the entire procedure behaviour. Finding link correlations or process causation between various variables is crucial for successful data mining and analytics.

D. Data cleaning and quality evaluation

Since it influences how sometimes the best predictive machine learning algorithm could produce a false model since process quality of data cannot be reliably ensured. Therefore, process data requires to be cleared and its accuracy assessed before employing methods of machine learning for model creation, it is necessary to address how to handle missing data and outliers as well as the problem of different sample rates across process variables [25].

8. CONCLUSION

Businesses in the process sector can automate and optimise manufacturing processes, resulting in higher productivity and less waste, by implementing machine learning

algorithms. By seeing probable equipment breakdowns or bottlenecks, predictive analytics can improve resource allocation and enable proactive maintenance. Additionally useful for improved inventory control and production planning, machine learning models can help with demand forecasting.

Machine learning-based data mining and data analytics are essential for the process industries because they provide a mechanism to mine the vast amounts of data that are created by the sector for valuable information. Businesses can employ machine learning algorithms to automate production processes, save costs, boost productivity, and improve product quality. Businesses in the process industry must optimise their operations and maintain competitiveness by utilising machine-learning in data-mining and analytics. Industry, academia, and government organisations must collaborate to address the opportunities and challenges this sector faces if it is to continue to expand and prosper.

References

[1] B. Zhao, H. Zhou, G. Li, and Y. Huang, "ZenLDA: Large-scale topic model training on distributed data-parallel platform," Big Data Min. Anal., vol. 1, no. 1, pp. 57–74, 2018, doi: 10.26599/BDMA.2018.9020006.

[2] N. Yu, Z. Li, and Z. Yu, "Survey on encoding schemes for genomic data representation and feature learning-from signal processing to machine learning," Big Data Min. Anal., vol. 1, no. 3, pp. 191–210, 2018, doi: 10.26599/BDMA.2018.9020018.

[3] Faiz, M., Fatima, N., Sandhu, R., Kaur, M., & Narayan, V. (2023). IMPROVED HOMOMORPHIC ENCRYPTION FOR SECURITY IN CLOUD USING PARTICLE SWARM OPTIMIZATION. Journal of Pharmaceutical Negative Results, 2996-3006.

[4] J. Wang et al., "Relationship between Health Status and Physical Fitness of College Students from South China: An Empirical Study by Data Mining Approach," IEEE Access, vol. 8, pp. 67466–67473, 2020, doi: 10.1109/ACCESS.2020.2986039.

[5] Y. Luo and Y. Xiang, "Application of data mining methods in internet of things technology for the translation systems in traditional ethnic books," IEEE Access, vol. 8, pp. 93398–93407, 2020, doi: 10.1109/ACCESS.2020.2994551.

[6] Rajesh, R., Koh, S. C. L., & Ganesh, K. (2012). Modelling Optimization and Computing 2012.

[7] Mall, P. K., Narayan, V., Pramanik, S., Srivastava, S., Faiz, M., Sriramulu, S., & Kumar, M. N. (2023). FuzzyNet-Based Modelling Smart Traffic System in Smart Cities Using Deep Learning Models. In S. Pramanik & K. Sagayam (Eds.), Handbook of Research on Data-Driven Mathematical Modeling in Smart Cities (pp. 76-95). IGI Global. https://doi.org/10.4018/978-1-6684-6408-3.ch005

[8] Choudhary, S., Narayan, V., Faiz, M., & Pramanik, S. (2022). Fuzzy approach-based stable energy-efficient AODV routing protocol in mobile ad hoc networks. In Software Defined Networking for Ad Hoc Networks (pp. 125-139). Cham: Springer International Publishing.

[9] Faiz, M., Daniel, A.K. (2022). Wireless Sensor Network Based Distribution and Prediction of Water Consumption in Residential Houses Using ANN. In: Misra, R., Kesswani, N., Rajarajan, M., Veeravalli, B., Patel, A. (eds) Internet of Things and Connected Technologies. ICIoTCT 2021. Lecture Notes in Networks and Systems, vol 340. Springer, Cham. https://doi.org/10.1007/978-3-030-94507-7_11

[10] S. S. Gill, G. S. Lehal, and R. Malhotra's 2019 article "A Comprehensive Review of

Machine Learning for Big Data Analysis"

[11] S. Ledesma, M. A. Ibarra-Manzano, E. Cabal-Yepez, D. L. Almanza-Ojeda, and J. G. Avina-Cervantes, "Analysis of data sets with learning conflicts for machine learning," IEEE Access, vol. 6, no. c, pp. 45062–45070, 2018, doi: 10.1109/ACCESS.2018.2865135.

[12] Faiz, M., & Daniel, A. K. (2022). Threats and challenges for security measures on the internet of things. Law, State and Telecommunications Review, 14(1), 71-97.

[13] A. Ishaq et al., "Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques," IEEE Access, vol. 9, pp. 39707–39716, 2021, doi: 10.1109/ACCESS.2021.3064084.

