Early screening of heart disease through classification on imbalanced survey data and identifying influential features via explainability methods

Paul Ma
p32ma@uwaterloo.ca
University of Waterloo
Waterloo, ON, Canada

Abstract

Given the relative ease with which survey data may be collected and accessed, any possibility of using such data to identify individuals at risk of heart disease, or indeed any illness, could have positive implications for early detection and prevention. However, survey data is often limited by lack of reliability and precision, and the issue of class imbalance must also be addressed. Cost-sensitive Logistic Regression and cost-sensitive Decision Tree achieved close to optimal performance given the experimental scope while maintaining reasonably low training time complexity. LIME and DiCE explainability methods are also leveraged to uncover potentially important features that may be critical to early screening of heart disease, overall identifying age, self-reported poor general health, and history of stroke as significant indicators.

Introduction

Despite the ever-increasing efforts aimed at preventing, detecting, and treating heart disease, it was still the second most common cause of death in 2020, and the most common cause of hospitalization, according to Statistics Canada.[1] Consequently, any technology designed to improve the existing standards for diagnostics and treatment has a potential to play a crucial role in saving lives, particularly as relates to preventive measures and early diagnosis. The importance of the latter cannot be overstated as it regularly determines the effectiveness and the success of treatment or preventative measures.

The question remains, however, how to efficiently conduct a regular assessment on a large group of people. One approach that potentially allows rapid screening of a larger population is through the use of surveys. Although surveys do suffer from a number of flaws including issues of bias[2], there still exists the possibility of building prediction models on any collected data. These models, in turn, may subsequently be used as another tool to address the issues posed by various complexities associated with early diagnosis, such as lack of evident symptoms or misinterpretation of the early signs of disease. For instance, if such a model suggests that those with certain prevalent lifestyle habits, or those belonging to particular age groups or income classes are more susceptible to having heart disease, appropriate strategies can be developed to reach out to these demographics and target preventative care towards individuals who may be at risk.

The broad aim of this project is to build binary classifiers to effectively predict whether a survey respondent may be at risk for heart disease, given their responses to some simple and generic survey questions, and determine which features may be most impactful in making such predictions using Local Interpretable Model-agnostic Explanations (LIME) and Diverse Counterfactual Explanations (DiCE) explainability methods. However, given that only a small minority of respondents are expected to be at risk for heart disease, it is critical to take into account the unequal class distribution. As such, the primary goal is to evaluate three popular strategies for conducting imbalanced classification - data resampling, algorithm-level accounting for misclassification costs and ensemble methods.

The baseline models that will be explored are Logistic Regression, Naive Bayes, Linear Support Vector Machine (SVM), Neural Network and Decision Tree, with Random Forest and Bagging Classifier as the ensemble variants of the latter. Cost-sensitive variants of these models will be used to evaluate algorithm-level techniques on imbalanced classification, with Synthetic Minority Oversampling Technique (SMOTE) used as a means to assess data-level resampling. Model performance will be evaluated on convergence time and F1 score.

The chosen dataset is sourced from Kaggle[3] and consists of a consolidated subset of telephone health survey data from 2011 to 2015[4]. The survey is conducted by the Centers for Disease Control and Prevention (CDC) on an an-

[1] https://www150.statcan.gc.ca/n1/daily-quotidien/220124/dq220124a-eng.htm

[2] https://www.forbes.com/sites/serenitygibbons/2019/04/27/why-your-customer-surveys-are-probably-inaccurate/?sh=6c4bda2665bf

[3] Heart Disease Health Indicators Dataset by Alex Teboul: https://www.kaggle.com/alexteboul/heart-disease-health-indicators-dataset

[4] Original dataset hosted on Kaggle, sourced from the Centers for Disease Control and Prevention: https://www.kaggle.com/cdc/behavioral-risk-factor-surveillance-system

nual basis in order to collect data on potential health risk and early preventative care behaviours as relates to a number of health-related issues, including chronic illnesses, diseases and injuries. It is known as the Behavioral Risk Factor Surveillance System (BRFSS).

Related Work

Imbalances in the class distribution of a dataset can have severe impact on classification performance if left unaddressed. One significant issue is that accuracy no longer becomes a reliable evaluation metric, as high classification accuracy may be trivially achieved by predicting the majority class for all or most instances (Chicco and Jurman 2020; Zhu 2020; Boughorbel, Jarray, and El-Anbari 2017). The impact this has on a classifiers ability to correctly predict the minority class increases significantly with the severity of the imbalance. As such, it is beneficial to look at both precision and recall when evaluating classification performance.

Two metrics that are commonly used in imbalanced learning that incorporate both precision and recall are F1 score and Matthews Correlation Coefficient (MCC), though a great deal of contention exists as to which is the most suitable classification metric (Chicco and Jurman 2020; Zhu 2020; Chawla 2009). Issues with F1 primarily arise when the positive class is the majority, or when the positive class is equally or less important than the negative class. However, if neither of these criteria are met and false negatives are of no great concern, F1 is simultaneously as useful as MCC whilst being simultaneously more interpretable.

Outside of choices of evaluation metrics, there is the execution of imbalanced learning itself. To that end, (Krawczyk 2016) considers there to be three main classes of strategies, namely algorithm-level strategies, data-level strategies and hybrid strategies that incorporate the benefits of both while making up for their shortfalls.

(Krawczyk 2016; Leevy et al. 2018; Ganganwar 2012) detail a number of the more popular branches of each imbalanced learning strategy, as well as some relatively novel approaches at time of writing. Among the most popular algorithm-level methods are cost-sensitive training, ensemble learning and boosting. Data-level strategies primarily concern subsampling (oversampling or undersampling) methods that address the distribution imbalance by either creating new observations or removing observations according to some criteria, though another less explored branch (due to potential concerns regarding excessive computational costs) exists focusing on feature selection. Hybrid approaches typically involve careful combination of both data and algorithm-level approaches which introduces a degree of complication, though performance gains appear to be promising when this is successful (Leevy et al. 2018; Santos et al. 2018).

Concerning subsampling methods, a number of caveats exist to their usage that may impact classification performance. (Santos et al. 2018) highlights the importance of not applying subsampling to the data prior to training as this may result in data leakage, which is likely to lead to overoptimistic classification performance. To prevent this, any subsampling should be performed in tandem with cross-validation. Additionally, it should be noted that oversampling methods will inherently impact training complexity due to increase in the training set size, as well as introduce the possibility of overfitting depending on the oversampling strategy (Krawczyk 2016; Leevy et al. 2018).

