CLUSTERING AND GRAPH ANALYSIS OF PATIENT READMISSIONS ON DIABETES 130-US HOSPITALS FOR YEAR 1999-2008

Adrià Matas Bosch (s232775), Luca D'Este (s233231)
Jesús Díaz Pereira (s233142), Ioannis Vlasakoudis (s232755)

Technical University of Denmark

ABSTRACT

This study analyzes hospital readmissions among diabetic patients from 130 US hospitals (1999–2008) using clustering and graph-based methods. Techniques like hierarchical clustering and k-means identified patient subgroups, evaluated through metrics such as silhouette score and Davies-Bouldin Index. Graph-based community detection, incorporating Random Forest-derived features, revealed patterns in medication use, diagnoses, and readmission rates. While community-aware features modestly enhanced predictive modeling, results highlight the potential of these methods for healthcare analytics and the need for further refinement to improve insights into readmission drivers.

GitHub Repository

1. INTRODUCTION

Diabetes has become a rapidly spreading global epidemic that poses significant challenges to public health globally [1]. Furthermore, diabetes remission is a hotly disputed concept in contemporary endocrinology [2]. Due to its burden on the healthcare system, preventing unplanned readmissions is critical not only for improving patient outcomes but to optimizing healthcare resources.

This study aims to analyze a large dataset of hospital records to identify patterns associated with early readmissions of diabetic patients. We hypothesize that graphs and clustering can provide key insights that could help model complex interactions within the data, such as relationships between patient profiles, clinical outcomes, and readmission rates. Additionally, we explore the potential of incorporating graph-derived features into predictive models that could help design targeted interventions to improve patient outcomes and reduce healthcare costs.

2. MATERIALS AND METHODS

2.1. Data Description, Exploration and Processing

The initial dataset encompasses ten years (1999–2008) of clinical care records from 130 US hospitals and integrated delivery networks. Each row represents a hospital stay for a diabetic patient, including laboratory results, medications, and stays of up to 14 days [3]. The target variable is early readmission within 30 days post-discharge.

To streamline the analysis, the dataset was reduced in size, and the target variable, $readmitted$, was balanced across classes. We performed exploratory data analysis (EDA) to understand the dataset's structure, identifying low-variance categorical variables and highly skewed numerical distributions. We also analyzed the mean values of numerical variables across classes to assess their correlation with the target.

For preprocessing, missing numerical values were imputed with column means and categorical values with the most frequent category. Numerical features were scaled, and categorical variables, including diagnoses, were one-hot encoded. Diagnostic features were standardized using ICD-9 codes.

2.2. Clustering

To try to characterize the dataset, different types of clustering have been tried.

The first method used is Hierarchical clustering, which is a method suitable for clustering a dataset with mixed features (categorical and numerical). Out of the two Hierarchical clustering techniques, Agglomerative (Bottom-up approach) will be used for this method, where each data point starts as its own individual cluster and after many iterations, in which the two most similar clusters merge into one cluster, the process concludes with the predetermined number of clusters.

Another idea was to try K-prototypes [4], an extension of the more famous k-means that is also suited for this scenario where some features are categorical and some continuous. Besides this K-means using PCA after feature selection and processing has been used.

Thanks to DTU.

The methods have been evaluated by means of scores like the Silhoutte score [5], that measures how similar an object is to its own cluster compared to other clusters and for which a higher value (tending to 1) is better, the Davies-Bouldin Index [6], which measures the average similarity ratio of each cluster with its most similar cluster and for which a lower value is hoped and the Dunn Index, which measures how compact a cluster is and so a higher score is better. Also several plots have been used to understand if these clusters could characterize the situation.

For the feature selection of these two clustering methods, a Chi-Square [7] test was performed, in order to identify the features highly associated with the target variable (patients readmitted). This approach reduces the dimensionality of the dataset, retaining only the most relevant features. Since the Chi-Square test requires categorical data, numerical features were transformed by binning continuous variables into discrete categories. Then the test was performed with a significance value of $p = 0.05$, resulting in nine features being retained for clustering.

