Adele Collin and Kyla Gabriel
12/15/2023

Harvard Medical School
BMI 715

Predicting Coronary Heart Disease diagnosis in NHANES dataset

Coronary heart disease (CHD) is the leading cause of death in the United States with more than 18.2 million adults affected (National Heart, Lung and Blood Institute, 2022). This medical condition is characterized by the accumulation of lipids on the arterial walls causing them to narrow and restrict cardiac blood flow (Goodwill et al., 2017). Because of its detrimental consequences, researchers are interested in early prevention methods using machine learning algorithms (Hassan et al., 2022). We aim to predict CHD using hematological data from the NHANES (National Health and Nutrition Examination Survey) dataset taken from 2013-2014. Current literature supports that hematological indices play a key role in the diagnosis and prognosis of CHD patients (Budzianowsk, 2017) in addition to the well-studied impact of lipid accumulation such as triglycerides (Gotto et al., 1998) and cholesterol (Grundy et al., 1986). Consequently, we were wondering: can blood cell counts and small metabolites blood concentration be predictive of past CHD diagnosis? To answer this question, we aim to build a predictive model of past CHD diagnosis (variable MCQ360C in NHANES database) using the following variables: cholesterol, triglycerides and glucose blood concentrations, white blood cells and red blood cells counts, and lymphocytes, eosinophil, and basophil percents. Later on, we wish to explore the socio-economic correlation between CHD diagnosis and monthly family income. Previous studies highlighted a demographic transition of the disease (Janati et al., 2011). Consequently, we were curious to know if our dataset also displayed such correlation and if it is the case if CHD diagnosis has been more frequent in upper or lower socio-economic demographics.

Classification model: can previous CHD diagnosis be predicted using hematological data?

Before building the model, we proceeded to preliminary cleaning of the database. First, we delete the rows where the proportion of NA values was more than 60% and the columns where the proportion of NA was more than 30%. Those thresholds allow for a trade-off by selecting the top 50% of brown and columns in terms of completeness. The remaining NA values are replaced by the median to have a minimal influence over the original distribution. Then we eliminate one variable out of each pair that had a Spearman coefficient superior to 0.3 to ensure a limited collinearity between variables. This threshold has been determined so that it would eliminate the top 50% of the most correlated pairs. Then we notice by plotting boxplots that some variables were constant all along the dataset and, consequently, were impossible to use as predictive variables. Therefore, we remove all variables that have equal first and third quartile, considering those variables too prone to encourage the use of outliers by the model. These pre-processing steps enabled us to narrow down the 414 variables in the original dataset to 33 good quality variables. Then, we chose our dependent and predictive variables from this shortlist, based on scientific literature. We use this 3-step approach of pre-processing, selection and fine-tuning to get the best quality predictors for our model.

Once the predictive and dependent variables are selected, we plot them against each other and compute the correlation matrix to ensure that the correlation or collinearity between them is limited (see Figure 1). None was above our threshold of 30% and there is no visible correlation on the scatter plots, therefore we decided that the independence hypothesis can be considered met by our selection. We plot all the 9 variables to visualize their distribution and noticed that all of them had a significant number of outliers. To eliminate those errors we decided to implement the Interquartile Range Method (Vinhuta et al., 2018). We proceeded to do that before scaling with the z-score so that the outliers do not influence the new distribution.

Since the dependent variable is binary and the predictive ones continuous, we decide to use a logistic regression. The only assumption for this model is the independence of variables and samples, the former has been verified at the precedent step and the latter is verified since each sample corresponds to one respondent (NHANES, 2023). Since all the selected variables have been proven to be predictive of CHD diagnosis, we keep them all for our first model. The dataset is separated between a train and a test set to compute prediction scores. The train set is balanced using over and under sampling to optimize the chances for our model to capture the patterns regarding CHD-diagnosed patients. Regarding the results, the deviance is smaller for our model (5269) than for the null model (6110). Our model fits better than just the intercept. To assess the significance of such a difference, we perform a chi-square goodness of fit test which works in a similar way than the Fischer statistic for linear regression. The p-value is rounded to 0 which is less than the chosen level of significance (0.05) so this model is significantly better than the sole intercept to predict the classes of CHD diagnosis. Regarding individual coefficients, most of them have p-values which are by far inferior to 0.05 which means that CHD diagnosis changes significantly with each of them. Only the eosinophils percent (LBXEOPCT) has a p-value of 0.8 which is greater than the significance threshold of 0.05. The red blood cell count p-value is close to the threshold (p-value of 0.02) but remains significant.

