Analyzing Key Factors and Drug Therapy Impact on 6-Month Readmission for Heart Failure Patients through Statistical Learning methods

1. Introduction

Heart failure is a pathological condition that consists in the inability of the heart to pump an adequate amount of blood to vascularize tissues and districts of the organism. It can be caused by various abnormalities, including myocardial dysfunction, cardiomyopathy, and coronary artery disease. Severe cases of heart failure, evaluated based on ejection fraction, require yearly hospitalization [1]. Hospitalization for heart failure is common among older adults and often leads to readmission rates of 25% within 30 days and approximately 50% within 6 months [2][3]. However, readmission rates vary between hospitals. Facilities that offer improved care, including follow-ups and self-care education during and after hospitalization, have lower readmission rates compared to others. Identifying the factors contributing to readmission is crucial for patient well-being, reducing healthcare costs, and optimizing healthcare system performance. These factors can be used to implement targeted interventions, develop personalized care plans, and address specific risk factors. Statistical models are essential as they provide insights into complex relationships and accurate predictions, helping optimize resource allocation and improve patient outcomes [4]. In light of these considerations, the primary objective of this study was to analyze the key factors contributing to readmission within 6 months for heart failure patients, by defining a *classification* task with particular focus on drug therapy. To achieve this, a high-quality retrospective heart failure dataset was used. It is important to emphasize that the results of the study are intrinsically biased to the population considered. Nonetheless, the systematic analysis method employed holds the potential to be readily applied to other datasets, making it a versatile approach with broader implications.

2. Material and Methods

2.1. Data description

The employed dataset derived from a public repository on PhysioNet [5] included 2008 patients admitted in a hospital in Sichuan (China) between 2016 and 2019. For each patient, 165 descriptors are provided regarding (i) the type of heart failure and other cardiac indicators frequently adopted to characterize the severity of the myocardial dysfunction (including Ejection Fraction (EF) measures, New York Heart Association (NYHA) Functional Classification [6] and Killip Grade [7]), (ii) baseline clinical characteristics measured on the admission day, (iii) the patient's level of consciousness according to the Glasgow Coma Scale (GCS), (iv) subsequent hospital admissions and mortality. It is important to note that many of these variables exhibit missing information, leading to multiple NaN values in numerous columns, sometimes exceeding 60% of the patients. Therefore, careful consideration and appropriate handling of missing data was necessary. In addition, medications administered during hospitalization were recorded in a separated dataset. It included 18 different drug entities, some of which are essentially the same drug but differentiated by the method of administration, belonging to 7 primary categories (diuretics, inotropes, vasodilators, LDL cholesterol reducers, Angiotensin-converting enzyme (ACE) inhibitors, Angiotensin II receptor blocker (ARB) and Chinese traditional medications). Altogether the drug related data includes 12654 records related to 2004 patients, resulting in a therapy composed of one or more drugs for each patient, except for 4 subjects for whom treatment information was not provided. The rationale employed for managing missing information holds significant relevance in this study and will be discussed in detail in the next paragraph.

2.2. Data preparation

The initial preliminary step in the development of this study was data processing. The key operations involved in this phase will be discussed in the following sections. Firstly, it was necessary to extract

therapy information, format it into a usable form, and integrate it into the main dataset. From the additional dataset the set of drugs administered to each patient was extracted and reported in the main dataset with a one-hot-encoding approach, both including each specific medication (one column per drug) and drug categories. This was done to have access to therapy information at both levels during the feature selection phase. This operation resulted in the creation of 25 binary columns, having as baseline the absence of information regarding the administered therapy.

Subsequently, the information regarding hospital re-admissions and mortality was handled. Considering the study's focus on readmission *within* 6 months after hospitalization, all the information available in the original dataset regarding readmission at 28 days, 3 months, 6 months, and the return to emergency department within 6 months were aggregated in a single *Outcome* feature, to be later used as target variable. Then, still considering as objective the readmission *after discharge*, all patients dead during hospitalization were excluded. Furthermore, it was initially intended to classify patients who died after discharge within the 6-month period as a separate outcome from surviving without readmission and being readmitted. However, due to the limited size of this population, it was deemed insufficient to generate statistically significant insights for model development. Consequently, these patients were excluded from the dataset *[Fig.1, Appendix A]*.

Afterwards, a rigorous investigation was conducted on the information contained in both categorical and numerical variables, and on the evaluation of the optimal method for handling missing information. At this point, since some of the following steps leverage the variables distribution to operate on the dataset, 20% of the dataset has been randomly taken out for final testing, by means of a stratified splitting. The same processing approach implemented on the training dataset will be applied to this held-out portion.

