2023 IEEE Symposium Series on Computational Intelligence (SSCI)
Mexico City, Mexico. December 5-8, 2023

Unveiling Precision Medicine with Data Mining: Discovering Patient Subgroups and Patterns

1st Nasim Sadat Mosavi
algoritmi research center
University of Minho
Guimaraes, Portugal
0000-0002-6153-2524

1nd Manuel Filipe Santos
algoritmi research center
University of Minho
Guimaraes, Portugal
0000-0002-5441-3316

Abstract— **Data mining techniques, prominently clustering, assume a pivotal role in fortifying precision medicine by facilitating the revelation of patient subgroups that share common attributes. By harnessing clustering for the analysis of data behavior within the realm of precision medicine, distinctive disease patterns, and progression dynamics are unveiled, thereby contributing to the formulation of precisely tailored treatment strategies. This paper aims to present the outcomes derived from a clustering analysis applied to diverse clinical datasets encompassing critical facets such as vital signs, laboratory exams, medications, sepsis, Glasgow Coma Scale, procedures, interventions, diagnostics, and admission/discharge records. This compilation of datasets pertains to a cohort of seventy patients. The resultant analysis uncovers intrinsic patterns and relationships residing within intricate datasets. Executed following the rigorous CRISP-DM methodology, this discovery study identified three distinct clusters that group similar data characteristics, encompassing both categorical and numerical clinical data, and resulted in three major groups: patients with stable health conditions, recovery stage, and at risk. This pivotal outcome catalyzes future endeavors, including classification tasks aimed at identifying new patients within specific classes, thereby advancing the horizons of precision medicine**.

Keywords—data mining, precision medicine, intelligence, clinical decision-making, clustering

I. Introduction

In the era of advanced healthcare technologies and personalized treatment strategies, the intersection of data mining and medicine has emerged as a transformative force. [1],[2].Data mining techniques, particularly clustering analysis, have taken center stage in revolutionizing the landscape of clinical research and practice. One of the most promising applications of data mining in healthcare is the pursuit of precision medicine, a paradigm that aims to tailor medical interventions to individual patient characteristics. This introduction delves into the pivotal role of data mining, with a specific focus on clustering analysis, in unlocking the potential of precision medicine [3],[4].

Data mining, often described as the process of extracting hidden knowledge and patterns from large datasets, has found extensive utility across diverse domains. In healthcare, it has transcended traditional boundaries to reshape the way medical information is harnessed and interpreted [5], [6].

Precision medicine, the embodiment of personalized healthcare, relies on the precise understanding of patient heterogeneity and the identification of patient subgroups sharing common attributes. This is precisely where the power of data mining, especially clustering analysis, comes to the forefront [7]–[11].

Clustering analysis is a data mining technique that involves the categorization of data points into distinct groups based on shared characteristics. It serves as a vital tool for partitioning complex datasets into meaningful segments, thereby unraveling intricate relationships and patterns that might otherwise remain hidden [12]. The application of clustering analysis in the realm of precision medicine holds immense promise, as it allows for the identification of patient subgroups that exhibit similar disease profiles, treatment responses, and outcomes. By uncovering these subgroups, clustering analysis facilitates the development of tailored treatment regimens, diagnostic approaches, and prognostic models [13].

The concept of precision medicine encompasses a departure from the traditional "one-size-fits-all" approach to healthcare. Instead, it embraces a patient-centric model that recognizes the unique genetic, physiological, and environmental factors contributing to an individual's health and disease. Clustering analysis serves as a cornerstone in this transformative shift, offering a data-driven means to stratify patients based on multifaceted attributes. This not only enhances the understanding of disease heterogeneity but also paves the way for targeted interventions that optimize therapeutic efficacy and minimize adverse effects[14], [15].

As we embark on this exploration of data mining's role in precision medicine, it becomes evident that the amalgamation of these two disciplines has the potential to reshape the landscape of healthcare delivery. The subsequent sections of this paper delve into the intricate interplay between data mining techniques, clustering analysis, and precision medicine. Through a comprehensive analysis of clinical datasets encompassing vital signs, laboratory exams, medications, disease severity indicators, and more, we unveil the dynamic patterns and patient subgroups that remain hidden within the vast expanse of medical data. This journey not only holds implications for our understanding of disease but also lays the foundation for the design of individualized treatment strategies that hold the promise of better patient outcomes and improved quality of care [16].

