Heart Disease ADS Report

Authors: Tina Shi (ts3875) & Erin Kong (ek3631)

1 Background

1.1 Purpose and Goals

The heart disease ADS from Kaggle aims to analyze medical and lifestyle factors to predict heart disease. Its goals are rooted in the healthcare industry, where the early detection of heart disease will be able to significantly reduce the costs of treatment and hopefully extend lifespan. The ADS wants to be able to accurately identify risk factors and develop a predictive model. However, fairness should be balanced to avoid biases and disparities in healthcare across demographic groups for protected characteristics such as age, sex, and race.

1.2 Tradeoffs

Our ADS involves different tradeoffs for different stakeholders, depending on competing priorities. While a complex model may be able to perform and predict better than a simple model, they are harder to explain. Healthcare professionals and patients alike prioritize accuracy, but a patient may not be able to fully understand a complex model. Thus, this presents a tradeoff between performance and explainability.

Ensuring fairness across demographic groups may reduce the model's overall accuracy. Minority groups might prioritize fairness, while the majority group would be affected with a decrease in accuracy. In this case, we face a tradeoff between accuracy and fairness.

In healthcare, minimizing false negatives is important to avoid missing a diagnosis. However, doing so could increase the false positive rate. Healthcare providers and patients may want to prioritize false negative rate to avoid missing diagnoses. However, patients may also want to prioritize false positive rate to avoid unnecessary tests and treatments associated with a false diagnosis. This creates a tradeoff between false positive rate and false negative rate.

2 Input and Output

2.1 Data Source

The data used for this ADS is sourced from the 2020 Center for Disease Control and Prevention (CDC) annual survey data and includes responses from over 400,000 adults. The survey was conducted via telephone across all 50 states, the District of Columbia, and three U.S. territories. The original dataset consisted of over 300 features which were reduced to the 18 most relevant features by Kaggle user Kamil Pytlak.

The dataset contains 319,795 rows and has no null values. The reduction in the number of rows compared to the number of survey responses indicates that some individuals were excluded from the dataset. It is revealed that the user Kamil Pytlak excluded responses that were missing data. The dataset has since been updated in 2022 based on the 2022 CDC survey, but the ADS we are exploring uses the 2020 data. One important note is that while there are no missing values, the binary classification for heart disease is heavily imbalanced.

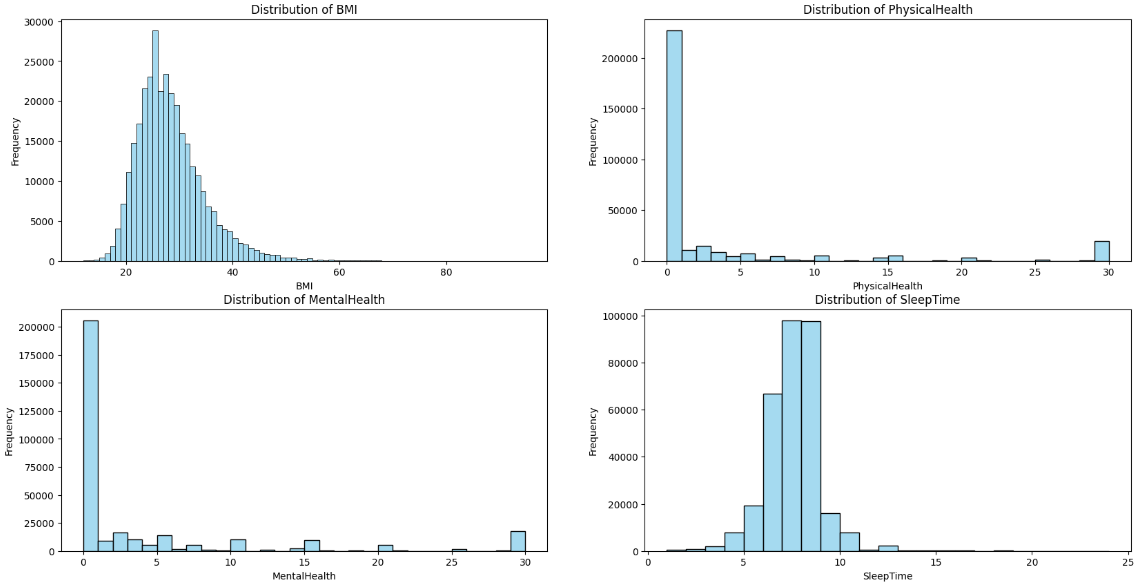
2.2 Feature Overview

Each row in the dataset corresponds to one survey response, with the following features:

Feature	**Datatype**	**Description**
HeartDisease (Target)	Categorical (Binary)	Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)
BMI	Numerical (Float)	Body Mass Index (BMI)
Smoking	Categorical (Binary)	Smoked at least 100 cigarettes ever (5 packs = 100 cigarettes)
AlcoholDrinking	Categorical (Binary)	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week)
Stroke	Categorical (Binary)	Ever been diagnosed with a stroke
PhysicalHealth	Numerical (Integer)	Number of days in the past 30 days where physical health was not good
MentalHealth	Numerical (Integer)	Number of days in the past 30 days where mental health was not good
DiffWalking	Categorical (Binary)	Serious difficulty walking or climbing stairs
Sex	Categorical (Binary)	Male or Female
AgeCategory	Categorical (Ordinal)	14 age categories, grouped in 5-year intervals
Race	Categorical (Nominal)	Imputed race/ethnicity value
Diabetic	Categorical (Nominal)	Ever been diagnosed with diabetes
PhysicalActivity	Categorical (Binary)	Doing physical activity or exercise in the past 30 days, other than regular job
GenHealth	Categorical (Nominal)	Perception of their own general health
SleepTime	Numerical (Integer)	Average hours of sleep in a 24 hour time period
Asthma	Categorical (Binary)	Ever been diagnosed with asthma
KidneyDisease	Categorical (Binary)	Ever been diagnosed with kidney disease (excluding kidney stones, bladder infection, or incontinence)
SkinCancer	Categorical (Binary)	Ever been diagnosed with skin cancer

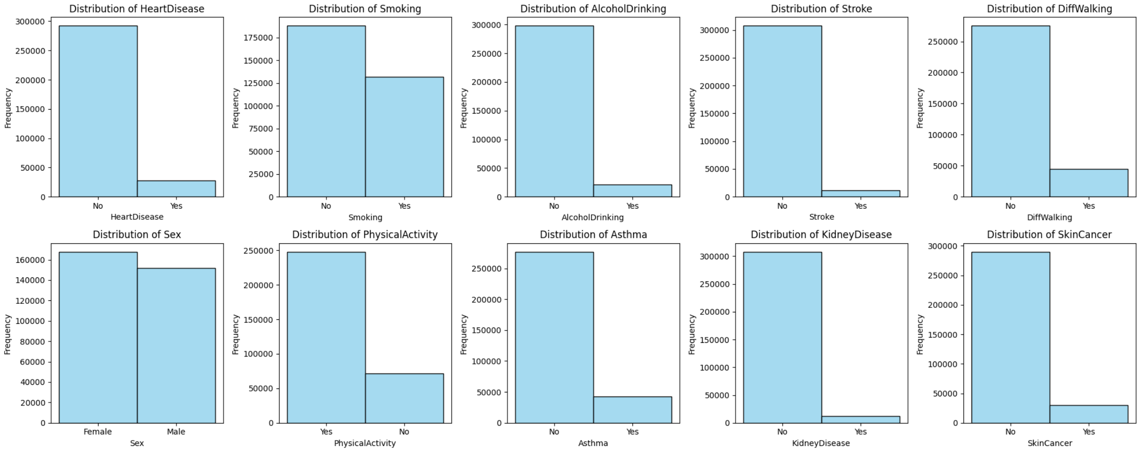
Note once again that there are no null values in this 2020 dataset.

