Hypertension Classification Based on NHANES

SI 670 Team 18: Xinye Jiang, Chen Xie, Xun Wang

1. Introduction

Hypertension (also called high blood pressure) is a common disease with the growth of age, but it could also be caused by unhealthy dietary habits, obesity, smoking, high daily alcohol consumption or high pressure brought by a fast-paced lifestyle nowadays. Long-term hypertension could damage blood vessels, thus raise the risk of stroke, cardiovascular disease and a series of complications. According to research, around 60 million Americans and 1.13 billion individuals worldwide are bothered by hypertension. Despite the importance of high blood pressure, people have a poor understanding of its pathogenesis and symptoms. Then the study of hypertension can raise people's prevention awareness of this disease and help people make proper precaution strategies for future uncertainty.

The general goal of our project is to predict whether a person is a hypertensive patient or not according to physiological indices and to identify factors that are most relevant to hypertension. We intend to accomplish this task by performing machine learning classification methods such as Random Forests, Logistic Regression and KNN (K-Nearest Neighbors) on the NHANES dataset.

2. Methods

NHANES (National Health and Nutrition Examination Survey) is a recurring cross-sectional study that aims to reflect the health and nutritional status of the US population comprehensively. The survey includes (1) interviews of demographic, dietary, socioeconomic, and other health-related questions, (2) physical examinations consisting of medical, dental, and physical measurements, as well as (3) laboratory tests of a series of biochemical metrics. The dataset can be found on the website of Centers for Disease Control and Prevention (CDC) [1].

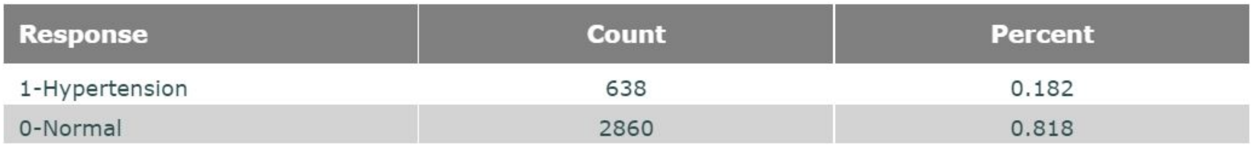
2.1 Data Preprocessing

Our project joins around 110 available variables from 18 datasets in the demographics, examination and laboratory data folders of NHANES 2015-2016 by their unique respondent identity. We limit our analysis to the non-pregnant respondents who are at least 20 years old. As missing data with unknown distribution exists in many variables and many observations, imputation might introduce invalid information into our dataset and thus worsen the prediction performance. Therefore, we simply drop the observations with missing values. In the next step, we split the data into training and test sets, use a min-max scaler to scale the numeric variables and one-hot encode the categorical variables in the training set. After applying the same transformation on the test set, we obtain a dataset with 3498 complete records and 186 explanatory variables. The percentages of training and test sets in the whole dataset are 75% and 25% respectively.

2.2 Lable Assignment

According to the conventional standard, we assume that people have hypertension if (1) they have average systolic blood pressure (SBP) >= 140 mm Hg or average diastolic blood pressure (DBP) >= 90 mm Hg or (2) they have one-time SBP >= 160 mm Hg or one-time DBP >= 100 mm Hg. We assign label 1 to hypertension and label 0 to others based on 4 consecutive measurements of the blood pressure (BP) [2]. We could observe from Table 1 that our binary response is highly imbalanced that minority class (hypertension) only accounts for 18% of the whole data.

Table 1. Response



Response	Count	Percent
1-Hypertension	638	0.182
0-Normal	2860	0.818

2.3 Parameter Tuning

We plan to use complex classification methods, like SVM (Support Vector Machine), KNN (K-Nearest Neighbors), Random Forests, Gradient Boosting, and also classifiers that could generate linear decision boundaries such as Logistic Regression and Ridge to make predictions about the presence of hypertension in individuals. The prediction performance of different models could help us to understand the structure of the data.

