Heart Disease Data Science/Machine Learning Capstone

Prepared by: Frank F. Fassos

For: Coding Nomads

March 17, 2025

Abstract:

Cardiovascular diseases are the primary cause of death throughout the world (approximately 20.5 million deaths), & carry large health care costs, with cost estimates over $400B for the period of 2018-2019. While preventive and medical interventions are effective in reducing cardiovascular disease in developed nations, screening to diagnose heart disease before it become clinically apparent can reduce both the global burden of the actual disease and the cost to nations.

The purpose of this report was to use an existing dataset from the Centre for Disease Control (CDC) and use the data to train a model and evaluate the model's ability to accurately predict patients with heart disease.

Balanced Random Forest yielded a recall of 0.79, f1 score of 0.33, and a receiver operating characteristics area under the curve of 0.83. While the 0.83 value shows a good predictive ability, a higher ROC-AUC is needed for use in clinical settings. As a result, it is the recommendation of the author not to pursue the use of the model for predicting heart disease in its present form.

Background:

Cardiovascular diseases are the primary cause of death throughout the world (approximately 20.5 million deaths) (World Health Report, 2023), and carry large health care costs & disability adjusted life years (Vaduganathan et al., 2022).

Direct and indirect costs of cardiovascular diseases in the United States were estimated to be $407.3 B for the period of 2018-2019 (Tsao et al., 2023).

Therapeutic interventions include preventive and medical. Preventive measures include lifestyle modification with an improved diet, an increase in physical exercise, and screening. While medical interventions include pharmaceutical treatment that reduce the incidence and severity of cardiovascular disease, the majority of the benefit takes place in developed nations. While in developing nations pharmaceutical products are underused (Adhikary et al., 2022).

Screening or diagnosing heart disease before it requires clinical intervention is important in reducing both global burden to individuals, to health care systems and to national budgets.

The purpose of this report was to use an existing dataset from the Centre for Disease Control (CDC) and use the data to train a model and evaluate the model's ability to accurately predict patients with heart disease. An ML predictive model would be instrumental as an early warning system against heart disease, the world's number one killer.

Materials and Methods:

Software:

Anaconda Navigator, an open-source data science & artificial intelligence platform & Jupyter Lab, a web-based interactive development environment (IDE)), were used for generating notebooks for this project. The software was run on a Hewlett-Packard laptop.

Data File Source:

A .csv data file (here) was downloaded from www.Kaggle.com and the data were used for analysis in this project. The dataset was sourced from the Centre for Disease Control and had nearly 320,000 patient entries.

File Conversion:

The initial file ("heart-2020-cleaned.csv") had "Yes" and "No" binary answers and was converted to ("heart-2020-converted.csv"), where "Yes" and "No" were replaced by 1 and 0, respectively.

Data Analysis:

Following reading of the converted data file, exploratory data analysis consisted of .info(), .describe(), .head() and .tail() functions to review the data.

The .isna().sum() function was used to determine if there were any missing data from the dataset.

The .value_counts() function was used to tally up the number of patients of the parameter columns. Results were plotted as bar charts with only two columns (for binary results) and donut graphs. Some parameters (race, general health, age, sleep time, BMI (body mass index) and days of physical/mental health) required multiple columns in the histogram. Donut graphs were plotted as percent of total values.

Further exploratory analysis included logistic regression, confusion matrices, chi square and Cramer's V calculations and a correlation heatmap.

Model Training and Performance Evaluation:

For training the model, data were divided into 70% (training) and 30% (testing), and precision, recall, and f1 score, were calculated, to evaluate the performance of the classification model.

SMOTE (Synthetic Minority Over-Sampling Technique) was used for generating synthetic samples for the imbalanced dataset.

Balanced Random Forest Classifier was used to deal with the imbalanced dataset, by addressing the bias towards the majority class.

ROC-AUC (Receiver Operating Characteristic – Area Under the Curve) score was used to evaluate the classification model's performance and how it distinguishes between positive and negative classes.

Results:

Six Jupyter notebooks (herein, Books 1-6) are part of this report.

Book 1 involved exploratory data analysis. Upon initial inspection of the data with the info() function, all columns were composed of binary data (Yes, No), except for BMI (body mass index), Physical Health, Mental Health and Sleep Time. Mean values for the non-binary data are listed below:

Column Name	Mean ± SD
BMI	28.33±6.36
Physical Health	3.37±7.95
Mental Health	3.90±7.96
Sleep Time	7.10±1.44

The dataset comprised of n=319,794 patients, and there were no missing data or blanks.

