10.48047/jocaaa.2024.33.05.27

A Research based on Predicting Software Project Outcomes and Skill Gaps using Machine Learning

Ms. Renu
Assistant Professor, Department of Computer Science and Engineering, Dronacharya College of Engineering, Gurugram, Haryana
renu@ggnindia.dronacharya.info

Dr. Rajat Kumar
Associate Professor, Department of Computer Science and Engineering, Dronacharya Group of Institutions, Greater Noida, Uttar Pradesh
rajat.kumar@gnindia.dronacharya.info

Abstract

The data processing sector's growth has led to increased reliance on data-driven business choices and enterprise-scale data models. However, the accuracy of these choices depends on the quality of the data used in analysis. This paper proposes a novel framework for data cleaning methods, including missing value identification, domain-specific outlier identification, detailed generic outlier reduction, and dimensionality reduction. The framework achieves approximately 99 percent accuracy when compared to a benchmark dataset. Software project management is a critical component in contemporary research and practice, with high stakes and budgetary restrictions. This study uses a unique multiregression model to forecast software project success, producing an output of almost 98.3% accuracy. The study also produces a recommendation system for an ideal team design matrix based on different team skills, allowing for more timely management and a lower risk of project failure. Human competencies are still crucial for project success, and this work proposes an innovative framework to profile projects, identify available employees, profile them based on training requirements, and propose training recommendations. The framework's results are highly acceptable and have achieved almost 99.1% accuracy.

Keywords: Software Engineering, Intelligent Recommender System, ML in Software Management, Project Risk Prediction

Introduction

This research aims to improve the success rate of software projects by combining non-conventional organizational metrics with deep learning methods. Software metrics are essential for planning work items, monitoring productivity, and other applications [1]. They are similar to the four roles of management: planning, organizing, controlling, and improving the organization. Technical difficulties, employee skill development, social aspects, and overall satisfaction with the organization significantly impact project success or failure. The research focuses on the need to build a new set of software metrics due to the high number of multi-billion-dollar projects delivered without 100% deliverables and satisfactory closures [2]. The proposed framework uses a multi-regression model to forecast project success and recommends ideal team matrix templates based on skills. The research also profiles projects using clustering methods and identifies the best-suited employees for a specific project. The research aims to develop a more accurate and reliable approach to predict software project success or failure, helping organizations identify potential issues early in the project lifecycle

10.48047/jocaaa.2024.33.05.27

and take proactive measures to mitigate risks [3]. This research can contribute to the growth of the software industry and improve the success rate of software projects. This research aims to extract the Stack Overflow Developer Survey's parametric dataset, reduce data dimensionality, develop a new machine learning algorithm to identify the best starting, operating, and finishing points, and develop a recommendation system using project success and failure circumstances [4]. The proposed solution architecture includes a large set of data on employee and organizational aspects, focusing on skillsets, social aspects, mental health, and project aspects. The metric parameters are mapped and a prediction framework is built using these metrics [5]. After identifying success and failure points, a recommendation system is built to avoid failure conditions. The research is divided into three phases: extraction of the dataset, building parameters, identifying boundary conditions, formulating the final metric, adopting the proposed metric and applying machine learning conditions, and designing adaptive thresholds for the recommendation system [6]. The ultimate goal is to minimize failure chances and improve skill development.

