Predicting Heart Disease Using the BRFSS Dataset

Aslı Yıldırım

Akdeniz University

201908080039@ogr.akdeniz.edu.tr

***Abstract*— This study analyzes the Behavioral Risk Factor Surveillance System (BRFSS) 2015 dataset to develop a predictive model for heart disease. The analysis includes data exploration, addressing missing values, handling class imbalance, and building classification models to assess the most significant factors in heart disease prediction. Several models, including Logistic Regression, Random Forest, and Support Vector Machines (SVM) etc., were trained and evaluated.**

***Keywords*— Heart Disease Prediction, BRFSS, Classification, Machine Learning, Public Health, Risk Factors, Class Imbalance.**

I. Introduction

The BRFSS 2015 dataset provides a set of health-related variables that can be used to predict the circumstances of heart disease. This report focuses on looking at the dataset to understand and improve machine learning models' ability to accurately predict heart disease and evaluate their performance based on key metrics like accuracy, precision, recall, etc.

II. Data Exploration and Preprocessing

A. Dataset Overview

The dataset contains 253680 instances and 22 features. The target column (HeartDiseaseorAttack) shows whether or not an individual has heart disease. The key features include blood pressure, cholesterol, and lifestyle factors. The dataset doesn't include any missing data.

B. Exploratory Data Analysis (EDA)

The correlation analysis reveals that GenHlth, Stroke, HighBP, and Age show the strongest positive correlations with HeartDiseaseorAttack. Moderate associations are observed for Diabetes and PhysHlth.

In contrast, Fruits and Veggies shows negligible correlations, indicating minimal direct predictive power in this dataset. But when looking at the important predictors of the model this would be change.

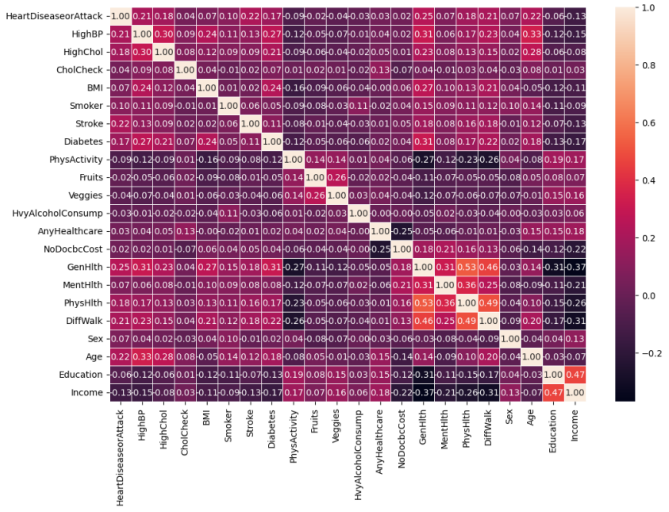

Fig. 1 Feature Correlation Heatmap for Heart Disease Prediction

Some numerical columns, like BMI, has some values that are unusual. Their portion amount is big so just removing these unusual numbers, which might hide important information.

The dataset shows the model is not very significant to predict the minority class because of the imbalance, with 9.41% of belong a diagnosis of heart disease (positive class), while the remaining 90.58% belong to the non-heart disease group (negative class).

```
HeartDiseaseorAttack
0.0    90.581441
1.0     9.418559
Name: proportion, dtype: float64
```

HeartDiseaseorAttack	count
0.0	229787
1.0	23893

Fig. 2 Feature Shows percentage and numbers for HeartDiseaseorAttack

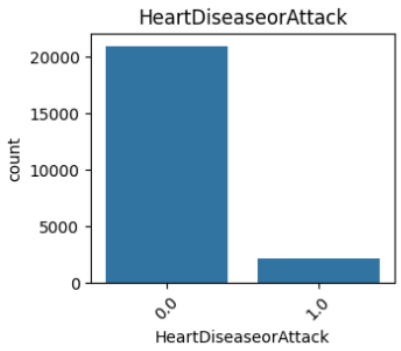

Fig. 3 Bar chart of HeartDiseaseorAttack

C. Data Preprocessing

Numerical features are standardized to ensure that the models are not biased by different feature scales. There are no missing values and all categorical columns were already numerical but because of this, the distinction between numerical and categorical columns was taken after looking at the unique numbers in the columns. In the resources, the unique number for that is generally either 10-12. So with that the data has three (BMI,MentHlth,PhysHlth) numerical columns

