Heart Disease Risk Identification via Supervised Machine Learning

Ian Sindelar

ABSTRACT

As the leading cause of death in the United States [1], heart disease is a salient issue to address in the field of disease prevention. Various data mining approaches have the potential to lend valuable insights into both identification of key risk factors as well as prediction of likelihood of disease development on an individual basis. However, previous machine learning approaches have tended to focus on the latter goal of prediction rather than the former goal of identification [1, 2]. Additionally, accuracy of prediction can often be inflated by training and testing prediction models using imbalanced datasets with low rates of occurrence. In this paper, a ranking of coronary heart disease (CHD) risk factors from 15 different health-related attributes was identified from a dataset consisting of patients. Subsequently, three different prediction models were compared in prediction accuracy. Each model was trained and tested on both imbalanced as well as balanced data, using F1 score to assess the suitability. Ultimately, it was found that age was the biggest contributor to ten-year CHD outcome of the 15 given attributes, and the highest prediction accuracy and F1 score was obtained from a multiple linear regression trained on balanced datasets. Regarding naïve Bayes models, the implementation of the complement naïve Bayes approach outperformed the standard Gaussian version, especially when dealing with imbalanced data.

1. INTRODUCTION

In 2020, coronary heart disease killed 696,962 people in the US alone, more than any other cause of death [3]. As with any disease, various attributes may contribute to an individual's risk factor. In the case of CHD, the Center for Disease Control (CDC) lists key risk factors to be "high blood pressure, high low-density lipoprotein (LDL) cholesterol, diabetes, smoking and secondhand smoke exposure, obesity, unhealthy diet, and physical inactivity" [4]. However, even given this knowledge, there lies further value in determining which of these factors carry more weight than the others, as well as assessing whether additional common risk factors contribute to risk of CHD even more so than those provided.

In 2021, the total amount of funding directed toward the various forms of heart disease was estimated to exceed $2 billion [5]. Given the vast sums involved with fighting this global health crisis, proper allocation of funds towards high-impact risk factors may be sensible even if those risk factors seem to be insurmountable at first glance. Accurate identification of the individual impact of each relevant risk factor may allow researchers to better consider the health implications of what may come from progress in combating those individual factors.

1.1 Related Work

Previous papers have approached CHD prediction via machine learning analyses.

R. Detrano et al. [6] examine the predictive effects of coronary calcium levels related to heart disease as analyzed by logistic regression. Over 6,000 samples were collected, including 17 additional risk factors. Coronary calcium levels were collected during the course of the study. In determining the coincidence of this single variable with CHD, the various other risk factors were controlled for as to not confound the results. It was found that there is indeed a relationship between the two, with CHD incidence increasing alongside coronary calcium levels. However, as this work was restricted to assess a single attribute, the simultaneous comparison of multiple attributes may lend significantly more information as to which attributes are the most valuable in predicting CHD.

In an approach designed to assess the coincidence of 13 different health-related attributes with CHD, S. Ashtekar at al. [1] utilize logistic regression hybridized with artificial neural networks to train a model in CHD likelihood prediction. This combination resulted in a notable improvement over either of the individual methods, approaching a 92.30% measure of precision when using the hybrid approach. The authors utilized various technologies in analyzing the data, including Jupyter Notebook, scikit-learn's Logistic Regression API, and Keras for Feed-Forward Neural Network modeling. The dataset included 14 total attributes (including CHD outcome) and over 300 samples. Though a high level of prediction precision was achieved, there may be value in utilizing alternative attributes as well as a larger sample size.

2. PROPOSED WORK

This paper will approach a set of 15 health-related attributes using decision tree induction, naïve Bayes, and linear regression in order to determine both the predictive power of these methods as well as which of these attributes have the greatest predictive power in determining likelihood of CHD.

2.1 Dataset and Tools

The dataset for this paper was sourced from Kaggle and tracks residents of Framingham, Massachusetts [7]. It is comprised of 4,238 samples with 15 attributes in addition to the measure of CHD outcome. Table 1 below describes the included attributes coupled with their descriptions.

Table 1. Dataset Attribute Descriptions

Col. No.	Attribute	Description
1	male	1 = male, 0 = female
2	age	Numeric age in years
3	education	Level of education 1-4, not specified
4	currentSmoker	1 = current smoker, 0 = not a current smoker
5	cigsPerDay	Number of cigarettes smoked per day on average
6	BPMeds	1 = using blood pressure meds, 0 = not using BP meds
7	prevalentStroke	1 = patient previously had a stroke, 0 = has not had a stroke
8	prevalentHyp	1 = patient was hypertensive, 0 = was not hypertensive
9	diabetes	1 = patient had diabetes, 0 = did not have diabetes
10	totChol	Total cholesterol level (mg/dL)
11	sysBP	Systolic blood pressure
12	diaBP	Diastolic blood pressure
13	BMI	Body mass index
14	heartRate	Heart rate (BPM)
15	glucose	Glucose level (mg/dL)
16	TenYearCHD	10 year risk of coronary heart disease, 1 = positive, 0 = negative

Relevant code and analysis utilized Python3 [8] in a Jupyter Notebook setting [9]. The Pandas [10] and NumPy [11] libraries were of particular assistance in this project. Extensive use was made of scikit-learn[12] in model training and evaluation. Equations incorporated into this paper were written using LaTeX.