The fit is evaluated using the sensitivity score over the test set. Given the medical context, it is important to get as few false negatives as possible, i.e. patients that are suffering from CHD but who are not detected. The score is 0.58 which means that 58% of the total patients that are positive for CHD are correctly classified. This score is good, but the class imbalance probably limits the capacity of the model to correctly predict the minority class. Regarding the fit plot, we choose a binned residuals plot (see Figure 1) to check the repartition of residuals (Gelman et al., 2000) and assess the quality of fit for the whole range of expected CHD scores. There are strong positive residuals for the low expected CHD scores and strong negative residuals for the scores around 0.5. This correlates with our previous analysis: a significant number of patients are false positives (low CHD scores predicted higher) and the positive patients are often wrongly classified as negative (middle-ground patients predicted lower). Also, numerous points are outside of the confidence interval which shows higher-than-expected residuals and a looser fit than by chance alone.

Since the p-value of two coefficients are above or close to their significance level, we decide to run a stepwise AIC-based variable selection algorithm to select the most relevant ones. This stepwise approach is chosen for its flexibility. The number of variables small enough to avoid overfitting. In the end, the eosinophil percent is discarded but not the red blood cell count. Since the fitting of the logistic regression is based on maximum likelihood estimation, we choose to compare the two models based on their log-likelihood. The first model has a log-likelihood of 2634.66 and the second 2634.7 which means that the second model has a better fit than the first one. To answer the question around statistical significance, we follow-up with a likelihood ratio test. The null hypothesis is that the more complex model (initial) is not significantly better than the simpler model (secondary). The alternative hypothesis is that the more complex model provides a significantly better fit. The p-value of the test is 0.79 which is greater than the significance level of 0.05 so we cannot reject the null hypothesis and the more complex model does not have a significantly better fit. Thus, for the sake of parsimony, we keep the simpler secondary model.

Hypothesis testing: is previous CHD diagnosis correlated with socio-economic status?

Because of the known association between social determinants of health and cardiac history, we were curious to know if our dataset also displayed such correlation. Our null hypothesis is that there is no significant correlation between CHD diagnosis and socio-economic demographics. Our alternative hypothesis is that there is a statistically significant correlation between CHD diagnosis and socio-economic demographics. We first visualize the distributions of the monthly family income and CHD diagnosis, then fix the data imbalance of CHD diagnosis by oversampling the minority class. After visualizing and determining that the income variable was non-parametric, we test our data using a Wilcoxon rank sum. With the resulting W test statistic of 101882 and p-value of 2.2e-16 (alpha set at 0.05), we can reject the null hypothesis and conclude that there is a statistically significant difference between the two groups.

Conclusions

The primary goal of this project is to predict if an individual has CHD based on metabolite and blood cell clinical factors using machine learning. With our model, we see that each predictor variable is significant to our model apart from eosinophils. We have also determined that the initial complex model does not provide a significantly better fit than the simpler one we got after stepwise variable selection. Consequently, the secondary nested model is to be preferred for any further analysis. When evaluating the model’s sensitivity, we find a multitude of false positive errors with many CHD-positive patients being misclassified as negative which is an important limitation and aspect that is necessary to assess in future steps. When investigating the possible correlation between CHD diagnosis and socioeconomic status, we reject the null hypothesis and find that there is a correlative relationship between CHD diagnosis and the socio-economic factor of monthly family income.

Contributions

Adele and Kyla both worked on the write-up analysis and equally contributed to the introduction on this page. The individual contributions are as follows: the classification model was performed by Adele while hypothesis testing was performed by Kyla. Data processing and analysis from both collaborators took place throughout the entire project.