Explainable AI is a broad and rapidly growing area of research, driven in large part due to the rise in popularity of complicated and uninterpretable black-box models (Sahakyan, Aung, and Rahwan 2021; Du, Liu, and Hu 2019; Burkart and Huber 2021). The exact intuition behind how such predictions are made by these models is often extremely difficult, if not impossible, to understand. There are suggested to be five main categories of explanations (Sahakyan, Aung, and Rahwan 2021): feature importance, feature interaction, decision rules, simplified models and counterfactuals. LIME came to prominence with the surge of explainable AI and is an example of a model-agnostic, local, post-hoc technique that produces explanations in the form of the first of these categories, feature importance (Ribeiro, Singh, and Guestrin 2016; Sahakyan, Aung, and Rahwan 2021).

Methodology

The Heart Disease Health Indicators Dataset, as consolidated from the BRFSS 2015 dataset, contains 253,680 responses from individuals who participated in the BRFSS surveys from 2011 to 2015. The number of features have been reduced down to 21 and include both binary and ordinal features (see Table A.2 in the Appendix for full feature descriptions). There is a single target binary feature, which indicates whether the respondent has heart disease or is at risk of heart disease. The dataset is highly imbalanced at a ratio of approximately 90:10, with 23,893 respondents having reported suffering from heart disease or heart attack.

Despite the lack of continuous features present in the dataset, categorical features have previously been shown to possess predictive power in determining coronary heart disease risk (Wilson et al. 1998). Survey data tends towards the categorical, so it is useful to be able to make predictions using such data. Though early detection benefits significantly from a more diverse collection of data (Ng et al. 2016), it is nonetheless worthwhile to explore given the relative ease by which survey data may be obtained.

Libraries leveraged to complete this project include: numpy, pandas, seaborn, imblearn, sklearn, lime and dice-ml.

For reasons of brevity, when considering whether a respondent reported as being diagnosed with heart disease or having suffered from a heart attack, this may simply be referred to as heart disease from here on.

Supervised Learning Models

A brief summary of the chosen classifier models will be outlined here.

Naive Bayes is an algorithm developed on the Bayes' theorem and a key independence assumption between features given the target labels, and is primarily used for binary classification tasks (though multi-class classification and regression are also possible). Due to this independence assumption, classification is theoretically more robust to class imbalance, though overall classification performance is somewhat reliant on the degree to which the assumption holds in actuality.

Logistic Regression is an extension of linear regression, but rather than fitting a straight line to the data, it fits a logistic function which takes the form of an S-shaped curve with a range from 0 to 1. This allows Logistic Regression to be used in predicting probabilities, as well as simultaneously being well-suited to binary classification.

SVM separates data into two classes via a hyperplane, which is synonymous to the most preferable decision boundary. Simply put, the algorithm takes in data and tries to classify it, by converting it into higher dimension linearly separable data using a kernel function and drawing a line (or hyperplane) between the two classes. Unfortunately, training complexity with non-linear kernels may scale at least quadratically (possibly cubicly) with the number of training observations. As such, a Linear SVM is chosen for this experiment, optimized via stochastic gradient descent for further training time reduction.

Decision Trees are trained top-down by systematically choosing features by which remaining examples may be split. How these splits are chosen is based on estimated information gain until the full tree is completed. It is an intrinsically explainable model and easily grasped for human understanding.

Random Forest and Bagging Classifier can both be described as classifiers consisting of multiple decision trees, though they differ in one crucial aspect. Splits in Random Forest are performed on a subset of the features that are drawn with replacement at each split, whereas in the Bagging Classifier, all features are used to build each tree.

Neural Networks are modelled on multiple layers of perceptrons with activation functions and training is conducted by updating the edge weights between layers through backpropagation. Though they are powerful and, by the universal approximation theorem, are able to theoretically approximately represent any continuous function given at least a single hidden layer, they are notoriously difficult to train, in terms of both input data requirements and training time complexity.

Imbalanced Learning Strategies

The two imbalanced learning strategies explored in this project include an algorithm-level strategy and a data-level strategy, namely cost-sensitive training and Synthetic Minority Oversampling Technique (SMOTE).

Cost-sensitive training broadly involves accounting for misclassification costs by adjusting the importance of classification by the class that is being predicted (Mienye and Sun 2021; Yang et al. 2009; López et al. 2012; Ling and Sheng 2008). This is often done through class weights, whereby more important classes are assigned a larger weight and less important classes a smaller weight. A common strategy for determining class weights is to assign them according to the inverse of the class distribution, which is the strategy adopted in this experiment for those classifiers where cost-sensitive training is feasible.

SMOTE is a data augmentation strategy that involves oversampling the minority class to address inherent imbalances in the data set (López et al. 2012; Fernández et al. 2018; Chawla et al. 2002). Rather than naively duplicating observations in the minority class, SMOTE attempts to synthesize ”new” observations from those already existing minority class examples. This is done through the use of k nearest neighbours, whereby random individuals in the minority class are selected and their k nearest neighbours (also in the minority class) are found. One of these k neighbours is then also selected at random, and a new synthetic example is created through interpolation between the two selected minority class examples.

Explainability Methods

Local Interpretable Model-agnostic Explanations (LIME) and Diverse Counterfactual Explanations (DiCE) are both model agnostic local post-hoc explainability methods chosen for the purposes of identifying what features are considered most important, according to our heart disease classifiers.

Broadly, LIME generates explanations by building a weighted interpretable surrogate model (such as linear regression) which is fit to a particular instance or observation using perturbations of that instance and the accompanying black box predictions (Ribeiro, Singh, and Guestrin 2016; Sahakyan, Aung, and Rahwan 2021). The fitted surrogate model weights can then be interpreted, and theoretically can be used as a local explanation for how important features are to the prediction.

Rather than through interpretable surrogate models, DiCE generates its explanations using counterfactuals, which consist of a set of perturbed features for an instance that would result in a change in prediction (Wachter, Mittelstadt, and Russell 2017; Chou et al. 2022). These counterfactuals are generated through determinantal point processes (Mothilal, Sharma, and Tan 2020), which is a discrete probabilistic modeling paradigm which arose from the fields of quantum physics and random matrix theory that allows for a wide variety of applications (Kulesza 2012), including efficient, high quality subset selection subject to constraints on diversity (in this case, as applies to subsets of counterfactuals).

