Technical Audit of Heart Disease ADS

Guilherme Albertini and Martin Keenan

May 11, 2023

1 Background

Automated Decision Systems (ADS) have become widespread in the medical domain. Despite their success in optimizing patients' medical outcomes, the societal biases prevalent in these systems are often ignored. In this report we document a technical audit of machine learning model that could hypothetically serve as an ADS. The model was published by Wessam Salah Walid in a public Kaggle notebook titled "Heart Disease Prediction (EDA Modelling)(Acc 95%)" [1].

This ADS aims to predict whether a patient will have heart disease given the status of other key risk factors. This would enable health care providers to get a head start in screening for and mitigating the indicators that may eventually compound and lead to this potentially fatal outcome. The ADS further attempts to identify which factors have the most direct influence on the likelihood of heart disease, though this is not explicitly answered in the analysis. Mr. Walid states that his goal is to "draw conclusions about the factors that contribute to heart disease" and predict the presence of heart disease among patients in the dataset. While this model is not used in any productionized system, the typical stakeholders for such a model would be a health system or set of providers who use the ADS to help diagnose patients, as well as the patients themselves who expect that the ADS will provide them with an accurate and fair assessment.

Our technical audit is inspired by the framework summarized by Koshiyama et al. [2]. This framework breaks down the deployment of automated decision systems into five stages: (i) data and task setup; (ii) feature pre-processing; (iii) model selection; (iv) post-processing and reporting; and (v) productionizing and deploying. With the exception of the last stage, since this model is not used in any deployed system, we analyze each with respect to explainability, robustness, and fairness. In Section 2, we discuss the data used to train the ADS as well as the system's output. In Section 3, we analyze the feature pre-processing and model selection and training algorithms used by Mr. Walid. In section 4, we evaluate the system's reported accuracy metrics in addition to our own fairness metrics. Finally, in Section 5 we conclude with our thoughts on how well the ADS performs and how we think it could be improved.

2 Input and Output

The dataset originates from the records collected by the CDC Behavioral Risk Factor Surveillance System (BRFSS), which conducts annual telephone surveys to gather data on the health status of U.S. residents [3]. This yearly health-related survey collects information on over 400,000 adults regarding their health-related behaviors, conditions and use of preventive services. The dataset contains eighteen features, a binary target variable for the prevalence of heart disease, several sensitive features (race, sex, and age) and other features such as BMI and alcohol usage which serve as key indicators of heart disease. Table 1 displays the name, datatype, and definition for each of these. There are no missing values for any feature in the dataset.

Dataset Features		
Name	**Datatype**	**Definition**
HeartDisease	Binary	Reported having coronary heart disease (CHD) or myocardial infarction (MI).
Asthma	Binary	Reported having asthma.
SkinCancer	Binary	Reported having skin cancer.
Stroke	Binary	Reported having a stroke.
KidneyDisease	Binary	Reported having kidney disease.
Diabetic	Categorical	Reported having diabetes.
BMI	Float	Body Mass Index.
PhysicalHealth	Float	Number of days during the past 30 days that the respondent's physical health was not good.
MentalHealth	Float	Number of days during the past 30 days that the respondent's mental health was not good.
GenHealth	Categorical	The respondent's classification of their health (e.g. "very good").
PhysicalActivity	Binary	Reported physical activity or exercise during the past 30 days other than their regular job.
DiffWalking	Binary	Has serious difficulty walking or climbing stairs.
Smoking	Binary	Smoked at least 100 cigarettes in entire life.
AlcoholDrinking	Binary	Adult men having more than 14 and adult women having more than 7 drinks per week.
SleepTime	Float	Average number of hours of sleep in a 24-hour period.
Sex	Binary	Male or female (Male=1).
Race	Categorical	Race/Ethnicity.
AgeCategory	Categorical	Fourteen-level age category.

Table 1: Datatypes and definitions for features in the dataset.

The distributions of binary features are shown in Figure 1. The prevalence of heart disease in the dataset is about 8.5%. Other major health conditions occur at around the same frequency. The behavioral features show that most respondents did some form of physical activity in the past month, about half are smokers, and about 14% reported difficulty walking. The sensitive feature of sex also shows that the dataset is split almost evenly between males and females.

