Datathon 2

Nitya Kuruvila, Shuyang Wu, Gwen Eagle (Team 17)

Dalla Lana School of Public Health, University of Toronto

CHL 5230: Applied Machine Learning for Health Data

Dr. Zahra Shakeri

October 3, 2023

Introduction

The overall objective of this paper is to explore the relationship between demographic, lifestyle, medical factors and cardiovascular diseases. Cardiovascular diseases (CVD) are the leading cause of death worldwide; accounting for 32% of deaths globally in 2019 (WHO, 2021). Firstly, this paper will identify the association between key patient data, including health metrics and demographic characteristics, and hypertension, one of the most significant risk factors for cardiovascular disease (Wu et al., 2015). Globally, approximately 47% of coronary artery diseases and 54% of strokes can be attributed to hypertension every year (Wu et al., 2015). Next, this paper will explore the relationship between hypertension and mortality. Given the high prevalence and global impact of CVD, it is important to understand how health and lifestyle factors contribute to hypertension and subsequently mortality. The findings of this paper can improve knowledge and understanding regarding CVD and help inform healthcare initiatives to reduce the prevalence of CVD and its associated consequences.

Data Engineering Process

The analyzed dataset comes from the Faisalabad Institute of Cardiology and Allied Hospital, Pakistan and consists of medical, demographic and lifestyle data from 229 individuals with class III or IV left ventricular systolic dysfunction. This dataset consists of 105 females and 194 males ranging from ages 40 to 95. We initiated our data processing by identifying missing values within the dataset. We then looked at the summary statistics, for all features and frequency counts for binary features. Binary features with less than 10% variability were removed from further analysis due to the limited information they provided. Following this, we generated a correlation heatmap to assess the collinearity between features. This preliminary analysis serves as a reference for feature selection in our predictive models. Next, we constructed box plots, to visualize the distribution of each feature and identify outliers. We created a pair plot, looking at patients with and without hypertension, to visualize relationships between features and our prediction target, hypertension.

For KNN classification, we began by splitting the data into labels and features. Hypertension status was defined as the label and features used were age, anaemia, creatinine phosphokinase (CPK), diabetes, ejection fraction, serum sodium, sex, and smoking. Mortality status and follow-up duration were excluded as we were looking at classifying hypertension status in patients prior to mortality. In addition, serum creatinine and platelet count were excluded as studies have shown that both these factors are increased by smoking therefore they are accounted for by the inclusion of smoking status (Eid et al., 2022; Ghahremanfard et al., 2015). The data was then split, 80% for the training set and 20% for the testing set. All features were scaled to improve the accuracy of the model. Testing was also done to determine the optimal K-value for the model.

For the logistic regression model, the dataset was prepared by identifying 'death event' as the outcome variable while the remaining variables were classified as predictors. These predictors included hypertension status, age, hemoglobin level, CPK, diabetes, heart ejection rate, platelet count, serum creatinine, blood sodium, smoking, gender and follow-up time. The data set was then split into the training (20% of data) and test model (80% of data). The data was standardized and transformed to a normal distribution.

Analysis

To conduct this analysis we first used the K-Nearest Neighbors algorithm (KNN) to determine if hypertension status could be effectively classified using the aforementioned features. We chose this approach for two reasons. Firstly, the hypertension status, used to classify the data, was known and secondly, we were unaware if the data followed a certain distribution. Given that the KNN classification model is a non-parametric supervised learning model, it could be used effectively to accomplish the intended outcome. A classification report and confusion matrix were used to evaluate the model.

Next, we used logistic regression to determine the association between mortality and hypertension, as well as the other predictors. The logistic regression model was built using L2 regularization to discourage large coefficients and prevent overfitting, while still maintaining all features within the model due to potential collinear features, and used C = 1.0 for moderate regularization. The 'liblinear' solver was used due to the small sample size and binary outcome. Logistic regression was used as the outcome of interest, mortality, is categorical. Additionally, this model provides the probability estimate of the outcome, mortality, given the predictor variables. A confusion matrix and class report were performed to evaluate the model and logistic regression curves were plotted for each feature. A summary of the logistic regression model was obtained for model interpretation.