The distribution of each numerical feature (BMI, PhysicalHealth, MentalHealth, SleepTime) is shown below:

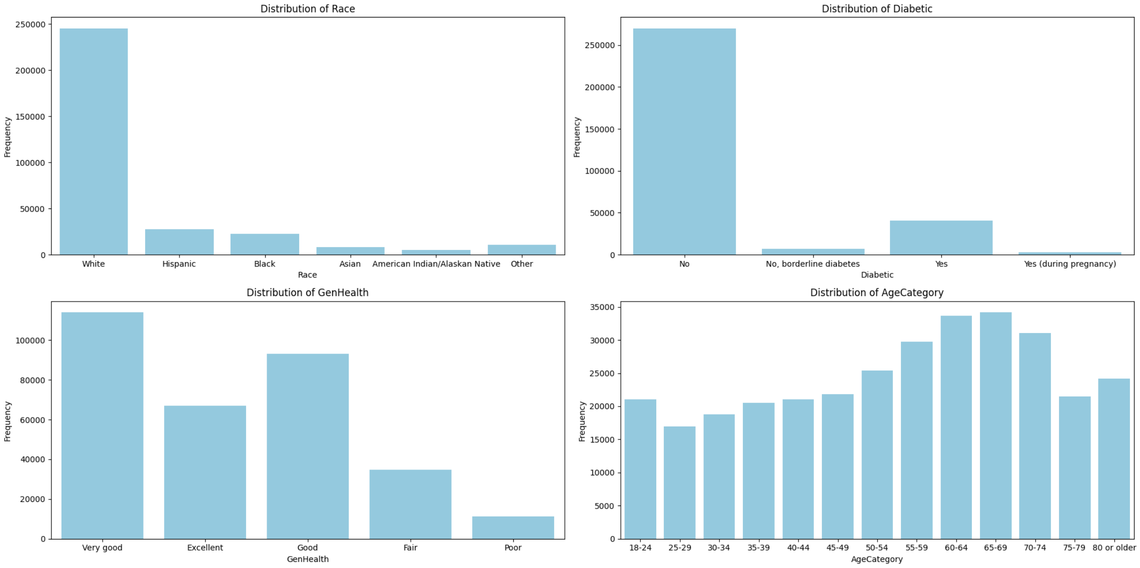


We notice that the distribution for these features are right-skewed. An overwhelming majority of the respondents reported having zero days in the last 30 days where they felt their mental or physical health was not good. The distribution of BMI follows a distribution closely to what we expected; most respondents fall in the healthy to obese BMI range of 20-30. For sleep, most respondents reported a standard 7-8 hours, with some having 6 hours of sleep.

The counts of each categorical feature are shown below. First, the binary categories (HeartDisease, Smoking, AlcoholDrinking, Stroke, DiffWalking, Sex, PhysicalActivity, Asthma, KidneyDisease, SkinCancer):



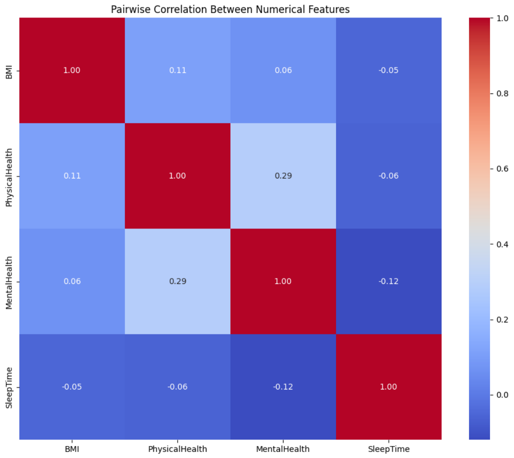
Next, the non-binary categories (Race, Diabetic, GenHealth, AgeCategory):



Looking at the distribution of age groups, most of the respondents lie towards the older end, primarily in their 60s and 70s. A majority of the respondents responded they do not consider themselves heavy alcohol drinkers and have never been diagnosed with a stroke, kidney disease, or skin cancer. To a lesser extent, a majority of the respondents reported having no difficulty walking and never having been diagnosed with asthma.