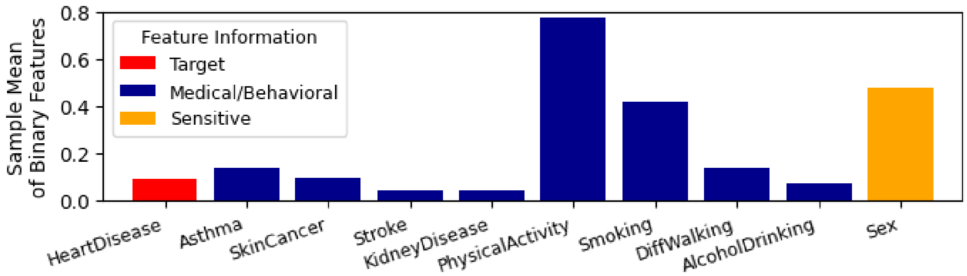

Figure 1: The proportion of the dataset with positive values for each of the binary features.

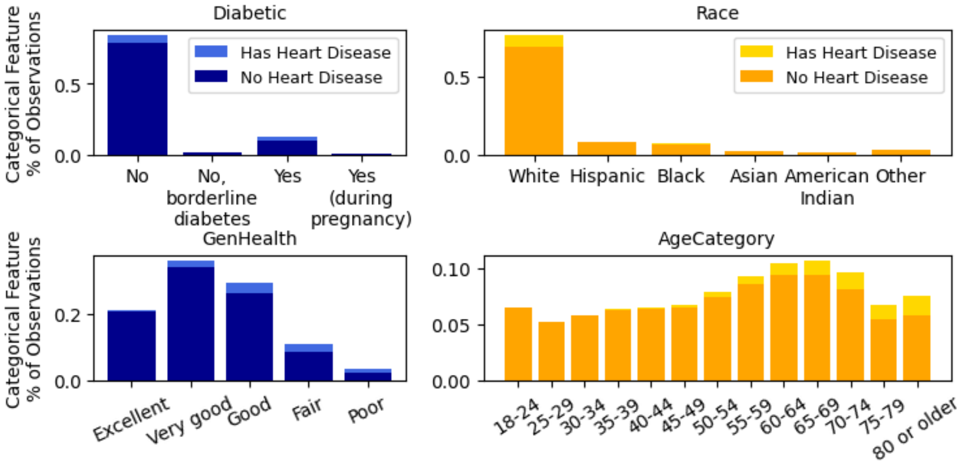

Figure 2: Distribution of values for each of the categorical features in the dataset. Sensitive features are shown in orange/yellow.

The distributions of nonbinary categorical features are displayed in Figure 2. Diabetes occurs among the patients at a similar rate as the other major health conditions. The majority of patients also classify their health as “good” or “very good”. The two sensitive categorical features are race and age. The dataset contains mostly white patients and age appears to be skewed towards older patients. It also appears that age is an important predictor of heart disease, as case rates are generally increasing with age.

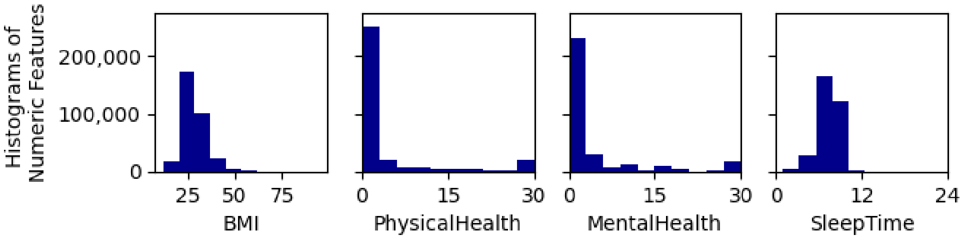

Figure 3: Histograms of numeric features in the dataset.

Figure 3 displays the distribution of numeric features. The features for BMI and sleep time both seem to roughly follow a Gaussian distribution. The physical and mental health features, which are self-reported scores, have peaks at the minimum score of zero and the maximum score of thirty. The pairwise correlations between these and the binary features are shown in Figure 4. The features that have the strongest correlation with the target are “Stroke,” “PhysicalHealth,” “PhysicalActivity,” and “DiffWalking.” There are also strong correlations between non-target features, such as “MentalHealth” and “PhysicalHealth,” and “DiffWalking” and “PhysicalActivity.”

The designer optimizes for accuracy in predicting whether an adult has heart disease. In addition, he compares the accuracy metrics generated across different models in using majority class label (“hard voting”) rules. This means that the predicted class label (i.e., heart disease or no heart disease) for a particular sample was assigned based on the majority class label produced by the models individually. In cases of a tie, the voting classifier selected the class label based on the most recently generated label. The accuracy metric produced by this classification scheme simply characterizes how well the ensemble was able to classify a data point correctly based on a reference value (i.e., whether a patient was predicted to have heart disease given the actual status of disease). Formally, this is the number of true positives (TP) and true negatives (TN) divided by the sum of the TP, TN, FP, and FN. The generated precision, recall, and F1 scores were, rather dubiously, not dissected.

