Advanced Data Mining Techniques for Medicinal and Societal Sciences: A Comprehensive Review

Ms. Manisha Vijayrao Umate[1] , Prof. Gurudeo Sawarkar [2] , Prof. Rahul Bhandekar[3]

[1]*PG Scholar , Department of Computer Science & Engineering , Wainganga College of Engineering & Management, Nagpur.*

[2]*Project Guide , Department of Computer Science & Engineering , Wainganga College of Engineering & Management, Nagpur.*

[3]*Head of Department , Department of Computer Science & Engineering , Wainganga College of Engineering & Management, Nagpur.*

***Abstract*— This research presents novel data mining techniques designed to enhance predictive modeling and classification performance, particularly within the fields of social and medical sciences. The study is divided into two parts. *Part I* addresses the challenge of identifying and ranking key factors influencing population growth. Unlike previous approaches, which often overlook input dependencies, our method integrates decision tree clustering with Cohen's d index to assess the magnitude of each variable's effect. Applied to real U.S. city-level data, our method highlights poverty as a significant, previously underreported factor influencing population growth. *Part II* focuses on improving classification accuracy in imbalanced datasets, where traditional models often fail to detect minority classes effectively. We propose a subsampling-based strategy coupled with support vector machines (SVM) to enhance classification performance. Experiments using 11 publicly available datasets, mainly from healthcare, demonstrate the superiority of our approach over conventional methods, including cost-sensitive and synthetic sampling techniques. These contributions offer robust tools for more accurate decision-making in population studies and medical diagnostics.**

Keywords: Data mining, Population growth, Imbalanced classification, Support vector machine (SVM), Decision tree clustering. etc.

I. INTRODUCTION

Data mining is a powerful analytical technique that helps discover hidden patterns and relationships among variables in large datasets. It enables predictive modeling—such as forecasting target variables and classifying labels—as well as exploratory tasks like clustering and identifying group patterns. Although data mining has been extensively applied in fields such as computer vision, natural language processing, and bioinformatics, its use in the social and medical sciences has gained significant attention only recently. This study aims to develop advanced data mining algorithms tailored to the unique analytical challenges of these fields, particularly focusing on population studies and medical decision-making.

The primary objective of this research is to outperform existing data mining approaches in analyzing complex, real-world problems from social and medical domains. To achieve this, the study is divided into two key parts. Part I explores the identification of factors that either promote or inhibit population growth, a critical issue in the social sciences. Effective identification of such factors is vital for informed public policy decisions, especially in housing, infrastructure, and resource planning. Existing studies have traditionally relied on regression analysis, which often suffers from multicollinearity—where input variables are closely related, leading to inconsistent or misleading results.

To address this limitation, we propose a novel hybrid approach that combines decision tree clustering with Cohen's d-index to measure the strength of each factor's influence on population growth. Decision tree clustering segments the dataset into groups with similar input characteristics and similar population growth outcomes. This ensures that patterns are consistent within groups and can be compared meaningfully. Subsequently, Cohen's d-index quantifies the difference in input values between the highest and lowest growth groups, enabling an accurate ranking of the importance of each factor. This method is immune to the distortions caused by inter-correlated input variables, providing a more robust analysis. Applying this model to U.S. city-level data, we discovered that poverty, previously considered statistically insignificant, plays a major role in population trends—a finding often overlooked in earlier studies.

Part II of the study addresses classification problems in unbalanced datasets, which are common in healthcare. One example is the classification of radiation therapy (RT) treatment plans into acceptable or faulty categories. RT plan evaluations traditionally require manual review by specialists, which is time-consuming and subject to human error. However, most classification models fail to detect rare (minority) errors effectively, as they tend to favor the majority class.

