Project 1 - Final Report

Team 24: Hongyan Chen, Shuyang Wu

Dataset: CHL5230 dataset

Introduction:

According to statistical results, cardiovascular diseases (CVDs) stand as the primary worldwide contributor to mortality, claiming approximately 17.9 million lives annually (WHO, 2021). And from 2000 to 2019, there was a 3% rise in age-specific mortality rates for diabetes. In 2019, an estimated 2 million fatalities were attributed to diabetes and diabetes-related kidney disease (WHO, 2023). We can easily see that cardiovascular diseases (CVDs) and diabetes are major public health challenges worldwide. They impose a significant economic and healthcare burden on society, so identifying individuals at risk of developing these chronic conditions is crucial for effective management, early intervention, and improved patient outcomes. This research paper is going to explore risk prediction models through machine learning technique for identifying those at risk of CVD and diabetes[3-5], evaluates their strengths and limitations, and interprets the models with real-world patient datasets. We will develop our research based on the following two research questions: 1) using health factors to accurately estimate an individual's risk of developing cardiovascular disease, and 2) using medication usage data to create predictive models that effectively estimate the likelihood of an individual developing diabetes. Answering these research questions is vital for providing healthcare professionals and policymakers with valuable tools for risk assessment and preventive measures. Early identification of individuals at high risk of CVD or diabetes allows for timely interventions, lifestyle modifications, and targeted healthcare delivery, ultimately reducing the burden on healthcare systems and improving patient outcomes.

Methods:

The research analysis datasets come from Primary Care Electronic Medical Record data in the lab of Professor Aziz Guergachi, Toronto Metropolitan University. The dataset called C-Change Analysis Data contains 976148 observations in total of individual patients and 68 columns in total, which are all health-related statistical factors, including information about the patient's physiological data, health condition, medication treatment, and medical history. At the same time, the dataset includes patients in all age groups, from lower than 20 to higher than 80 years old. Using this dataset, we sought to explore the ability of the combination of physical condition variables(age, TC, HDL, LSL and sBP), lifestyle factors (smoking status) and medication usage factors(Any_AntiHTN, DM_Station, DM_ACE/ARB, ACE+ARB, ACE-or ARB, Diuretic_usage, CCB_usage, CCB_usage, Thiazide_usage, ACEI_usage, ARB_usage, Warfarin_usage, Antiplatelet_usage) to predict risk of CVD and diabetes[6-8]. We began by performing descriptive data analyses to ensure sufficient data quality for machine learning analysis. We identified missing values within the dataset. We then dealt with missing values with methods including mean imputation, deleting rows, and setting missing values to zero. We then looked at the summary statistics for all features and frequency counts for binary factors in order to get a general overview of all columns in the dataset. By looking at the critical statistics like mean and median, we are able to know the approximate distribution of variables. As well, we created box plots and histograms for all variables in the dataset to visualize distributions further and detect outliers at the same time. Our goal here was to ensure that there was no significant bias with the dataset that might impact our final result.

After focusing on the whole dataset, we started to focus on several specific features according to each research question. These predictive variables are chosen based on the literature review and collected

information. Then, we performed correlation heatmaps to assess the collinearity between features and response variables. This also served as a reference for future feature selection in our predictive models. Next, we also created a pair plot to visualize relationships between features and our prediction target.

For the prediction of CVD risks, we fitted a Linear Regression Model with multiple predictive features, which are sex, age, smoking status, total cholesterol, high-density lipoprotein, low-density lipoprotein, systolic blood pressure, and usage of any anti-hypertension medicine. At the same time, the response variable is the percentage risk of developing CVD diseases. Then, we built a linear regression model and trained the data with the fitted model. Lastly, we evaluated model performance, including calculating confidence intervals for regression coefficients, variance, and residual analyses. We used the trained model to make predictions for the risk of CVD on the test dataset.

For diabetes prediction based on medication usage, we used a Logistic Regression Model on input variables, including usage of statin medications, ACE, ARB, diuretic, CCB, Tiazide medications, warfarin, and antiplatelet medications with Diabetes Mellitus Hemolobin A1c as the response variable which indicates the level of control of an individual's diabetes's condition. We pre-processed the data to ensure the variables were in the proper format for logistic regression modeling. To improve the stability of the model, we examined multicollinearity among independent variables and deleted some factors(ACE+ARB, AVE or ARB, ACEI, ARB) to interpret regression coefficients and enhance predictive capability correctly. We also normalized our data using scaler transformation from the Scikit library in order to ensure all variables were normalized with a standard deviation (SD) of 1 and a mean of 0. This prevents our LR model from being disproportionately impacted by variables with different scales. Then, we used the trained LR model to predict the test dataset (X_test) and assess the model's performance with a confusion matrix and heat map. Lastly, we Perform a statistical summary analysis of the LR model and visualize the relationships between different features and diabetes.