2.2 Decision Tree Induction

Information gain will be the primary target in determining which attributes are most strongly associated with CHD outcome. Information gain works within decision trees to sort attributes based on their discriminative power [13].

Information gain can be modeled by the following series of equations,

$$Info(D) = -\sum_{i=1}^{m} p_i log_2(p_i) \qquad (1)$$

$$Info_A(D) = \sum_{i=1}^{m} C_i Info(C_Y, C_N) \qquad (2)$$

$$Gain(A) = Info(D) - Info_A(D) \qquad (3)$$

where:

- p_i represents the proportion of objects of its class in dataset D.
- C_i represents the proportion of objects in its class of a particular attribute A.
- C_Y and C_N represent the counts of positive and negative (respectively) values of objects in their classes associated with an attribute and the outcome.

Equation 1 calculates the entropy of the dataset with regard to the positive and negative outcomes. Equation 2 calculates the entropy of a particular attribute, summing the entropy of each bin of that attribute. Finally, the information gain calculated in equation 3 will determine how strongly that particular attribute discriminates between positive and negative outcomes, with larger values resulting in more pure classes.

2.3 Naïve Bayes

Naïve Bayes models are a straightforward way to attain classification capabilities from dataset in which attributes are independent or can be treated as such [14]. The premise of this technique revolves around Bayes' Theorum, given by,

$$P(X|C) = \frac{P(C|X)P(X)}{P(C)} \qquad (4)$$

where the probability of the occurrence of X given the occurrence of some condition C is calculated.

Equation 4 can be applied to the dataset in the form of the follow equation,

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \qquad (5)$$

where the product of probabilities of a positive outcome occuring given each class within an attribute is calculated. Once the probabilities of a positive outcome occuring given a class of an attribute has been calculated, the outcome of an individual sample can be predicted by finding the product of the probabilities of the classes which pertain to that sample.

2.4 Multiple Linear Regression

In contrast to naïve Bayes, linear regression models form predictions based on continuous values rather than perform classification based on categorical variables. The discriminative nature of linear regression is typically preferred over the generative classification of naïve Bayes [15]. The prediction accuracies of these two methods will thus be compared.

The essence of this model lies in the minimization of the sum of squared errors between the calculated regression and the individual observations [16]. The math for this approach is readily available at the given source, which results in a regression of the following form,

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p \qquad (6)$$

Determination of the coefficients, which will be performed in this case by the scikit-learn library, provides a predictive model for calculation of continuous variables.

2.4 DATA CLEANING

The dataset is rather complete but requires some wrangling before it can be used in the desired downstream applications. Missing values will require filling. For classifier models, continuous data will be binned as to minimize the effects of noise and prevent overfitting.

An initial assessment of the dataset indicated a positive outcome rate of ~15%, a significant imbalance in outcomes. Models trained on the unbalanced dataset will be compared to models trained on balanced slices of the dataset.

The native dataset contains 645 missing values. Rather than remove these entries entirely, the missing values were replaced with attribute means as to minimize their impact on training the models. Missing categorical values were replaced with their attribute mean rounded to the nearest int, as they require manifestation as an existing value.

Variables with greater than four possible values were then binned into discrete categories. These attributes and their quartiles were identified, followed by mapping those values to their appropriate quartile bin. This binning is useful for both decision tree induction as well as naïve Bayes down the line.

Finally, it was observed that the outcome values existed in binary numerical format. Though this is acceptable for many models, the decision tree induction model that will be deployed was written for ‘yes’/’no’ outcomes, and a conversion of these values is quite simple. This conversion adds a degree of human legibility to the dataset and will be trivial to convert to numeric values when required downstream.

3. EVALUATION

Determination of success will be approached in various manners. Since the dataset is relatively small in terms of data processing, efficiency is not a key concern, though all processes should be able to run in a reasonable amount of time.

3.1 Decision Tree Induction

The aspect of information gain of the decision tree is the target, as this will provide a useful ranking of attribute impact on outcomes. Clear separation between weights determined to be on the high and low ends of the distribution should be observed. This should result in a metric that allows straightforward ranking and comparison between the impacts that the tested factors have on CHD outcome.

The previously-cleaned dataset was loaded and processed using a series of algorithms that reflected the equations necessary to calculate information gain.

Ultimately, the values reflected in Table 2 were obtained.

