RDS Final Report: Heart Disease Prediction

Ray Chen (yjc464)
Nancy Wen (nw1334)

Background

We propose to build a nutritional label for an ADS system that predicts heart disease (https://www.kaggle.com/nareshbhat/eda-classification-ensemble-92-accuracy/notebook#Model-Evaluation) using a healthcare dataset from Kaggle (https://www.kaggle.com/nareshbhat/health-care-data-set-on-heart-attack-possibility). We chose this ADS because we wanted to work on a healthcare-related system. We believe that it is important to provide a nutritional label for healthcare ADS systems in order to prevent bias and increase trust in the system.

When we inspected the dataset, we saw that it contained 'sex' as a senstive attribute. We are particularly interested in exploring any potential differences when the ADS is applied to patients of different sex. The ADS system claims 90% overall accuracy, which seems very good, but we are concerned there will be differences in accuracy depending on gender, especially since the dataset is highly imbalanced (32% female and 68% male). We know that minimizing average error tends to fit the majority population (in this case, sex=Male), resulting in lower accuracy for the underrepresented class (sex=Female). If such differences exist, then that means there exists algorithmic biases in this ADS system. We would also like to explore whether we can use mitigation techniques, such as the generation of synthetic data, to create a model that can generalize better across all subpopulations.

Input and Output

The dataset is derived from the Cleveland Heart Disease Dataset from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Heart+Disease), which was collected by medical researchers. The data contains 303 rows. The original UCI dataset contains up to 76 attributes, but all published experiments including the ADS system that we are evaluating only uses a subset of 14 of them:

1) age
2) sex
3) chest pain type (4 values)
4) resting blood pressure
5) serum cholesterol in mg/dl
6) fasting blood sugar > 120 mg/dl
7) resting electrocardiographic results (values 0,1,2)
8) maximum heart rate achieved
9) exercise-induced angina
10) old peak = ST depression induced by exercise relative to rest
11) the slope of the peak exercise ST segment
12) number of major vessels (0-3) colored by fluoroscopy
13) thal: 0 = normal; 1 = fixed defect; 2 = reversible defect

14) target: 0= less chance of heart attack 1= more chance of heart attack

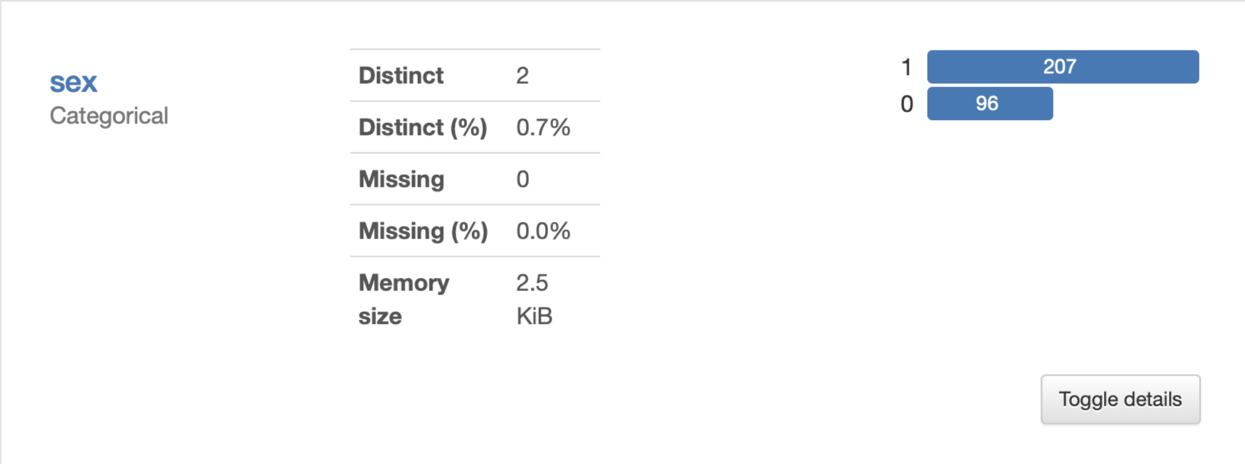
The sensitive attribute is ’sex’. The dataset is imbalanced by gender: there are 96 female patients and 207 male patients (32% female and 68% male).

In the table below, we have listed out all of the features, as well as their datatype, number of distinct values and number of missing values.