The number of patients with binary responses were plotted as bar graphs or histograms (when there were more than 2 options), and donut graphs (% of total). The numbers and percentages are summarized in Tables 1-5.

Table 1: Tabular representation of total numbers of patients answering Yes and No to the columns indicated.

	Total No. of Patients (% of total) "No" Responses	Total No. of Patients (% of total) "Yes" Responses
Heart Disease	292,422 (91.4%)	27,373 (8.6%)
Smoking	187,887 (58.8%)	131,908 (41.2%)
Alcohol Drinking	298,018 (93.2%)	21,777 (6.8%)
Stroke	307,726 (96.2%)	12,069 (3.8%)
Difficulty Walking	275,385 (86.1%)	44,410 (13.9%)
Diabetic	269,653 (86.9%)	40,802 (13.1%)
Physical Activity	71,838 (22.5%)	247,957 (77.5%)
Asthma	276,923 (86.6%)	42,872 (13.4%)
Kidney Disease	308,016 (96.3%)	11,779 (3.7%)

Skin Cancer	289,976 (90.7%)	29,819 (9.3%)

Table 2: Tabular representation of total numbers of male and patients in the dataset.

	Female	Male
Gender	167,805 (52.5%)	151,990 (47.5%)

Table 3: Tabular representation of total numbers of patient’s ethnic races.

	Total No. of Patients (% of total)
White	245,212 (76.7%)
Black	22,939 (7.2%)
Asian	8,068 (2.5%)
American Indian / Alaskan Native	5,202 (1.6%)
Hispanic	27,446 (8.6%)
Other	10,928 (3.4%)

Table 4: Tabular representation of total numbers of patient’s General Health.

	Total No. of Patients (% of total)
Excellent	66,842 (20.9%)
Very Good	113,858 (35.6%)
Good	93,129 (29.1%)
Fair	34,677 (10.8%)
Poor	11,289 (3.5%)

Table 5: Tabular representation of patient age distribution.

	Total No. of Patients (% of total)
25-29	16,955 (6.2%)
30-34	18,753 (6.8%)
35-39	20,550 (7.5%)
40-44	21,006 (7.7%)
45-49	21,791 (7.9%)
50-54	25,382 (9.2%)
55-59	29,757 (10.8%)
60-64	33,686 (12.3%)
65-69	34,151 (12.4%)
70-74	31,065 (11.3%)
75-79	21,482 (7.8%)
80 +	0 (0%)

Sleep Time, BMI, Physical Health, Mental Health were not tabulated because the means ± SD were already summarized in the first paragraph of the Results section.

Table 6: Tabular representation of correlations to heart disease (Chi-square, p-values, Cramer’s V values).

	Chi-Square	p-value	Cramer’s V
Smoking	3,713.03	< 1 x 10^{-10}	0.1078
Alcohol Drinking	328.65	1.89 x 10^{-73}	0.0321
Stroke	12,386.49	< 1 x 10^{-10}	0.1968
Difficulty Walking	12,951.15	< 1 x 10^{-10}	0.2012

Gender	1,568.31	< 1 x 10^{-10}	0.0700
Diabetic	10,959.86	< 1 x 10^{-10}	0.1851
Physical Activity	3,199.01	< 1 x 10^{-10}	0.1000
General Health	21,542.18	< 1 x 10^{-10}	0.2395
Age	19,299.92	< 1 x 10^{-10}	0.2457
Race	844.31	2.99 x 10^{-180}	0.0514
Asthma	548.85	2.24 x 10^{-121}	0.0400
Kidney Disease	6,739.23	< 1 x 10^{-10}	0.1452
Skin Cancer	2,783.64	< 1 x 10^{-10}	0.0933

The binary dataset results were converted to 0 and 1 values (from No and Yes, respectively) in book 2 and stored in a new .csv file ("heart_2020_converted.csv") for training/testing the model.

Correlation heatmap (book 2) was also performed, but the highest positive correlation was 0.48.

Notebooks 3,4,5 and 6 dealt with modelling.

In book 3, a RandomForestClassifier was used and the data were scaled and split for training and testing in a 70:30 split, respectively. The results are listed below.

Model accuracy	0.91
Disease State	
Precision	0.37
Recall	0.12
F1 Score	0.18

The results did not improve even after dropping some columns.

Model accuracy	0.90
Disease State	
Precision	0.33
Recall	0.13
F1 Score	0.19

The results did not improve after Logistic Regression.