Discussion:

Within the dataset, there were 976147 observations. Sex and age variables are evenly distributed. Based on these findings, this data set is likely representative of the diversity among the patients.

- Prediction of CVD risks

According to the result of model fitting, we could see that the mean square error is 24.550, which indicates that the average squared differences between the observed actual outturn values of the dataset and the values predicted by the model are 24.550. And the R2 score of the linear regression model is 0.519, which is the coefficient of determination, measures the proportion of variance of the model that is explained by the independent variables. An R2 score of 0.519 (or 51.9%) is a relatively high R2 score, and it suggests that about 51.9% of the variability in CVD risk can be explained by predictive variables in the model, which is a large portion of the variation. Thus, it shows that these factors included in the model significantly impact CVD risk, so these variables should be included in the prediction model. Also, according to Figure 1 below, the scatter plot for residuals shows a random pattern, and the histogram of residuals shows an approximate normal distribution around zero, indicating that the model performs well on individual observations without essential errors. Besides, we performed Lasso regression and Ridge regression to check for any overfitting in the current model. According to the result of fitting, the mean squared errors of lasso regression and ridge regression are not getting lower compared to the current model, and the R2 score is not getting higher as well. Thus, we could conclude that there are not any redundant or unnecessary variables in our current model that need to be removed.

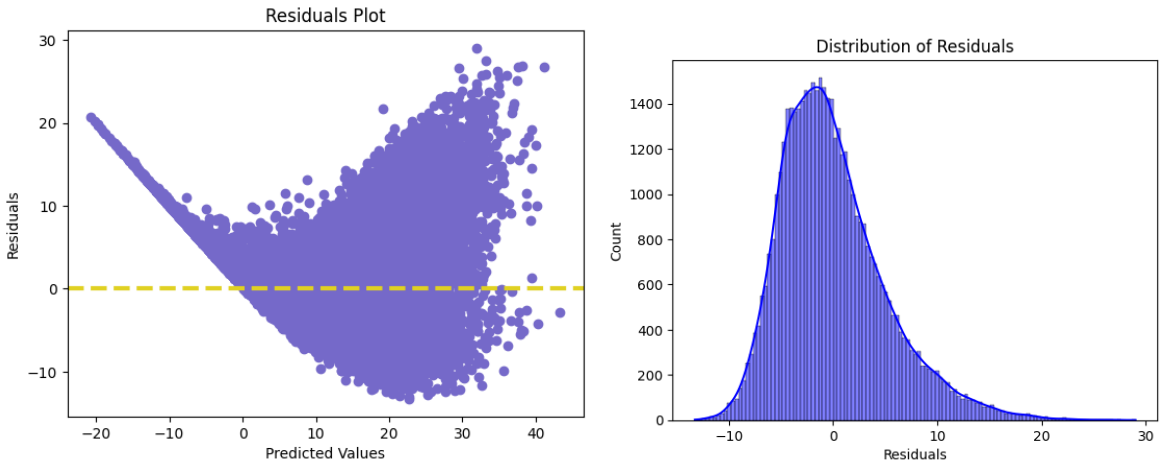

Figure 1

- Diabetes prediction based on medication usage

According to Figure 2, the result of confusion matric, and model summary, we could see that for samples predicted to be diabetic, only about 64% were truly diabetic. For samples expected to be non-diabetic, about 97% were truly non-diabetic. And the recall of the model is 0.35. This indicates that the model's accuracy is relatively low, especially for the prediction of diabetics. Lastly, the accuracy of the LR model is 0.96, which suggests that the model performs better in general, but there is space for improvement in the identification of diabetic patients.