Experimental Setup

With regards to data preparation, a number of modifications were made to features prior to training. Values for *GenHlth* were reversed for consistency with the negative to positive ordering of other features in the dataset. Additionally, *BMI* was converted to ordinal health risk categories *BMICat* (see Table 1) based on specifications outlined by Health Canada[5].

Following these modifications, min-max scaling was applied to all non-binary features in preparation for training

[5] https://www.canada.ca/en/health-canada/services/food-nutrition/healthy-eating/healthy-weights/canadian-guidelines-body-weight-classification-adults/body-mass-index-nomogram.html

with models sensitive to feature scaling. Typically, due to the risk of data leakage, it is considered best practice to apply any scaling measures solely on the training set. However, since domain knowledge of the possible survey responses provides the full range of values for all features, it was deemed safe in this instance. The data was then split into train and test sets with an 80:20 ratio. This was done via stratified sampling (using sklearn's StratifiedShuffleSplit class) due to the heavy class imbalance inherent in the data. The random seed for the train-test split was set to a constant for reproducibility and consistency across experiments.

BMI	Health Risk Classification (ordinal)
18.5	0 - Underweight
18.5 - 25	1 - Normal Weight
25 - 30	2 - Overweight
30 - 35	3 - Obese class I
35 - 40	4 - Obese class II
40	5 - Obese class IIII

Table 1: BMI to ordinal Health Risk categories conversion.

On concluding all preprocessing steps, models can be trained and evaluated. Three variants of each classifier were trained: a base classifier with no subsampling, a classifier with SMOTE oversampling, and a cost-sensitive classifier. Two exceptions to this are the Naive Bayes and Neural Network classifiers, as the means to set class weights for these do not currently exist in sklearn (and not strictly needed given the independence assumption in Naive Bayes).

In order to conduct SMOTE oversampling, pipelines were used to sequentially assemble all subsampling (using imblearn's SMOTE class) and estimators to be cross-validated together. This is particularly important for the application of any subsampling such as SMOTE due to the risk of data leakage (Santos et al. 2018). Training was performed through stratified 5-fold cross-validation (using sklearn's StratifiedKFold class) with hyperparameter tuning (using sklearn's RandomizedSearchCV) and F1 as the scoring metric. With regards to SVM classifiers, prediction probabilities for use with both LIME and DiCE were obtained through an additional step of performing probability calibration (using sklearn's CalibratedClassifierCV class). Any hyperparameters selected for tuning are detailed in Table A.1 in the Appendix.

Once the classifiers are fully trained, performance is evaluated on accuracy, precision, recall and F1 score, and explanations for model predictions are generated using LIME and DiCE. Generated counterfactual explanations from DiCE are limited to features subject to the following restrictions: firstly, features that are completely non-actionable should not be varied (this includes *Stroke*, *Diabetes*, *Sex* and *Age*); secondly, features that are subjectively deemed not immediately actionable (or for which initial steps towards action may not be realistic or reasonable depending on external factors) in the short to medium term are also not to be varied (this includes *AnyHealthcare*, *NoDocbcCost*, *GenHlth*, *MentHlth*, *PhysHlth*, *DiffWalk*, *Education* and *Income*).

Estimates for global feature importance can also be derived by DiCE through generating counterfactual explanations for a subset of examples and aggregating the results. For this experiment, a subset of 500 examples from the test set are used to acquire these global feature importance estimates.

Convergence time is also given some consideration, though only peripherally as the primary focus is on evaluating the predictive capability of classifiers on health survey data and identifying potentially important factors affecting heart disease risk.

Results

Data Exploration

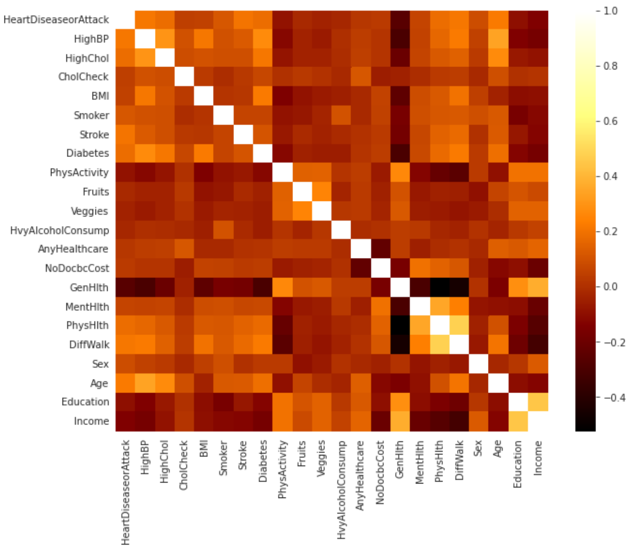


Figure 1: Heart disease dataset correlation matrix.

Initial exploration of the dataset revealed a number of strongly correlated features, as shown in Figure 1. These primarily concern features indicating respondents' self-reported health status (both physical and mental) in some time interval prior to being surveyed, namely *GenHlth*, *PhysHlth*, *MentHlth* and *DiffWalk*.

The correlation between responses regarding physical health is at first glance unsurprising, as well as any self-reported poor mental health (though any degree of causality would be difficult to establish solely with the data at hand). Tangentially, there appears to also exist strong positive correlation between self-reported *GenHlth* and Income class, as well as *Age* and *HighBP*, the implications of which may be worth future investigation.

With regards to the target feature of *HeartDiseaseorAttack*, the top five most correlated features in descending order are *GenHlth*, *Age*, *DiffWalk*, *HighBP* and *Stroke*.

Classifier Evaluation

As detailed in the methodology, training of all classifiers was performed on a 80:20 train-test split through 5-fold stratified cross-validation, with additional hyperparameter tuning.

The average training time (in seconds) per hyperparameter tuning iteration for individual training splits was recorded for each classifier, and the log-scaled times can be seen in Figure 2. Note that no training times were recorded for cost-sensitive variants of Naive Bayes and Neural Network classifiers as cost-sensitive variants of these classifiers were not trained.