About 78% of respondents said they engage in physical activity outside of their job. However, only 59% of respondents reported never having smoked at least 100 cigarettes. The sex distribution is roughly balanced, with 48% male and 52% female. However, our target variable, heart disease, is heavily imbalanced: 292,422 reported no heart disease and 27,373 reported yes.

Next, we looked at the pairwise correlation between numerical features.



We observed that there is a small positive correlation between physical and mental health. However, we don't think this is significant enough to suggest a strong relationship. Rather, there could be noise that creates this correlation.

2.3 Output

The system produces a binary output, "Yes" and "No".

The HeartDisease column of the dataset indicates whether the respondent reported as ever having heart disease. The ADS aims to predict this value given the respondent's other features. Thus, the output "No" means that the model predicts the respondent does not have heart disease. The output "Yes" means that the model predicts the respondent has heart disease.

3 Implementation and Validation

3.1 Data Cleaning and Preprocessing

The dataset used in the ADS consists of 319,795 responses and 18 features related to heart disease risk factors. It was already cleaned by dropping records with null values and selecting only the most relevant variables (from the original 400+ features). The ADS follows a standard exploratory data analysis (EDA). Visualizations were created to understand feature distribution and relationship. Categorical features were encoded using a OrdinalEncoder, which transforms categorical variables by encoding unique values as unique integers (different from one-hot encoding where columns are created for each unique value). Numerical features were also analyzed to understand their distribution.

3.2 Model Implementation

The ADS split the dataset into train and test sets using a 90/10 split. It then trained KNeighborsClassifier, LogisticRegression, XGBClassifier, and ExtraTreesClassifier models using default parameters. These models were then evaluated on accuracy, precision, recall, and F1-score.

3.3 Validation

The models were validated by using the training set (10% of the data) to test the model's performance. However, this split may not be representative of the whole dataset and we should be using cross-validation with multiple splits to reliably assess the model's performance. This model meets it goals of identifying risk factors for heart disease, but there are limitations with its procedure.

The dataset used is large and the selected features correlate with heart disease. However, it is severely limited by the class imbalance of the target variable; only 9% of respondents had heart disease. As a result, accuracy is a misleading metric: if the model were to simply predict "No" for every case, it would be correct 91% of the time, because 91% of the respondents do not have heart disease. Additionally, all models have low recall rate, meaning that they miss many of the actual heart disease cases. Since the class imbalance is not addressed when selecting a suitable model, the models are biased towards predicting "No", which is ineffective when we want to predict positive cases.

4 Outcomes

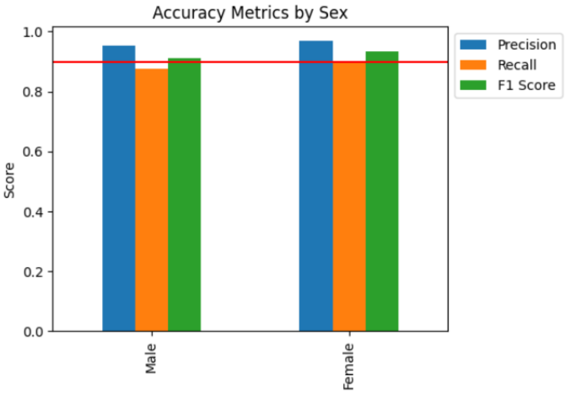
The evaluation metrics for each model in the ADS are shown below:

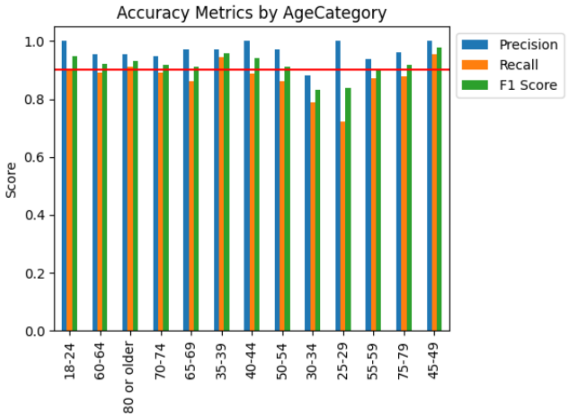
Model	Accuracy	Precision	Recall	F1-score
KNeighborsClassifier	0.9025	0.3203	0.0774	0.1247
LogisticRegression	0.9116	0.5424	0.0893	0.1533
XGBClassifier	0.9121	0.5580	0.0955	0.1631
ExtraTreesClassifier	0.8932	0.3046	0.1492	0.2003