In this experimental endeavor, we adhered to the esteemed standard of knowledge discovery by meticulously following the CRISP-DM methodology. CRISP-DM, which stands for Cross-Industry Standard Process for Data Mining, is a widely recognized and comprehensive methodology for guiding data mining and knowledge discovery projects. It provides a

978-0-7381-4408-5/23/$31.00 ©2023 IEEE

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

structured framework that assists in efficiently and effectively managing the various stages of a data-mining project. CRISP-DM is widely adopted due to its flexibility and applicability across diverse industries and problem domains. The methodology consists of six main phases (business understanding, data understanding, data preparation, modeling, evaluation, and deployment), each serving a specific purpose and contributing to the overall success of the data-mining project.

This insight is invaluable as it provides a clear and concise differentiation among the clusters based on a specific medical parameter, allowing for targeted analysis and potentially guiding further investigation into the underlying health conditions or anomalies associated with each cluster. It underlines the potential of clustering techniques to uncover meaningful and clinically relevant patterns in the data, which could have implications for medical diagnosis, treatment planning, and patient care [17].

While the fundamental concept of defining the Temporal Framework for Interrelating Clinical Events (TS) has been previously discussed in our publications, it's important to note that this latest version represents a substantial improvement and optimization. In this iteration, our integrated platform continues to employ the same data structure as detailed in our two papers, namely "Data Engineering to Support Intelligence for Precision Medicine in Intensive Care" and "Intelligent Decision Support System for Precision Medicine: Time Series Multi-variable Approach for Data Processing." This consistent data structure allows us to leverage the capabilities of descriptive analytics effectively, revealing intricate patterns and providing invaluable insights.

Our paper is thoughtfully structured to provide a coherent and logical presentation of our research journey. We have meticulously organized our content to align with our application and objectives, which are elaborated upon in Section II. In Section III, we delve into a detailed exploration of each dataset, unraveling the intricate variables associated with patients' clinical backgrounds. Subsequently, we offer a comprehensive overview of the pivotal data preparation steps in the following section.

As we progress, our paper navigates through the experimental trajectory with careful consideration. We engage in the application of advanced modeling techniques, unraveling insightful patterns and relationships embedded within the data. The fruits of our modeling endeavors undergo a stringent evaluation process, which we meticulously elucidate and discuss in the ensuing section. This thorough evaluation allows us to discerningly interpret the results, shedding light on their implications and significance in the context of precision medicine.

Our deliberate organizational approach ensures a seamless and logical flow that guides readers through essential stages of our research journey. By methodically traversing the application, objectives, experimental process, data preparation, modeling, evaluation, and discussion phases, we present a comprehensive narrative that facilitates a profound understanding of our data mining methodology and its far-reaching implications within the realm of precision medicine.

II. PPLICATION|BUSINESS UNDERSTANDING

The primary business objective of this endeavor is to harness the power of data mining, specifically clustering analysis, to unearth hidden relationships and patterns within clinical datasets. By doing so, we aim to identify distinct patient subgroups characterized by shared attributes.

This clustering of patients into meaningful groups serves as a crucial foundation for the broader goal of precision medicine – tailoring medical interventions to the unique characteristics of individual patients. The ultimate objective is to enhance clinical decision-making, optimize treatment regimens, and improve patient outcomes through personalized healthcare strategies.

From a data mining perspective, our objective revolves around applying clustering analysis to diverse clinical datasets encompassing a wide array of critical facets such as vital signs, laboratory exams, medications, disease severity indicators, and more. The aim is to partition this complex and multidimensional data into coherent clusters, each representing a distinct patient subgroup with similar attributes.

III. DATA UNDERSTANDING

As depicted in Table 1, the dataset comprises 10 distinct categories, each offering valuable insights into different aspects of patient health and medical interactions.