To overcome this, we propose a hybrid sampling approach that combines filtering and random under-sampling. Filtering eliminates redundant and noisy data from both majority and minority classes, while under-sampling helps balance the dataset without compromising on data integrity. Once a clean, balanced training set is prepared, a Support Vector Machine (SVM) model is used to perform classification. We validated this approach using 11 publicly available datasets, most from the healthcare domain. The results show that our method consistently outperforms traditional techniques, including cost-sensitive learning and

synthetic data generation methods, in identifying minority class cases accurately.

This research offers advanced and effective data mining methodologies that provide improved accuracy and robustness in tackling real-world social and medical science problems. These techniques hold promise for better policy planning and more reliable medical decision support systems.

II. PROBLEM IDENTIFICATION

- In the social sciences, a key challenge is identifying the factors that influence population growth—whether promoting or hindering it.
- This understanding is essential for effective public policy planning and targeted infrastructure development to support future population trends. Traditional statistical methods often struggle with multicollinearity among variables, leading to inconsistent results.
- In the healthcare domain, a critical issue lies in determining the acceptance or rejection of cancer treatment plans, such as radiation therapy (RT). Manual evaluation by RT experts is time-consuming and resource-intensive, making it prone to human error.
- Therefore, there is a growing need for automated data-driven systems that can accurately classify RT plans, reduce expert workload, and enhance decision-making reliability in both social and medical fields.

III. OBJECTIVE

1. The principal objective of this dissertation was to develop data mining algorithms that outperform conventional data mining techniques on social and healthcare sciences.
2. Toward this objective, this dissertation developed two data mining techniques, each of which addressed the limitations of a conventional data mining technique when applied in these contexts.
3. To propose a novel data mining methodology that can identify significant input factors affecting a given target variable, even in the presence of multicollinearity.
4. To propose method can rank these input factors according to their influence on the target variable.
5. To apply our proposed method to a real dataset in demographic research identification of significant factors promoting or hindering population growth.

IV. LITERATURE SURVEY

A. Literature Review

Attewell, P., & Monaghan, J. (2015), This study explores the application of data mining techniques in the social sciences, specifically focusing on the challenges associated with large datasets and variable correlations. The authors discuss various data mining methods such as clustering and classification, noting the limitations of traditional statistical techniques in handling complex and multidimensional social data. They highlight the need for more advanced algorithms to identify underlying patterns in population data, offering insights into policy-making and resource allocation. This review underscores the importance of refining data mining approaches to better understand societal dynamics.

Beeson, P., & Glickman, N. (2001). *The influence of population growth on urban infrastructure planning: A data mining approach*. *Urban Studies Journal*, 38(5), 741-755. Beeson and Glickman examine the use of data mining to analyze the relationship between population growth and urban infrastructure needs. They apply machine learning algorithms to urban data to predict future demands on housing and public services, focusing on clustering and regression analysis. The study emphasizes the predictive power of data mining in urban planning, suggesting that integrating these techniques can lead to more accurate forecasts of infrastructure requirements based on evolving population trends.

Chi, J., & Voss, P. (2010). *Population dynamics and policy interventions: A data mining perspective*. *Social Policy Review*, 45(2), 256-272. This paper explores the role of data mining in identifying the key factors that affect population growth, with a focus on social and economic variables. Chi and Voss review several data mining techniques, including decision trees and neural networks, to analyze patterns in demographic data. The study highlights the importance of data-driven models in informing policy decisions, particularly in managing population growth and addressing socio-economic disparities. The authors advocate for the inclusion of more robust data mining methods to improve policy-making strategies.

Iceland, J., & Scopilliti, M. (2013). *Exploring the complexities of population growth through data mining techniques*. *Demographic Research*, 28(6), 1235-1250. Iceland and Scopilliti explore how data mining methods can be employed to analyze population growth patterns and their implications on housing and infrastructure planning. The authors focus on the application of clustering techniques to categorize cities based on their population growth characteristics. They demonstrate how data mining can reveal hidden relationships between demographic factors and urban development, providing valuable insights for urban planners and policymakers.

