Prediction of Cardiovascular Disease and Stroke

Arisara (Pie) Vichitchoti, Jiayue(Amaris) Han, Yitian(Aaron) Liang, Zhaolin(Helen) Li
Data Science 3000
Professor Sophine Clachar

Abstract

Our project explored two datasets, the cardiovascular and the stroke datasets, to predict the risk of cardiovascular and stroke probabilities based on selected important risk factors. Age was a salient risk factor for cardiovascular, more specifically, stroke. We also targeted other relevant risk factors (i.e., hypertension, glucose level, gender) to predict the probability of the two diseases. Based on model trainings (i.e., RandomForest Regression/Classifier, Support Vector Machine, K-fold Cross-Validation), we found that the highest model accuracy in predicting cardiovascular disease was approximately 72%, and 76% for the stroke prediction. Our models showcased the corresponding patients' detailed information regarding age, gender, height, weight, other health information, and cardiovascular disease or stroke prevalence. **We built disease prediction models with these critical health factors with the prospect of facilitating health professionals in the field of medicine.**

Introduction

Problem:

We explored risk factors that influenced the prevalence of Cardiovascular Diseases (CVDs) and Stroke. Since Stroke is considered a type of cardiovascular disease, we investigated the similarities and differences of risk factors for CVDs and Stroke.

Motivation:

We aimed to analyze health data and create machine learning models that could help generalize the pattern of risk factors for CVDs, and, more specifically, Stroke. We chose CVDs and Stroke datasets for this project because, among all the diseases, CVD is the top health burden across the world. 393.11 million of the population suffered from CVDs in 2019 by living with a disability or even death [5]. Stroke, damage to the brain from interruption of its blood supply, is one CVD that is the third leading cause of death in the United States [6].

Objective:

We used the Cardiovascular Dataset and the Stroke Prediction Dataset (https://www.kaggle.com/sulianova/cardiovascular-disease-dataset & https://www.kaggle.com/fedesoriano/stroke-prediction-dataset) to predict CVDs and Stroke with selected risk factors. We focused on blood pressures, cholesterol levels, and a few other risk factors from the datasets to analyze their correlations with CVDs and Stroke. By using machine learning models, we wanted to predict if a patient will have a CVD/Stroke based on their high systolic blood pressure, high cholesterol, and glucose in their bodies.

Methodology

EDA

We checked the shape of the dataset, which contained 7000 rows and 13 columns. We then fixed the age column by dividing each age value by 365. We ensured that our systolic and diastolic blood pressure values remained positive by making them absolute numbers. We also created a new dataframe column to show the relationship between blood pressure and cardiovascular disease by making a function to put the blood pressure into categories. In addition, we visualized the relationship between blood pressure categories and the number of patients. Another visualization we did was to find the correlation between age, gender, weight, height, and cardiovascular disease. As for the Stroke dataset, we found the data to contain 5110 rows and 12 columns. The visualization for our EDA shows the correlation between gender and the presence of stroke.

Model training and evaluation

We used the **Random Forest Classifier (RFC)** to obtain the prediction accuracy on the presence of CVD: 1 stood for having CVD, and 0 stood for not having CVD. The precision and recall values produced by the classification report helped us evaluate prediction quality. We also used the **Random Forest Regression (RFR)** to predict the probability of having CVDs. For our **K-fold cross-validation**, we used this model to double-check the Random Forest Classifier model accuracy and found out that the accuracy score is the same for both the K-fold cross and random forest classifier. As for the Stroke dataset, all input values for each model were age, hypertension, heart_disease, and avg_glucose_level. Due to the similarity of variables and patterns between CVD and the Stroke datasets, we applied the same model to the CVD dataset. We applied an additional **Support Vector Machines model using GridSearchCV** and **hyper-parameter tuning** to increase the model accuracy. By tuning C, gamma, and kernel from a range of data, we filtered out the best set of hyperparameters to further increase model accuracy while avoiding overfitting and underfitting.