“Vital Signs” (43,9025 records and 108 biological variables), focusing on vital signs that play a crucial role in assessing a patient's overall condition.“Lab Results” (11,3320 records and 9 variables) provides information about various laboratory exams. “Procedures” (911 records and 6 variables) sheds light on the medical actions recommended and prescribed by healthcare professionals. In addition, “Sepsis (Gravity Score)” capturing data from 176 records and 6 variables, this category gauges the severity of patients' conditions. “Glasgow Coma Scale”: containing 861 records and 6 variables, evaluates patients' consciousness levels, and “Diagnosis”: (124 records and 9 variables) focuses on recording signs, symptoms, and potential medical conditions. “Medication Prescriptions”: This category provides data on medications prescribed by clinicians, helping track patient treatment plans. “Intervention Actions”: Capturing information about various interventions, this category showcases actions taken to manage patients' health.

ICU Admission and Discharge (“Admin-Discharge”) houses data about patients' admissions and discharges from the Intensive Care Unit (ICU), facilitating comprehensive patient care management. Reference Dataset: Serving as a point of reference, this dataset includes episode and process numbers, wherein the episode number represents clinical events, while the process number signifies patient identity. In Table 1, datasets marked by: “|”have the time or date of the clinical event, and others with ||, include both (time and date). In addition, the symbol: “*” shows that data is associated with ICU. In this table “R” shows the number of records and “V” means the number of variables. Moreover, two variables include distinct values whether “Process Number” (DP) or “Episode Number” (DE).

TABLE1. DATA COLLECTION

<table>
<tr><td rowspan="10">Patient Data</td><td># *| vital sign</td><td>70DE, 439025R, 108V</td></tr>
<tr><td>|| lab result</td><td>69DE,113320R, 9V</td></tr>
<tr><td>*| procedure</td><td>63DP,911R, 6V</td></tr>
<tr><td>*|| sepsis</td><td>17DE,176R,6V</td></tr>
<tr><td>*|| galgw</td><td>49DE, 861R,6V</td></tr>
<tr><td>*| diagnosis</td><td>67DE,124R,9V</td></tr>
<tr><td>|| med prescription</td><td>70DE,35422R, 39V</td></tr>
<tr><td>*|| intervention</td><td>70DE,18674R, 4V</td></tr>
<tr><td>*admin-discharge -ICU</td><td>70DE, R,2V</td></tr>
<tr><td>process-episod number</td><td>70DEP,70R, 2V</td></tr>
</table>

IV. Data Preparation

This phase encompasses a series of pivotal data processing tasks, aimed at enhancing the quality and applicability of the collected information. These tasks revolve around critical aspects such as aggregating time-series data for vital indicators, thus addressing sporadic data registration issues encountered with biological sensors in the ICU. By adopting an hourly aggregation approach, we effectively mitigated the challenge of infrequent data updates. Furthermore, to glean meaningful insights from laboratory results, we engineered novel features to detect and analyse abnormal findings. In addressing gaps within the vital sign dataset, a meticulous strategy was employed: we judiciously populated missing cells by computing the average value from neighbouring cells preceding and following the voids. Similarly, within another dataset, a pragmatic approach was taken by eliminating missing cells.

As an integral part of our preparations, we meticulously fine-tuned all datasets to ensure they boasted fitting data types and pertinent features. To fulfill the core objective of our experiment—identifying akin data clusters—we conducted an array of data engineering tasks. This transformational process facilitated the independent alignment of records, divorcing them from the constraints of the date and time of clinical events [18], [19].

This strategic transformation yielded a unified timeline of clinical events, where each event is uniquely identifiable through a meaningful key. Notably, this achievement was twofold in significance. Firstly, it equipped us to seamlessly apply two distinct forms of analysis—clustering and temporal clustering. In the realm of conventional clustering, our focus was on comprehending the distinctive attributes characterizing each cluster. On the other hand, the temporal clustering (as future work) avenue empowered us to delve into the nuanced behavior of data throughout a patient's treatment day. By combining these methodologies, we stand poised to unravel profound insights into the multifaceted landscape of patient health trajectories. This integrated approach holds immense promise, not only in delineating data patterns and trends but also in nurturing a more comprehensive understanding of patients' medical journeys and treatment outcomes.

The data engineering phase harnessed a formula to craft a distinct key, uniquely pointing to each clinical event for individual patients. Illustrated in Figure 1, this key exhibits a structured composition: the initial number designates the event's day, followed by an abbreviation denoting the data type. Additionally, the third and fourth components encompass the patient's process number and episode number, respectively. The sequential order of events during a specific period is encapsulated within the event sequence, an ascending value ranging from one to n, culminating in the count of parallel events transpiring on the same day.