We select parameters for each model separately by applying 5-fold cross-validation on the training data. As our label is imbalanced, a macro score that weights more towards the minor class is applied. To lay more emphasis on recall without fully ignoring precision, we decide to tune parameters by a recall-oriented F_β score,

$$F_\beta = \frac{(1+\beta^2) \cdot Precision \cdot Recall}{(\beta^2 \cdot Precison) + Recall}, \beta = 2$$

rather than accuracy. The reason why we prefer F_β score to accuracy is that we care more about the misclassification of hypertensives than the false classification of non-hypertensives, i.e., we want hypertension to be detected as much as possible in this case. To be precise, the recall rate for hypertensives is defined as the number of the identified hypertensives over the total number of hypertensives and thus the emphasis on the recall can help control the percentage of undiagnosed hypertensives. Another reason is that due to the imbalance of response, the baseline dummy classifier could already reach an accuracy rate of over 0.82 on the test set, which is almost the best accuracy rate that other classifiers can get. So accuracy could not display the advantages of other classifiers well.

In general, classifiers actually output predictions based on estimated probabilities for individuals to be considered as hypertensives. The estimated probabilities can also be viewed as the risk of developing hypertension in the future. A probability value higher than 0.5 indicates hypertension and a lower one shows non-hypertension. In the next stage, we plan to modify the original classification threshold (also called decision boundary) 0.5 to identify more hypertensives.

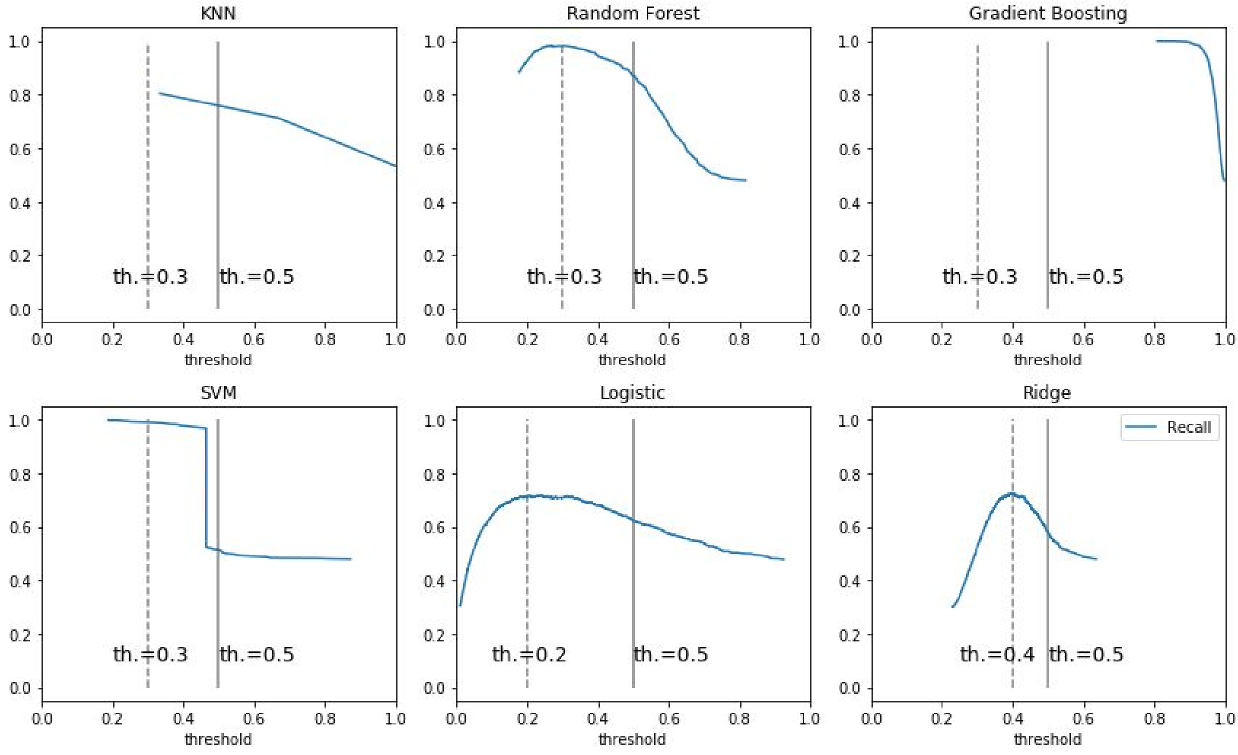
3. Evaluation and Analysis

3.1 Threshold Adjustment

In theory, when we lower the decision threshold, more persons will be classified as hypertensive patients, thus recall tends to increase while precision will decrease due to their trade-off relationship. Here we are trying to find the thresholds that optimize the macro F_β score for each model in Figure 1.