Regarding the number of clusters, different approaches were used for each clustering method. To determine the number of clusters for the Agglomerative method, a combination of the two evaluation scores mentioned above was used. After calculating the Silhoutte scores and Davies-Bouldin Indices for a reasonable range of clusters (up to 50), the optimal number of clusters based on these results was 18. For k-means the same scores have been tried and it showed the best number was 3 for both scores with PCA to 2 dimensions.

2.3. Graph Construction and Analysis

We used a Random Forest (RF) classifier with cross-validation to identify informative features for graph construction. Feature importances ranked variables by their contribution to readmission prediction, and those with importance below 0.02 were excluded.

Using the selected features, we constructed a graph where nodes represented patients and weighted edges reflected similarities. Feature importances scaled the feature space, and cosine similarity quantified patient alignment. A k-nearest neighbors (k-NN) approach sparsified the similarity matrix, retaining only the top k connections per patient.

Louvain and Spectral Clustering community detection techniques were applied to identify patient clusters within the graph [8, 9]. To assess the homogeneity of patient clusters, entropy values were calculated for the readmission status distribution within each community.

Additionally, a Chi-Square test evaluated the relationship between community structure and readmission status. Finally, we conducted a detailed profiling of the detected communities to explore their characteristics.

2.4. Predictive Modeling with Community-Aware Features

Community labels from the Louvain algorithm were treated as categorical features, and three community-based metrics such as Normalized Within-Module Degree, Participation Coefficient, and Community Association Strength, were added to the initial RF pipeline [10].

A Random Forest (RF) model was evaluated using stratified cross-validation (CV), with F1 score, accuracy, precision, and recall calculated to assess performance. Feature importance analysis confirmed the predictive value of community-aware features.

3. RESULTS

3.1. Feature Selection for Clustering

The results from the Chi-Square test can be seen in Figure 1. Among the 9 features remaining, some can be directly associated with Diabetes, such as the number of diagnoses and medications prescribed, insulin use, and glucose serum results.

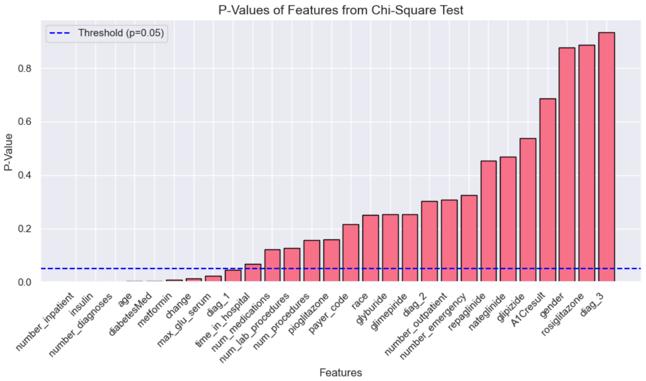


Fig. 1: Feature P-values

3.2. Clustering Techniques

For clustering, we applied Hierarchical clustering K-prototypes and K-means after processing. To evaluate the 3 clustering methods 3 metrics were used and these are the Silhouette Score, the Davies-Bouldin Index and the Dunn Index, and the measure the cohesion, the separation and the compactness of the clusters respectively.

Table 1: Comparison of Clustering Methods

Method	Clusters	Silhouette Score	Davies-Bouldin Index	Dunn Index
Hierarchical Clustering	18	0.4088	2.1376	**0.1146**
K-prototypes	5	-0.0333	29.5218	0.0022
K-means PCA selected features	3	**0.6295**	**0.4392**	0.0111
K-means PCA ICD-9	3	0.2679	1.3397	0.0085

In Figure 4, we can see the dendrogram of Hierarchical clustering and observe how the clusters are formed with each iteration. The height of the two merged clusters indicates the distance between them, thus the higher a merge is happening the less similar the two clusters are.

In the appendix, in Figure 5, we can see the clusters based on K-means on PCA on ICD-9 processed features.

After considering these measures, we conclude that Hierarchical Clustering is the selected method for further analysis, since it has the best combined score between them (see Figure 3).

3.3. Cluster Profiling

A comprehensive cluster profiling can yield interesting insights about the characteristics and patterns of the groups formed from clustering and help us understand what distinguishes a cluster from another.