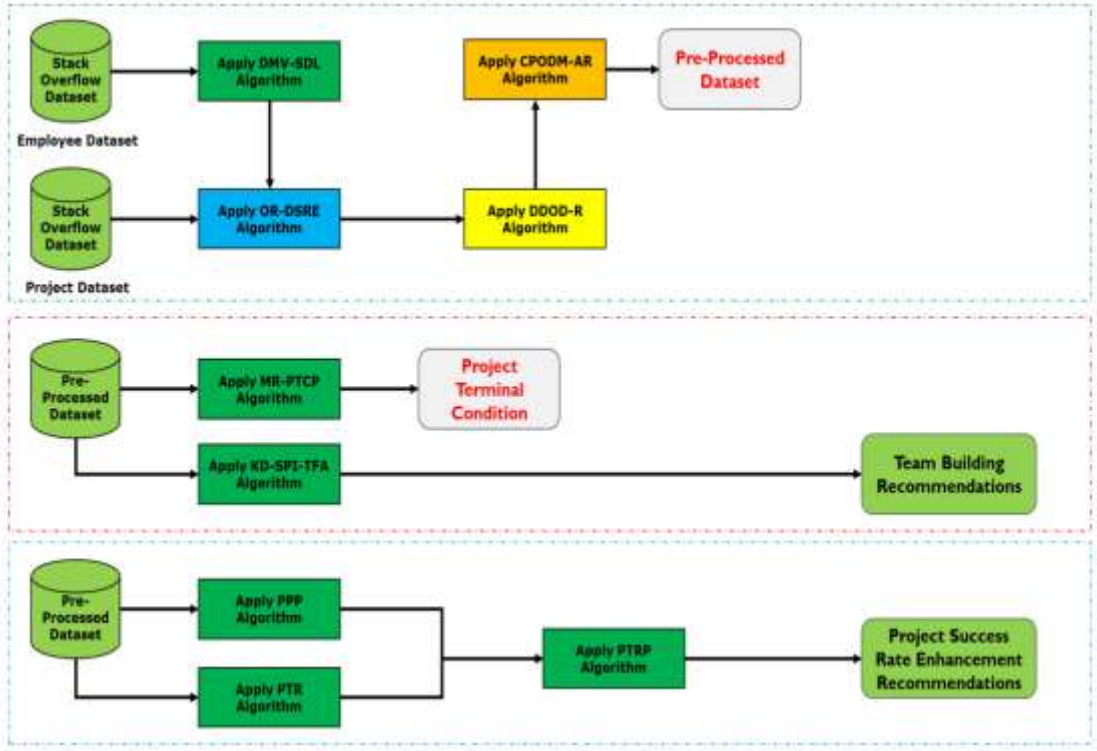

Fig. 1 Basic building and block diagram

Software cost estimation is crucial for project management and has led to the development of various methodologies to improve the estimating process [7]. However, the validity of these methods has been questioned due to the numerous biases involved. This paper uses a multiple comparisons algorithm to rank cost estimation models, identifying those with significant differences in accuracy and clustering them into nonoverlapping groups [8]. The study compares 11 prediction models across six datasets in a large-scale comparison, demonstrating the advantages and valuable data gained through a comprehensive comparison of different approaches [9]. Global Software Development (GSD) has been successful due to its use of

10.48047/jocaaa.2024.33.05.27

Software Project Management (SPM). This work aims to define and categorize studies on SPM techniques for GSD, identify their strengths and weaknesses, and examine their use in the industry [10]. A systematic mapping study was conducted, examining 84 articles and identifying the most commonly reported methods and estimating strategies for GSD. A case study on aviation safety-critical software development used the COCOMO II approach to estimate the needed effort. This study identified a Magnitude of Relative Error (MRE) of 31% and suggested ways to improve effort estimating accuracy in safety-critical software projects [11]. The Intelligent Recommender and Decision Support System (IRDSS) was presented to help scrum masters better evaluate an impending software project[12].

This study focuses on the use of simulation in software development to reduce the risk of overestimation of software development effort. Using ARENA R® and historical data from a software business, the study suggests replacing subjective estimations with continuous simulation [13]. This simulation can accurately estimate software development efforts and the level of risk taken on each overall effort estimate, helping in risk management decisions. The study also proposes the use of fuzzy inference rules for semi-automatic estimating to mitigate the negative aspects of expert judgment-based estimation [14]. The results show that when fuzzy inference rules are included, the expert judgment-based approach's estimate accuracy improves by 39.35 percent. The study also examines the Extreme Learning Machine (ELM) model and compares it to existing literature-based methods for estimating effort. The results suggest that the ELM model provides the best results for evaluating software design effort [15].

Research Methodology

The data processing sector's growth has led to increased reliance on data-driven business choices and enterprise-scale data models[16]. This paper proposes a novel framework for data cleaning methods, including missing value identification, domain-specific outlier identification, detailed generic outlier reduction, and dimensionality reduction, which achieves approximately 99 percent accuracy. The study also uses a unique multiregression model to forecast software project success, producing an output of almost 98.3% accuracy [17]. It also produces a recommendation system for an ideal team design matrix based on different team skills, allowing for more timely management and a lower risk of project failure [18]. The framework also profiles projects, identifies available employees, profiles them based on training requirements, and proposes training recommendations. The research aims to improve the success rate of software projects by combining non-conventional organizational metrics with deep learning methods. It extracts the Stack Overflow Developer Survey's parametric dataset, reduces data dimensionality, develops a new machine learning algorithm, and develops a recommendation system using project success and failure circumstances [19].