Figure 1 below shows the plots of the macro F_β score against thresholds for each of the six classifiers. It can be observed that Random Forests, Logistic and Ridge reach the highest macro F_β score at points 0.3, 0.2 and 0.4 respectively. While KNN, Gradient Boosting, SVM do not display an obvious pattern for us to choose the optimal thresholds. As the adjusted thresholds should not be too high or too low, we assign these classifiers the same threshold 0.3 as the one of Random Forests.

Figure 1. Fbeta score(beta=2)-Threshold



3.2 Evaluation

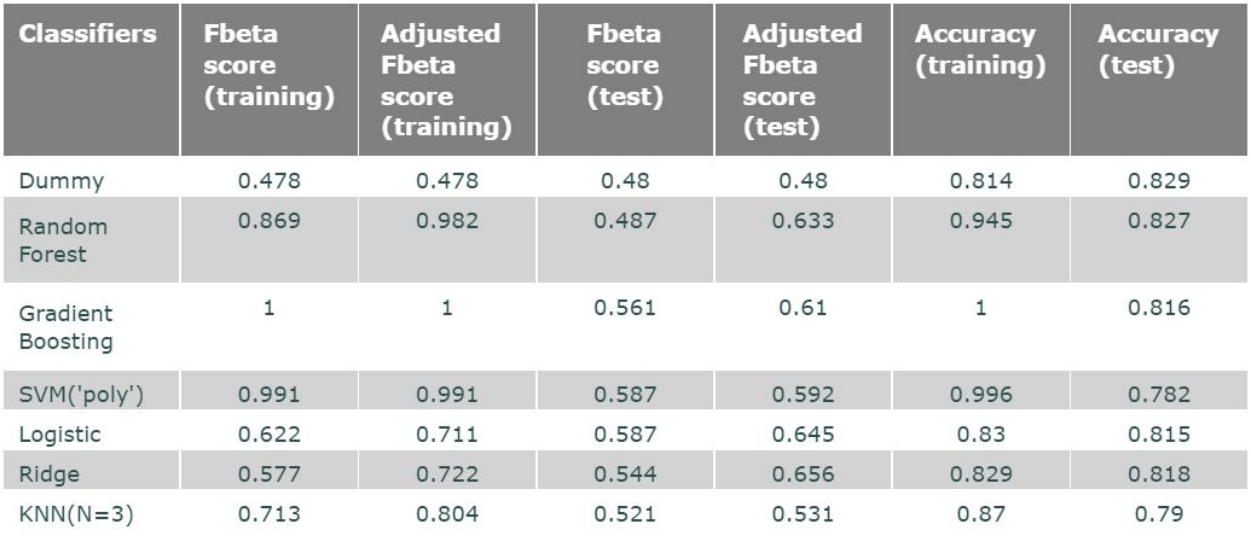
Applying the fitted models to the test dataset with and without the adjusted thresholds, we obtain the predicted probabilities of having hypertension for individuals and then we calculate the corresponding macro F_β scores before and after threshold adjustment for each classifier.

Comparing the scores on the training and test sets listed in Table 2, we see that those complex algorithms, i.e. Random Forests, SVM with polynomial kernel and especially Gradient Boosting, tend to generate very overfitting results that their macro F_β scores on the training set are much higher than those on the test set. The training macro F_β score of Gradient Boosting even reaches 1, while the test score is not that good. This is possibly due to the fact that our dataset is high dimensional and complicated methods tend to give high-variance and low-bias estimates which have merely average performance on new data.