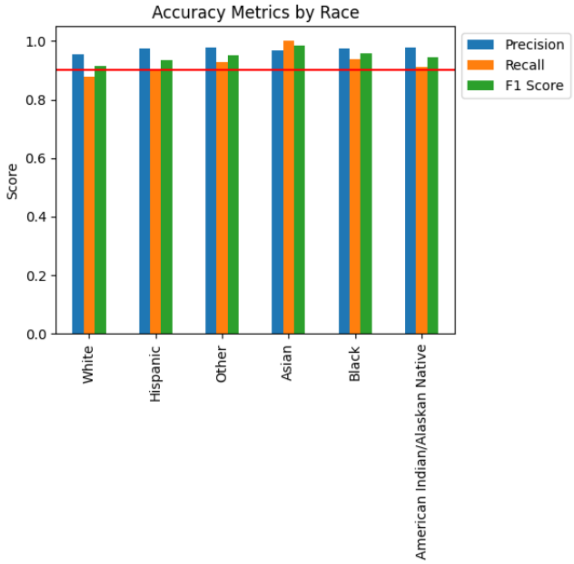
Given that 91.44% of the dataset does not have heart disease, there is a significant class imbalance in the HeartDisease target variable. All models perform worse than simply predicting "No" for all cases when evaluating the accuracy metric. This was expected, given that accuracy is unreliable in nature if there is a class imbalance.

In this medical context, the ADS prioritized selecting the model with the highest recall (maximizing the rate of positive cases we correctly identify), which is crucial in this setting. The model chosen is the ExtraTreesClassifier, which performs well with imbalanced datasets like ours.

First, we will compare the model's accuracy across different subgroups. We will use precision, recall, and F1-score as our accuracy metrics. Recall is particularly valuable in a healthcare setting because we want to catch as many positive cases as possible. Precision allows us to assess how accurate our positive predictions are. The F1-score acts as a balanced measure between recall and precision. The plots below show these metrics across subgroups of the sensitive attributes sex, age, and race. A red horizontal line is shown at 0.9 to represent our chosen baseline score of 90%.







Accuracy metrics are similar between males and females. However, when comparing subgroups by age, we start to see variations. Most notably, accuracy metrics are much lower for the 25-29 and 30-34 age categories. This could be a result of the fact that these two categories having less samples (less than 20,000) compared to other age groups (more than 20,000). A smaller sample size means our model cannot predict well for these age groups.

Race subgroups are generally comparable across all accuracy metrics. Asians, however, exceed well above the 90% baseline across all three metrics. Interestingly, whites overwhelmingly dominate the dataset, but have the lowest accuracy. This seems counterintuitive, but it could be explained by the greater variation in features within this larger group, making it harder for the model to predict accurately. In contrast, Asians may have less variation in their features, which makes it easier for the model to predict accurately.

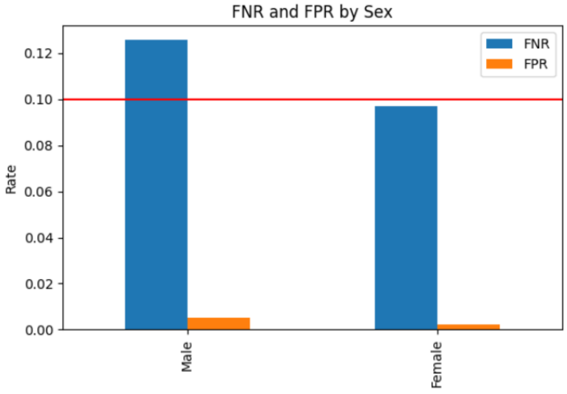
Next, we will compare how fair the model is across different subgroups using the fairness metrics of demographic parity, false negative rate, and false positive rate. The false negative rate should be low as possible because it represents how often we missed positive cases, which is undesirable in a healthcare setting. Similarly, the false positive rate should also be low because it represents how many cases we misclassify as positive, leading to a misallocation of resources on both patient and hospital sides.