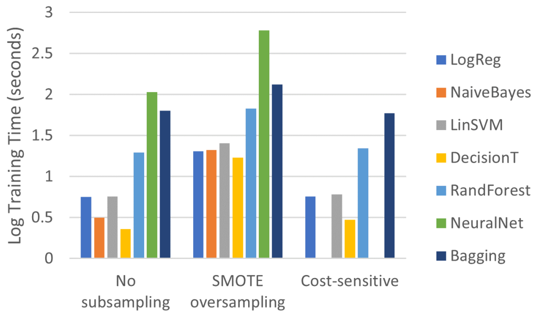


Figure 2: Average log-scaled training times (in seconds) per hyperparameter search iteration in a single split.

The Neural Network classifiers, despite being relatively shallow (no more than two hidden layers), were by far the slowest to train, particularly when coupled with SMOTE oversampling, taking on average 600 seconds per training iteration. For all classifier variants, the fastest classifiers to train per training iteration were the Decision Trees (with the base variant averaging 2.3 seconds per training iteration).

In terms of total training time, Naive Bayes achieved by far the lowest times (on average 19 seconds for baseline Naive Bayes, and 2 minutes 6 seconds for Naive Bayes with SMOTE), though this is unsurprising given that no hyperparameter tuning was performed for either the baseline or SMOTE variants.

The imbalanced learning strategy having the strongest impact on training times was SMOTE oversampling due to the obvious increase in input size, with cost-sensitive classifier variants seeing largely similar convergence speeds compared to the baseline models.

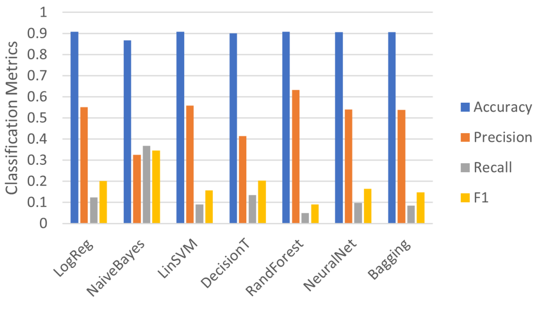


Figure 3: Classification metrics for models with no subsampling.

As shown in Figure 3, the F1 score, recall and precision of baseline classifiers with no imbalanced learning strategies vary across the board. Accuracy is universally high, reaching approximately 0.9 for all classifiers, with low recalls and F1 scores (unsurprising, given the unaddressed imbalance). The sole exception was Naive Bayes, which obtained an accuracy of 0.869, and saw the best performance in terms of average recall and F1 score, which were 0.368 and 0.345, respectively.

The Random Forest classifier performed extremely poorly in the absence of any imbalance-tackling measures, with the lowest recall and F1-scores of all the baseline classifiers, which were 0.049 and 0.091, respectively.

Average recall over all baseline classifiers was 0.136, and average F1 score was 0.187.

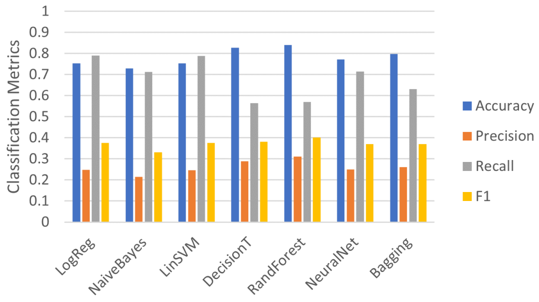


Figure 4: Classification metrics for models with SMOTE.

Classification performance with SMOTE improved for all classifiers over their baseline models (see Figure 4), with accuracy and precision declining overall in exchange for an increase in recall and F1 scores. One minor exception to this is Naive Bayes, whose SMOTE variant saw little to no improvement in average F1 score compared to the baseline classifier, though did see the same shift in emphasis from accuracy and precision to recall.

SMOTE Random Forest obtained the highest average F1 score of any model and imbalanced learning strategy, with a recall and F1 score of 0.57 and 0.401, respectively. Out of the remaining SMOTE classifiers, Logistic Regression notably achieved one of the highest overall recalls of 0.79, with an F1 score of 0.376.

Average recall over all SMOTE classifiers was 0.681, and average F1 score was 0.372.

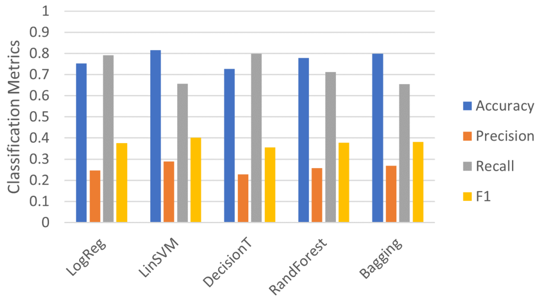


Figure 5: Classification metrics for cost-sensitive models.

Cost-sensitive classifiers achieved comparable gains in classification performance to those seen with SMOTE (see Figure 5), with similar observed trade-offs between accuracy, precision and recall.

Recall performance was relatively high with the cost-sensitive variants, with the cost-sensitive Decision Tree achieving the highest average recall of 0.799, and an F1 score of 0.356. Cost-sensitive Logistic Regression achieved highly comparable performance to its SMOTE variant, with a recall and F1 score of 0.791 and 0.376, respectively. The best performing cost-sensitive classifier in terms of F1 score was cost-sensitive Linear SVM, achieving an F1 score of 0.401 and a recall of 0.656.

Average recall over all cost-sensitive classifiers was 0.722, and average F1 score was 0.378.