Materials

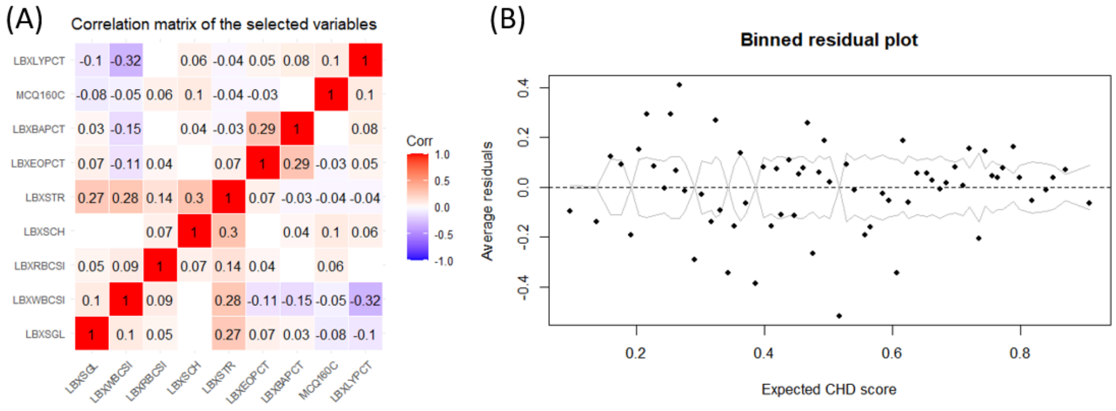

Figure 1: (A) Correlation matrix between the variables selected for prediction. The blank cases mean that the Spearman correlation coefficient is not significant. (B) Binned residual plot to analyze the goodness of fit between the average of residuals per bin and the expected CHD score.

References

Budzianowski, J., Pieszko, K., Burchardt, P., Rzeźniczak, J., & Hiczkiewicz, J. (2017). The role of hematological indices in patients with acute coronary syndrome. *Disease Markers*, *2017*, 1–9. https://doi.org/10.1155/2017/3041565

Coronary heart disease - what is coronary heart disease? | NHLBI, NIH. (2022, March 24). https://www.nhlbi.nih.gov/health/coronary-heart-disease

Gelman, A., Goegebeur, Y., Tuerlinckx, F., & Van Mechelen, I. V. (2000). Diagnostic checks for discrete data regression models using posterior predictive simulations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *49*(2), 247–268. https://doi.org/10.1111/1467-9876.00190

Goodwill, A. G., Dick, G. M., Kiel, A. M., & Tune, J. D. (2017). Regulation of coronary blood flow. In R. Terjung (Ed.), *Comprehensive Physiology* (1st ed., pp. 321–382). Wiley. https://doi.org/10.1002/cphy.c160016

Gotto, A. M. (1998). Triglyceride as a risk factor for coronary artery disease. *The American Journal of Cardiology*, *82*(8), 22–25. https://doi.org/10.1016/S0002-9149(98)00770-X

Grundy, S. M. (1986). Cholesterol and coronary heart disease: A new era. *JAMA*, *256*(20), 2849. https://doi.org/10.1001/jama.1986.03380200087027

Hassan, Ch. A. U., Iqbal, J., Irfan, R., Hussain, S., Algarni, A. D., Bukhari, S. S. H., Alturki, N., & Ullah, S. S. (2022). Effectively predicting the presence of coronary heart disease using machine learning classifiers. *Sensors*, *22*(19), 7227. https://doi.org/10.3390/s22197227

Janati, A., Matlabi, H., Allahverdipour, H., Gholizadeh, M., & Abdollahi, L. (2011). Socioeconomic status and coronary heart disease. *Health Promotion Perspectives; ISSN: 2228-6497*. https://doi.org/10.5681/HPP.2011.011

NHANES—About the national health and nutrition examination survey. (2023, May 31). https://www.cdc.gov/nchs/nhanes/about_nhanes.htm

Vinutha, H. P., Poornima, B., & Sagar, B. M. (2018). Detection of outliers using interquartile range technique from intrusion dataset. In S. C. Satapathy, J. M. R. S. Tavares, V. Bhateja, & J. R. Mohanty (Eds.), *Information and Decision Sciences* (Vol. 701, pp. 511–518). Springer Singapore. https://doi.org/10.1007/978-981-10-7563-6_53