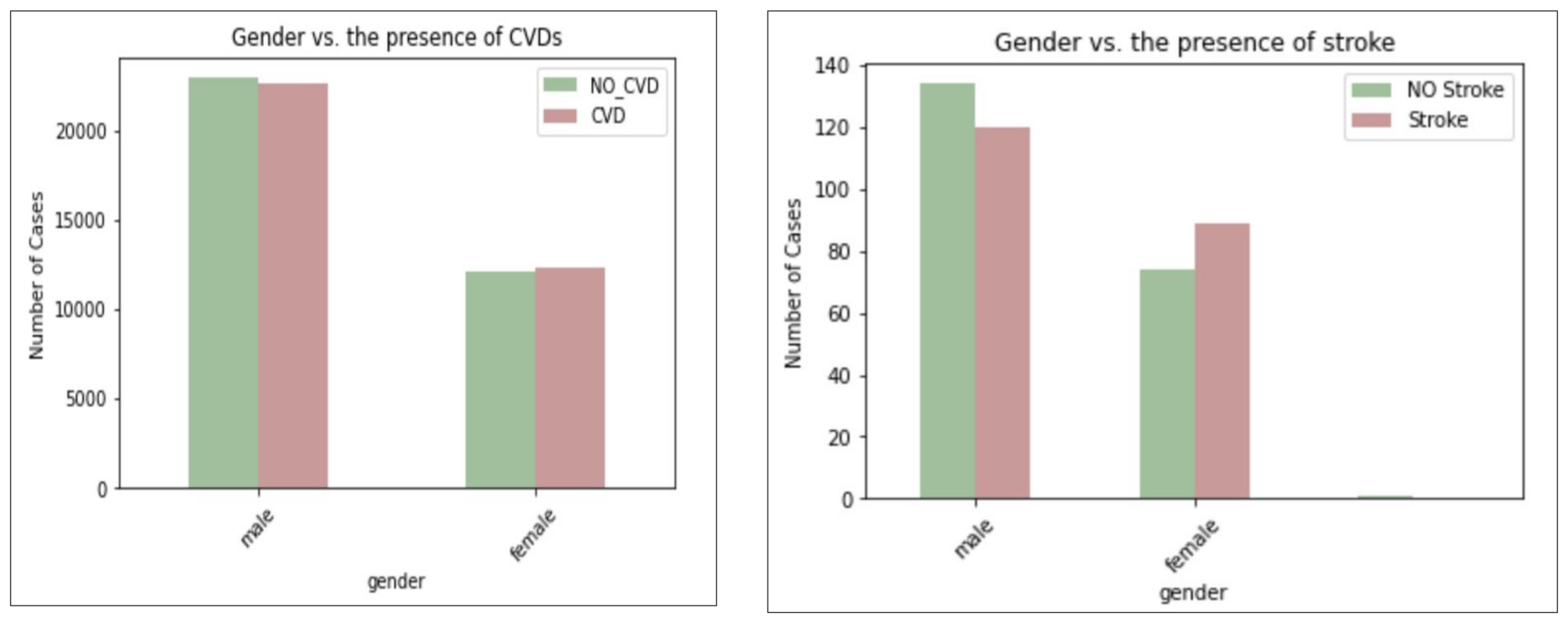


Figure 1A and 1B. There is no salient difference between genders regarding the risk of CVDs. However, females are more likely to have a stroke than their male counterparts.