In comparison, Logistic and Ridge have a tuning parameter that controls the regularization term of the estimators, making them grasp the general structure of the data and outperform other classifiers on the unseen dataset. Finally, Ridge and Logistic classifiers that generate linear boundaries have the best performances among all the algorithms, reaching 0.656 and 0.645 macro F_β scores.

KNN exhibits a relatively poor performance compared with other models. It is because KNN is established on the computation of distance and therefore is not that appropriate for datasets including categorical variables. KNN algorithm also has difficulties in handling high-dimensional data, like the dataset in this task.

Table 2. Fbeta score and accuracy



Classifiers	Fbeta score (training)	Adjusted Fbeta score (training)	Fbeta score (test)	Adjusted Fbeta score (test)	Accuracy (training)	Accuracy (test)
Dummy	0.478	0.478	0.48	0.48	0.814	0.829
Random Forest	0.869	0.982	0.487	0.633	0.945	0.827
Gradient Boosting	1	1	0.561	0.61	1	0.816
SVM('poly')	0.991	0.991	0.587	0.592	0.996	0.782
Logistic	0.622	0.711	0.587	0.645	0.83	0.815
Ridge	0.577	0.722	0.544	0.656	0.829	0.818
KNN(N=3)	0.713	0.804	0.521	0.531	0.87	0.79

In general, macro F_β scores of our selected classifiers all perform better than the baseline dummy classifier with the most frequent strategy. Adjusted scores improve greatly for most classifiers on both the training and test datasets, proving that the adjustment of thresholds is effective for this recall-oriented task.

4. Related Work

A volume of literature applies machine learning techniques to predict hypertension based on clinical data. In these related work, the most common factors used in the prediction models are age, sex, glucose status, smoking, overweight, lack of physical activity, salt intake, stress, family history, etc., and we could extract most of the related variables in NHANES datasets [4]. Other similar works also utilize NHANES data and supervised machine learning models to identify patients with cardiovascular disease or diabetes. The labels of such diseases are also assigned according to questionnaire answers and laboratory results [5].

The threshold adjustment method in our project is similar to the alteration of the decision boundary in one of the related work [6]. This paper focuses on predicting diabetes utilizing an ensemble model that combines Logistic Regression, KNN, Random Forests, Gradient Boosting and SVM together. The decision boundary is altered to achieve a better recall rate for diabetics by sacrificing some recall rate of the non-diabetics. Biological-related and medical-related classification problems that pay more attention to recall may use this decision threshold adjustment technique, for example, in cancer prediction [7].

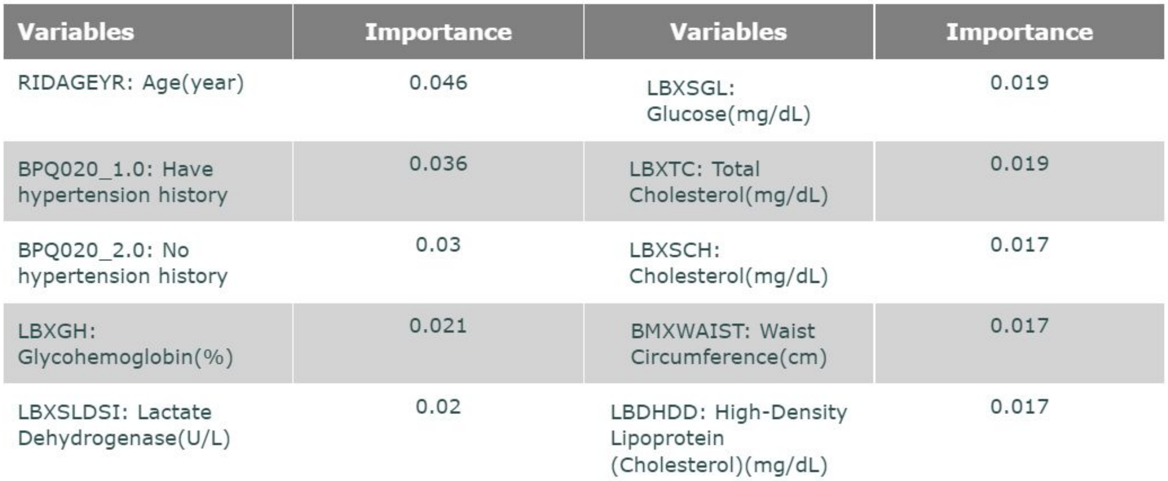
5. Discussion and Conclusion

As the dataset is imbalanced and we want to identify more hypertensive patients, a recall-oriented macro F_β score rather than accuracy is applied as a metric to evaluate the model. We also adjust the classifier threshold to further help detect more hypertension cases. After comparison, the threshold adjustment turns out to work well for our target.