This study examines the effectiveness of Software Project Management (SPM) techniques in Global Software Development (GSD) and their application in the industry. A systematic mapping study was conducted, examining 84 articles and identifying the most commonly reported methods and estimating strategies [20]. A case study on aviation safety-critical software development used the COCOMO II approach to estimate the needed effort,

10.48047/jocaaa.2024.33.05.27

identifying a Magnitude of Relative Error (MRE) of 31%. The Intelligent Recommender and Decision Support System (IRDSS) was presented to help scrum masters evaluate an impending software project [21]. The study also suggests using simulation in software development to reduce the risk of overestimation of software development effort. The study also proposes the use of fuzzy inference rules for semi-automatic estimating to mitigate the negative aspects of expert judgment-based estimation [22]. The Extreme Learning Machine (ELM) model was compared to existing literature-based methods for estimating effort, with the ELM model providing the best results for evaluating software design effort [23].

Results and Analysis

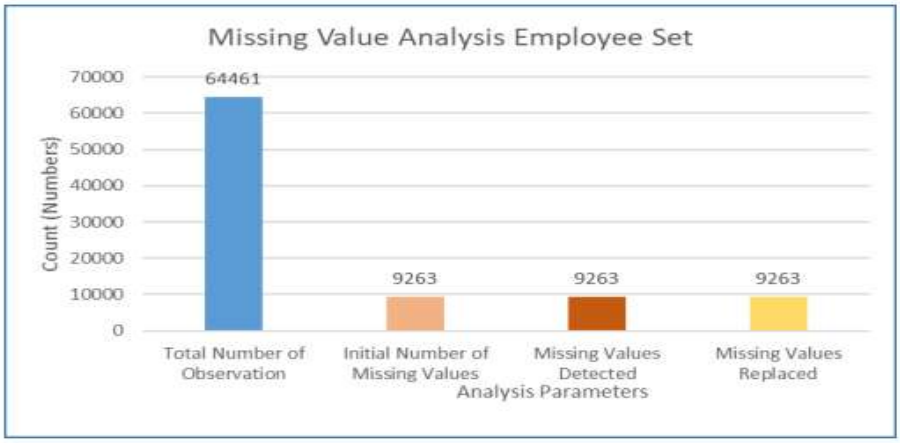

Fig.2 Employee Dataset Missing Value Analysis

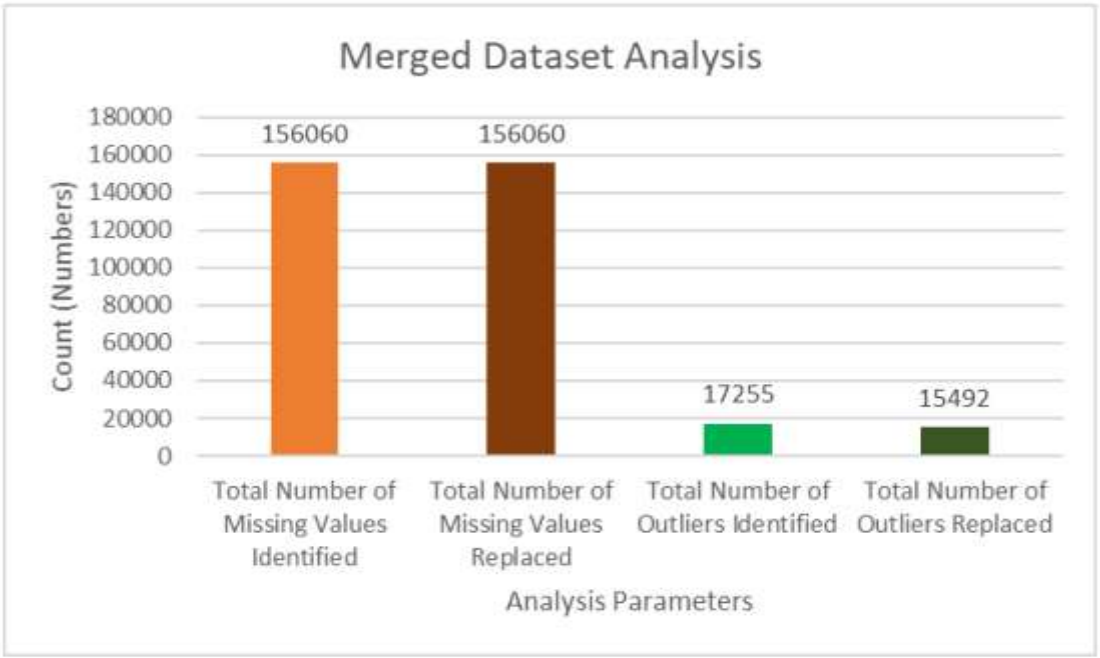

Fig.3 Merged Dataset Missing Value & Outlier Analysis

10.48047/jocaaa.2024.33.05.27

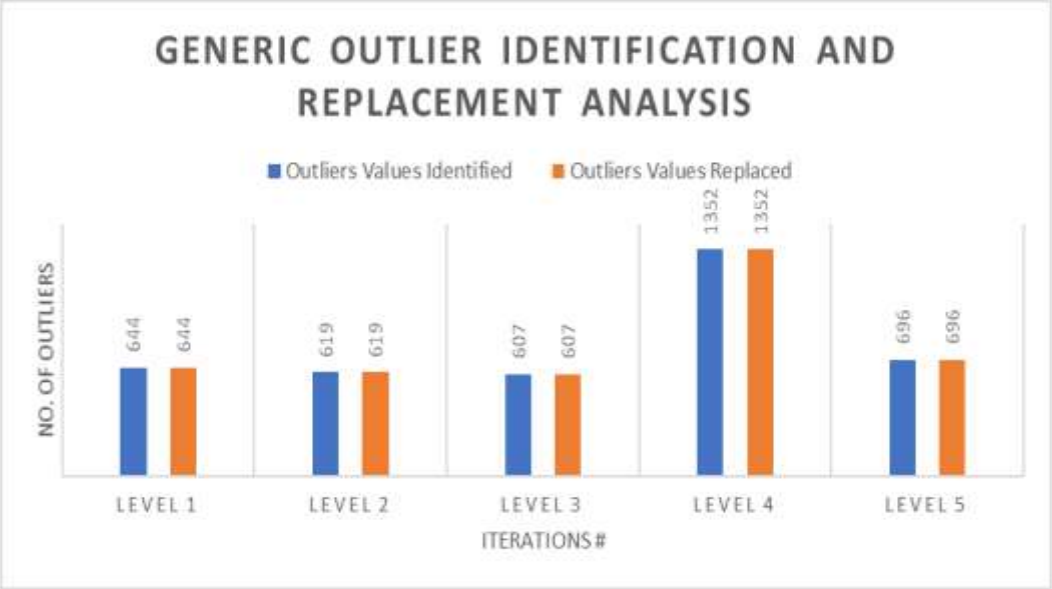

Fig. 4 Generic Outlier Identification & Replacement Analysis

Table-1 Attribute Ranking Analysis

Rank	Attribute Number	Attribute Name
Class Variable	0	CS
1	13	TI
2	6	SKILLS_UP
3	4	EXP
4	3	AGE
5	2	JC
6	5	SKILLS_NOW
7	7	JS
8	12	MI
9	8	JCHA
10	11	CI
11	10	DUR
12	9	PID
13	1	ID

10.48047/jocaaa.2024.33.05.27

Table-2 Final Attribute Reduction Analysis

Iteration #	List of Attributes	Classification Accuracy	Time Complexity (msec)
1	13,6,4,3,2,5,7,12,8,11,10,9,1	66	188
2	13,6,4,3,2,5,7,12,8,11,10,9	92	152
3	13,6,4,3,2,5,7,12,8,11,10	97	143
4	13,6,4,3,2,5,7,12,8,11	98	101
5	13,6,4,3,2,5,7,12,8	97	99
6	13,6,4,3,2,5,7,12	96	97
7	13,6,4,3,2,5,7	94	96
8	13,6,4,3,2,5	93	95
9	13,6,4,3,2	92	91
10	13,6,4,3	92	87
11	13,6,4	71	76
12	13,6	69	71
13	13	66	70