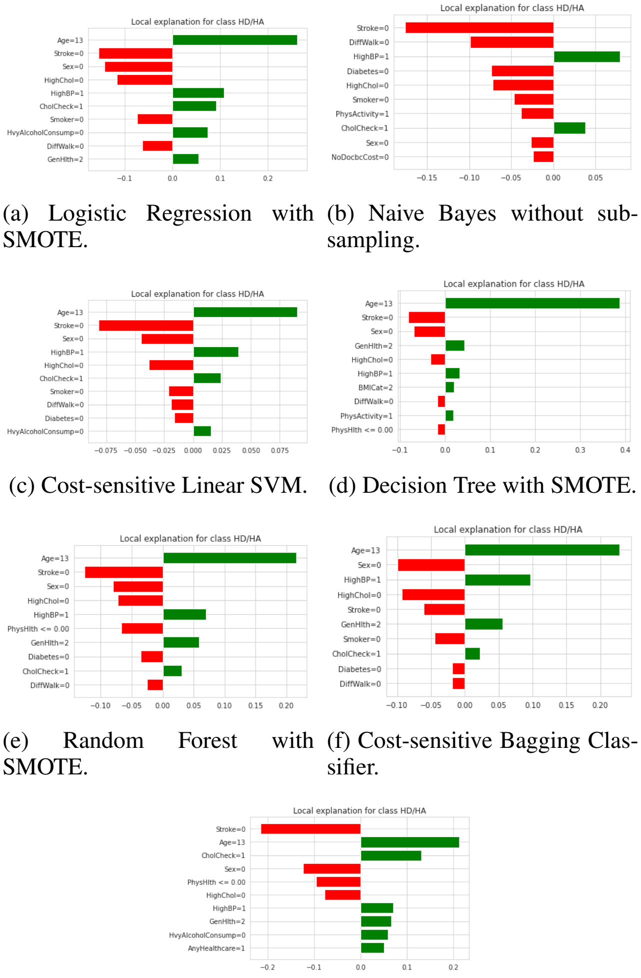
Feature Importance and Model Explainability

LIME and DiCE explanations were generated for a number of test instances following training and classification evaluation. Figure 6 shows a set of example local explanations generated for a single test instance, with Figure 7 showing generated counterfactual explanations (that were deemed most reasonable out of a maximum of 5 generated counterfactuals) for the same instance (additionally showing the actual feature values and label for said instance).

The instance in question is for a survey respondent with heart disease, for whom a number of features would indeed intuitively appear to place them at higher risk of such. This includes presence of high blood pressure and their high age category, though a number of features would seem to indicate an otherwise relatively healthy lifestyle, such as their being physically active and their incorporation of fruits and vegetables into their diet. Three models misclassified this respondent, namely the Naive Bayes, SMOTE Random Forest and SMOTE Neural Network classifiers.

For this particular instance, *Age* was the feature with the largest weight magnitude amongst all classifiers (with the SMOTE Decision Tree in Figure 6d attributing to it the largest magnitude weighting overall), with the exception of Naive Bayes (Figure 6b) and SMOTE Neural Network (Figure 6g). Both of these attributed the most weighting to *Stroke*, though the magnitude difference with *Age* is significantly smaller with the Neural Network. In fact, the Neural Network had the greatest degree of prediction uncertainty with this test instance at almost exactly 50%, though others also had relatively low prediction certainty, with only Naive Bayes with prediction confidence of over 60% at approximately 98% (despite also misclassifying this test instance, SMOTE Naive Bayes had a lower prediction confidence of 72%).

DiCE counterfactual explanations (Figure 7) were unable to be generated (given the restrictions on features to be varied) for both Naive Bayes and SMOTE Decision Tree. Amongst the other classifiers, the most sensible generated counterfactuals were retained, but despite this, a number of oddities do appear. For example, the counterfactual explanations for cost-sensitive Linear SVM would seem to suggest that ceasing vegetables in their daily diet and taking up heavy alcohol consumption would lead to a change in heart disease risk; whilst those for the cost-sensitive Bagging Classifier, though reasonably suggesting that a lowering of blood pressure would result in a no heart disease prediction, also recommend an increase in BMI health risk category from 2 to 4 (from Overweight to Obese class II).



(a) Logistic Regression with SMOTE.

(b) Naive Bayes without subsampling.

(c) Cost-sensitive Linear SVM.

(d) Decision Tree with SMOTE.

(e) Random Forest with SMOTE.

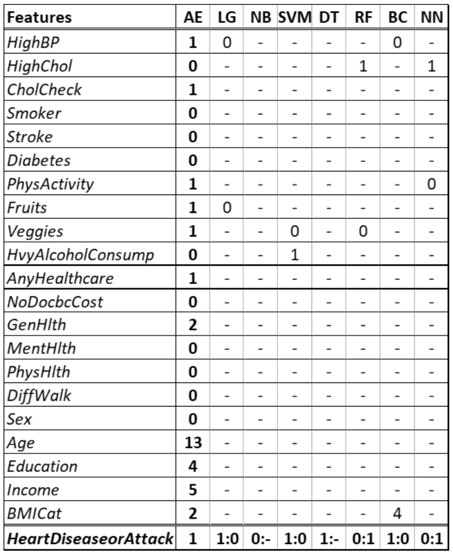
(f) Cost-sensitive Bagging Classifier.

(g) Neural Network with SMOTE.

Figure 6: Local explanations generated by LIME for chosen test instance (classifier variants with best F1 scores).

The three most globally important features as estimated by DiCE amongst the base classifiers are universally *Stroke*, *GenHlth* and *Age*, in that order, with the sole exception of Naive Bayes, for which the important features were *Stroke*, *DiffWalk* and *HighBP* (see Table 2). Comparing this to Naive

[6] Header abbreviations: AE = Actual Example, LG = Logistic Regression (SMOTE), NB = Naive Bayes (Base), SVM = Linear SVM (Cost-sensitive), DT = Decision Tree (SMOTE), RF = Random Forest (SMOTE), BC = Bagging Classifier (Cost-sensitive), NN = Neural Network (SMOTE). For the *HeartDiseaseorAttack* label, the value to the left of the colon refers to the original prediction, and the value to the right is the opposite class the counterfactuals were generated for.



Features	AE	LG	NB	SVM	DT	RF	BC	NN
HighBP	**1**	0	-	-	-	-	0	-
HighChol	**0**	-	-	-	-	1	-	1
CholCheck	**1**	-	-	-	-	-	-	-
Smoker	**0**	-	-	-	-	-	-	-
Stroke	**0**	-	-	-	-	-	-	-
Diabetes	**0**	-	-	-	-	-	-	-
PhysActivity	**1**	-	-	-	-	-	-	0
Fruits	**1**	0	-	-	-	-	-	-
Veggies	**1**	-	-	0	-	0	-	-
HvyAlcoholConsump	**0**	-	-	1	-	-	-	-
AnyHealthcare	**1**	-	-	-	-	-	-	-
NoDocbcCost	**0**	-	-	-	-	-	-	-
GenHlth	**2**	-	-	-	-	-	-	-
MentHlth	**0**	-	-	-	-	-	-	-
PhysHlth	**0**	-	-	-	-	-	-	-
DiffWalk	**0**	-	-	-	-	-	-	-
Sex	**0**	-	-	-	-	-	-	-
Age	**13**	-	-	-	-	-	-	-
Education	**4**	-	-	-	-	-	-	-
Income	**5**	-	-	-	-	-	-	-
BMICat	**2**	-	-	-	-	-	4	-
HeartDiseaseorAttack	**1**	**1:0**	**0:-**	**1:0**	**1:-**	**0:1**	**1:0**	**0:1**

Figure 7: DiCE counterfactuals for chosen test instance.[6]

bayes with SMOTE which attained a slightly lower F1 but higher recall, a shift in importance can be observed, with the most important features being *GenHlth* (0.338), *Stroke* (0.316), and *HighBP* (0.236).