Using the recall-oriented macro F_β score and adjusted threshold, Ridge and Logistic classifiers that generate linear boundaries and naturally apply regularizations show the best performances among all the algorithms. Complex models such as Random Forests, SVM with polynomial kernel and Gradient Boosting, tend to overfit the training set and obtain mediocre macro F_β scores on the test set. KNN has poor performance on this dataset as it has difficulties in dealing with categorical variables and high-dimensional data.

We also use Random Forests to compute the feature importance and show the 10 most important variables that associate with hypertension in Table 3. Related variables are age, hypertension history, cholesterol, basic body measures and diabetes-related factors including blood glucose and glycohemoglobin. Age plays an extremely important role in predicting hypertensives. As the rise in blood pressure with age may induce many diseases, elderly people should take extra care of their health status. Another point worth mentioning is that three of these factors are relevant to cholesterol, indicating that cholesterol is highly linked with high blood pressure. People should pay much attention to these indices in their daily life to avoid developing hypertension.

Table 3. Top 10 Important Features in Random Forest



Variables	Importance	Variables	Importance
RIDAGEYR: Age(year)	0.046	LBXSGL: Glucose(mg/dL)	0.019
BPQ020_1.0: Have hypertension history	0.036	LBXTC: Total Cholesterol(mg/dL)	0.019
BPQ020_2.0: No hypertension history	0.03	LBXSCH: Cholesterol(mg/dL)	0.017
LBXGH: Glycohemoglobin(%)	0.021	BMXWAIST: Waist Circumference(cm)	0.017
LBXSLDSI: Lactate Dehydrogenase(U/L)	0.02	LBDHDD: High-Density Lipoprotein (Cholesterol)(mg/dL)	0.017

For future work, we might try to include more related laboratory factors like those from urine routine test to enhance the prediction performance. Another proposal is to add data from other years to enrich the dataset to compare the temporal trend of the percentage of hypertensives in the population. We also would like to try deep learning techniques such as Artificial Neural Network (ANN) model [8] or ensemble models [6] or add regularization terms to classifiers to see if they can produce better performances.

6. References

[1] Data Source (NHANES DATA) https://wwwn.cdc.gov/nchs/nhanes/ContinuousNhanes/Default.aspx?BeginYear=2015

[2] Ciemins EL, Ritchey MD, Joshi VV, Loustalot F, Hannan J, Cuddeback JK. Application of a Tool to Identify Undiagnosed Hypertension, United States, 2016. MMWR Morb Mortal Wkly Rep 2018; 67: 798–802.

[3] F. Pedregosa, G.Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.

[4] D. LaFreniere, F. Zulkernine, D. Barber and K. Martin. Using machine learning to predict hypertension from a clinical dataset. 2016 IEEE Symposium Series on Computational Intelligence (SSCI), Athens, 2016, pp. 1-7.

[5] A. Dinh, S. Miertschin, A.Young. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. BMC Med Inform Decis Mak 19, 211, 2019.

[6] J. Semerdjian, S. Frank. An Ensemble Classifier for Predicting the Onset of Type II Diabetes. ArXiv e-prints, 2017.1708.07480.

[7] J.J. Chen, C.Tsai, H. Moon, H. Ahn, J.J. Young, C. Chen. The Use of Decision Threshold Adjustment in Classification for Cancer Prediction, 2005.

[8] S. Sakr, R. Elshawi, A. Ahmed, WT. Qureshi, C. Brawner, S. Keteyian. Using machine learning on cardiorespiratory fitness data for predicting hypertension. The Henry Ford Exercise Testing (FIT) Project. PLoS ONE 13(4), 2018.