Figure 1 serves as an illustrative example, underscoring an event linked to an episode number (20016701) attributed to a patient identified by process number 859785. This informative data point delineates the event's occurrence on the fifth day within the ICU context, constituting the eleventh clinical transaction.

The ramifications of these transformative endeavors reverberate across each dataset, engendering a newly introduced variable denoting the Time Series (TS). This pivotal inclusion of the Time Series facet endows medical practitioners with a versatile toolset, replete with filters facilitating a comprehensive review of diverse clinical events, along with their associated details, confined within a unified temporal framework. This innovative approach equips medical professionals with a holistic lens through which to navigate and discern intricate patient trajectories, auguring well for informed decision-making and enhanced patient care.

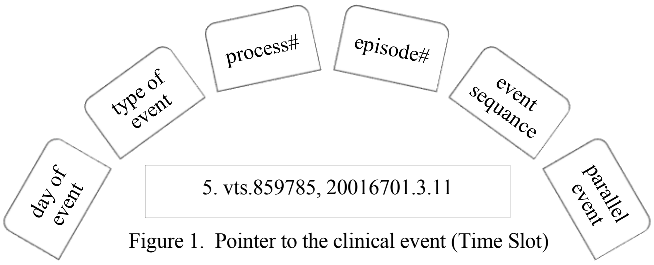

Figure 1. Pointer to the clinical event (Time Slot)

To create a unique table for clustering, the following steps were performed on each dataset:

- Extracted the day of the clinical event.
- Grouped the rows by taking the mean of numerical variables and the mode of categorical variables.
- selected the top 20 most frequent categorical variables to overcome memory limitations.
- Applied the merge function based on the common columns 'Process number' and 'TS'.
- Eliminated columns with more than 50% missing cells and deleted rows containing NaN values.

As a result of these data processing steps, a consolidated dataset was created with 147,100 rows and 17 columns. The table below provides a list of the categorical and numerical columns, along with a description of each variable.

- BLD_PULS_RATE_ART_ABP: The pulse rate of the arterial blood pressure, which measures the number of heart beats per minute while considering the arterial blood pressure.
- PRESS_BLD_ART_ABP_SYS: This variable refers to the systolic arterial blood pressure, which is the higher value measured during a heartbeat and represents the pressure in the arteries when the heart contracts.
- PRESS_BLD_ART_ABP_MEAN: This variable represents the mean arterial blood pressure, which is the average pressure in the arteries over a cardiac cycle.
- PRESS_BLD_ART_ABP_DIA: This refers to diastolic arterial blood pressure, which is the lower value measured during a heartbeat and represents the pressure in the arteries when the heart is at rest.
- Temperature: This represents the body temperature in degrees Celsius.
- PULS_OXIM_SAT_O2: the pulse oximetry oxygen saturation level, which measures the percentage of oxygen-saturated hemoglobin in the blood.
- ECG_HEART_RATE: This represents the heart rate measured using electrocardiography (ECG) in beats per minute.
- lab_Result: This variable could represent a laboratory test result, but without more context, referring to the Lab Exam Code.
- Glasgow Coma Scale (GCS): This might represent a numerical value associated with the Glasgow Coma Scale (GCS), which assesses a patient's level of

consciousness. A higher value could indicate a better level of consciousness.

In addition, categorical variables include Intervention_code, Medication_code, Diognostic_code, Lab_condition, Local_code (procedure), Zona_code(procedure), Lenght_of_stay(day) and Lab_exam_code.

V. Modeling

In the modeling phase, we employed the K-Prototypes algorithm to identify homogeneous data segments. K-Prototypes combine the strengths of both K-Means and K-Modes, making them well suited for effectively handling datasets with mixed data types. This hybrid approach is particularly valuable in scenarios like healthcare, where the data often comprises a blend of continuous and categorical variables. To fine-tune the number of clusters, we utilized the Elbow Method, a common technique to pinpoint the "elbow point" in the within-cluster sum of squares plot [20]. By analysing the plot shown in Figure 2 we identified the most suitable number of clusters, which in this case was found to be n=3 clusters.