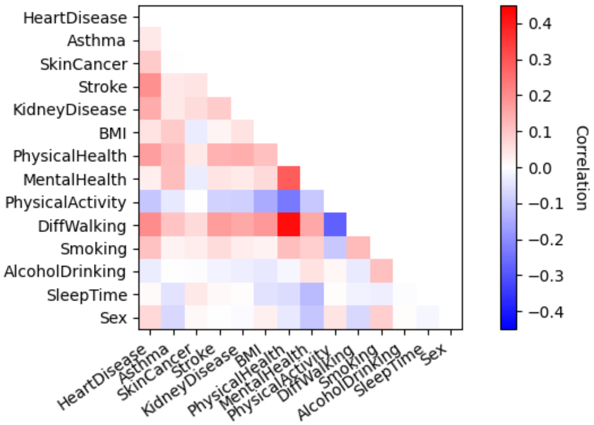

Figure 4: Pairwise correlations between features in the dataset.

3 Implementation and Validation

3.1 Analyzing Encoding with t-SNE

Proper encoding can often help to prevent bias in modeling. If categorical features are not properly encoded, the model may assign different weights to different categories based on their numerical representations, which could lead to biased predictions. A fairly common mistake is to just assign a unique numerical value to each category of a feature, disregarding any order it holds (for ordinal variables) and, similarly, enforce integer orderings for nominal features that don't naturally present any order. For example, in encoding [fair, poor, fair, excellent, poor] into [1, 2, 1, 3, 2] using LabelEncoder, the imposed ordinality suggests that the average of fair and excellent is poor. Observe that only "AgeCategory" and "GenHealth" have a natural ranking for categorical features. Unfortunately, the author decides to use LabelEncoder to encode all of the categorical features without the regard to any natural ranking orders (or lack thereof) and the weights that can be arbitrarily assigned to each feature value.

We adjust encoding (using both OneHotEncoder and OrdinalEncoder) to address these issues and compare the t-SNE visualization from the original to adjusted encoding schemes. The result? Data is more separable using the proper encoding scheme, most likely leading to more accurate modeling downstream. Note that t-SNE (t-Distributed Stochastic Neighbor Embedding) is a dimensionality reduction technique that is commonly used for visualizing high-dimensional data in two or three dimensions. The basic idea behind t-SNE is to find a low-dimensional representation of the data that preserves the relationships between the points in the high-dimensional space as much as possible. We clearly see that the naive method used by the designer does not lead clear indications of separability of the feature space in Figure 5. When properly encoding categorical features, we start to see separability in Figure 6.

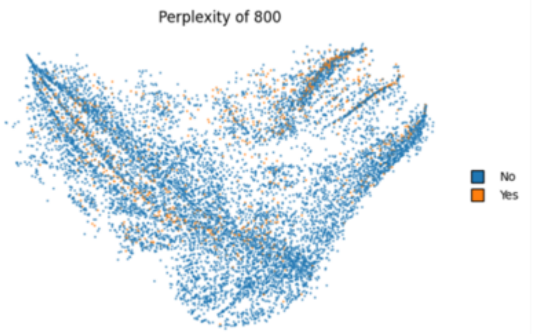

Figure 5: By only using the LabelEncoder for categorical features, the t-SNE Visualization across perplexity values still suggests this data is hardly separable. A visualization using perplexity 800 is displayed above.

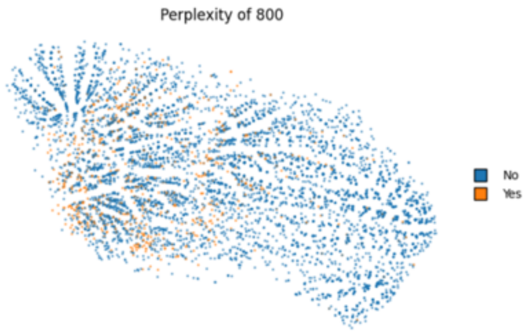

Figure 6: Using both LabelEncoder and OrdinalEncoder (for GenHealth and AgeCategory features), the t-SNE Visualization across several perplexity values produced more separable data. A visualization using perplexity 800 is displayed above.

The perplexity parameter determines the width of the Gaussian kernel that is used to compute the pairwise similarities between the high-dimensional data points. Specifically, it determines the number of nearest neighbors that are considered for each point when computing the similarities. A higher perplexity value means that more neighbors are considered, which can result in a smoother and more global representation of the data. A lower perplexity value means that fewer neighbors are considered, which can result in a more local and detailed representation of the data.