For this, we are going to examine only clusters with a sufficient number of patients in them, since there are clusters with less than 5 patients. Starting with the most important feature, the distribution of 'readmitted' is shown in Figure 7.

The focus, with regard to the remaining features, is on 'metformin,' 'insulin,' 'change,' and 'diabetesMed', where there are some interesting patterns that appear across the clusters. The analysis starts with 'metformin' and 'insulin', two very important drugs in the treatment for diabetes. Their distribution across the clusters can be seen in Figures 8 and 9.

Continuing to the feature 'change', which indicates if there was a change in diabetic medications, we can see the distribution in Figure 11, if the patients inside these clusters had either a change in diabetic medication or no change at all. Last but not least, the distribution for the feature 'diabetesMed' is shown in Figure 10.

As for numerical features, we highlighted the average number of medications, which can be seen in Figure 12.

3.4. Feature Importance for Model Selection

Feature importances from the Random Forest model were used to define the relationships between the nodes of the graph. We show them in the appendix, in Figure 14.

3.5. Comparison of Community Detection Methods

Metrics comparing Louvain and Spectral Clustering methods are summarized in Table 2. Chi-squared test results indicate differences in the distribution of 'readmitted' values across communities for each method.

Table 2: Metrics Comparison of Community Detection Methods

Method	# Communities	Modularity	Entropy	χ^2	(p-value)
Louvain	17	0.73	1.5	82.9	5.8×10^{-6}
Spectral Clustering	5	0.60	1.5	0.06	0.99

3.6. Community Profiling and Skewed Communities

The distribution of 'readmitted' values within Louvain communities is presented in Figure 15 in the appendix. Communities with a dominant 'readmitted' value frequency exceeding 40% were classified as skewed, resulting in 9 communities.

Figure 16 shows the most frequent primary diagnosis per community. Diseases of the circulatory system is the most frequent value in the majority of communities, but a few of them are characterized by other primary diagnoses such as diseases of the musculoskeletal system and connective tissue, or Endocrine, nutritional and metabolic diseases. Regarding the age, it can be observed in Figure 17 how the most predominant value in the dataset is 8 which represents the eldest people. Again, there are a few communities characterized by less common values, such as 7, 6 and 3. This last one is formed by very young people.

We have also analyzed the most important numerical features that have the greatest impact on our target. Figure 18 shows visualization of box-plots of 'num_lab_procedures', highlight their distributions within communities. Other numerical feature distributions can be seen in the appendix as well.

3.7. Evaluation of Community-Based Features

Community-based metrics, including Normalized Within-Module Degree, Participation Coefficient, and Community Association Strength, were included in the machine-learning pipeline. Model performance metrics, evaluated using stratified cross-validation, are presented in Table 3.

Table 3: Model Performance Metrics (Mean and Std Dev)

Metric	Mean	Std Dev
Accuracy	0.4088	0.0273
Precision	0.4058	0.0269
Recall	0.4087	0.0271
F1 Score	0.4055	0.0264

Feature importance analysis, including community-based features, is illustrated in Figure 2.

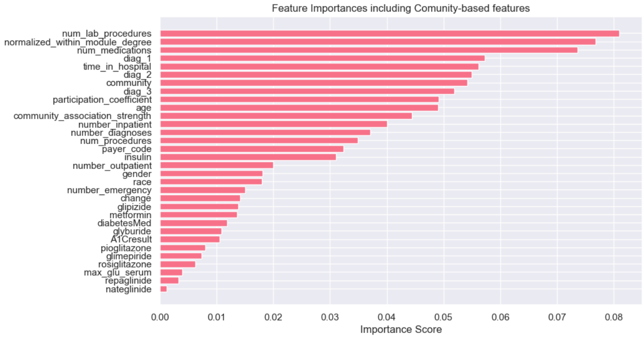


Fig. 2: Feature importances including community-based metrics.

4. DISCUSSION

The clustering results produced some outcomes that were below our expectations overall. This is related mostly to the distribution of 'readmitted'. We can observe that most patients are grouped into 7 clusters out of the total 18. Then, these 7 clusters were examined further.