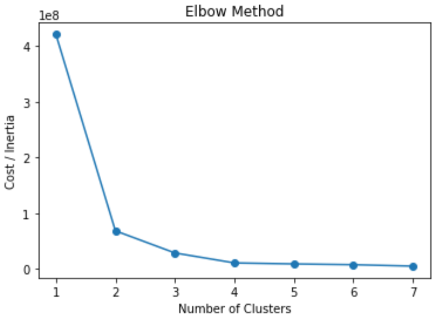

Figure 2. Elbow technique

This process helps in obtaining meaningful and interpretable clustering results that can assist in uncovering patterns and insights within complex datasets. The KPrototype algorithm was employed to discern and categorize patients into three distinct clusters, each sharing similar characteristics. These clusters were formed based on an analysis of various medical attributes, allowing us to group patients with comparable health profiles into coherent categories.

VI. Evaluation and Discussion

Assessing the results of K-Prototypes clustering involves a combination of quantitative metrics and visualizations In this phase we used Silhouette Score, PCA visualization, cluster characteristics, and domain knowledge, and also analyzed the business impact.

A. Silhouette score

The Silhouette score serves as a metric to gauge the degree of cohesion within a cluster and the separation from other clusters. Ranging from -1 to 1, a higher Silhouette score, such as our achieved value of 0.3450, signifies well-defined clusters, where instances within a cluster are more similar to each other than to those in neighboring

B. clusters Visualizations- PCA

Visualizations Principal Component Analysis (PCA) visualization is a technique used to reduce the dimensionality of data while retaining as much variance as possible. It understands the patterns and relationships between data points in a lower-dimensional space. This allows us to explore the distribution of data points, identify clusters, and understand the relationships between them in a more compact and interpretable manner. Each data point in the scatter plot represents a row from the original dataset. The position of the point on the plot corresponds to its values along the first and second principal components. While PCA doesn't capture categorical relationships, it can still provide a sense of how well-separated the clusters are based on numerical variables. often used as an exploratory technique to guide further analysis and interpretation [21].

Figure 3 depicts the PCA visualization, highlighting key aspects of cluster separation and distribution. Notably, the clusters exhibit distinct boundaries with no observable overlap, ensuring a certain level of separation. However, it is worth noting that some data points from cluster 0 and cluster 2 appear closely situated, implying a potential resemblance between these clusters. Additionally, the demarcation between clusters 1 and 2 seems less defined, suggesting the likelihood of instances where these clusters share proximity.

In terms of cluster distribution, a significant majority of data points tend to concentrate within a specific range spanning from -120 to 40. Interestingly, cluster 1 and Cluster 0 demonstrate a more compact grouping, whereas Cluster 2 showcases a comparatively broader spread across the visualization space. These observations collectively underline the need for further exploration and fine-tuning of cluster boundaries to achieve a comprehensive understanding of the underlying patterns and relationships within the data.

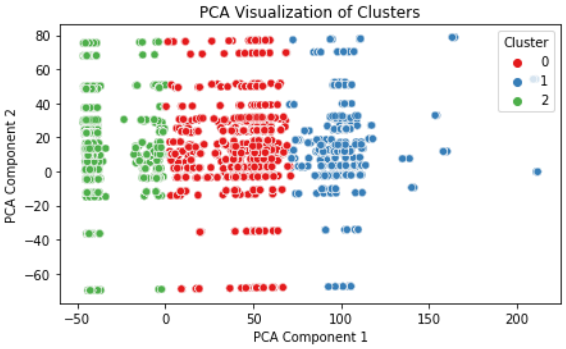

Figure 3. PCA visualization

C. Cluster Characteristics

The general purpose of cluster characteristics is to provide a clear and concise summary of the key attributes or features that define each cluster formed through clustering analysis. Cluster characteristics help in understanding the distinct patterns, trends, and behaviors exhibited by different groups of data points within the dataset. They play a crucial role in interpreting and explaining the results of clustering and can serve several purposes such as insight generation (offering insights into the average or typical values of variables for each cluster), helping to identify the inherent patterns and differences among the clusters. Moreover, Profile Identification: (identify the unique profiles or attributes of data points within each cluster, which can lead to meaningful interpretations of the data). In addition, Comparison and Contrast: help in understanding the similarities and differences in terms of numeric and categorical variables, aiding in hypothesis generation and further analysis. Finally, insights from cluster characteristics can inform decision-making processes, such as identifying target segments for personalized marketing, tailoring healthcare interventions, or optimizing resource allocation.