Table 2. Information Gain of Attributes

Attribute	Information Gain
age	0.03483
sysBP	0.02528
prevalentHyp	0.02135
diaBP	0.0142
male	0.00558
diabetes	0.00531
education	0.00525
totChol	0.00509
BMI	0.00471
BPMeds	0.00432
glucose	0.00276
cigsPerDay	0.00221
prevalentStroke	0.00202
heartRate	0.0004
currentSmoker	0.00027

3.2 Prediction Models

Prediction accuracy will be the key indicator of success in both the naïve Bayes and linear regression models. Given that the base rate of CHD occurrence in this dataset is rather low, any test of accuracy will need to account for this disparity as apparently high accuracy can be achieved by a poor predictive model. As a demonstration of this effect, a population with a rate of occurrence of 10% of a positive outcome can trivially achieve 90% accuracy by simply assuming that the outcome is always negative, even though this is a poor method of prediction.

In order to combat this effect, a second supporting measure of predictive power known as the F1 score is often employed, defined as

$$F_1\,score = 2 \times \frac{precision \times recall}{precision + recall} \qquad (7)$$

Where precision is a measure of the number of true positives out of all the positives, and recall is a measure of true positives out of all the positive elements [17]. A high F1 score (approaching 1) is ideal, which indicates high precision and recall combined as one metric.

Since both imbalanced and balanced data were used for training and testing, the accuracy and F1 scores of both levels of parity will be compared for each prediction method.

3.2.1 Naïve Bayes Classifier

The previously-cleaned dataset was imported and converted to category codes for compatibility. The imbalanced dataset was then split into 80% train, 20% test partitions and used to train the 'GaussianNB' model from scikit-learn. Resulting accuracy and F1 scores were as follows:

```
accuracy: 0.8372641509433962
F1 score: 0.11538461538461539
```

Though the accuracy was seemingly rather good, the quite poor F1 score indicates the poor suitability of this model. In order to alleviate this anticipated issue, the dataset outcomes were brough to parity by shuffling the dataset and removing excess CHD-positive outcomes until balance was achieved. The model was again trained and assessed on the split partitions, resulting in the following scores:

```
accuracy: 0.5697674418604651
F1 score: 0.3018867924528302
```

Though the F1 score saw a modest gain, its nonetheless low value coupled with the rather weak predictive power of this model indicates poor suitability for this dataset given its current form.

The 'ComplementNB' model of scikit-learn is an implementation of naïve Bayes that is intended to alleviate some of the assumptions made with the standard model [12]. Using the same form as the 'GaussianNB' model, scores from the unbalanced data were as follows:

```
accuracy: 0.6721698113207547
F1 score: 0.27604166666666663
```

Training and testing on a balanced dataset resulted in an outperformance of the 'GaussianNB' model:

```
accuracy: 0.6046511627906976
F1 score: 0.5887096774193549
```

Finally, in a measure to subdue the effects of randomness present in the removal of entries used to bring the dataset to parity, the shuffling and removal of CHD-negative rows was run 100 times prior to the model training. Mean scores were as follows:

```
accuracy: 0.6058139534883722
F1 score: 0.5816314786874224
```

3.2.2 Multiple Linear Regression

Because linear regression performs better with data that has not been binned, the native dataset was imported, followed by a previously-implemented replacement of missing values. In this state, the dataset was split into 80% train, 20% test partitions and used to train the 'Linear Regression' model of scikit-learn. It is important to note that linear regression outputs continuous values while the expected outcome values consist of a numeric binary. Model outcomes predicted from the input test data were simply rounded to the nearest integer and compared to their reciprocal outcomes from the output test data. Resulting accuracy and F1 scores were as follows:

```
accuracy: 0.8679245283018868
F1 score: 0.034482758620689655
```

Bearing similarity to the aforementioned naïve Bayes model training, the high accuracy but low F1 score indicates a potential need for balancing the dataset. With outcome parity achieved in the same way as that of the naïve Bayes training, evaluation scores were as follows:

```
accuracy: 0.6782945736434108
F1 score: 0.6937269372693726
```

The massive improvement to F1 score places it in an acceptable position, while the prediction accuracy retained a significant portion of its prediction power, though some expected loss was observed.

To reduce the role of randomness in balancing the data and selecting the train and test partitions, a function was created to perform 100 such shuffling occurences, each used to train and test a model with 100 different randomized train and test splits. From this function, mean prediction accuracy and F1 score were determined to be the following:

```
accuracy: 0.6646889986077095
F1 score: 0.6628178675107267
```

4. DISCUSSION

Based on relative information gains, the attribute determined to hold the greatest positive correlation with CHD outcome was shown to be age, followed by measures of cardiovascular health. Somewhat surprisingly, it was shown that being a male carried more association with a CHD-positive result than diabetes did. Table 2 delineates the rankings of information gain of each attribute.