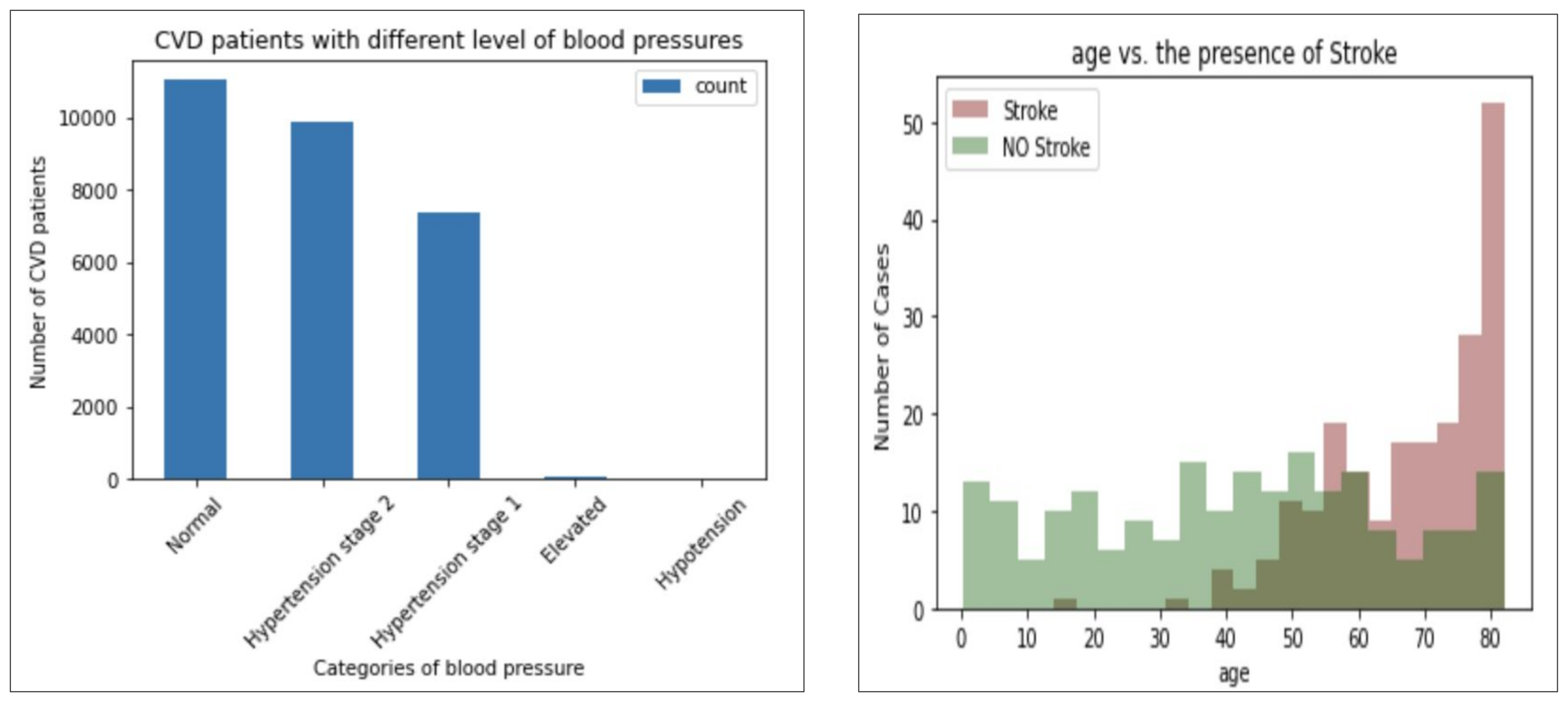


Figure 2A. Individuals with hypertension are more likely to have CVDs. Most stroke patients have normal blood pressure, suggesting that other risk factors are more influential in having CVDs.
Figure 2B. Stroke cases increase as age increases above approximately 50 years old.