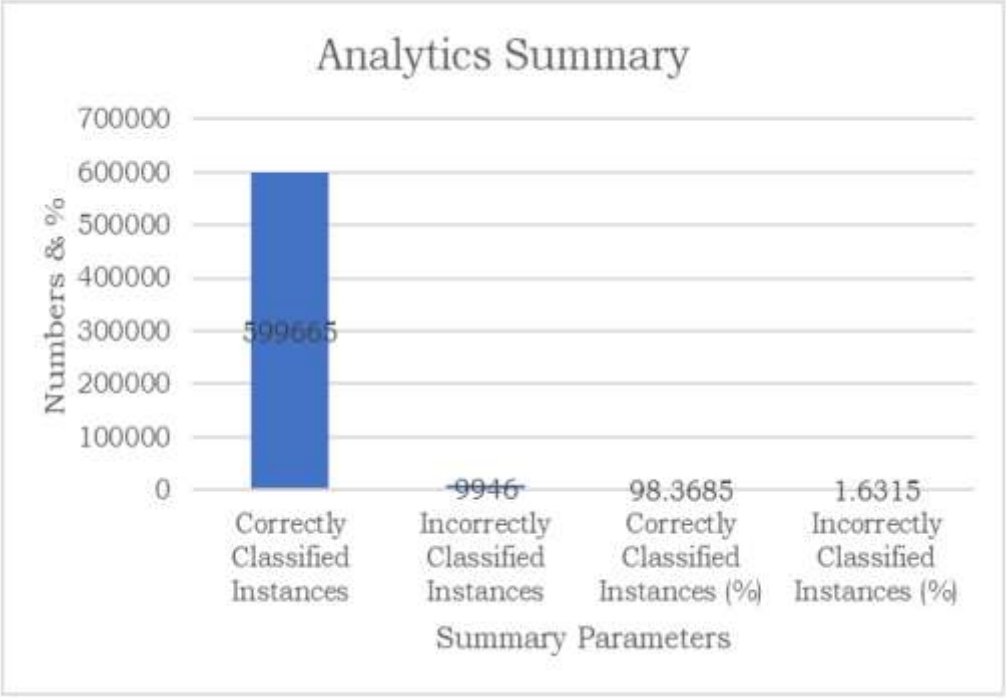

Fig. 5 Predictive Classification Results (A)

10.48047/jocaaa.2024.33.05.27

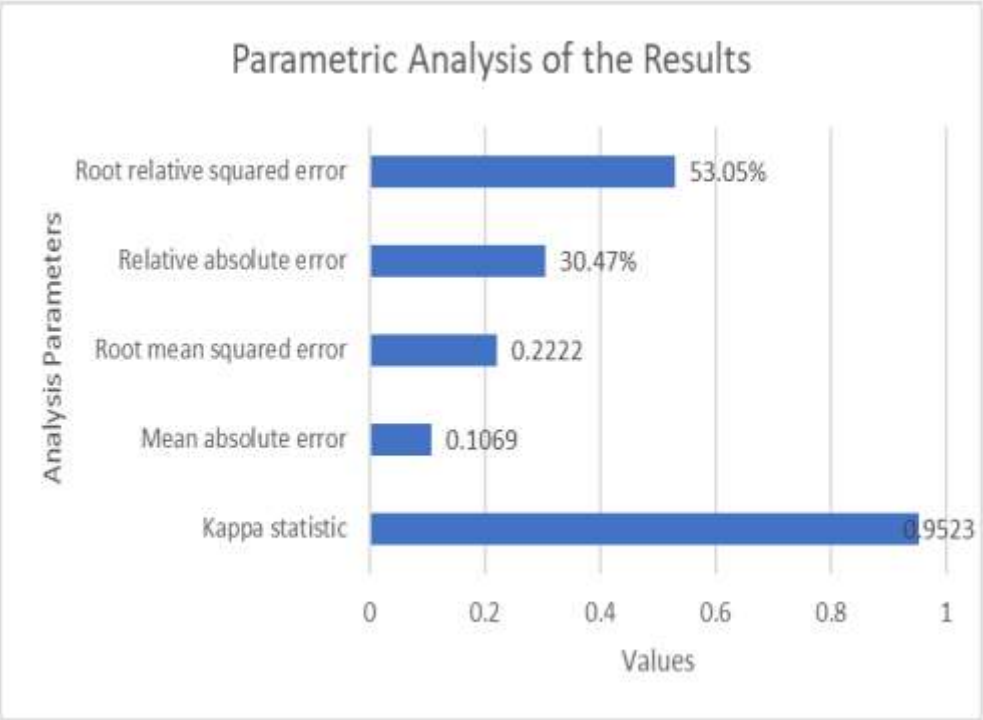

Fig. 6 Predictive Classification Results (B)