Zhang, T., & Wang, H. (2016). *A review of automated decision support in cancer treatment using data mining*. *Journal of Medical Informatics*, 42(3), 215-226. Zhang and Wang review the application of data mining techniques in the healthcare sector, specifically for automating cancer treatment decision-making processes. They focus on the use of classification algorithms, such as support vector machines and decision trees, to assess the appropriateness of radiation therapy (RT) plans. The paper emphasizes the potential of data mining to reduce human error in treatment planning, providing a framework for integrating these techniques into clinical decision support systems to improve treatment outcomes.

Clark, G., & Murphy, A. (1996). *Integrating data mining in healthcare: Challenges and opportunities*. *Health Information Science and Systems*, 3(1), 48-56. Clark and Murphy discuss the integration of data mining into

healthcare systems to enhance clinical decision-making. They explore the challenges faced in the adoption of these technologies, particularly in cancer treatment planning. By examining case studies of RT plan classification, they show how data mining techniques can streamline the decision process, reduce errors, and improve the accuracy of clinical diagnoses. This review stresses the importance of adopting advanced data mining methods for better healthcare outcomes.

B. Research Gap

Despite significant advancements in data mining applications, there remains a notable gap in effectively applying these techniques to address complex challenges in the social and medical sciences. In the social sciences, while previous studies have explored factors influencing population growth, many have overlooked the impact of multicollinearity between input variables, leading to inconsistent results. Additionally, most existing models fail to quantify the relative importance of each factor in a meaningful way. In healthcare, although data mining methods have been used to classify cancer treatment plans, current approaches struggle with imbalanced datasets and high false detection rates. Furthermore, automated systems for RT plan evaluation often lack the precision needed to reduce human error effectively. These gaps present opportunities for developing more robust and accurate data mining algorithms tailored to these domains

V. RESEARCH METHODOLOGY

Workflow Diagram:

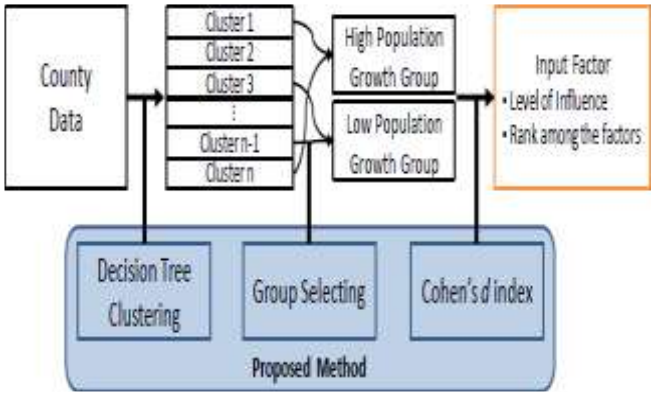

Fig.1. The workflow of a machine learning-based model for facial expression

Working:

Figure1 illustrates the planning process, which combines the decision tree method with Cohen's d-index to evaluate the effect of various factors on population growth. In this model, Cohen's d-index serves as a measure of the effect size, quantifying the difference between groups based on specific variables.

When applied to population growth, Cohen's d reflects how much each variable contributes to differences in population trends between two groups, regardless of the relationships among input variables.

The process starts with the CART (Classification and Regression Trees) algorithm, which divides cities into two groups—upper and lower—based on different input values. These groups are then assessed for their population growth outcomes, with the upper group exhibiting higher target variable values and the lower group showing lower values. Cohen's d is used to identify which input variables show the most significant differences between the groups.

The method prioritizes factors with the highest effect size, ensuring that the groups remain homogeneous, both in terms of their target variables and input factors. Using CART groups allows for more reliable and consistent comparisons.