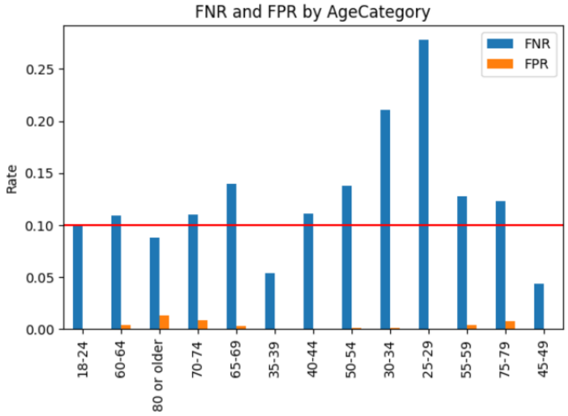
First, the demographic parity difference for each sensitive feature is shown in the table below:

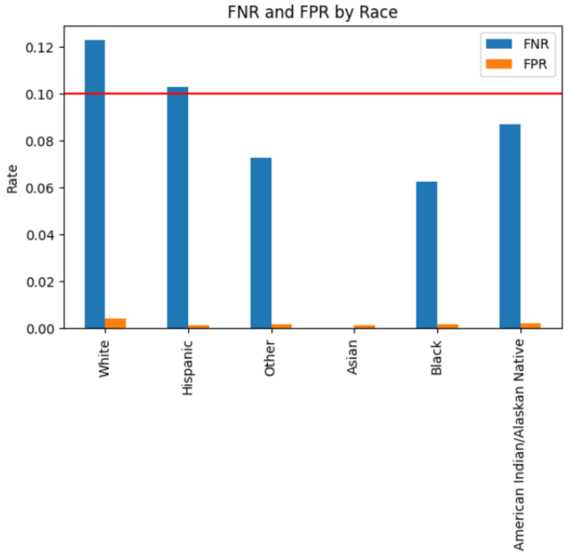
Sensitive Attribute	**Demographic Parity Difference**
Sex	0.0337
AgeCategory	0.2179
Race	0.0465

We notice that the difference in demographic parity is relatively low for the sensitive attributes sex (3.37%) and race (4.65%), indicating that the model predicts positive outcomes almost equally across these groups. However, the demographic parity difference is significantly higher for age, with a difference of 21.79%. This suggests that the model is much more likely to predict heart disease for some age groups than others, indicating that the model is sensitive to age. This could reflect the real world trend where age is correlated with an increased risk of heart disease.

The graphs below show the false negative rate and false positive rate across the subgroups of analysis. A red horizontal line is shown at 0.1 to represent a 10% baseline score.







The first observation is that false positive rates are low across all subgroups, meaning that we do not misclassify positive outcomes often. However, the false negative rate is considerably higher than false positive rates for all subgroups. This indicates that we are not catching a lot of positive cases, which is not ideal in a healthcare setting.

For the sex subgroups, males have a slightly higher false negative rate, about 2% higher than females, which is likely not significant. When analyzing age groups, the results become much more concerning. The 25-29 and 30-34 age groups have considerably higher false negative rates compared to other age groups, at 25% and 20% respectively. This puts these age groups at a disadvantage, as they are much more likely to be incorrectly predicted as not having heart disease, when they actually do.

Looking at race subgroups, Asians stand out with near-zero false positive and false negative rates, suggesting that the model predicts extremely well for this group. In contrast, whites are the only subgroup with a false negative rate much higher than the baseline 10%, raising fairness concerns since Whites are the majority of the dataset.