The majority of the features in these clusters were balanced and well distributed, except of 'metaformin', 'insulin', 'change' and 'diabetesMed'.

Starting with 'metformin' and 'insulin', we can notice that the clusters are different regarding the distribution of these two variables, especially for 'insulin'. As for 'metformin', we can say that the clusters are divided into two groups, one with patients that haven't been prescribed metformin and the other with patients that had no change in the dosage of this drug.

Then, continuing to 'change', these large clusters clearly distinguish themselves from each other, since each cluster contains patients with change or no change for their medication. As for the feature 'diabetesMed', all the patients in the clusters have been prescribed diabetes medication, except of cluster 4, which includes patients with no medication.

Last but not least, for the numerical features the distributions are also similar between the clusters. There are some minor differences in the average number of medications, where the highest is in cluster 2 with 21.25 and the lowest in cluster 13 with 13.18.

While most clusters showed balanced distributions for many features, the analysis revealed some patterns in the treatment and medication usage among patients. In order to achieve better and more meaningful results, the emphasis should be placed on refining the clustering process and even try more clustering techniques.

The application of community detection techniques yielded mixed outcomes. The Louvain method demonstrated stronger chi-squared results compared to Spectral Clustering, suggesting better performance in capturing meaningful structures in the data. However, the entropy values, approximately 1.5 across Louvain communities, indicated a relatively balanced distribution of readmission outcomes, highlighting the lack of strong separations between clusters. This suggests that, while patterns exist, the communities may not fully align with distinct readmission profiles.

Profiling of skewed communities revealed some notable insights. For instance, most communities were predominantly associated with circulatory system diseases, while a few had distinct characteristics, such as one community of younger patients (age group 3), primarily suffering from endocrine, nutritional, and metabolic diseases. This community, distinct from others, highlights how age and disease type can define specific patient subgroups. Additionally, the distribution of *num_lab_procedures* varied significantly, with one community (Community 2) receiving fewer procedures, potentially reflecting differences in healthcare delivery or patient needs.

Community-aware features, when incorporated into the Random Forest algorithm, offered some improvement in understanding factors influencing readmission. However, the overall predictive performance remained modest, emphasizing the need for further refinement in feature engineering and modeling approaches to enhance interpretability and utility.

This study underscores the potential of graph-based methods in healthcare analytics but also highlights their limitations. Further work is needed to improve segmentation quality and to explore how these methods can better inform targeted interventions for reducing readmissions.

5. CONCLUSION

This work explored the application of community detection and graph-based feature engineering to analyze readmission patterns in diabetic patients. While the Louvain algorithm captured some meaningful structures and highlighted distinct patient subgroups, the results were inconclusive in providing clear separations for predictive modeling. Community-aware features contributed modestly to understanding readmission drivers, but predictive performance requires further enhancement.

Future research should focus on refining clustering techniques, integrating additional data sources. Such improvements promise to develop targeted interventions and optimize care delivery for high-risk groups.

6. REFERENCES

[1] Huai Zhang, Xiao-Dong Zhou, Michael D Shapiro, Gregory YH Lip, Herbert Tilg, Luca Valenti, Virend K Somers, Christopher D Byrne, Giovanni Targher, Wah Yang, et al., "Global burden of metabolic diseases, 1990–2021," *Metabolism*, vol. 160, pp. 155999, 2024.

[2] Sanjay Kalra, Arbinder Singal, and Tejal Lathia, "What's in a name? redefining type 2 diabetes remission," *Diabetes Therapy*, vol. 12, pp. 647–654, 2021.

[3] Beata Strack, Jonathan P DeShazo, Chris Gennings, Juan L Olmo, Sebastian Ventura, Krzysztof J Cios, and John N Clore, "Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records," *BioMed research international*, vol. 2014, no. 1, pp. 781670, 2014.

[4] Z. Huang, "Clustering large data sets with mixed numeric and categorical values," 1997.

[5] Ketan Rajshekhar Shahapure and Charles Nicholas, "Cluster quality analysis using silhouette score," in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, 2020, pp. 747–748.

[6] Junwei Xiao, Jianfeng Lu, and Xiangyu Li, "Davies bouldin index based hierarchical initialization k-means," *Intelligent Data Analysis*, vol. 21, no. 6, pp. 1327–1338, 2017.