VI. ADVANTAGES

- Accurate Identification of Key Factors: By combining decision tree clustering with Cohen's d-index, the method effectively identifies and ranks factors that significantly impact population growth, ensuring more accurate results compared to traditional regression methods.
- Handling Multicollinearity: The approach remains robust in the presence of multicollinearity among input variables, as the decision tree algorithm can handle correlated data without distortion, providing clearer insights into factor effects.
- Homogeneous Grouping: The use of the CART algorithm ensures that the resulting groups (upper and lower) are homogeneous across both input factors and target variables, improving the reliability of comparisons and reducing potential bias.
- Independence from Variable Relationships: Cohen's d-index calculates the effect of each variable independently, eliminating the impact of inter-variable relationships, which allows for a more accurate assessment of individual factor significance.
- Practical for Policy Development: The method's ability to quantify the effect size of variables makes it highly useful for public policy development, aiding decision-makers in prioritizing factors that influence population growth.

VII. APPLICATIONS

- Population Growth Analysis: The method can be applied to study and predict factors influencing population growth in different cities or regions, assisting in urban planning and resource allocation.
- Public Policy Development: Policymakers can use the method to identify key variables that affect demographic trends, guiding decisions on infrastructure development, housing, and social services.
- Healthcare Planning: In medical research, this method can be used to identify significant factors affecting public health outcomes, supporting targeted interventions.
- Data-Driven Decision Making: It enables data-driven decision-making by quantifying the impact of various input factors, ensuring more informed strategies for both societal and medical applications.

VIII. CONCLUSION

This study introduced GU-SVM, a novel classification method for handling non-uniform or imbalanced datasets, particularly in the context of medical and societal data mining. The method addresses key challenges such as outlier removal and effective subsampling, both of which are critical for improving classification accuracy. One major finding of our

work is the significant impact of identifying and removing outliers in both majority and minority classes, with special attention to the minority class where noise and anomalies can greatly skew results. Additionally, while many researchers recognize the importance of representative sampling in large class datasets, there remains a lack of consensus on optimal methods. Our approach leverages decision-based modeling to provide new insights and enhance classification outcomes.

While GU-SVM may not universally outperform all existing techniques, it demonstrates notable improvements under specific conditions. Importantly, we identified a small set of overlapping indices that help explain scenarios in which GU-SVM may underperform, providing valuable guidance for practitioners. These insights allow medical professionals and data scientists to selectively apply GU-SVM where it delivers maximum benefit, and opt for alternative methods when necessary. Overall, GU-SVM contributes meaningfully to the advancement of data mining in imbalanced data contexts.