Regarding the categorical variables, the main focus was to find the optimal method for mapping them to numerical values, the only feasible inputs for a model, while minimizing information loss. Features that presented an inherent ordering relative to an increasing degree of intensity or severity were mapped to progressive integer numbers. This was done to preserve the information regarding their relative ordering that would have been lost by using a straightforward one-hot encoding, which was indeed the method employed over the other features. Therefore, the remaining variables were binarized using a more traditional approach, but care was taken to determine the most appropriate value to be encoded as baseline (all zeros instance) on a case-by-case basis.

On the other hand, the analysis of numerical data focused on the detection and handling of noisy values and potential errors in the data collection process. In this study, having limited possibility to interact with domain experts and considering the high dimensionality of the feature space, it has been decided to develop an automatic pipeline for outliers detection. To this purpose, two approaches were explored: namely the Robust Z-Score and a quantile-based method, both tailored to selectively identify the truly extreme values with the highest likelihood of being unreliable. The former method is based on the same rationale of the traditional z-score but uses the median and median absolute deviation (MAD) instead of the mean and standard deviation, therefore the identification of outliers still relies on the choice of a specific threshold. The latter identifies as outliers the values outside the chosen lower and upper quantiles. Both approaches are valid and conservative when their parameters are appropriately set. However, as they are automatic methods, they are still prone to errors. To enhance the robustness of the approach, an ensemble of the two methods was also investigated. The rationale behind this was that if a value was identified as an outlier by both methods, the likelihood of it being an error was considerably high. Consequently, such values were deemed unreliable and treated as missing data. An important preliminary step in this process is data understanding. Indeed, as the good practice suggests, before blindly removing the outliers, an accurate analysis of variable distributions, perhaps stratifying over the target, has been performed to ensure the rationale applicability.

The subsequent step involved addressing missing data and determining the most suitable imputation technique on a case-by-case basis. Initially, columns with missing information exceeding 30% of the samples and samples with more than 40% of missing features were deemed insufficiently informative and excluded from further analysis, as their imputation would have yielded meaningless results. Then, the resulting missing data were imputed column by column, preserving data distributions: the distribution of each variable in the *training set* was estimated using KDE (Kernel Density Estimation [8]) stratifying over the target, the missing values have been sampled from *that distribution* and assigned to missing values both in the *train set* and in the *test set*. This solution allowed the use of training distributions to impute test missing values, making this method suitable for a hypothetical hidden test set too.

Finally, through visual inspection of the obtained dataset stratified over the target, we were able to analyze data distributions, with the majority displaying either Gaussian or exponential patterns. In order to improve the distribution of the exponential data, we applied a logarithmic transformation to re-distribute data closer to Gaussianity and enhance the performance of machine learning algorithms that rely on the assumption of normality.

2.3. Feature selection

Once the dataset was cleaned and prepared for the task, it was possible to analyze its content and perform feature selection. Variables were separated between categorical and numerical variables to differentiate the feature selection process in the two cases. Each numerical variable was split based on the target values, and the distribution of the two groups was evaluated using the Kruskal-Wallis test [9]. Only variables with a p-value < 0.05 were considered statistically significant and retained. To prepare these variables for modeling, a Standard Scaler was used for standardization. Furthermore, the correlation among the selected numerical variables was computed, excluding those with a correlation coefficient higher than 0.75. On the other hand, for categorical variables, the Chi-Squared test [10] was utilized. Similar to the Kruskal-Wallis test, the categorical variables were split based on the target values, and the distribution of each group was evaluated. Variables with a p-value < 0.05 were considered significant and retained. By applying these feature selection techniques, the most relevant variables for the modeling task were identified, while significantly reducing the dataset's dimensionality. The final set of variables, comprising both numerical and categorical features, will serve as the basis for subsequent modeling steps, enhancing the efficiency and interpretability of the machine learning models.

2.4. Model development

After establishing the working dataset, the model design phase followed. The model families that have been included in this investigation are Decision Trees, Random forests, K-Nearest Neighbors (KNN), Logistic Regression, Gaussian Naive Bayes classifiers, Linear discriminant analysis (LDA), Quadratic Discriminant Analysis (QDA), Support-Vector Classifiers (SVC), Adaboost classifiers and Multi-Layer Perceptron classifiers (MLP). The hyperparameter tuning of all these models has been performed using automatic search procedures (both Grid search and Random search) within a set of parameters tailored to each specific model and employing K-fold cross-validation. Once the optimal hyperparameters were determined, each model was re-trained using the complete training dataset. Finally, the performances of the models were evaluated on the held-out test set.