In fact, extending our observations to the cost-sensitive and SMOTE variants of the classifiers, feature importance rankings appear to change considerably from their baselines. Table 2 shows the five most important global features for the best performing variants of each classifier (according to F1 score).

Overall, a larger degree of importance is shifted towards *Age*, giving it the highest level of global importance for all but baseline Naive Bayes and SMOTE Neural Network classifiers. Additionally, it appears that as imbalanced learning strategies are applied, significantly less emphasis is placed on *Stroke* as a predictor. In fact, the SMOTE Neural Network is the sole non-baseline classifier still prioritising *Stroke* as a predictor, according to DiCE (though the relative weighting does decrease by more than half from its baseline counterpart at 0.753).

Model	Highest importance global features
Logistic Regression (SMOTE)	*Age*: 0.53, *GenHlth*: 0.369, *Stroke*: 0.185, *HighBP*: 0.102, *MentHlth*: 0.094
Naive Bayes (Base)	*Stroke*: 0.642, *DiffWalk*: 0.438, *HighBP*: 0.335, *Diabetes*: 0.303, *HighChol*: 0.28
Linear SVM (Cost-Sensitive)	*Age*: 0.51, *GenHlth*: 0.421, *Stroke*: 0.129, *PhysHlth*: 0.102, *HighChol*: 0.091
Decision Tree (SMOTE)	*Age*: 0.556, *GenHlth*: 0.371, *MentHlth*: 0.139, *BMICat*: 0.098, *Sex*: 0.093
Random Forest (SMOTE)	*Age*: 0.584, *GenHlth*: 0.4, *PhysHlth*: 0.314, *Stroke*: 0.178, *Education*: 0.139
Bagging Classifier (Cost-Sensitive)	*Age*: 0.468, *GenHlth*: 0.331, *HighBP*: 0.16, *Stroke*: 0.144, *PhysHlth*: 0.143
Neural Network (SMOTE)	*Stroke*: 0.353, *Age*: 0.272, *PhysHlth*: 0.208, *GenHlth*: 0.168, *MentHlth*: 0.124

Table 2: DiCE - Top five most important global features by classifier (variant with best F1 score).

Discussion

With regards to feature importance, both DiCE and LIME appear to strongly identify *Age*, *GenHlth* and (to a more varying extent) *Stroke* as critical to the predictions of most of the non-baseline classifiers, with the notable exception of the SMOTE Neural Network, particularly given its relative complexity as a model in comparison to the other evaluated classifiers.

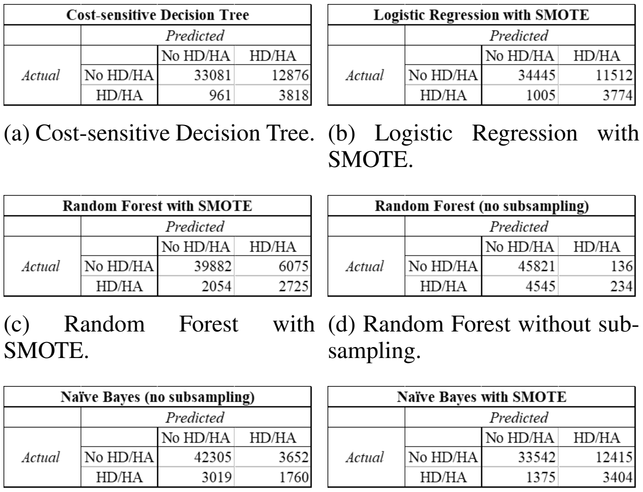
There does appear to be some degree of consensus between LIME and DiCE as to which features might tentatively be actionable venues for targeting heart disease risk; for example, the DiCE explanations (Figure 7) for both SMOTE Logistic Regression and SMOTE Random Forest suggest addressing *HighBP* and *HighChol*, both being amongst the more heavily weighted actionable features of their respective LIME explanations (Figures 6a and 6e). It should, however, be noted that the DiCE counterfactual explanations were often found to be nonsensical (as seen with Linear SVM in Figure 7). This is possibly relating to the restrictions imposed upon which features are allowed to be varied, but may also point towards the need for additional parameter tuning, or other limitations regarding to counterfactual explanation generation.

The results of the classifier performance evaluations themselves show a clear need for considering appropriate strategies when training on imbalanced data, as well as clearly exemplifying the pitfalls of relying on accuracy in such conditions. The baseline Random Forest classifier is emblematic of this, as can be seen in Figure 8d. The classifier is able to achieve over 90% accuracy simply by labeling almost all examples as not having heart disease, whilst simultaneously misclassifying over 95% of those examples belonging to the heart disease class.

This is even the case regarding the Naive Bayes classifier, which was the best performing classifier out of all the baseline variants. Despite its relatively high F1 score, the recall performance of the baseline Naive Bayes classifier is rather disappointing, failing to correctly classify just over 63% of respondents with heart disease in the test set (see Figure 8e). Comparing this to the SMOTE Naive Bayes classifier in Figure 8f which reduced this misclassification percentage to about 29%. This gain in recall comes at the cost of precision and accuracy, but given that (from a public health perspective) it is likely more important to highlight risk factors or identify individuals at elevated risk in order to take potential preventative measures, it seems reasonable to accept such a trade-off[7].

Though there is a stark difference in performance between the baseline classifiers and those incorporating imbalanced learning strategies, it is not entirely clear which of these strategies is optimal in this particular classification problem.