Lastly, we will further analyze the ADS by training new models using the same seed and 10 other seeds and a more robust 80/20 train test split. We will continue to use the same accuracy and fairness metrics to analyze: precision, recall, F1-score, demographic parity difference (DPD), false negative rate (FNR), and false negative rate (FNR). A table of the mean and standard deviation (std dev) for each metric, by subgroup, is shown below:

Metric	**Mean**	**Std Dev**	**Re-train with the same seed**
Precision	0.982	0.017	0.7428
Recall	0.925	0.018	0.5507
F1 Score	0.953	0.013	0.6325
FNR	0.075	0.018	0.4493
FPR	0.001	0.001	0.0179
DPD (Sex)	0.0365	0.0023	0.0360
DPD (AgeCat)	0.2088	0.0054	0.1822
DPD (Race)	0.0671	0.0087	0.0502

Across the different models we trained, we observe that all fairness measures have improved compared to the model trained in the ADS. Precision increased from 30.46% to 98.2%, recall increased from 14.92% to 92.5%, and F1 score increased from 20.03% to 95.3%. This demonstrates that our new models predicts much better then the ADS model and excels on all fairness metrics tested. Zooming in, both the false negative rate and false positive rate have decreased significantly. Demographic parity remains comparable to the previous model.

5 Summary

The data used was largely appropriate for this ADS, as the goal is to predict heart disease. Thus, using features known to be correlated with heart disease, such as age, sex, race, and other conditions is relevant. Reducing the dataset to these relevant features from the original 618 features (source) is a reasonable choice to improve model efficiency. The original dataset consisted of 401,958 responses in which some were dropped for containing missing values, resulting in 319,795 responses to be used losing more than 82,000 samples. We must also consider the context of data collection. The survey was conducted throughout the year of 2020 where, due to the COVID-19 pandemic, interviews shifted from in-person to phone calls. This shift could introduce bias, as the type of people that would accept a phone interview might not represent the entire population, especially those with more severe cases of heart disease would not be able to conduct a phone interview. This would contribute to the class imbalance in our dataset.

During preprocessing, it's important to consider the dropped samples. While they were dropped for missing information, if the sensitive attributes of sex, age, and race were present, the dropped responses could have still been useful for training our model and assessing accuracy and fairness on those subgroups. Therefore, the decision to drop these rows might have unnecessarily reduced the overall sample size.

While the data used in the ADS is relevant, we recommend that this ADS not be deployed in the real world. The ADS acknowledges the class imbalance and recognizes that recall is a more appropriate metric than accuracy in this situation. However, the awareness was not extended when training models. For example, the model with the highest recall rate was ExtraTreesClassifier which scored a concerningly low 14.92% recall rate. When we trained the model using different seeds and a 80/20 train-test split, we observed significant improvements in both accuracy and fairness. Even when re-training with the same seed as the ADS but using our chosen 80/20 train-test partition, performance improved, though to a lesser extent than with other seeds. Based on these findings, we suggest that the ADS also use the more robust 80/20 train test split as well as using different seeds.

In our analysis of the ADS, we used accuracy using precision, recall, and F1 score, as well as fairness using demographic parity, false negative rate, and false positive rate. Precision is important for healthcare providers concerned with the accuracy of positive predictions, as they want to avoid allocating resources to those who don't actually have the disease. Recall, on the other hand, is critical for patients who want their heart disease to be correctly identified for an earlier treatment. The F1 score balances these two metrics and is useful for decision makers who need to consider both false positives and false negatives. In terms of fairness, demographic parity is important for patients, as group membership should not influence predictions. False negative rate is especially critical to patients because a missed diagnosis could result in delayed or no treatment. Finally, false positive rate is important for both patients and healthcare systems. Overdiagnosis (false calls) can cause unnecessary emotional and financial stress for patients, while also diverting valuable healthcare resources away from individuals who actually have the disease.

In conclusion, the ADS attempts to predict the risk of heart disease through relevant features and machine learning models. However, it does not adequately address the limitations of class imbalance in training its model. Our analysis shows that accuracy and fairness can be improved significantly through model tuning. Given the importance of a healthcare setting and the gravity of making incorrect precisions, we recommend that this ADS be updated to effectively predict and mitigate biases across sensitive subgroups before being deployment for use in the real world.