Variable name	***Variable description***	***Datatype***	***Type***	***Num of distinct values***	***Num of missing Values***
age	Age in years	int	continuous	41	0
sex	1 = male; 0 = female	int	categorical	2	0
cp	Chest pain type	int	categorical	4	0
trestbps	resting blood pressure (in mm Hg on admission to the hospital)	int	continuous	49	0
chol	serum cholesterol in mg/dl	int	continuous	152	0
fbs	fasting blood sugar > 120 mg/dl; 1 = true; 0 = false	int	continuous	2	0
restecg	resting electrocardiographic results	int	categorical	3	0
thalac	maximum heart rate achieved	int	continuous	91	0
exang	exercise induced angina (1 = yes; 0 = no)	int	categorical	2	0
oldpeak	ST depression induced by exercise relative to rest	float	continuous	40	0
slope	the slope of the peak exercise ST segment	int	categorical	3	0
ca	number of major vessels (0-3) colored by fluoroscopy	int	categorical	5	0
thal	3 = normal; 6 = fixed defect; 7 = reversible	int	categorical	4	0

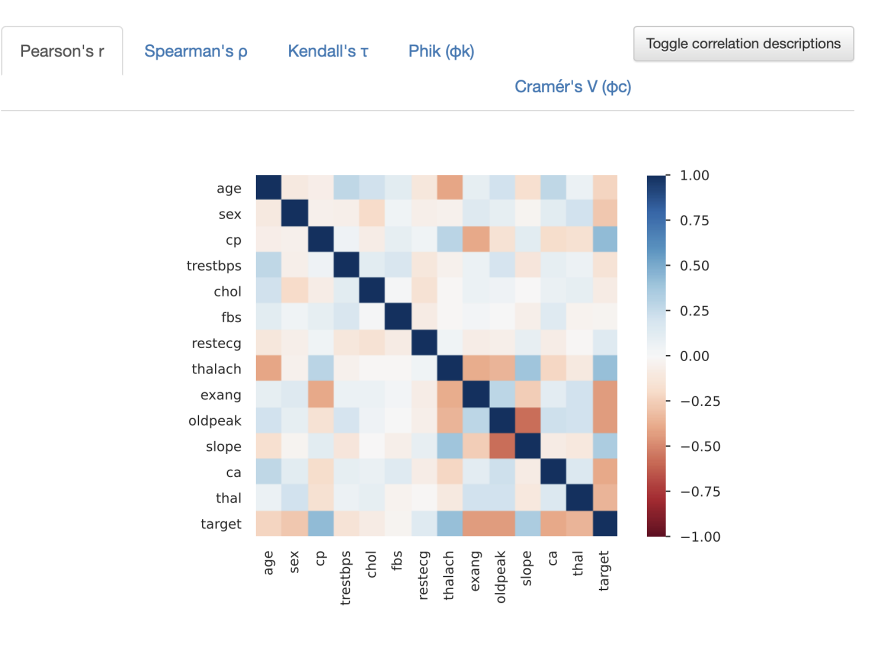
	defect				
target	1 or 0	int	categorical	2	0

Using the pandas profiling library, we were able to generate the following statistics and graphs about the dataset. Here is a subset of the graphs (the full report can be found in the colab notebook). From the data profiling, we see that there is no missing data for any of the columns. We also see that the values for the 'sex' feature is imbalanced: there are fewer women then men in the dataset (96 versus 207). The profile report also includes pairwise correlation between the features.