Overall, the classification performances of both Logistic Regression and Bagging Classifier remained relatively stable between cost-sensitive and SMOTE variants (see Figures 4 and 5). Linear SVM, Decision Tree and Random Forest variants, though largely retaining similar F1 scores across the board, saw more fluctuation in recall, precision and accuracy. Figures 8a, 8b and 8c show the confusion matrices for three of the best performing classifiers in terms of F1 score. Evaluation at this stage appears to boil down to which of precision or recall is a higher priority, as well which strategy is more time and training efficient (in which case, cost-sensitive learning appears to be more suitable here).



Cost-sensitive Decision Tree			
		Predicted	
		No HD/HA	HD/HA
Actual	No HD/HA	33081	12876
	HD/HA	961	3818

(a) Cost-sensitive Decision Tree.

Logistic Regression with SMOTE			
		Predicted	
		No HD/HA	HD/HA
Actual	No HD/HA	34445	11512
	HD/HA	1005	3774

(b) Logistic Regression with SMOTE.

Random Forest with SMOTE			
		Predicted	
		No HD/HA	HD/HA
Actual	No HD/HA	39882	6075
	HD/HA	2054	2725

(c) Random Forest with SMOTE.

Random Forest (no subsampling)			
		Predicted	
		No HD/HA	HD/HA
Actual	No HD/HA	45821	136
	HD/HA	4545	234

(d) Random Forest without subsampling.

Naïve Bayes (no subsampling)			
		Predicted	
		No HD/HA	HD/HA
Actual	No HD/HA	42305	3652
	HD/HA	3019	1760

(e) Naive Bayes without subsampling.

Naïve Bayes with SMOTE			
		Predicted	
		No HD/HA	HD/HA
Actual	No HD/HA	33542	12415
	HD/HA	1375	3404

(f) Naive Bayes with SMOTE.

Figure 8: Confusion matrices for classifiers of interest.

With that in mind, classifier performance under this experimental setup does appear to plateau at an F1 score of around 0.4. As such, there are number of avenues that may be worth investigating, including: appropriateness of F1 as an evaluation metric, the choice of imbalanced learning strategy, and further considerations for the limitations of this type of survey data.

Firstly, as concerns F1 score, given the aforementioned potential need to prioritise either precision or recall over the other, it may be more appropriate to use an alternative evaluation metric, such as F2 score. Secondly, given that both SMOTE oversampling and cost-sensitive learning resulted in comparable performance, it may be worth considering if other types of subsampling (e.g., undersampling with Tomek links) or some combination of cost-sensitive learning with subsampling might yield any improvement. Finally, and possibly most importantly, it should be acknowledged that the plateau in classification performance may also be related to the high likelihood that much of the variance in the class distribution is explained by factors not present, or obtainable, from the survey data in question. Given this, it may be beneficial to investigate the possibility of incorporating other external data in detecting heart disease or attack risk.

Conclusion

The primary goals of this project were to explore and evaluate a number of different classifiers and imbalanced learning strategies in their potential ability to detect heart disease or heart attack risk using survey data, as well as leverage explainability methods to identify influential features. Most classifiers performed significantly worse in the absence of either SMOTE oversampling or cost-sensitive learning, with the arguable exception of Naive Bayes which achieved relatively decent results out of the box, though still seeing large gains in recall when training with SMOTE.

The two best performing models in terms of either pure F1 score or recall were cost-sensitive SMOTE Random Forest, and cost-sensitive Decision Tree. All F1 scores for non-baseline classifiers were comparable across the board, thus overall effectiveness for the given classification task should be evaluated on alternative metrics, such as recall and training time complexity. In that regard, more complicated models such as the Neural Network, Random Forest and Bagging Classifier, given their need for excessive training times and hyperparameter tuning, do not provide sufficient, if any, improvement to classification performance for this task. This can also be extended to the comparison between SMOTE oversampling and cost-sensitive learning, as the former results in significant increases in training times, thus both cost-sensitive Logistic Regression and Decision Tree appear to be the best overall classifier variants.

Explanations generated by LIME and DiCE for non-baseline classifiers suggest (in descending order of importance) that respondent age, general health status and, to varying extents, history of stroke are the most important features in predicting heart disease or heart attack history. Significantly, the sole exception to this is the SMOTE Neural Network classifier, for which DiCE estimates gave higher weightings to history of stroke, age and recent history of poor physical health, in that order.

Possible areas for further work include: exploring feature selection and engineering techniques for addressing imbalanced classification performance; investigating different imbalanced learning strategies such as alternative subsampling methods and combining both cost-sensitive learning with subsampling methods; conducting a more rigorous examination and evaluation of LIME and DiCE explainability methods (as well as other alternatives); exploring possibilities for improving heart disease risk detection by combining classification on survey data with other strategies or datasets.

[7] Whether or not preventative measures taken on the basis of such health-risk screening would be more resource-effective compared to other possible venues is outside the scope of this project, but it may be worth consideration.

References

Boughorbel, S.; Jarray, F.; and El-Anbari, M. 2017. Optimal classifier for imbalanced data using matthews correlation coefficient metric. *PloS one* 12(6):e0177678.

Burkart, N., and Huber, M. F. 2021. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research* 70:245–317.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16:321–357.

Chawla, N. V. 2009. Data mining for imbalanced datasets: An overview. *Data mining and knowledge discovery handbook* 875–886.

Chicco, D., and Jurman, G. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21(1):6.

Chou, Y.-L.; Moreira, C.; Bruza, P.; Ouyang, C.; and Jorge, J. 2022. Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. *Information Fusion* 81:59–83.

Du, M.; Liu, N.; and Hu, X. 2019. Techniques for interpretable machine learning. *Communications of the ACM* 63(1):68–77.

Fernández, A.; Garcia, S.; Herrera, F.; and Chawla, N. V. 2018. Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research* 61:863–905.

Ganganwar, V. 2012. An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering* 2(4):42–47.

Krawczyk, B. 2016. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence* 5(4):221–232.

Kulesza, A. 2012. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning* 5(2-3):123–286.

Leevy, J. L.; Khoshgoftaar, T. M.; Bauder, R. A.; and Seliya, N. 2018. A survey on addressing high-class imbalance in big data. *Journal of Big Data* 5(1):1–30.

Ling, C. X., and Sheng, V. S. 2008. Cost-sensitive learning and the class imbalance problem. *Encyclopedia of machine learning* 2011:231–235.