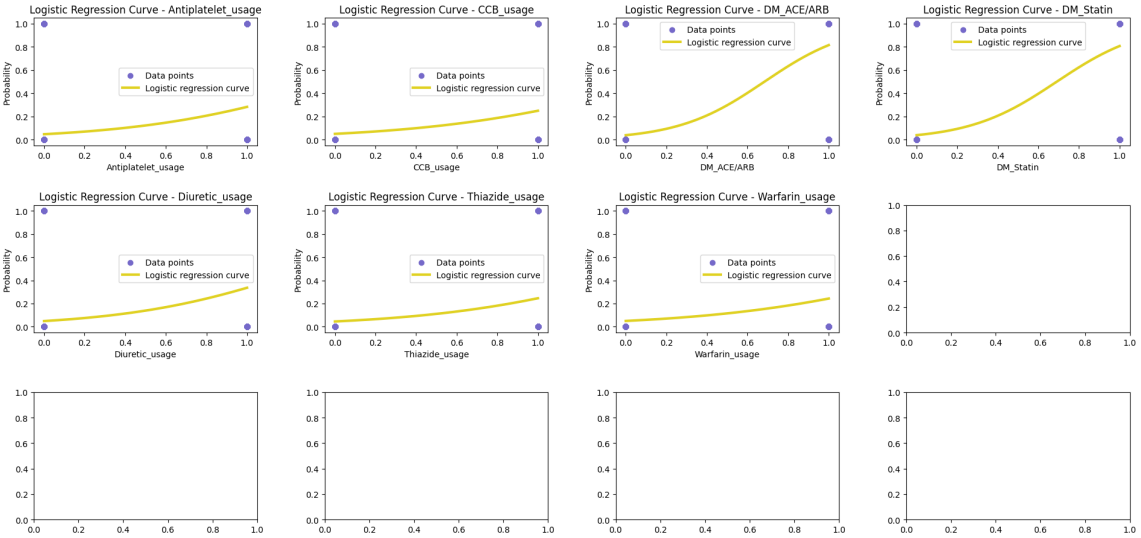

Figure 2

- Future Planning

For the future direction of project development, we plan to add or restrict some variables in the Linear Regression Model for CVD risk to reduce the residual sum of squares and correct the pattern shown in the residual plot. And we could find an appropriate method to deal with outliers in the dataset, including filtering and dropping methods. Besides, we could also tune regularisation parameters in logistic regression for diabetes prediction.

Result:

For CVD risk prediction, we could easily conclude that these factors included in the linear regression model significantly impact CVD risk, and we are able to make an accurate prediction of CVD risk based on these predictive variables of patients.

For diabetes prediction, the result of the model suggests that all factors are associated with the risk of diabetes, and there exists a significant positive correlation between the usage of statin medications and ACE or ARB and the risk of diabetes.

Our work shows that we could predict the risk of CVD and diabetes with some accuracy based on our input variables. However, further work is required to fine-tune our model to obtain a more significant accuracy that would be clinically relevant.

Individual Contribution:

Hongyan Chen and Shuyang Wu all contributed to the project design. All members equally contributed to coding and writing the report.

Code Link:

hongyan627/5230-project1 (github.com)

Reference:

1) World Health Organization. (2023, April 5). Diabetes. https://www.who.int/news-room/fact-sheets/detail/diabetes
2) Roglic, Gojka. WHO Global report on diabetes: A summary. International Journal of Noncommunicable Diseases 1(1):p 3-8, Apr–Jun 2016. | DOI:10.4103/2468-8827.184853
3) Sanz, J.A.; Galar, M.; Jurio, A.; Brugos, A.; Pagola, M.; Bustince, H. Medical diagnosis of cardiovascular diseases using an interval-valued fuzzy rule-based classification system. *Appl. Soft Comput.* **2014**, *20*, 103–111.
4) Sun Y, Zhang D. Machine learning techniques for screening and diagnosis of diabetes: a survey[J]. Tehnički vjesnik, 2019, 26(3): 872-880.
5) Vijiyarani, S.; Sudha, S. Disease Prediction in Data Mining Technique—A Survey. *Int. J. Comput. Appl. Inf. Technol.* **2013**, *2*, 17–21
6) Nabel E G. Cardiovascular disease[J]. New England Journal of Medicine, 2003, 349(1): 60-72.
7) Anderson K M, Odell P M, Wilson P W F, et al. Cardiovascular disease risk profiles[J]. American heart journal, 1991, 121(1): 293-298.
8) Pai P Y, Muo C H, Sung F C, et al. Angiotensin receptor blockers (ARB) outperform angiotensin-converting enzyme (ACE) inhibitors on ischemic stroke prevention in patients with hypertension and diabetes—a real-world population study in Taiwan[J]. International Journal of Cardiology, 2016, 215: 114-119.
9) Wayne A. Larsen & Susan J. McCleary (1972) The Use of Partial Residual Plots in Regression Analysis, Technometrics, 14:3, 781-790, DOI: 10.1080/00401706.1972.10488966
10) Lloyd-Jones DM, Leip EP, Larson MG, D'Agostino RB, Beiser A, Wilson PW, Wolf PA, Levy D. Prediction of lifetime risk for cardiovascular disease by risk factor burden at 50 years of age. **Circulation**. 2006; 113: 791–798.