In this phase, to further increase the performance of the models after the initial attempts, some additional techniques were included in the pipeline. Firstly, a powerful ensemble comprising the top three models has been devised. Secondly, to address the issue of class imbalance, the inclusion of Synthetic Minority Over-sampling Technique (SMOTE) has been implemented on the minority class after having down sampled the majority class in order to add only 30% of synthetic data rather than 50%. Additionally, some experiments have been performed using the Maximum Relevance Minimum

Redundancy (MRMR) technique to perform a further and enhanced feature selection by measuring mutual information with the target and between each pair of features in conjunction, while checking the effectiveness of the set of predictors in terms of model performances.

2.5. Explainability

To get some insight about the decision process learned by the models, some explainability techniques were employed. The simplest linear models, as Logistic Regression, were the first investigated given their higher interpretability. These models enable the disclosure of the feature importance by simply analyzing the learned coefficients of the linear combination. While, for more complex models, the following explanatory framework have been used: SHapley Additive exPlanations (SHAP [11]), Local Interpretable Model-agnostic Explanations (LIME [12]), Partial dependence plots and feature shuffling.[1]

3. Results

Based on the steps described above, the data preparation process started from the integration of data from the original files resulting in a dataset comprising 2008 patients and 191 variables (190 excluding the patient ID). Then, the aggregation in the "outcome" variable of columns related to readmissions and the removal of dead patients further refined the dataset to 184 variables and 1951 patients. At this point, by performing the split operation, a train and a test dataset comprising respectively 1556 and 390 patients were obtained. Finally, processing the categorical variables and performing a preliminary drop of the insufficiently informative ones, the number of columns was reduced to 153. Afterwards, the methods for outlier identification described in the previous section were evaluated in comparison. By examining the ranges that the two methods and their ensemble defined as boundaries for value legitimacy, it was evident that the combined approach yielded the most favorable results. Consequently, based on this criterion, a total of 19 outliers were identified in the train dataset, along with 7 outliers in the test set. These data were further handled as missing values *[Fig.3, Appendix A]*. Then, removing both columns and samples deemed as insufficiently informative according to the criteria explained in the previous section led to the drop of an additional column, whereas all rows were kept. Subsequently, it was observed that the distribution of 7 of the features followed an exponential pattern. Consequently, according to the rationale explained in the previous section, a logarithmic transformation was applied to these variables *[Fig.2, Appendix A]*.

Next, regarding feature selection, applying the Kruskal-Wallis test to numerical variables and the Chi-Squared test for categoricals, only 36 numerical variables out of 89 (53 excluded) and 15 categoricals out of 61 (46 excluded) resulted statistically significant for the outcome. By removing the covariates that resulted to be correlated among each other, another 6 numericals variables were excluded. The results of this phase were found to be entirely consistent with medical knowledge and common sense. For instance, the Mean Arterial Pressure was deemed redundant due to the availability of systolic and diastolic values, while the latter two were retained as they did show correlation, but below the defined threshold, which might be consistent with the altered blood pressure control and the increased myocardium stiffness in patients with heart failure. Moreover, by comparing the model performance with and without the feature selection pipeline the effectiveness of the procedure was evident, obtaining an improvement up to 20% in terms of prediction accuracy. It's important to note that the original dataset contained variables that were strongly correlated with the target variable, such as information about re-hospitalizations and emergency re-admissions. Including these variables as predictors in the model would have significantly improved its performance. However, it would have also introduced a sort of cheating deviating from the study's objective. Hence, the results presented in this paper are based on a dataset that only includes variables available at the time of hospital discharge. Therefore, in conclusion, after further excluding from the working dataset the

[1] Details about the explainability techniques can be found in the notebook ***03. Explainability.***

patient's ID and retaining the *Outcome* as a separated target vector, a subset of 43 predictors was selected and employed for model training. Carrying out the aforementioned model design procedure and specifically using the MRMR technique, a further reduction of the set of predictors was performed (22 features kept) and the following results were obtained[2]:

	Train Accuracy	Test Accuracy	Parameters
Logistic Regression	0.659	0.638	C=1.0, max_iter=1000
Random Forest	0.871	0.615	max_depth=8, min_samples_leaf=8, min_samples_split=5
SVC	0.768	0.646	C=1.0, kernel='rbf'
LDA	0.661	0.6435	//
MLP	0.737	0.641	alpha=0.8, hidden_sizes=(10, 5)
Ensemble (best 3)	0.73	0.659	*Soft ensemble*

At this point, it has been tried to use the trained model to get some insights about the correlation between the predictors and the target. Through SHAP explainability technique over the best ensemble, it emerged that some blood values were the most relevant indicators for re-hospitalization within 6 months. In particular, a high D-dimer concentration, seemed to reduce the probability of being readmitted, as well as high red blood cells, sodium and prothrombin activity. Vice versa, a high eosinophil and basophil count, as well as a high NYHA classification and CCI score, resulted to be effective predictors of readmission *[Fig.4 and 5, Appendix A]*. Regarding medications, after dimensionality reduction, only a limited number of variables were retained, with some primary categories completely excluded. The final set of predictors consisted solely of diuretic singular therapies, namely Hydrochlorothiazide tablet, Furosemide injection, and Furosemide tablet. The first one had a positive influence on the outcome, while Furosemide had a negative influence *[Fig.4, Appendix A]*. It's worth mentioning that the other explainability techniques, also over other models, were in accordance with what presented above *[Fig.6, Appendix A]*.