[7] James E Wert, Charles O Neidt, and J Stanley Ahmann, "Chi square.," 1954.

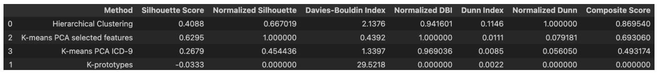
[8] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, pp. P10008, 2008.

[9] Ulrike Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, pp. 395–416, 2007.

[10] Bogumił Kamiński, Paweł Prałat, François Théberge, and Sebastian Zajac, "Predicting properties of nodes via community-aware features," *Social Network Analysis and Mining*, vol. 14, no. 1, pp. 117, 2024.

A. APPENDIX

Include supplementary information such as additional tables, figures, or extended results.



	Method	Silhouette Score	Normalized Silhouette	Davies-Bouldin Index	Normalized DBI	Dunn Index	Normalized Dunn	Composite Score
0	Hierarchical Clustering	0.4088	0.667019	2.1376	0.941601	0.1146	1.000000	0.869540
2	K-means PCA selected features	0.6295	1.000000	0.4392	1.000000	0.0111	0.079181	0.693060
3	K-means PCA ICD-9	0.2679	0.454436	1.3397	0.969036	0.0085	0.056050	0.493174
1	K-prototypes	-0.0333	0.000000	29.5218	0.000000	0.0022	0.000000	0.000000

Fig. 3: Combined scores of clustering techinques

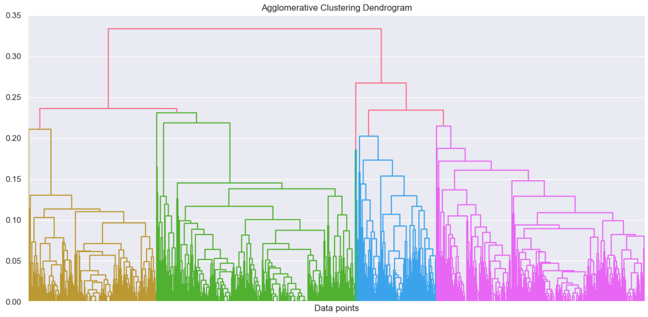


Fig. 4: The Agglomerative dendrogram

In Figure 5, as mentioned in the results section, we can see the plots of the clusters in the case of ICD-9, comparing them to the pca data used and the true labels.

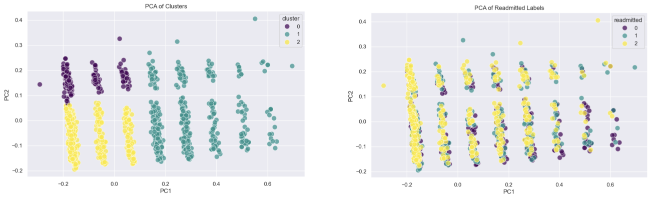


Fig. 5: Example of K-means clusters and compared to the original labels

An interesting plot to better characterize and explain PCA, we can plot,in Figure 6, the importance of each feature on PCA components.

```
                      Dim 1     Dim 2
number_inpatient  -0.084945  0.253910
number_diagnoses  -0.998563 -0.041441
age               -0.176832 -0.080687
max_glu_serum      0.038454  0.015268
metformin         -0.061652 -0.040930
insulin           -0.122539  0.956887
change            -0.077749  0.725744
diabetesMed       -0.053912  0.583659
diag_1            -0.138742  0.178377
```