3.2 SHAP Values

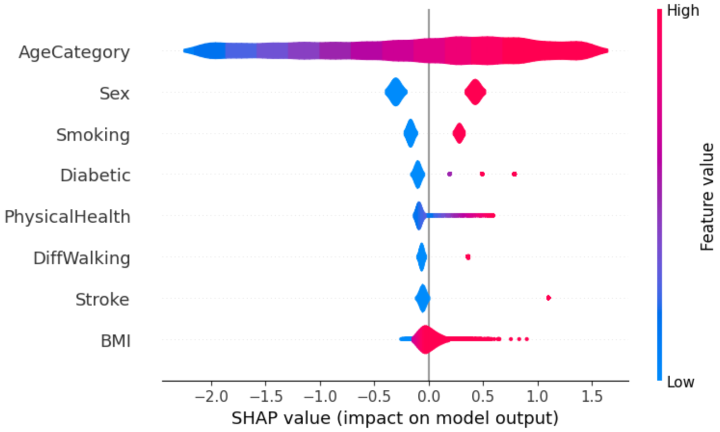

Figure 7: Feature importances for the Logistic Regression Model.

A SHAP (SHapley Additive exPlanations) summary plot graph provides insights into the feature importance and impact on the predictions made by a machine learning model. It helps to explain the contribution of each feature in the final prediction for individual instances or the overall model behavior. The importance is measured by the absolute mean SHAP value, which represents the average impact of a feature on the predictions across all instances. Higher absolute mean SHAP values indicate more influential features. Longer bars signify larger effects, both positive (towards a Heart Disease prediction) and negative (towards a No Heart Disease prediction), while shorter bars indicate less influential features. We observe that most of the author's models deem "AgeCategory", "DiffWalking", and "Sex" as features having the highest average impact on the predictions of Heart Disease across all instances; specifically, we see that "AgeCategory" is an overall key indicator of heart disease.

3.3 Sampling Techniques

In addition to the model analyzed in this report, the programmer also trained two other versions of the model where he oversampled and undersampled the dataset based on the target class. However, we omit analysis of the outcomes of these models because we believe his approach to sampling was flawed. Instead of only oversampling the training data, he oversampled both the training data and the test data and then reported the model's accuracy on the new oversampled test set. This means that the evaluation metrics are not comparable to the metrics reported for the model in this audit (inasmuch as likely being inflated due to leakage and overfitting) and are not generalizable to real-world use. An appropriate use of oversampling would have trained a model on oversampled training data, then evaluated results on the untouched test dataset.

4 Outcomes

The programmer decided to optimize and evaluate his model for accuracy. The accuracy of his final ensemble model on the test set is 0.916. His model seems to be successful, but his evaluation ignores the fact that the target class is heavily imbalanced. In fact, a simple baseline model that always predicts zero would have an accuracy of 0.92, which is slightly better than 0.916.

However, the fact that the programmer decided to optimize on accuracy does not mean accuracy is the most appropriate metric. Given that we are dealing with imbalanced data, it is also useful to look at the precision and recall, to see how confident the system is when it predicts "yes" and what proportion of the true heart disease cases it is able to detect. The precision and recall of the model are 0.56 and 0.06, respectively. These metrics are more promising than accuracy because they show that the model has some usefulness for predicting the positive class. In comparison, the baseline prediction of all zeros would have zero precision and zero recall.

It is also important to evaluate the error rates. The model's false negative rate (FNR) and false positive rate (FPR) are 0.939 and 0.005, respectively. Because of the medical context, it is particularly important to pay attention to the FNR. If a patient receives a false positive, it may result in increased medical care, which could be a time and financial burden, but it would not have any serious health consequences. However if a patient receives a false negative, then they will not be aware that they have a serious life-threatening issue and may not receive the medical care they need. Ideally the model's FNR of 0.939 would be closer to zero to avoid so many false negatives. If the programmer were to retrain his model, we would suggest that he focus more on lowering the FNR. One possible solution would be to use a probabilistic model like logistic regression instead of the hard-voting ensemble classifier, so that he could use an ROC curve to choose a score threshold that gives a desirable FNR.

In terms of fairness, we can see how the evaluation metrics differ across the sensitive features age and sex. Table 2 shows evaluation metrics broken down by race. It is clear from this table that the system seems to vastly underperform for Asians. The selection rate for Asians is almost zero, which leads to zero precision and recall for this group, as well as a FNR of one. This is not acceptable because the true prevalence of heart disease among Asians in the test set is 3%. For the rest of the races, the FNR is between approximately 0.92 and 0.95, with the lowest for American Indians and highest for Hispanics. In addition to decreasing overall FNR, we would also suggest that the programmer increase parity between the error rates across races to make this system fairer.