The clustering results based on the mean values of the numerical data variables reveal distinct health patterns and conditions among the identified clusters.

- Cluster 0

BLD_PULS_RATE_ART_ABP: 90.35
BLD_ART_ABP_SYS: 105.22
BLD_ART_ABP_MEAN: 73.59
BLD_ART_ABP_DIA: 58.59
PULS_OXIM_SAT_O2: 91.59
ECG_HEART_RATE: 91.60
Temperature:e 33.49, and lab_Result: 86.62,
Glasgow Coma Scale: 12.51

Cluster 0 exhibits moderate values across various vital signs, such as blood pressure, temperature, and heart rate, which are generally within a healthy range. The Glasgow Coma Scale suggests a reasonable level of consciousness. The relatively higher laboratory result may indicate the presence of specific health conditions but not at a critical level. Overall, this cluster could represent patients with stable and relatively normal health conditions.

- Cluster 1:

BLD_PULS_RATE_ART_ABP: 92.11
BLD_ART_ABP_SYS: 106.93
BLD_ART_ABP_MEAN: 74.57
BLD_ART_ABP_DIA: 59.36
PULS_OXIM_SAT_O2: 92.21
ECG_HEART_RATE: 93.31
Temperature: 33.87, and lab_Result: 5.98
Glasgow_Coma Scale: 12.72

Cluster 1 is characterized by slightly elevated vital sign values, including blood pressure, heart rate, and oxygen saturation, compared to Cluster 0. However, the laboratory result indicates a relatively low value, suggesting that these patients may have healthier laboratory values. The Glasgow Coma Scale remains consistent with a normal consciousness level. This cluster might represent patients with well-controlled health conditions or those in a stable state of recovery.

- Cluster 2:

BLD_PULS_RATE_ART_ABP: 93.03
BLD_ART_ABP_SYS: 106.20
BLD_ART_ABP_MEAN: 74.16
BLD_ART_ABP_DIA: 58.97
PULS_OXIM_SAT_O2: 92.27
ECG_HEART_RATE: 94.80
Temperature: 33.81, and lab_Result: 146.12
Glasgow_Coma Scale: 12.74

Cluster 2 stands out with elevated vital signs, particularly in blood pressure and heart rate. The lab_Result value is significantly higher, indicating potential health concerns or abnormalities in laboratory results. Despite the higher laboratory result, the Glasgow Coma Scale suggests a normal level of consciousness. This cluster might represent patients with more acute or severe health conditions requiring closer medical attention or intervention.

The bar chart in Figure 4 shows a distinct trend among the clusters, highlighting a significant pattern in the laboratory status of the data points

.

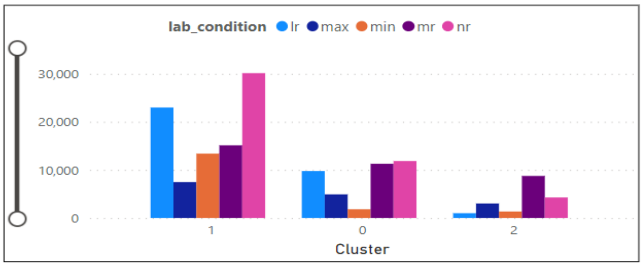

Figure 4. Laboratory conditions

- Cluster 0:Lab Condition: Normal

Cluster 0 is characterized by moderate vital signs and normal laboratory conditions. This cluster represents patients with relatively stable and typical health conditions. The combination of normal lab conditions and vital signs suggests that patients in this cluster are generally healthy and not currently experiencing any significant health issues.

- Cluster 1:Lab Condition: Normal

Similar to Cluster 0, Cluster 1 exhibits slightly elevated vital signs but with normal lab conditions. Patients in this cluster might have certain health factors that lead to slightly elevated vital signs, but their laboratory results are within a normal range. This suggests that patients in Cluster 1 are likely in a stable condition and may be in the process of recovery or managing their health conditions effectively.

- Cluster 2:Lab Condition: More than Maximum

Cluster 2 stands out with elevated vital signs and abnormal lab conditions that are categorized as "more than maximum." Patients in this cluster are experiencing significantly elevated laboratory values, which could indicate potential health complications or acute issues. The elevated vital signs further emphasize the severity of their health conditions. Patients in Cluster 2 may require immediate medical attention and intervention to address abnormal lab conditions and stabilize their health.