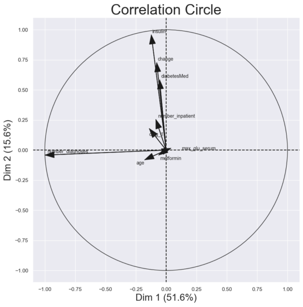


Fig. 6: Correlation Circle

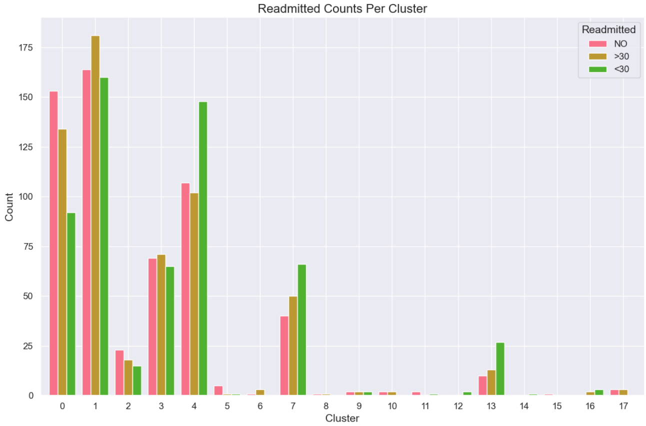


Fig. 7: Readmitted category by cluster

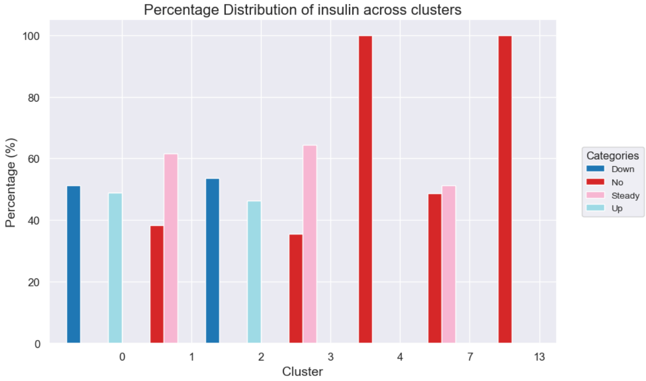


Fig. 8: Distribution of Insulin across clusters

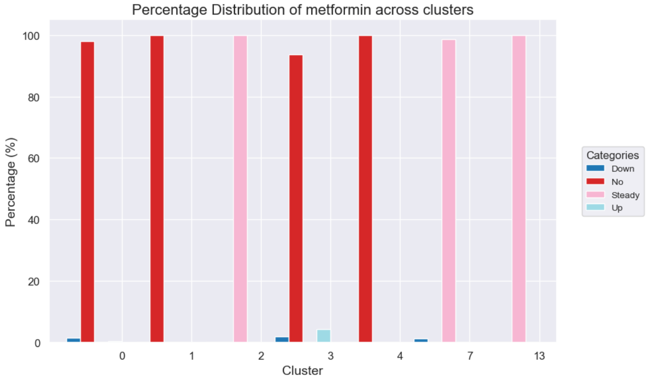


Fig. 9: Distribution of Metformin across clusters

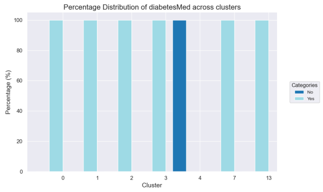


Fig. 10: Distribution of DiabetesMed across clusters

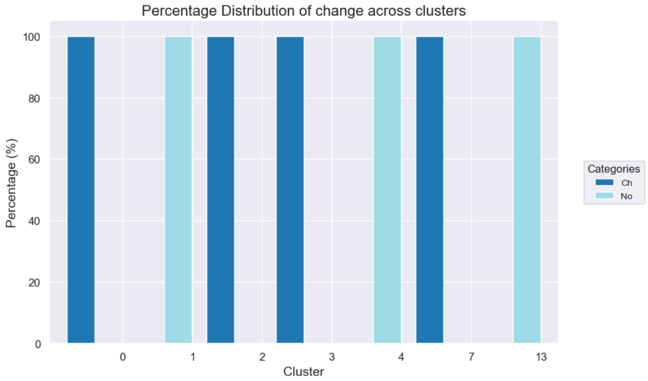