Disease State	
Precision	0.53
Recall	0.11
F1 Score	0.18

Using an IsolationForest improved the recall results, but not enough for the model to have predictive value.

Disease State	
Precision	0.27
Recall	0.33
F1 Score	0.30

The reason for these poor results was that the data were unbalanced heavily towards the non-disease state.

To correct for the imbalance, downsampling of the majority class (non-disease state) was used to match the minority class (disease state) (book 4). The data were split 70:30 (training:testing) as before. The results improved and are listed below.

Model accuracy	0.74
Disease State	
Precision	0.73
Recall	0.77
F1 Score	0.75

Following normalization and scaling with StandardScaler, model accuracy decreased, but recall had a considerable increase.

Disease State	
Precision	0.19
Recall	0.88
F1 Score	0.31

In an attempt to improve the results a SMOTE (Synthetic Minority Over-sampling Technique) was used (book 5) to generate synthetic samples so that the imbalance in the dataset could be corrected. However the results for recall remained at 0.12 with the RandomForestClassifier. However after repeating with a BalancedRandomForestClassifier, recall increased to 0.80. There were no further improvements with XGBoost or Logistic Regression.

Finally, when Ensemble and BalancedRandomForestClassifier were used without SMOTE a recall of 0.79 & a ROC-AUC (received operating characteristic area under the curve) of 0.83 were achieved (book 6). The ROC-AUC of 0.83 indicates that the model has a strong (but not excellent) predictive ability in identifying high risk patients.

Conclusions:

The ROC-AUC of the model was 0.83, suggesting that while the model has good predictive ability in discriminating between positive and negative classes, but not excellent. Since we are dealing with clinical diagnosis, the importance of finding positive instances of disease for treating cannot be understated.

These findings suggest that this model should no longer be pursued as a predictive model for heart disease.

Limitations of datasets for this disease condition in general, include the possibility that individuals have asymptomatic heart disease and do not know it. Put another way, an individual who is 30 years old, works out and lives an otherwise normal life but has plaque buildup on his coronary arteries, technically has heart disease.

Using echocardiograms (herein, “echo”), or stress echo (to detect ischemia) could be used on people who are classified as not having heart disease, thereby generating a new dataset. Stratifying data based on age and/or severity of condition could be useful in “cleaning” the data. Echocardiography is very useful in identifying asymptomatic heart disease. Echo is a non-invasive procedure and offers real-time clinical diagnosis of structural abnormalities, heart performance, arterial and valvular diseases.

Future improvements may include trying using threshold tuning to improve sensitivity or sensitivity or using neural networks (deep learning) for deciphering more complex patterns in early heart disease screening.

References:

Adhikary D, Barman S, Ranjan R, Stone H. A Systematic Review of Major Cardiovascular Risk Factors: A Growing Global Health Concern. *Cureus* 2022:14(10): e30119.

Hinton W, McGovern A, Coyle R, Han TS, Sharma P, Correa A, Ferreira F, de Lusignan S. Incidence and prevalence of cardiovascular disease in English primary care: a cross-sectional and follow-up study of the Royal College of General Practitioners (RCGP) Research and Surveillance Centre (RSC). *BMJ Open* 2018;8:e020282.

Tsao CW, Aday AW, Almarzooq ZI, Anderson CAM, Arora P, Avery CL, Baker-Smith CM, Beaton AZ, Boehme AK, Buxton AE, Commodore-Mensah Y, Elkind MSV, Evenson KR, Eze-Nliam C, Fugar S, Generoso G, Heard DG, Hiremath S, Ho JE, Kalani R, Kazi DS, Ko D, Levine DA, Liu J, Ma J, Magnani JW, Michos ED, Mussolino ME, Navaneethan SD, Parikh NI, Poudel R, Rezk-Hanna M, Roth GA, Shah NS, St-Onge M-P, Thacker EL, Virani SS, Voeks JH, Wang N-Y, Wong ND, Wong SS, Yaffe K, Martin SS. Heart Disease and Stroke Statistics—2023 Update: A Report From the American Heart Association. *Circulation.* 2023;147:e93–e621.

Vaduganathan M, Mensah GA, Turco JV, Fuster V, Roth GA. The Global Burden of Cardiovascular Diseases and Risk: A Compass for Future Health. *J Am Coll Cardiol*. 2022;80:2361 – 2371.

World Heart Federation. WORLD HEART REPORT 2023: Confronting the World’s Number One Killer. 2023; 1-48.