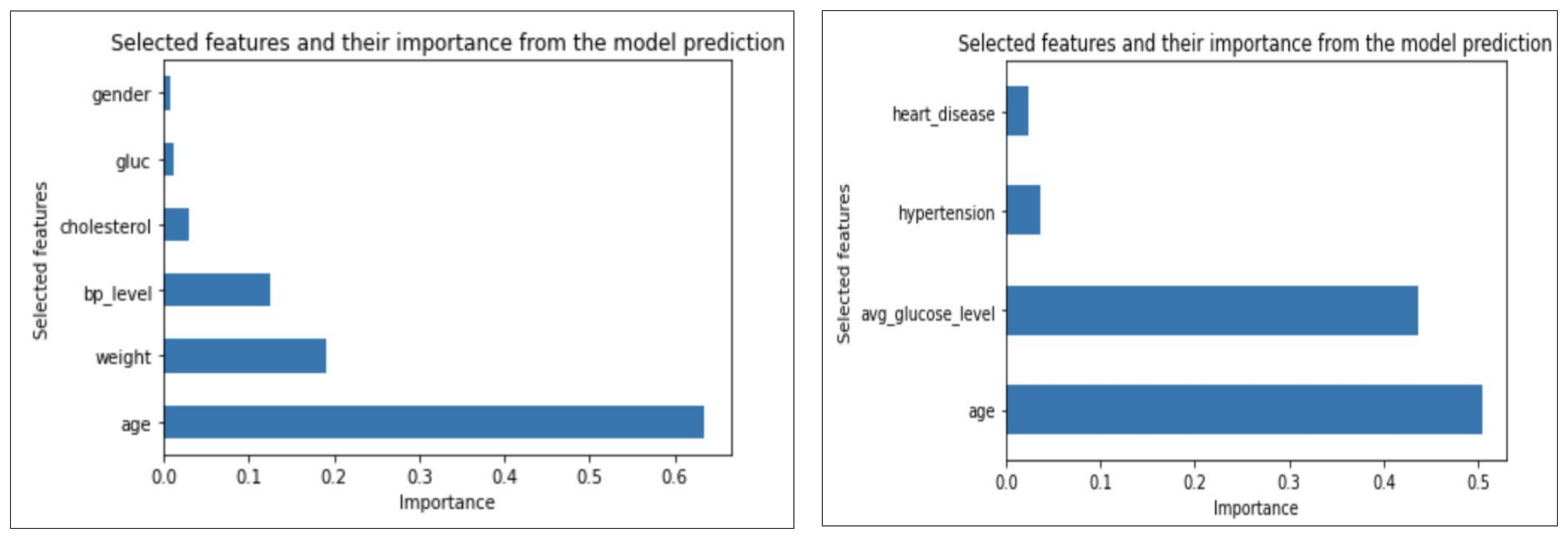
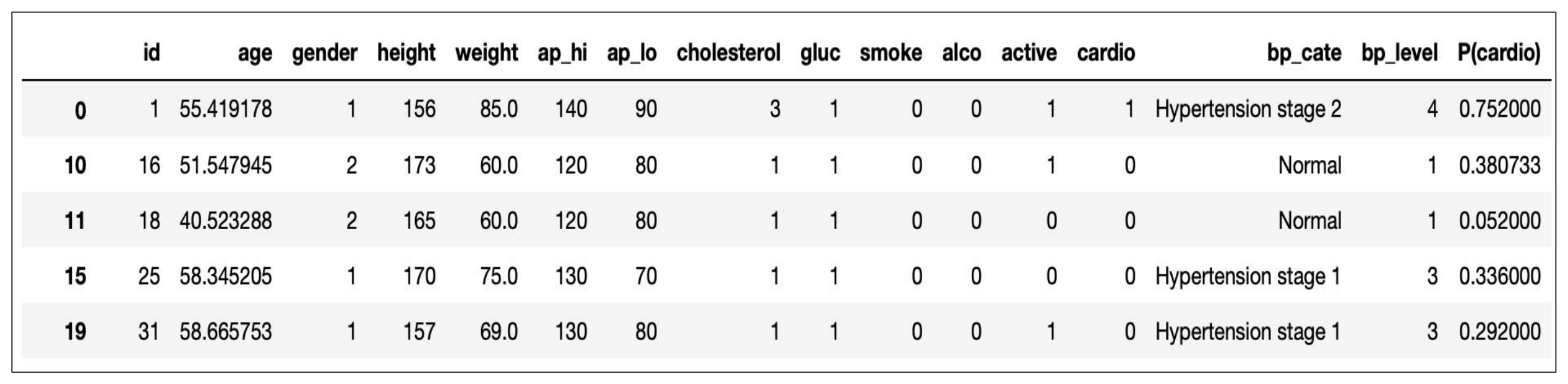


Figure 3A and 3B. Via the performance of RandomForestClassifier, age is the most important feature in CVDs prediction. As for the stroke prediction, age and avg_glucose_level are both important features.