In addition, considering the mode of categorical variables presented in Table 2, Cluster 0 is characterized by specific intervention codes, prc_LOCAL_code, prc_ZONA_code, lab_Exam_code, and medication . Patients in this cluster seem to have consistent patterns of medical interventions and procedures, as well as localized and zonal information. The diagnostic code U071 could suggest a specific condition or disease that is being managed.

Cluster 1 shares similarities with Cluster 0 in terms of intervention code, prc_LOCAL_code, prc_ZONA_code, and medication usage. However, patients in this cluster have different lab_Exam_code and Diognostic_code, suggesting variations in the specific medical tests and diagnostic procedures performed.

Cluster 2 shares similar patterns in terms of intervention code, prc_LOCAL_code, prc_ZONA_code, and medication usage. However, patients in this cluster have a different lab_Exam_code, lab condition, and Diognostic_code. The lab condition "mr" (more than maximum) could indicate a more complex or serious medical test result. The commonality in other features suggests that patients in this cluster might have a specific condition that is being addressed with various interventions and diagnostic procedures.

TABLE2. BEHAVIOR OF CATEGORICALS

Categorical	***Closter 0***	***Cluster 1***	***Cluster 2***

Lab Exam	lab9	lab8	lab18
Diagnostic	U071	8271	8271
Medication	997	997	997
Procedure/Zon	Prcz3	Prcz3	Prcz3
Procedure/location	prcl6	prcl6	prcl6
Intervention	int9	int9	int9
Length of stay	8	8	8

In summary, as depicted in Table 3, the clustering analysis unveils three distinctive health profiles within the patient population. This discovery establishes a cluster-based health profile, where Cluster 0 (highlighted in green) denotes individuals with generally stable and normal health conditions. Cluster 1 (depicted in blue) likely encompasses patients with controlled health conditions and a positive recovery trajectory. Cluster 2 (indicated in red) identifies patients with potentially more acute health issues, evident from elevated vital signs and abnormal laboratory results. These insights offer valuable guidance to healthcare professionals, empowering them to tailor interventions, treatment plans, and monitoring strategies according to the unique characteristics of each cluster.

TABLE3. CLUSTER-BASED HEALTH PROFILE

Cluster 0	*Cluster 1*	*Cluster 2*
Stable 🟩	Recovery 🟦	Attention 🟥

CONCLUSION

In conclusion, the seamless integration of data mining and clustering analysis within the realm of precision medicine has yielded a wealth of insights that hold the potential to revolutionize the landscape of healthcare. The revelations from this study serve as a cornerstone for the development of personalized treatment strategies, informed clinical decision-making, and ultimately, enhanced patient outcomes. By harnessing the power of precision medicine, we are not only advancing the boundaries of medical science but also shaping the future trajectory of healthcare delivery. Furthermore, the meticulous analysis of clustered patient data has illuminated the distinct disparities in health status and condition among the identified clusters, providing a comprehensive overview of cluster-based health profiles. Clusters 0 and 1 exhibit patient instances with relatively stable and typical laboratory results, indicative of a standard health state. In contrast, cluster 2 encapsulates instances where laboratory measurements surpass the established upper limits, suggesting potentially acute health issues. As we reflect on the outcomes of this study, we recognize the potential for future enhancements and refinements of the clustering experiment. Specifically, there is an opportunity to optimize the current clustering methodology, fine-tuning its parameters for even more accurate and insightful results. Moreover, the application of temporal clustering holds promise for monitoring and observing the dynamic health condition of patients during their stay in the intensive care unit (ICU). This avenue of exploration could unveil deeper insights into the progression of health status over time, enabling more precise interventions and tailored care strategies. In essence, this research not only underscores the transformative potential of data mining and clustering analysis in precision medicine but also emphasizes the ongoing pursuit of continuous refinement and innovation to shape the future of healthcare in a patient-centric and data-driven manner.

ACKNOWLEDGMENT

The work has been supported by FCT- Fundacao para a Ciencia e Technologia within rhe R&D Units Project Scope: UIDB/00319/2020

REFERENCES

[1] E. Shirzad, G. Ataei, and H. Saadatfar, “Applications of data mining in healthcare area: A survey,” *Eng. Appl. Sci. Res.*, vol. 48, no. 3, pp. 314–323, 2021.