Fig. 11: Distribution of Change across clusters

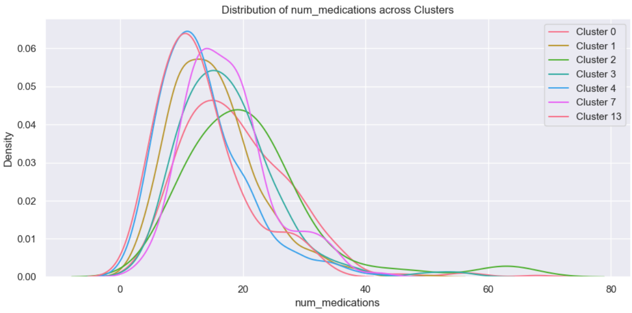


Fig. 12: Distribution of the Number of Medications across clusters

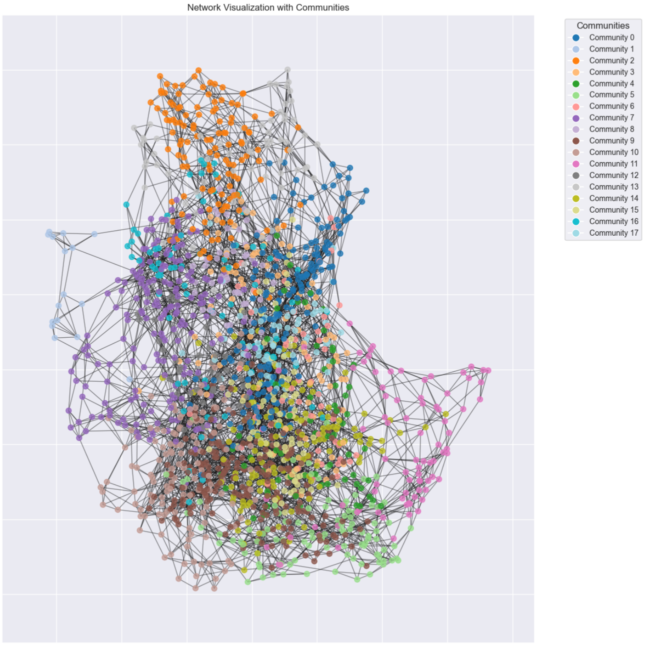


Fig. 13: The Louvain Graph Network

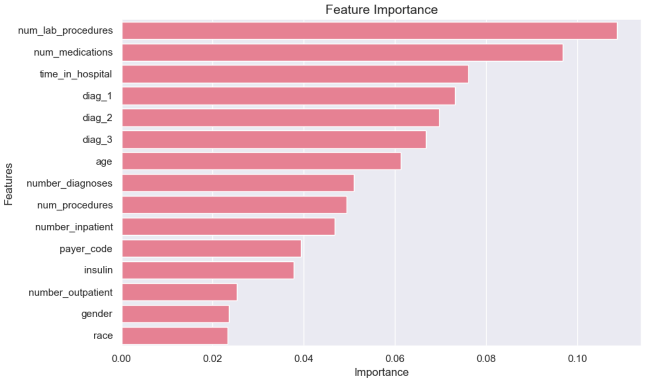


Fig. 14: Feature importances derived from a Random Forest model with cross-validation.

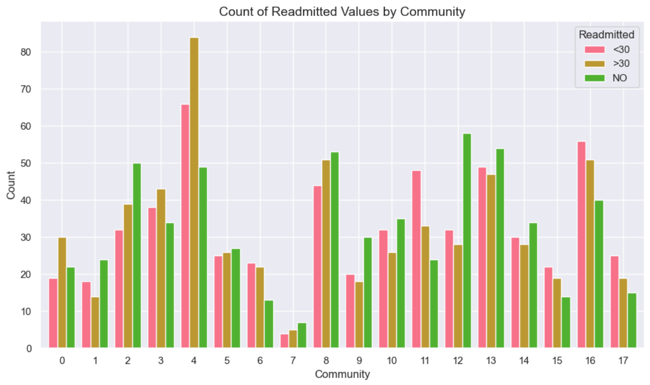


Fig. 15: Distribution of ’readmitted’ categories across Louvain communities.