4. Discussion and conclusions

Ultimately, the performance of the models is not particularly noteworthy, but explainability has yielded satisfactory results. It remarked on the utility of the NYHA classification and the impact that comorbidities can have on future readmissions (Charlson Comorbidity Index (CCI) is a scoring system that considers comorbidities to predict mortality risk and high white blood cells counts can be sign of an ongoing infection or can be linked to a variety of disorders). Moreover, the different influences of the two diuretic drugs on the outcome can highlight a higher effectiveness in the long term of a medication with respect to the other, which can have significant implications on medical practice. Altogether, the analysis revealed that the model learned to predict readmission based on reasonable aspects. Still, a further discussion with clinicians would provide more robust conclusions.

As possible future developments, the information regarding the administration times of medications might be included in the analysis to verify whether it affects patient outcomes. Another intriguing aspect would be to replicate the analysis on an independent dataset, preferably from a different population, and potentially perform model re-tuning. This would allow for a comparison of the results and highlight any divergent statistical conclusions regarding correlations.

[2] Only the top 5 and relative ensemble of the best 3 models is reported for sake of conciseness, additional details can be found in the notebook.

5. Bibliography

[1] McDonagh TA, Metra M, Adamo M, Gardner RS, Baumbach A et al.; ESC Scientific Document Group. *2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure*. Eur Heart J. 2021 doi: 10.1093/eurheartj/ehab368.

[2] Roger VL, Go AS, Lloyd-Jones DM, Benjamin EJ, Berry JD, et al.; American Heart Association Statistics Committee and Stroke Statistics Subcommittee. *Heart disease and stroke statistics--2012 update: a report from the American Heart Association*. Circulation. 2012. doi: 10.1161/CIR.0b013e31823ac046.

[3] Bergethon KE, Ju C, DeVore AD, Hardy NC, Fonarow GC, et al.; *Trends in 30-Day Readmission Rates for Patients Hospitalized With Heart Failure: Findings From the Get With The Guidelines-Heart Failure Registry*. Circ Heart Fail. 2016. doi: 10.1161/CIRCHEARTFAILURE.115.002594.

[4] Ross JS, Mulvey GK, Stauffer B, Patlolla V, Bernheim SM, et al.; *Statistical models and patient predictors of readmission for heart failure: a systematic review*. Arch Intern Med. 2008 doi: 10.1001/archinte.168.13.1371.

[5] Zhang, Z., Cao, L., Zhao, Y., Xu, Z., Chen, R., Lv, L., & Xu, P.. *Hospitalized patients with heart failure: integrating electronic healthcare records and external outcome data (version 1.3)*. PhysioNet. 2022. doi: https://doi.org/10.13026/5m60-vs44.

[6] Russell SD, Saval MA, Robbins JL, Ellestad MH, Gottlieb SS, Handberg EM, Zhou Y, Chandler B; HF-ACTION Investigators. *New York Heart Association functional class predicts exercise parameters in the current era*. Am Heart J. 2009 doi: 10.1016/j.ahj.2009.07.017.

[7] Killip T 3rd, Kimball JT. *Treatment of myocardial infarction in a coronary care unit. A two year experience with 250 patients*. Am J Cardiol. 1967 doi: 10.1016/0002-9149(67)90023-9.

[8]Węglarczyk, S.; Kernel density estimation and its application. *ITM Web of Conferences*. Vol. 23. EDP Sciences, 2018.

[9] McKight, Patrick E., and Julius Najab. Kruskal-wallis test. *The corsini encyclopedia of psychology*. 2010.

[10] Zibran, MF.; Chi-squared test of independence. *Department of Computer Science, University of Calgary, Alberta, Canada*. (2007)

[11]Nohara, Y., Matsumoto, K., Soejima, H., & Nakashima, N.; Explanation of machine learning models using shapley additive explanation and application for real data in hospital. *Computer Methods and Programs in Biomedicine*, 2022.

[12] Kumarakulasinghe, N. B., Blomberg, T., Liu, J., Leao, A. S., & Papapetrou, P.; Evaluating local interpretable model-agnostic explanations on clinical machine learning classification models. *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*. 2020.