Findings

The initial cleaning found no missing data in the dataset and all binary features had greater than 10% variability. Outliers were identified by the box plots and supported by the summary statistics for age, CPK, ejection fraction, blood platelets, serum creatinine and serum sodium. The highest correlations identified by the correlation heat map were 0.45 and 0.53 between smoking and sex and follow-up time and mortality respectively. No significant observations were made based on the pair plot.

Evaluation of the KNN classification model (k=14) determined an accuracy score of 0.68. The precision values and recall values for non-hypertensive and hypertensive individuals were 0.67, 1.00, 1.00 and 0.10 respectively. The number of true positive, true negative, false positive and false negative samples, determined by the confusion matrix, were 2, 39, 0 and 19 respectively. Based on these values, it suggests that the model can correctly classify individuals without hypertension however, cannot effectively classify individuals with hypertension.

The logistic regression model determined an accuracy score of 0.78. The precision values and recall values for no death event and death event were 0.75 and 0.94, and 0.88, and 0.56, respectively. The number of true positive, true negative, false positive and false negative samples, determined by the confusion matrix, were 14, 33, 2 and 11, respectively. These results suggest that the model can correctly predict those who will not experience death (high specificity), but will struggle to predict those who will experience death (low sensitivity). The predictors whose odds ratio values reached statistical significance at the 5% level were serum creatinine (*1.81*, 95% CI:1.36 , 2.78 , *p*=0.000), ejection fraction (0.93, 95% CI: 0.90, 0.96), p=0.000), age (*1.05*, 95% CI: 1.02, 1.08 , p =0.003), and time (*0.98*, 95% CI:0.97, 0.99, p =0.000). These results suggest that higher serum creatinine and older age are statistically significant risk factors for a death event, whereas higher ejection fraction and longer follow-up time are protective factors against a death event. Hypertension was not a statistically significant predictor (*0.90*, 95% CI: 0.45, 1.82, *p*= 0.775).

Conclusion

Both the KNN model and the logistic regression model had high specificity, accurately identifying patients who did not have hypertension or experience a death event, respectively. Thus, these models may have use in diagnostic confirmation of these outcomes rather than for population level screening. Contrary to the literature, hypertension was not a statistically significant predictor of mortality, which underscores the need for model improvement prior to real-world application. Given our findings, we would suggest that health interventions focus on serum creatinine, ejection fraction and age when aiming to reduce mortality in individuals with left ventricular systolic dysfunction.

References

Eid, H. A., Moazen, E. M., Elhussini, M., Shoman, H., Hassan, A., Elsheikh, A., Rezk, A., Moursi, A., Atef, M., & Kabil, A. (2022). The Influence of Smoking on Renal Functions Among Apparently Healthy Smokers. *Journal of Multidisciplinary Healthcare*, *15*, 2969–2978. https://doi.org/10.2147/JMDH.S392848

Ghahremanfard, F., Semnani, V., Ghorbani, R., Malek, F., Behzadfar, A., & Zahmatkesh, M. (2015). Effects of cigarette smoking on morphological features of platelets in healthy men. *Saudi Medical Journal*, *36*(7), 847–850. https://doi.org/10.15537/smj.2015.7.11026

WHO. (2021, June 11). *Cardiovascular diseases (CVDs)*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Wu, C.-Y., Hu, H.-Y., Chou, Y.-J., Huang, N., Chou, Y.-C., & Li, C.-P. (2015). High Blood Pressure and All-Cause and Cardiovascular Disease Mortalities in Community-Dwelling Older Adults. *Medicine*, *94*(47), e2160. https://doi.org/10.1097/MD.0000000000002160

Individual Contributions

Everyone contributed equally to all parts.

Slideshow

Github link: https://github.com/Kiara-Wu/17-CHL5230-F23