[2] I. Yoo *et al.*, “Data mining in healthcare and biomedicine: A survey of the literature,” *J. Med. Syst.*, vol. 36, no. 4, pp. 2431–2448, 2012.

[3] A. Arjun, A. Srinath, and B. R. Chandavarkar, “Predictive analytics and data mining in healthcare,” in *2021 12th International Conference on Computing Communication and Networking Technologies, ICCCNT 2021*, 2021.

[4] A. A. Toor, M. Usman, F. Younas, A. C. M. Fong, S. A. Khan, and S. Fong, “Mining massive e-health data streams for IoMT enabled healthcare systems,” *Sensors (Switzerland)*, vol. 20, no. 7, pp. 1–24, 2020.

[5] D. J. Hand, “Principles of data mining,” *Drug Saf.*, vol. 30, no. 7, pp. 621–622, 2007.

[6] M. S. Chen, J. Han, and P. S. Yu, “Data mining: An overview from a database perspective,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 8, no. 6. 1996.

[7] N. S. Mosavi and M. F. Santos, “Internet of things for precision intensive medicine,” *Procedia Comput. Sci.*, vol. 201, no. C, pp. 732–737, 2022.

[8] N. S. Mosavi and M. F. Santos, “To what extent healthcare analytics influences decision making in precision medicine,” *Procedia Comput. Sci.*, vol. 198, no. 2021, pp. 353–359, 2021.

[9] N. S. Mosavi and M. F. Santos, “How prescriptive analytics influences decision making in precision medicine,” *Procedia Comput. Sci.*, vol. 177, pp. 528–533, 2020.

[10] N. S. Mosavi and M. F. Santos, “Characteristics of the Intelligent Decision Support System for Precision Medicine (IDSS4PM),” pp. 1–8.

[11] N. Sadat Mosavi and M. Filipe Santos, “Adoption of Precision Medicine; Limitations and Considerations,” 2021, pp. 13–24.

[12] N. Cozzoli, F. P. Salvatore, N. Faccilongo, and M. Milone, “How can big data analytics be used for healthcare organization management? Literary framework and future research from a systematic review,” *BMC Health Serv. Res.*, vol. 22, no. 1, 2022.

[13] H. Nouraei, H. Nouraei, and S. W. Rabkin, “Comparison of Unsupervised Machine Learning Approaches for Cluster Analysis to Define Subgroups of Heart Failure with Preserved Ejection Fraction with Different Outcomes,” *Bioengineering*, vol. 9, no. 4, 2022.

[14] J. Gil *et al.*, “Data mining analyses for precision medicine in acromegaly: a proof of concept,” *Sci. Rep.*, vol. 12, no. 1, pp. 1–14, 2022.

[15] N. S. Mosavi and M. F. Santos, “ScienceDirect ScienceDirect To what extent healthcare analytics influences decision making in To what extent healthcare analytics influences decision making in precision medicine precision medicine,” *Procedia Comput. Sci.*, vol. 198, no. 2021, pp. 353–359, 2022.

[16] T. Hulsen *et al.*, “From big data to precision medicine,” *Front. Med.*, vol. 6, no. MAR, pp. 1–14, 2019.

[17] T. J. Loftus *et al.*, “Phenotype clustering in health care: A narrative review for clinicians,” *Front. Artif. Intell.*, vol. 5, 2022.

[18] N. S. Mosavi and M. F. Santos, “Intelligent Decision Support System for Precision Medicine: Time Series Multi-variable Approach for Data Processing,” *Int. Jt. Conf. Knowl. Discov. Knowl. Eng. Knowl. Manag. IC3K - Proc.*, vol. 3, pp. 231–238, 2022.

[19] N. S. Mosavi and M. F. Santos, “Data Engineering to Support Intelligence for Precision Medicine in Intensive Care.”

[20] H. Humaira and R. Rasyidah, “Determining The Appropiate Cluster Number Using Elbow Method for K-Means Algorithm,” 2020.

[21] K. Younes *et al.*, “Application of Unsupervised Learning for the Evaluation of Aerogels’ Efficiency towards Dye Removal—A Principal Component Analysis (PCA) Approach,” *Gels*, vol. 9, no. 4, 2023.