	id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio	bp_cate	bp_level	P(cardio)
0	1	55.419178	1	156	85.0	140	90	3	1	0	0	1	1	Hypertension stage 2	4	0.752000
10	16	51.547945	2	173	60.0	120	80	1	1	0	0	1	0	Normal	1	0.380733
11	18	40.523288	2	165	60.0	120	80	1	1	0	0	0	0	Normal	1	0.052000
15	25	58.345205	1	170	75.0	130	70	1	1	0	0	0	0	Hypertension stage 1	3	0.336000
19	31	58.665753	1	157	69.0	130	80	1	1	0	0	1	0	Hypertension stage 1	3	0.292000

Table 1. This table showed a portion of a record with the probability of having CVDs. It also displayed individuals' ID numbers, age, gender, and other health-related information, facilitating professionals to understand the risk factors and corresponding CVDs' prediction.

Related Work

Stroke: Online Calculator can Predict your Risk
https://newsroom.uvahealth.com/2020/08/13/stroke-online-calculator-can-predict-your-risk/#:~:text=Doctors%20can%20predict%20patients'%20risk,waist%2C%20a%20new%20study%20finds.

The article introduces a computer model designed by medical professionals and data scientists. This study found that stroke risk increased consistently with metabolic syndrome severity even in patients without diabetes. We think it is related to the bigger picture of our project. The calculator introduced by this article will calculate individuals' chance of having heart disease, which gave us insights to think about how data science could help analyze and predict health outcomes.

Results and Evaluation

After cleaning the invalid data for the prediction model of the CVD dataset, our group used RandomForestRegression, RandomForestClassifier, and K-fold Cross-Validation to test the model based on the CVD dataset. The RandomForest Classifier model's accuracy in predicting the presence of cardiac was approximately **65%**. The f-1 score for both CVD and no CVD was 0.65, meaning that our group's model had a low rate of false predictions. According to the important-feature analysis, age, weight, and bp_level were the three most influential factors of CVD prediction. The K-fold Cross-Validation model had a 0.65 accuracy with a standard deviation of 0.01, indicating that the fluctuation of model accuracy was slight and the model fit relatively well.

As for the Stroke dataset, after cleaning the invalid data, we tested the model with RandomForestClassifier, RandomForestRegression, K-fold Cross-Validation, and SVM model. The RandomForestClassifier model's accuracy in predicting the presence of strokes was approximate 72%. The f-1 score for both CVD and no CVD was around 0.72, indicating high precision and recall from the model. The test data for RandomForestRegression showed that the mean squared error of this model was 0.183, which was relatively low because it means that the model regression line was close to the data points of the test. Then we continued to test the model with the SVM model; the f-1 score for both CVD and no CVD improved to 0.76. Stroke prediction based on the classification report also improved. After a few hyperparameter tweaks, we found that the best mean test score was 0.7536, and the best mean test score was 0.7884. These results, albeit not perfect, demonstrated a moderate to high SVM model prediction of the existence of strokes. Furthermore, since the difference between the best mean and best test scores was minor, the model was unlikely to be underfitting or overfitting. After hyperparameter tunning, the accuracy climbed to **77.5%**, up from 76.1% before.

Impacts

Our team looked at the relationship between risk variables and illness in two data sets. The occurrence of cardiovascular disease may be influenced by age and body weight, whereas stroke may be influenced by age. As the aging population is growing, our methodology has the potential to benefit a large population in need. Cardiovascular diseases, especially stroke, are severe and widespread issues that society must address well. Our approach has a high degree of accuracy in predicting the presence of cardiovascular disease with potential risk factors. **We built clinical models that can assist in identifying high-risk individuals and enhance preventive care and treatment efforts by identifying the traits most related to these diseases**

Conclusions

Our project explored the preprocessed Cardiovascular Dataset and the Stroke Prediction Dataset to predict the presence of CVDs and Stroke using the health index of patients (i.e. gender, age, height, weight, hypertension, cholesterol level, bmi, and habits). The project was completed under three phases: data sources searching, datasets EDA, and prediction models building. Under the EDA, both datasets were cleaned by dropping the N/A rows, converting to reasonable data types, and categorizing cluttered data. In addition, we explored the correlation between risk factors and the presence of diseases in both datasets. The graphs indicated that age and weight could affect the presence of CVDs, while the age could affect the presence of stroke. Phase 2 provides us a direction of selecting valid risk factors for the predicting models.