III. Model Development and Evaluation

A. Class Imbalance Handling

The potential impact of class imbalance on model performance would be that the model would not learn minority class or data very well but the model would be able to learn dominant class. This could yield wrong predictions because the goal normally would be finding the minority class or related features so wrong predictions about the minority could be more costly for fixing.

To handle the class imbalance in the dataset, both the Synthetic Minority Over-sampling Technique (SMOTE) and Random UnderSampling were used. This two has opposite way fixing that.

SMOTE increases the number of minority class samples by generating synthetic instances, which helps maintain all available data and improves minority class recognition. However, it may introduce noise by creating less realistic samples, especially when the feature space is sparse. On the other hand, Random UnderSampling reduces the number of majority class instances, leading to faster training times and simpler models, but there is has a cost for potentially discarding valuable information. Considering these trade-offs, both techniques were applied to the dataset and were chosen as the better option.

HeartDiseaseorAttack	count
0.0	229787
1.0	229787

dtype: int64

Fig. 4 SMOTE result

HeartDiseaseorAttack	count
0.0	23893
1.0	23893

dtype: int64

Fig. 5 Under-sampling result

B. Model Building

Four classification models were developed and evaluated for the heart disease prediction task:

Logistic Regression: Serves as a strong baseline model due to its simplicity, low computational cost, and interpretability in binary classification tasks.

Random Forest Classifier: A method that handles non-linear relationships effectively. It is less sensitive to outliers and missing values, and it provides measures of feature importance, helping interpretability.

Artificial Neural Network (ANN): Capable of learning complex and non-linear patterns in the data, especially beneficial when relationships among features are not easily captured by traditional models.

XGBoost: Implemented because It shows good performance on large data sets and complex relationships. I tried, XGBoost to for learning how accuracy will improve because XGBoost sequentially corrects errors from previous trees (unlike Random Forest's independent trees), which often leads to higher accuracy.

Normally, SVC added to this list but because calculation of model is complex takes to much time (with grid search even the fitting is not finishing) and even changing kernel to the simpler one (linear) was not effective so It is changed to ANN but for ANN, the chosen library needed to be Keras library instead of KerasClassifier because of library conflictions related to the colab.

Hyperparameters were tuned using GridSearchCV, RandomizedSearchCV, and Keras Tuner, and early stopping was applied to prevent overfitting.The parameters taken from library websites etc. and changed the gap or numbers considering to the results.

C. Model Evaluation

The results of SMOTE significantly better than Random UnderSampling for all models which used.

The best model is other than XGBoost is Random Forest. Random Forest outperformed Logistic Regression and ANN because it is robust to outliers, naturally prioritizes important features (e.g., GenHlth), and because the dataset size is not very big, it gives better results where ANN typically requires more data or features to be effective.

Logistic Regression

```
Accuracy: 0.7822
Precision: 0.7659
Recall: 0.8122
F1-Score: 0.7884
AUC-ROC: 0.8553

Classification Report:
              precision    recall  f1-score   support

         0.0       0.80      0.75      0.78     46000
         1.0       0.77      0.81      0.79     45915

    accuracy                           0.78     91915
   macro avg       0.78      0.78      0.78     91915
weighted avg       0.78      0.78      0.78     91915
```

ANN

```
Accuracy: 0.8888
Precision: 0.8770
Recall: 0.9042
F1-Score: 0.8904
AUC-ROC: 0.9607

Classification Report:
              precision    recall  f1-score   support

         0.0       0.90      0.87      0.89     46000
         1.0       0.88      0.90      0.89     45915

    accuracy                           0.89     91915
   macro avg       0.89      0.89      0.89     91915
weighted avg       0.89      0.89      0.89     91915
```