The metrics by sex are shown in Table 3. The precision for females and males is similar, but recall for males is about twice as high (0.077 versus 0.038). Part of this difference in recall can likely be attributed to males having a selection rate that is almost three times as high as females. Males also have a significantly lower FNR than females, at 0.923 compared to 0.962. In addition to race, we would also suggest that the programmer attempt to have more parity between the FNR for males and females in his ADS.

Race	Accuracy	Selection Rate	FNR	FPR	Precision	Recall
American Indian	0.902	0.013	0.918	0.005	0.643	0.082
Asian	0.966	0.001	1.000	0.001	0.000	0.000
Black	0.925	0.011	0.928	0.006	0.481	0.072
Hispanic	0.950	0.005	0.950	0.002	0.519	0.050
Other	0.920	0.010	0.931	0.005	0.571	0.069
White	0.910	0.010	0.939	0.005	0.565	0.061

Table 2: Evaluation metrics broken down by race.

Sex	Accuracy	Selection Rate	FNR	FPR	Precision	Recall
Female	0.933	0.005	0.962	0.002	0.547	0.038
Male	0.897	0.014	0.923	0.007	0.559	0.077

Table 3: Evaluation metrics broken down by sex.

Table 4 shows the metrics broken down by the intersection of sex and race. This allows us to detect any disparities within the groups shown in the previous tables. This table reveals that the difference in FNR between males in females is not an issue for American Indians, but is mainly an issue among Blacks, Hispanics, White, and Other races. Furthermore, the FNR difference between males and females is largest among Blacks and Hispanics. In particular, Black men have the lowest FNR of any intersectional group, while Black women have one of the highest FNRs. If the programmer were to update his model, we would suggest to first try to create parity across race and sex separately, then look at the intersectional metrics to see if there are any significant disparities between intersectional groups.

5 Summary

This dataset was useful for training a heart disease ADS as it contained a very large sample of adults along with some important health features. However, if this were to be deployed publicly it would be important to know who the target population is. For example if it were used internationally, it would be better for the data to include non-Americans to prevent the model from being biased. There are also other important predictors of heart disease (e.g. height, weight, blood pressure, etc.) that are not included in the dataset or may violate privacy if included in a public dataset. Therefore, researchers who have secure access to sensitive health data would likely be able to train better predictive models.

Overall, we believe that the current methodology should not focus solely on accuracy. It should instead focus on a balance of precision, recall, and false negative rate, so that the system detects as many instances of heart disease

		Accuracy	Selection Rate	FNR	FPR	Precision	Recall
Race	Sex						
American Indian	Female	0.904	0.010	0.921	0.002	0.833	0.079
	Male	0.900	0.017	0.915	0.009	0.500	0.085
Asian	Female	0.975	0.000	1.000	0.000	0.000	0.000
	Male	0.959	0.002	1.000	0.002	0.000	0.000
Black	Female	0.934	0.005	0.957	0.003	0.533	0.043
	Male	0.912	0.020	0.896	0.012	0.459	0.104
Hispanic	Female	0.949	0.002	0.980	0.001	0.500	0.020
	Male	0.951	0.008	0.913	0.004	0.524	0.087
Other	Female	0.929	0.008	0.948	0.005	0.444	0.052
	Male	0.911	0.011	0.918	0.004	0.667	0.082
White	Female	0.931	0.005	0.962	0.002	0.545	0.038
	Male	0.887	0.015	0.924	0.007	0.572	0.076

Table 4: Evaluation metrics broken down by race and sex.

as possible, within reason. We would recommend the author to update his approach to feature encoding and oversampling to improve the overall performance of his model. We would also recommend that he optimize his model to achieve a low false negative rate, and to also increase false negative rate parity among sensitive groups, perhaps by setting different score thresholds for each.

References

[1] Wessam Salah Walid, “Heart Disease Prediction (EDA Modelling)(Acc 95%),” 2022, available at https://www.kaggle.com/code/wessamwalid/heart-disease-prediction-eda-modelling-acc-95.

[2] Koshiyama, Adriano and Kazim, Emre and Treleaven, Philip and Rai, Pete and Szpruch, Lukasz and Pavey, Giles and Ahamat, Ghazi and Leutner, Franziska and Goebel, Randy and Knight, Andrew and others, “Towards algorithm auditing: a survey on managing legal, ethical and technological risks of AI, ML and associated algorithms,” 2021.

[3] CDC Behavioral Risk Factor Surveillance System, “Personal Key Indicators of Heart Disease,” 2020.