López, V.; Fernández, A.; Moreno-Torres, J. G.; and Herrera, F. 2012. Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. open problems on intrinsic data characteristics. *Expert Systems with Applications* 39(7):6585–6608.

Mienye, I. D., and Sun, Y. 2021. Performance analysis of cost-sensitive learning methods with application to imbalanced medical data. *Informatics in Medicine Unlocked* 25:100690.

Mothilal, R. K.; Sharma, A.; and Tan, C. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 607–617.

Ng, K.; Steinhubl, S. R.; DeFilippi, C.; Dey, S.; and Stewart, W. F. 2016. Early detection of heart failure using electronic health records: practical implications for time before diagnosis, data diversity, data quantity, and data density. *Circulation: Cardiovascular Quality and Outcomes* 9(6):649–658.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Sahakyan, M.; Aung, Z.; and Rahwan, T. 2021. Explainable artificial intelligence for tabular data: A survey. *IEEE Access* 9:135392–135422.

Santos, M. S.; Soares, J. P.; Abreu, P. H.; Araujo, H.; and Santos, J. 2018. Cross-validation for imbalanced datasets: avoiding overoptimistic and overfitting approaches [research frontier]. *ieee ComputatioNal iNtelligeNCe magaziNe* 13(4):59–76.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.* 31:841.

Wilson, P. W.; D'Agostino, R. B.; Levy, D.; Belanger, A. M.; Silbershatz, H.; and Kannel, W. B. 1998. Prediction of coronary heart disease using risk factor categories. *Circulation* 97(18):1837–1847.

Yang, F.; Wang, H.-z.; Mi, H.; Cai, W.-w.; et al. 2009. Using random forest for reliable classification and cost-sensitive learning for medical diagnosis. *BMC bioinformatics* 10(1):1–14.

Zhu, Q. 2020. On the performance of matthews correlation coefficient (mcc) for imbalanced dataset. *Pattern Recognition Letters* 136:71–80.

Appendix

Model	Hyperparameters
Logistic Regression	’solver’: [’newton-cg’, ’lbfgs’, ’liblinear’], ’C’: loguniform(1e-3, 1e4)
Naive Bayes	None
Linear SVM	’max iter’: list(range(5,10)), ’alpha’: 10.0**-np.arange(1,7)
Decision Tree	’max depth’: list(range(5,15)), ’min samples split’: [2, 5, 10, 20, 50], ’min samples leaf’: [1, 3, 5, 10, 20, 50]
Random Forest	’max depth’: list(range(5,15)), ’min samples split’: [2, 5, 10, 20, 50], ’min samples leaf’: [1, 3, 5, 10, 20, 50], ’n estimators’: [10, 20, 50, 100]
Bagging Classifier	’max depth’: list(range(5,15)), ’min samples split’: [2, 5, 10, 20, 50], ’min samples leaf’: [1, 3, 5, 10, 20, 50], ’n estimators’: [10, 20, 50, 100]
Neural Network	’activation’: [’logistic’, ’tanh’, ’relu’], ’solver’: [’lbfgs’, ’sgd’, ’adam’], ’hidden layer sizes’: [(20,10), (30), (20,20)], ’alpha’: 10.0 ** -np.arange(1, 7)

Table A.1: Hyperparameters chosen for tuning for all classification models.

Feature	Description
HeartDiseaseorAttack	Binary target feature [0:1]. Indicates whether the respondent has experienced coronary heart disease or myocardial infarction.
HighBP	Binary feature [0:1]. Indicates whether respondent has been diagnosed with high blood pressure by a healthcare professional.
HighChol	Binary feature [0:1]. Indicates whether respondent has been diagnosed with high cholesterol by a healthcare professional.
CholCheck	Binary feature [0:1]. Indicates whether respondent has had their cholesterol checked in within the last five years.
BMI	Ordinal feature. Indicates respondents self-reported Body Mass Index (BMI).
Smoker	Binary feature [0:1]. Indicates whether respondent has smoked at least 100 cigarettes in their lifetime.
Stroke	Binary feature [0:1]. Indicates whether respondent has a history of stroke.
Diabetes	Ordinal feature [0:2]. 0 indicates no diabetes or diabetes only during pregnancy; 1 indicates prediabetes or borderline diabetes; 2 indicates diabetes
PhysActivity	Binary feature [0:1]. Indicates whether respondent has reported doing any physical activity or exercise outisde of work in the past 30 days.
Fruits	Binary feature [0:1]. Indicates whether respondent consumes fruit at least once a day.
Veggies	Binary feature [0:1]. Indicates whether respondent consumes vegetables at least once a day.
HvyAlcoholConsump	Binary feature [0:1]. Indicates whether respondent drinks more than 14 alcoholic drinks per week if they are male, or more than 7 alcoholic drinks per week if they are female.
AnyHealthcare	Binary feature [0:1]. Indicates whether respondent possesses any type of healthcare coverage.
NoDocbcCost	Binary feature [0:1]. Indicates whether respondent has opted not to see a doctor in the last 12 months due to healthcare costs.
GenHlth	Ordinal feature [1:5]. Indicates self-reported general health, with 1 being ”Excellent” and 5 being ”Poor”.
MenthHlth	Ordinal feature [0:30]. Indicates number of self-reported days in which respondent experienced mental health issues in the last 30 days.
PhysHlth	Ordinal feature [0:30]. Indicates number of self-reported days in which respondent experienced physical health issues in the last 30 days.
DiffWalk	Binary feature [0:1]. Indicates whether respondent has serious difficulty walking or climbing stairs.

Sex	Binary feature [0:1]. Indicates biological sex of respondent. 0 is female, 1 is male.
Age	Ordinal feature [1:13]. Indicates age category of respondent. Each category represents a 5 year increment in increasing order, with the sole exception of categories 1 and 13, where 1 indicates the range between 18-24, and 13 indicates the range from 80 and over.
Education	Ordinal feature [1:6]. Ordered categories from 1 to 6, where 1 indicates having never attended school or attended kindergarten only, and 6 indicates having attended college or university for 4 years or more.
Income	Ordinal feature [1:8]. Ordered categories from 1 to 8, where 1 indicates a household income of less than \$10,000 and 8 indicates a household income of \$75,000 or more.

Table A.2: Descriptions of all heart disease dataset features.