Random Forest

```
Accuracy: 0.8724
Precision: 0.8842
Recall: 0.8568
F1-Score: 0.8702
AUC-ROC: 0.9491

Classification Report:
              precision    recall  f1-score   support

         0.0       0.86      0.89      0.87     46000
         1.0       0.88      0.86      0.87     45915

    accuracy                           0.87     91915
   macro avg       0.87      0.87      0.87     91915
weighted avg       0.87      0.87      0.87     91915
```

XGBClassifier

```
Accuracy: 0.9444
Precision: 0.9863
Recall: 0.9011
F1-Score: 0.9418
AUC-ROC: 0.9828

Classification Report:
              precision    recall  f1-score   support

         0.0       0.91      0.99      0.95     46000
         1.0       0.99      0.90      0.94     45915

    accuracy                           0.94     91915
   macro avg       0.95      0.94      0.94     91915
weighted avg       0.95      0.94      0.94     91915
```

Fig. 4 Model's results (For SMOTE)

Performance Metrics

-Accuracy: 0.8888 → Model correctly predicts heart disease presence 88.9% of the time.

-Precision: 0.8770 → When predicting heart disease, 87.7% are true cases.

-Recall: 0.9042 → Captures 90.4% of actual heart disease cases.

-F1-Score: 0.8904 → Balanced measure of precision and recall.

-AUC-ROC: 0.9607 → Near-perfect score discrimination between classes.

Classification Report:

-Class 0 (No Heart Disease): High precision (0.90) → Few false positives.

-Class 1 (Heart Disease): High recall (0.90) → Few false negatives.

Precision-Recall Trade-off:

Recall (0.90): This means the model is good for minimizing false negatives (missed heart disease cases).

Precision (0.88): This means diagnosed cases are likely true positives, avoids unnecessary treatments.

Trade-off:

The model slightly favors recall over precision, meaning it prioritizes catching as many true heart disease cases as possible.

IV. Public Health Implications & Discussion

According to the Random Forest model, the most two important predictors of heart disease were:

1- HighBP (21.8%) – High blood pressure
2- HighChol (17.4%) – High cholesterol

The model achieved high accuracy and recall, which is good but because topic is related to the health, this is not the best case. And although HighBP and HighChol most significant in the model if we look at the correlation matrix that is slightly different from the model (the best was GenHlth there). This means that the relationships between variables and heart disease may be non-linear or influenced by complex relations, which are captured better by the model than by simple correlations.

Policy Recommendations:

This predictive modeling can assist healthcare providers in identifying high-risk patients and with that rationing resources more efficiently but which I mentioned because this kind of modeling is related to health, it needs more data, and with that It will get more accurate results for the bigger scale of people.

Looking at the results, early screening programs should target people with high blood pressure and high cholesterol with could considering age factor (age is not first but It is a significant predictor).

Limitations and future work:

Although handled with resampling techniques, some skewness in the target distribution could still affect generality which doesn't perfectly reflect real-world variability,

potentially affecting model robustness and the model was trained on the 2015 BRFSS dataset. Demographics, lifestyle patterns etc. may have shifted since then. Therefore, future validation is needed on more recent data to ensure model relevance with that deep learning (ann) could get better result and doing models like XGBoost etc. could get better results.

REFERENCES

[1] https://www.geeksforgeeks.org/smote-for-imbalanced-classification-with-python/

[2] https://www.youtube.com/watch?v=GR-OW5asKlk

[3] https://developers.google.com/machine-learning/crash-course/overfitting/imbalanced-datasets

[4] https://scikit-learn.org/stable/modules/grid_search.html

[5] https://stackoverflow.com/questions/19018333/gridsearchcv-on-logisticregression-in-scikit-learn

[6] https://www.btkakademi.gov.tr/portal/course/keras-ile-derin-ogrenme-algoritmalari-37613

[7] https://www.kaggle.com/code/sammihuang999/comparing-logistic-regression-and-catboost-model