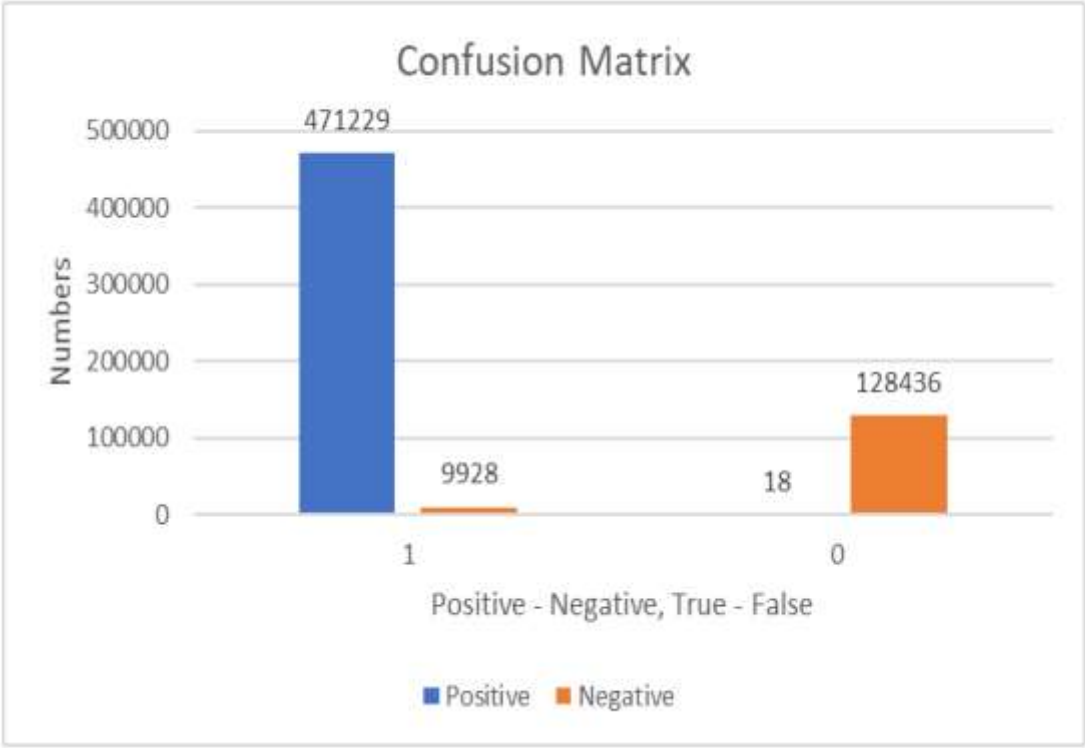

Fig. 7 Confusion Matrix

10.48047/jocaaa.2024.33.05.27

Table-3 Detailed Analytics

Class	1	2	Weighted Avg.
TP Rate	1.000	0.928	0.984
FP Rate	0.072	0.000	0.055
Precision	0.979	1.000	0.984
Recall	1.000	0.928	0.984
F-Measure	0.990	0.963	0.983
MCC	0.953	0.953	0.953
ROC Area	0.990	0.990	0.990
PRC Area	0.997	0.979	0.993

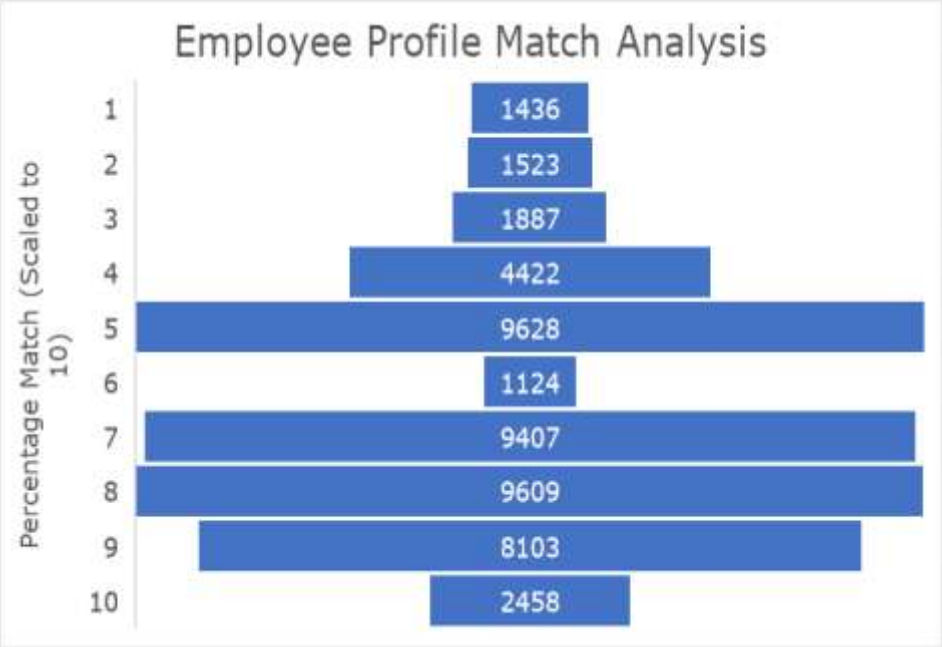

Fig. 8 Employee Profile Match Analysis

The study presents the results of a proposed framework and algorithms for anomaly reduction in a dataset provided by the Public 2020 Stack Overflow Developer Survey [24]. The dataset is divided into sub-sections and analyzed using a standard length domain moving averages approach. The results show that the DMV-DSL algorithm achieves 100% accuracy in missing value detection and replacement from the initial employee dataset. The merged dataset domain-specific outlier and missing value analysis shows 100% accuracy. The Domain Specific Rule engine algorithm eliminates un-realistic data in the datasets using a double clustering method [25]. The OR-DSRE algorithm shows 100% accuracy during the missing value analysis and nearly 90% accuracy during the domain-specific outlier detection process. The generic outlier removal outcomes show nearly 100% accuracy, with the iterative outlier identification and removal algorithm identifying all outliers within 5 iterations using the