Next, we built three different predicting models for each dataset. For the CVDs prediction, we built RFC, RFR, and K-fold cross validation using the selected factors (i.e. age, gender, glucose, bp level, weight and cholesterol). The accuracy rate of RFC is 65%, the mean squared error of RFR is 0.226, and the accuracy rate of K-fold is 65%. For the undersampled stroke prediction, we built RFC, RFR, K-fold cross validation, and SVM using the selected factors (i.e. age, hypertension, heart disease, and glucose). The accuracy rate of RFC is 72%, the mean squared error of RFR is 0.18, the accuracy rate of K-fold is 70%, and the accuracy rate of the SVM after applying the best hyperparameter is 77.5%.

Overall, we have reached our goals of using selected risk factors to predict the CVDs and stroke. **The most accurate model for the stroke dataset is the SVM, while there is no most accurate model for the CVDs dataset.** The accuracy of the RFC and the K-fold model is the same for the CVDs dataset. Although there is space to optimize the accuracy rates of the models, our project constructed clinical models that could help identify high-risk populations and improve prevention treatment strategies.

In the future work, we could construct the SVM model for the CVDs dataset to improve the accuracy. In addition, since the population of both the datasets are unknown, we could collect the data from local hospitals, such as Mass General Hospital, to generate more accurate predictions. Overall, we hope that the result of this project could more accurately predict the presence of CVDs and strokes based on the selected health index and provide tools for future researchers and clinical workers to identify high-risk patients.

Contact

Northeastern University
Boston, MA, 02115

References

1. Stroke: Online calculator can predict your risk. UVA Health Newsroom. (2020, August 13). Retrieved April 25, 2022, from https://newsroom.uvahealth.com/2020/08/13/stroke-online-calculator-can-predict-your-risk/#:~:text=Doctors%20can%20predict%20patients'%20risk,waist%2C%20a%20new%20study%20finds.
2. Tazin, T., Alam, M. N., Dola, N. N., Bari, M. S., Bourouis, S., & Monirujjaman Khan, M. (2021). Stroke Disease Detection and Prediction Using Robust Learning Approaches. Journal of Healthcare Engineering, 2021.
3. Amini, L., Azarpazhouh, R., Farzadfar, M. T., Mousavi, S. A., Jazaieri, F., Khorvash, F., Norouzi, R., & Toghianfar, N. (2013, May). Prediction and control of stroke by Data Mining. International journal of preventive medicine. Retrieved February 26, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3678226/
4. How AI is able to predict and detect a stroke. ReferralMD. (2021, May 12). Retrieved February 26, 2022, from https://getreferralmd.com/2019/10/how-ai-is-able-to-predict-and-detect-a-stroke/#:~:text=The%20importance%20of%20predicting%20an%20Ischemic%20stroke%20early.&text=treatment%2C%20found%3A,instead%20of%20to%20an%20institution
5. World Health Organization. (n.d.). Cardiovascular diseases. World Health Organization. Retrieved February 26, 2022, from https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1
6. U.S. Department of Health and Human Services. (n.d.). Cardiovascular disease and stroke - climate and human health. National Institute of Environmental Health Sciences. Retrieved February 27, 2022, fromhttps://www.niehs.nih.gov/research/programs/climatechange/health_impacts/cardiovascular_diseases/index.cfm
7. Roser, M., & Ritchie, H. (2021, September 25). Burden of disease. Our World in Data. Retrieved February 27, 2022, from https://ourworldindata.org/burden-of-disease
8. Ulianova, S. (2019, January 20). Cardiovascular disease dataset. Kaggle. Retrieved April 25, 2022, from https://www.kaggle.com/sulianova/cardiovascular-disease-dataset
9. Fedesoriano. (2021, January 26). Stroke prediction dataset. Kaggle. Retrieved April 25, 2022, from https://www.kaggle.com/fedesoriano/stroke-prediction-dataset