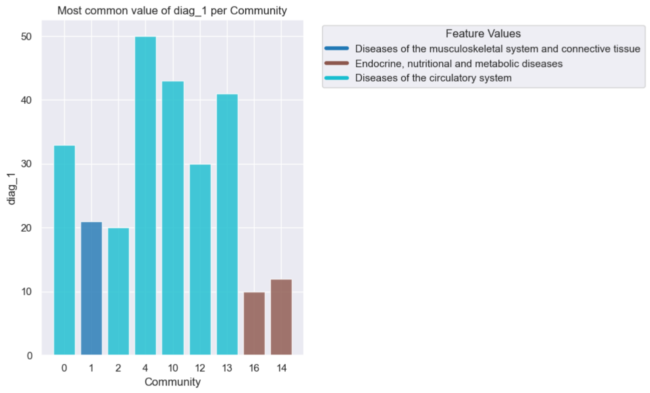


Fig. 16: Primary diagnoses across communities.

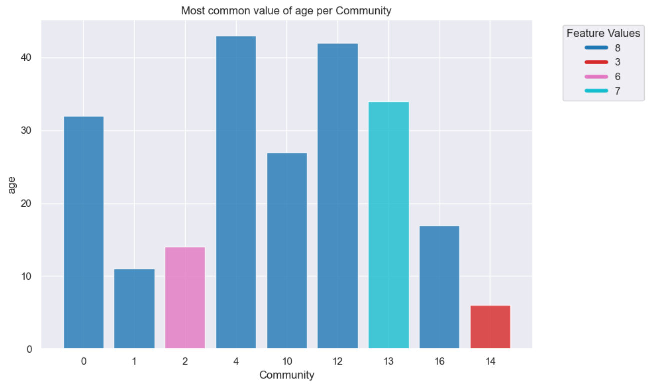


Fig. 17: Age distribution across communities.

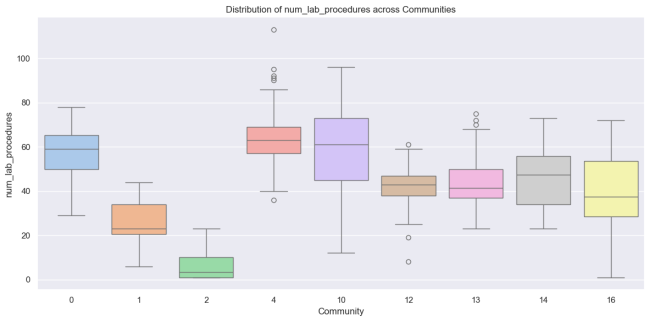


Fig. 18: Box plot of ’num_lab_procedures’ across communities.

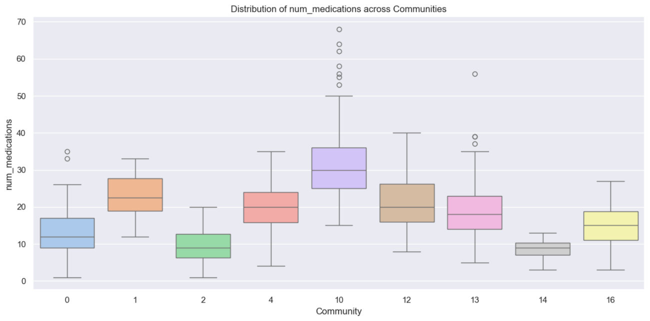


Fig. 19: Box plot of ’num_medications’ across communities.

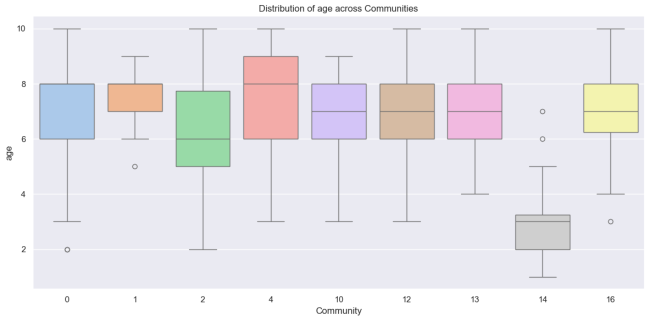


Fig. 20: Box plot of ’age’ across communities.

B. CONTRIBUTION

We want to emphasize that all group members contributed with the same effort and all reviewed each other’s production. In particular, Jesús and Adrià focused mainly on the processing, graph, and incorporating of community-aware features into a predictive Machine Learning model, while Ioannis and Lucas focused mainly on the cluster analysis.