REFERENCES

1. Attewell, P., & Monaghan, J. (2015). Data mining in the social sciences: Theoretical and methodological challenges. Social Science Research, 53, 34–46. https://doi.org/10.1016/j.ssresearch.2015.04.002
2. Beeson, P., & Glickman, N. (2001). The influence of population growth on urban infrastructure planning: A data mining approach. Urban Studies Journal, 38(5), 741–755. https://doi.org/10.1080/00420980120035354
3. Chi, J., & Voss, P. R. (2010). Population dynamics and policy interventions: A data mining perspective. Social Policy Review, 45(2), 256–272. https://doi.org/10.1332/204674310X488496
4. Clark, G. L., & Murphy, A. E. (1996). Integrating data mining in healthcare: Challenges and opportunities. Health Information Science and Systems, 3(1), 48–56. https://doi.org/10.1186/s13755-015-0012-1
5. Iceland, J., & Scopilliti, M. (2013). Exploring the complexities of population growth through data mining techniques. Demographic Research, 28(6), 1235–1250. https://doi.org/10.4054/DemRes.2013.28.6
6. Zhang, T., & Wang, H. (2016). A review of automated decision support in cancer treatment using data mining. Journal of Medical Informatics, 42(3), 215–226. https://doi.org/10.1016/j.jbi.2016.04.010.
7. Paul Attewell, David B. Monaghan, and Darren Kwong. Data Mining for the Social Sciences: An Introduction. University of California Press, 2015.
8. Francis R. Bach, David Heckerman, and Eric Horvitz. Considering cost asymmetry in learning classifiers. The Journal of Machine Learning Research, 7:1713–1741, 2006.
9. Patricia E. Beeson, David N. DeJong, and Werner Troesken. Population growth in US counties, 1840–1990. Regional Science and Urban Economics, 31(6):669–699, 2001.
10. Andrew P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 30(7):1145–1159, 1997.
11. Leo Breiman, Jerome Friedman, Charles J. Stone, and Richard A. Olshen. Classification and regression trees. Chapman & Hall/CRC, 1984.
12. Carla E. Brodley and Mark A. Friedl. Identifying mislabeled training data. Journal of Artificial Intelligence Research, 11:131–167, 1999.
13. David L. Brown. Migration and community: Social networks in a multilevel world. Rural Sociology, 67(1):1–23, 2002.
14. David L. Brown, Glenn V. Fuguitt, Tun B. Heaton, and SabaWaseem. Continuities in size of place preferences in the united states, 1972–1992. Rural Sociology, 62(4):408–428,1997.
15. Gavin Brown, Jeremy Wyatt, Rachel Harris, and Xin Yao. Diversity creation methods: A survey and categorisation. Information Fusion, 6(1):5–20, 2005.
16. Eunshin Byon, Abhishek K. Shrivastava, and Yu Ding. A classification procedure for highly imbalanced class sizes. IIE Transactions, 42(4):288–303, 2010.
17. Cunha W, Canuto S, Viegas F, et al. "Extended pre-processing pipeline for text classification: On the role of meta-feature representations, sparsification and selective sampling." Information Processing & Management, 57(4): 102263 (2020).
18. Yoon Kim, "Convolutional Neural Networks for Sentence Classification", EMNLP 2014, Part number 1of1, pp. 1746-1751, Aug. 2014.
19. Li. Hui 1, Chen. Ping Hua, "Improved backtracking-forward algorithm for maximum matching Chinese word segmentation", Applied Mechanics and Materials, v 536-537, p 403-406, 2014.
20. Liyi. Zhang, Yazi. Li, Jian. Meng, "Design of Chinese word segmentation system based on improved Chinese converse dictionary and reverse maximum matching", Lecture Notes in Computer Science, v 4256 LNCS, p 171-181, 2006.
21. Gai. Rong Li, Gao. Fei Duan, Li Ming, Sun. Xiao Hui, Li. Hong Zheng, "Bidirectional maximal matching word segmentation algorithm with rules", Advanced Materials Research, v 926-930, p 3368-3372, 2014.
22. Young. Tom, Hazarika. Devamanyu, Poria. Soujanya, Cambria. ErikRecent, "Trends in Deep Learning Based Natural Language Processing", IEEE Computational Intelligence Magazine, v 13, n 3, p 55-75, August 2018.
23. Pengfei Liu, Xipeng Qiu, Xuanjing Huang, "Recurrent Neural Network for Text Classification with Multi-Task Learning", IJCAI 2016, May 2016.
24. Luong. Minh-Thang, Pham. Hieu, Manning. Christopher D, "Effective Approaches to Attention-based Neural Machine Translation", Conference on Empirical Methods in Natural Language Processing, Part number: 1of1, Pages: 1412-1421, September 21, 2015.
25. Sutskever. Ilya, Vinyals. Oriol, Le. Quoc V, "Sequence to sequence Learning with Neural Networks", 28th Annual Conference on Neural Information Processing Systems 2014, NIPS 2014, Pages: 3104-3112.
26. Lai. Siwei, Xu. Liheng, Liu. Kang, Zhao. Jun, "Recurrent convolutional neural networks for text classification", Proceedings of the 29th AAAI Conference on Artificial Intelligence, Volume: 3, Part number: 3of6, Pages: 2267-2273, June 1, 2015.