10.48047/jocaaa.2024.33.05.27

DDOD-R algorithm [26]. The attribute reduction results are also presented, showing that the proposed framework and algorithms are effective in reducing anomalies in the dataset. The findings are presented in five segments, providing a comprehensive analysis of the proposed framework and algorithms [27]. The study focuses on the classification of a dataset using a regression method called MR-PTCP. The dataset is divided into five stages, with the first stage involving the classification of the dataset. The second stage involves the reduction of the dataset based on the rank of the attributes. The third stage involves the elimination of the highest-importance attributes [28]. The final analysis shows that the attributes identified till the fifth iteration are considered optimal. The results show that the accuracy of the proposed method during multiple regression is extremely high, but not compromised under overfitting constraints [29]. The confusion matrix provides a significant understanding of the defects in the algorithms in terms of true-negative and falsepositive. The confusion matrix obtained from this proposed method is used to compare the accuracy obtained from the proposed method with the parallel research outcomes.

Conclusion:

This study uses the Stack Overflow benchmark dataset and a synthetic dataset to analyze employee-related data. The DMV-SDL technique is used for employee-related analysis, reducing processing time. The OR-DSRE technique is used for domain-specific outlier imputation, merging the datasets. The DDOD-R method is applied for general outlier imputations. The framework achieves 99 percent accuracy in imputations. The study also provides a multi-purpose domain-specific data preprocessing framework for large-scale company data, increasing the reliability of data-driven corporate choices. The framework also provides a recommendation system for the best team composition based on a broad range of characteristics. The research also shows a methodology for predicting upskilling needs for software projects based on factors such as technicalities, interpersonal, and financial aspects. The findings are noteworthy and should be considered a standard for future research.

References:

1. B. Wang and W. Zhang, "Research on Edge Network Topology Optimization Based on Machine Learning," *2023 5th International Conference on Applied Machine Learning (ICAML)*, Dalian, China, 2023, pp. 41-46, doi: 10.1109/ICAML60083.2023.00018.
2. Bolognani S, Bof N, Michelotti D, Identification of power distributionnetwork topology via voltage correlation analysis [A]. // *IEEE Conference on Decisionand Control[C]*, Piscataway : IEEE, 2020 : 1659–1664. 11.
3. R. Salama, F. Al-Turjman, P. Chaudhary and S. P. Yadav, "(Benefits of Internet of Things (IoT) Applications in Health care - An Overview)," *2023 International Conference on Computational Intelligence, Communication Technology and Networking (CICTN)*, Ghaziabad, India, 2023, pp. 778-784, doi: 10.1109/CICTN57981.2023.10141452.
4. Mouha, R. A. (2021). Internet of Things (IoT). *Journal of Data Analysis and Information Processing*, 9 (2), 77–101.