[14] Y. Huang, Z. Cheng, Q. Zhou, Y. Xiang, and R. Zhao, "Data mining algorithm for cloud network information based on artificial intelligence decision mechanism," IEEE Access, vol. 8, pp. 53394–53407, 2020, doi: 10.1109/ACCESS.2020.2981632.

[15] Narayan, V., Awasthi, S., Fatima, N., Faiz, M., & Srivastava, S. (2023, January). Deep Learning Approaches for Human Gait Recognition: A Review. In 2023 International Conference on Artificial Intelligence and Smart Communication (AISC) (pp. 763-768). IEEE.

[16] Faiz, M., Daniel, A.K. A multi-criteria cloud selection model based on fuzzy logic technique for QoS. Int J Syst Assur Eng Manag (2022). https://doi.org/10.1007/s13198-022-01723-0

[17] A. L. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," IEEE Commun. Surv. Tutorials, vol. 18, no. 2, pp. 1153–1176, 2016, doi: 10.1109/COMST.2015.2494502.

[18] I. K. Nti, J. A. Quarcoo, J. Aning, and G. K. Fosu, "A mini-review of machine learning in big data analytics: Applications, challenges, and prospects," Big Data Min. Anal., vol. 5, no. 2, pp. 81–97, 2022, doi: 10.26599/BDMA.2021.9020028.

[19] S. Basodi, C. Ji, H. Zhang, and Y. Pan, "Gradient amplification: An efficient way to train deep neural networks," Big Data Min. Anal., vol. 3, no. 3, pp. 196–207, 2020, doi: 10.26599/BDMA.2020.9020004.

[20] W. Zhong, N. Yu, and C. Ai, "Applying big data based deep learning system to intrusion detection," Big Data Min. Anal., vol. 3, no. 3, pp. 181–195, 2020, doi: 10.26599/BDMA.2020.9020003.

[21] M. Faiz and A. K. Daniel, "FCSM: Fuzzy Cloud Selection Model using QoS Parameters," 2021 First International Conference on Advances in Computing and Future Communication Technologies (ICACFCT), Meerut, India, 2021, pp. 42-47, doi: 10.1109/ICACFCT53978.2021.9837347..

[22] J. Palmer, V. S. Sheng, T. Atkison, and B. Chen, "Classification on grade, price, and region with multi-label and multi-target methods in wineinformatics," Big Data Min. Anal., vol. 3, no. 1, pp. 1–12, 2020, doi: 10.26599/BDMA.2019.9020014.

[23] N. Fatima, L. Liu, S. Hong, and H. Ahmed, "Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis," IEEE Access, vol. 8, pp. 150360–150376, 2020, doi: 10.1109/ACCESS.2020.3016715.

[24] Sandhu R., Lakhwani K. 2021. Scientific Workflow Scheduling by Adaptive Approaches with Convex Optimization in Cloud Environment in Design Engineering (Toronto), ISSN: 0011-9342, Vol 2021, Issue 07. 1686-1712.

[25] Sandhu R., Lakhwani K. Enhanced Scientific Workflow Scheduling in Cloud System in International Conference on Communications and Cyber-Physical Engineering (ICCCE 21). ISSN: 1876-1100/ Vol no. 828/ pages-133-139. https://doi.org/10.1007/978-981-16-7985-8_14

[26] Mohseni, S., Yang, F., Pentyala, S., Du, M., Liu, Y., Lupfer, N., ... & Ragan, E. (2021, May). Machine learning explanations to prevent overtrust in fake news detection. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 15, pp. 421-431).

[27] Babu, S. Z., et al. "Abridgement of Business Data Drilling with the Natural Selection and Recasting Breakthrough: Drill Data With GA." Authors Profile Tarun Danti Dey is doing Bachelor in LAW from Chittagong Independent University, Bangladesh. Her research discipline is business intelligence, LAW, and Computational thinking. She has done 3 (2020).

[28] Narayan, Vipul, A. K. Daniel, and Pooja Chaturvedi. "E-FEERP: Enhanced Fuzzy based Energy Efficient Routing Protocol for Wireless Sensor Network." Wireless Personal Communications (2023): 1-28.

[29] Tyagi, Lalit Kumar, et al. "Energy Efficient Routing Protocol Using Next Cluster Head Selection Process In Two-Level Hierarchy For Wireless Sensor Network." Journal of Pharmaceutical Negative Results (2023): 665-676

[30] Paricherla, Mutyalaiah, et al. "Towards Development of Machine Learning Framework for Enhancing Security in Internet of Things." Security and Communication Networks 2022 (2022).

[31] Sawhney, Rahul, et al. "A comparative assessment of artificial intelligence models used for early prediction and evaluation of chronic kidney disease." Decision Analytics Journal 6 (2023): 100169.

[32] Mall, Pawan Kumar, et al. "Early Warning Signs Of Parkinson's Disease Prediction Using Machine Learning Technique." Journal of Pharmaceutical Negative Results (2022): 4784-4792.