CHL5230 Datathon #2
Final Report
Team 5: Xiao Yan, Amrit Tiwana, and Sashini Kosgoda

Using Machine Learning to Predict Heart Failure Mortality Based on Clinical Risk Factors

Introduction

The rise in heart failure cases is alarming as it leads to approximately 17 million deaths per year.[1] Heart failure occurs when the heart is unable to effectively meet the body's need for sufficient blood circulation. Previous research has identified lifestyle factors that increase the risk of heart failure including an unhealthy diet, sedentary lifestyle, cigarette smoking, and illicit drug and alcohol use.[2,3] While lifestyle-related risk factors are well researched, our understanding of clinical risk factors remains limited.

Using electronic medical record data from 299 heart failure patients in Pakistan, the objective of our analysis is to explore the implementation of a logistic regression machine learning algorithm. This algorithm aims to predict the likelihood of heart failure mortality based on a set of clinical features, specifically symptoms, body features, and values based on clinical laboratory tests. The use of logistic regression can not only predict patient mortality, but also identify the most important features contributing to mortality. The insights from this analysis have the potential to inform heart failure prevention strategies and improve the well-being of affected individuals.

Data Engineering Process

The mortality dataset for cardiovascular disease complications was used to analyze factors affecting heart failure.[4] We began by checking for missing values and outliers in the dataset. We visualized the distribution of numeric variables using histograms and boxplots and categorical variables using boxplots and barplots. Although the variables *creatinine phosphokinase, platelets, serum creatinine* and *serum sodium* contained outliers, we decided to keep outliers because they may indicate rare diseases or critical emergencies, which are important to keep in our analysis since our outcome of interest is mortality. Further, we explored descriptive statistics including the minimum, median, and maximum to identify any errors to correct for.

To prepare the dataset for analysis, we split the data into features and the outcome. All 12 features were kept for the purposes of the analysis since we wanted to explore the full model and then remove features after if any of them don't seem significant. The dataset was split with 80% of the data used to train the model and 20% used to test the model. Further, the features were all standardized to ensure they don't have large differences in range. A logistic regression algorithm was used to build and train a model which could predict the probability of death based on a specified set of features. To evaluate the model, we used a confusion matrix heat map and a classification report. Specifically, we used evaluation metrics including accuracy, precision, f1 score, and the pseudo R-squared. To further explore the significance of each feature to the model, a logistic regression model was created for each feature and a summary of the full logistic regression model was obtained.

Analysis

We choose to create a logistic regression model for this dataset to predict the death probability based on specific features since the outcome is a binary variable. From looking at boxplots of all the features we determined there were varying ranges, to minimize the effect of this all features were standardized. We also created individual logistic regression plots to examine how each feature influences the probability of death occurring for an individual. The logistic regression summary table was also useful to identify significant features by looking for p-values less than the significance level 0.05.

Findings

We found from the exploratory data analysis that there are no missing values in the dataset, and since we decided not to remove the outliers because of retaining the extreme cases, there was no need to clean the data. Thus, our analytic sample size is 299 observations.

Based on the evaluation metrics, the accuracy for the model is about 78%. The precision and F1 score is 75% and 84% for predicting patients who are alive and 88% and 68% for patients that had a fatal heart failure. In addition, the pseudo R-squared is 0.415, meaning that our logistic model captures about 41.5% of the total variation in the model. The variables *age*, *ejection fraction*, *serum creatinine* and *follow-up time* were statistically significant as they had a p-value less than 0.05 (see Appendix).

Conclusion

Our logistic regression model has an overall accuracy of 78%. Even though the precision for predicting dead patients is higher than predicting patients alive, the confusion matrix suggests that many patients who are actually dead are predicted as alive. We want to avoid this type of error because we do not want patients that have fatal heart failure to be predicted as alive. This is illustrated by a slightly higher recall and f1-score of predicting patients that are still alive than those that are dead, suggesting that our model performs better on predicting patients that are alive. We conclude that our model is not perfect in predicting patients that are alive and in accounting for total variance in the dependent variable, but is still powerful in terms of its overall accuracy as well as predicting patients that do not have a fatal heart failure.

There are some limitations in our research. We found that in the exploratory data analysis there were many outliers in the variables *serum creatinine* and *creatinine phosphokinase*. These outliers would largely decrease our model accuracy; however these extreme cases were kept in our model since we wanted to see their effect on the mortality of heart failure patients. Also, we found that only *serum creatinine*, *ejection fraction, follow-up time* and *age* are significantly related to *mortality*. This suggests that logistic regression may not be the best model for this dataset. Another model or a transformation may be needed to better interpret the feature variables.

Individual Contributions

Xiao Yan contributed to the findings, conclusion, and presentation. Amrit Tiwana contributed to the introduction, findings, and presentation. Sashini Kosgoda contributed to the data engineering process, analysis, findings, and presentation.

Source Code: GitHub **| Presentation**: Google Slides

Appendix

Table 1. Multivariable Logistic Regression Model

Variable	**Adjusted Odds Ratio (95% CI)**	***P* Value**
Age (Years)	1.05 (1.02 - 1.08)	0.003*
Anaemia (Yes vs. No)	0.99 (0.49 - 2.01)	0.983
Creatinine Phosphokinase (mcg/L)	1.00 (1.00 - 1.00)	0.212
Diabetes (Yes vs. No)	1.16 (0.58 - 2.30)	0.679
Ejection Fraction (Percentage)	0.93 (0.90 - 0.96)	<0.001*
High Blood Pressure (Yes vs. No)	0.90 (0.45 - 1.82)	0.775
Platelets (kiloplatelets/mL)	1.00 (1.00 - 1.00)	0.525
Serum Creatinine (mg/dL)	1.95 (1.36 - 2.78)	<0.001*
Serum Sodium (mEq/dL)	1.95 (0.87 - 1.01)	0.092
Sex (Male vs. Female)	0.59 (0.26 - 1.32)	0.197
Smoking (Yes vs. No)	0.99 (0.44 - 2.21)	0.974
Follow-up Time (Days)	0.98 (0.97 - 9.85)	<0.001*

*P<0.05

References

1. Chicco D, Jurman G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making. 2020;20(1). doi:10.1186/s12911-020-1023-5
2. Meijers WC, de Boer RA. Common risk factors for heart failure and cancer. Cardiovascular Research. 2019;115(5):844–53. doi:10.1093/cvr/cvz035
3. Nishimura M, Bhatia H, Ma J, Dickson SD, Alshawabkeh L, Adler E, et al. The impact of substance abuse on heart failure hospitalizations. The American Journal of Medicine. 2020;133(2). doi:10.1016/j.amjmed.2019.07.017
4. Ahmad T, Munir A, Bhatti SH, Aftab M, Raza MA. Survival analysis of heart failure patients: A case study. PLOS ONE. 2017;12(7). doi:10.1371/journal.pone.0181001