Performance of the 'GaussianNB' model was rather underwhelming in this application. When trained and tested on a balanced dataset, a prediction accuracy of only ~57% was achieved, hardly better than chance. Additionally, the very low F1 score of ~0.30 indicated poor precision and recall regarding returning accurate predictions.

When trained and tested on the unbalanced dataset, the 'ComplementNB' model had seemingly lower accuracy than that of the unbalanced 'GaussianNB' model, though the higher F1 score suggests that it is better suited to handle this dataset. This was further evinced by comparison of the two models when trained and tested on a balanced dataset, where the 'ComplementNB' outperformed 'GaussianNB' in both metrics, including a significant increase in F1 score. The improvements made to the 'ComplementNB' model were rather apparent in this exercise, though the drawbacks of naïve Bayes in general regarding assumption of attribute independence may have hindered the potential of this model.

Performance of the linear regression model on unbalanced data was similarly poor, with apparently high prediction accuracy but a low F1 score that indicated poor model suitability. However, training and testing on balanced data resulted in the highest of both metrics of the three tested models. With a prediction accuracy of ~66% and an F1 score of ~0.66, both values are satisfactory. Due to the large amount of variance of outcomes in individuals with similar medical backgrounds, it seems unlikely that an exceptionally high prediction accuracy could be attained from a dataset like this, though there very likely is room for improvement of the target metrics.

5. CONCLUSION

This paper explored the determination of which attributes hold the highest correlation with coronary heart disease outcomes in addition to forming prediction models capable of identifying at-risk individuals. The obtained ranking could plausibly be of use in determining which health risk factors are of highest priority in the fields of research as well as awareness campaigning. Multiple linear regression was determined to be the most powerful prediction model of the three tested, with both naïve Bayes

predictors underperforming. However, even in the case of the best-performing model, a prediction accuracy of ~66% with an F1 score of ~0.66 leaves much room for potential improvement. Though the pseudo-stochastic nature of medical disease outcomes perhaps caps the possible predictive power of any model, alternative machine learning techniques may be better suited for this dataset. Additionally, an ensemble method may provide better still performance. Despite potential room for improvement, the results obtained in this paper were satisfactory in achieving their goals.

REFERENCES

[1] S. Ashtekar, P. Kotkar, and S. Patil. A Hybrid Classification Method for Heart Disease Detection. In *International Journal of Applied Engineering Research ISSN 0973-4562 Volume 16, Number 8 pp. 685-689, 2021.*

[2] K. Amen, M. Zohdy, and M. Mahmoud. Machine Learning for Multiple Stage Heart Disease Prediction. *CSEIT, WiMoNe, NCS, CIoT, CMLA, DMSE, NLPD – 2020, pp. 205-223, 2020. CS & IT – CSCP, 2020.*

[3] Leading Causes of Death, *https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm, Nov. 2022.*

[4] Heart Disease and Stroke, *https://www.cdc.gov/chronicdisease/resources/publications/factsheets/heart-disease-stroke.htm, Nov. 2022*.

[5] Estimates of Funding for Various Research, Condition, and Disease Categories (RCDC), *https://report.nih.gov/funding/categorical-spending#/*, *Nov. 2022*.

[6] R. Detrano, M.D. et al. Coronary Calcium as a Predictor of Coronary Events in Four Racial or Ethnic Groups. *N Engl J Med 2008; 358:1336-1345, 2008.*

[7] Dileep, Naveen. Logistic Regression to Predict Heart Disease. *https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression*. *Nov. 2022.*

[8] Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. *Scotts Valley, CA: CreateSpace, 2009.*

[9] Kluyver, T. et al. Jupyter Notebooks – a publishing format for reproducible computational workflows. *In F. Loizides & B. Schmidt, eds. Positioning and Power in Academic Publishing: Players, Agents and Agendas. pp. 87–90, 2016*.

[10] McKinney et al. Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference, Volume 445, 2010*.

[11] Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. *Nature 585, pp. 357–362, 2020.*

[12] Pedregosa et al. Scikit-learn: Machine Learning in Python. *JMLR 12, pp. 2825-2830, 2011*.

[13] S. Tangirala. Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm. *(IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 11, 2020.*

[14] G. I. Webb. Naïve Bayes. *C Sammut and G I Webb (Eds) Encyclopedia of Machine Learning and Data Mining, Springer, 2017.*

[15] Changsung Kang & Jin Tian. A Hybrid Generative/Discriminative Bayesian Classifier. *FLAIRS Conference, pp. 562–567, 2006.*

[16] Eberly, L.E. Multiple Linear Regression. *Ambrosius, W.T. (eds) Topics in Biostatistics. Methods in Molecular Biology™, vol 404. Humana Press, (2007).*

[17] Chicco, D., Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics 21, 6 (2020).*