10.48047/jocaaa.2024.33.05.27

5. Bottou, Leon. Large-scale machine learning with stochastic gradient descent [A // *Proceedings ofCOMPSTAT2021 [C]*, Berlin : Springer, 2021, 177–186.
6. Singh, R. P., Javaid, M., Haleem, A., & Suman, R. (2020). Internet of things (IoT) applications to fight against COVID-19 pandemic. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 14 (4), 521–524.
7. Lee, I., & Lee, K. (2015). The Internet of Things (IoT): Applications, investments, and challenges for enterprises. *Business horizons*, 58 (4), 431–440.
8. Hassan, A., Prasad, D., Khurana, M., Lilhore, U. K., & Simaiya, S. (2021). Integration of internet of things (IoT) in health care industry: an overview of benefits, challenges, and applications. *Data Science and Innovations for Intelligent Systems*, 165–180.
9. Ahmad, M. O., & Siddiqui, S. T. (2022). The Internet of Things for healthcare: benefits, applications, challenges, use cases and future directions. In Advances in Data and Information Sciences: Proceedings of ICDIS 2021 (pp. 527–537). Singapore : SpringerSingapore.
10. Khujamatov, K. E., Reypnazarov, E. N., & Lazarev, A. P. (2021). Modern methods of testing and information security problems in IoT. *Bulletin of TUIT: management and communication technologies*, 4 (2), 4.
11. Hamid, S., Bawany, N. Z., Sodhro, A. H., Lakhan, A., & Ahmed, S. (2022). A Systematic Review and IoMT Based Big Data Framework for COVID-19 Prevention and Detection. *Electronics*, 11 (17), 2777.
12. Hireche, R., Mansouri, H., & Pathan, A. S. K. (2022). Security and privacy management in Internet of Medical Things (IoMT): A synpaper. *Journal of Cybersecurity and Privacy*, 2 (3), 640–661.
13. Dwivedi, R., Mehrotra, D., & Chandra, S. (2022). Potential of Internet of Medical Things (IoMT) applications in building a smart healthcare system: A systematic review. *Journal of oral biology and craniofacial research*, 12 (2), 302–318.
14. El Khatib, M., Hamidi, S., Al Ameeri, I., Al Zaabi, H., & Al Marqab, R. (2022). Digital Disruption and Big Data in Healthcare-Opportunities and Challenges. *ClinicoEconomics and Outcomes Research*, 563–574.
15. ikumar, K. S., Prathiba, S. B., Alazab, M., Gadekallu, T. R., Pandya, S., Khan, J. M., & Moorthy, R. S. (2022). FL-PMI: federated learning-based person movement identification through wearable devices in smart healthcare systems. *Sensors*, 22 (4), 1377.
16. C. J. Martínez and S. Galmés, "Analysis of the primary attacks on IoMT Internet of Medical Things communications protocols," *2022 IEEE World AI IoT Congress (AIIoT)*, Seattle, WA, USA, 2022, pp. 01-07, doi: 10.1109/AIIoT54504.2022.9817252.
17. T. R. N and R. Gupta, "A Survey on Machine Learning Approaches and Its Techniques:," *2020 IEEE International Students' Conference on Electrical,Electronics and Computer Science (SCEECS)*, Bhopal, India, 2020, pp. 1-6, doi: 10.1109/SCEECS48394.2020.190.
18. A. A. Eteng, S. Kamal, and A. Rahim, "*RFID in the Internet of Things*," pp. 135–151, 2018.
19. O. O. D. Ruas, "*Internet de las cosas en salud IoMT-Octubre 2020 Internet de las cosas y su aplicación en el sector de la salud. Internet of the things and their application in the sector of the*," no. October, 2020, doi: 10.13140/RG.2.2.23348.27520.
20. D. Koutras, G. Stergiopoulos, T. Dasaklis, P. Kotzanikolaou, D. Glynos, and C. Douligeris, "Security in IoMT Communications: A Survey," *Sensors (Basel).*, vol. 20, no. 17, pp. 1–49, 2020, doi: 10.3390/s20174828.
21. F. Qureshi and S. Krishnan, "Wearable hardware design for the internet of medical things (IoMT)," *Sensors (Switzerland)*, vol. 18, no. 11, 2018, doi: 10.3390/s18113812.

10.48047/jocaaa.2024.33.05.27

22. B. Ayten, O. Akmandor, and N. K. Jha, "*Smart Health Care*," no. january 2018, pp. 29–37, 2020.
23. A. Mosenia, "*Addressing Security and Privacy Challenges in Internet of Things*," arXiv, no. January, 2018.
24. S. U. Amin and M. S. Hossain, "Edge Intelligence and Internet of Things in Healthcare: A Survey," *IEEE Access*, vol. 9, pp. 45–59, 2021, doi: 10.1109/ACCESS.2020.3045115.
25. Y. Sun, F. P. W. Lo, and B. Lo, "Security and Privacy for the Internet of Medical Things Enabled Healthcare Systems: A Survey," *IEEE Access*, vol. 7, pp. 183339–183355, 2019, doi: 10.1109/ACCESS.2019.2960617.
26. T. Jadhav *et al*., "Predicting Urban Land Cover Using Classification: A Machine Learning Approach," *2023 IEEE 11th Region 10 Humanitarian Technology Conference (R10-HTC)*, Rajkot, India, 2023, pp. 450-454, doi: 10.1109/R10-HTC57504.2023.10461930.
27. S. Subramanian, B. Tseng, R. Barbieri and E. N. Brown, "Unsupervised Machine Learning Methods for Artifact Removal in Electrodermal Activity," *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Mexico, 2021, pp. 399-402, doi: 10.1109/EMBC46164.2021.9630535.
28. "IEEE Approved Draft Guide for Architectural Framework and Application of Federated Machine Learning," in *IEEE P3652.1/D6.1, July 2020* , vol., no., pp.1-70, 24 Sept. 2020.
29. "IEEE Draft Guide for Architectural Framework and Application of Federated Machine Learning," in *IEEE P3652.1/D6, April 2020* , vol., no., pp.1-70, 1 June 2020.