sex
Categorical

Distinct	2
Distinct (%)	0.7%
Missing	0
Missing (%)	0.0%
Memory size	2.5 KiB



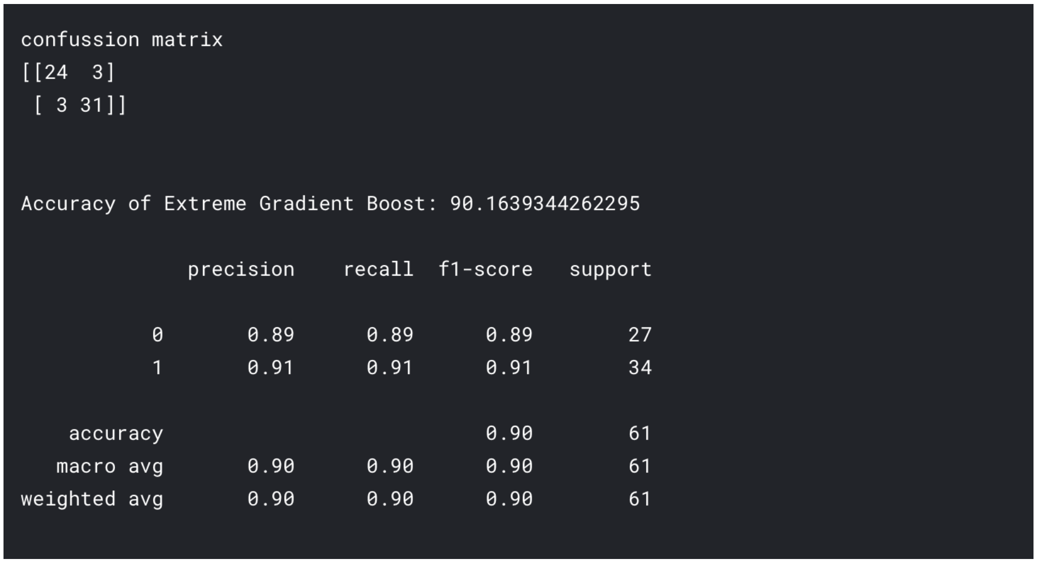
The target value of 1 indicates a diagnosis of heart disease (>50% diameter narrowing) and 0 indicates a diagnosis of no heart disease (<50% diameter narrowing). The target values are fairly balanced with 131 occurrences of a target value of 1 and 111 occurrences of a target value of 0 (54% and 46% respectively). The problem is formulated as a binary classification problem with two class labels.

Implementation and Validation

From the data profiling, we see that there is no missing data, and all the values fall within the expected range for each feature. In the ADS system, the creator also does exploratory data analysis, but he does no further data cleaning on the data. He does perform a preprocessing step by applying StandardScaler to all of the features, which normalizes the distribution of each feature.

The data is split into training and test set using a 80-20 split, resulting in 242 examples in the training set and 62 examples in the test set. We investigated the gender breakdown of the training and test set. The training set has 163 male examples and 79 female examples, and the test set has 44 male examples and 17 female examples.

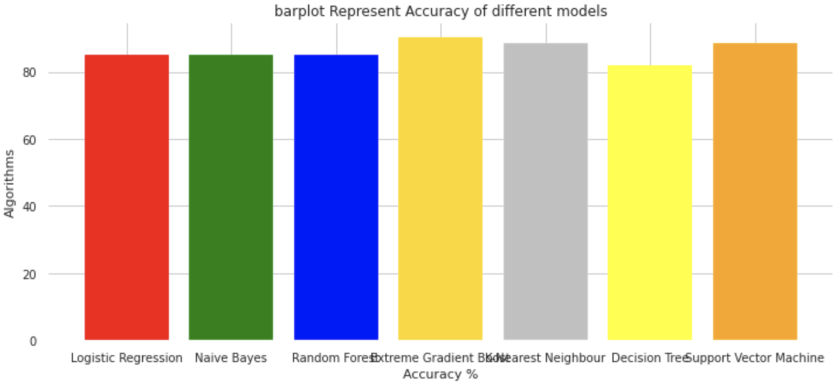
For model training, the creator of the notebook tried seven different algorithms including Logistic Regression, Naive Bayes, Random Forest Classifier, Extreme Gradient Boost, K-Nearest Neighbour, Decision Tree, and Support Vector Machine. He generated a confusion matrix and classification report (which includes precision, recall, and f1 score) on the test data for each of the models. For example, the image below shows the evaluation of the Extreme Gradient Boost model. The accuracy of the models ranged from 0.85 to 0.90, with the Extreme Gradient Boost model having the best performance.



```
confussion matrix
[[24  3]
 [ 3 31]]


Accuracy of Extreme Gradient Boost: 90.1639344262295

              precision    recall  f1-score   support

           0       0.89      0.89      0.89        27
           1       0.91      0.91      0.91        34

    accuracy                           0.90        61
   macro avg       0.90      0.90      0.90        61
weighted avg       0.90      0.90      0.90        61
```

In addition, the creator also plotted the accuracy of the different models and observed that the top three performing models were Extreme Gradient Boost (90.16), K Nearest Neighbors (88.5) , and Support Vector Machines (88.5). Then he decided to use an ensemble method called stacking to combine the prediction from the top three classifiers. The

StackingCVClassifier uses cross validation: the data is split into k folds, k-1 folds are used to fit the first level classifiers (EGB, KNN, SVC), and then the trained classifier is applied to the remaining 1 fold, and the resulting predictions are stacked and provided as input the to the second level classifier (SVC).



The resulting model results in the test accuracy of 90.2%. Using the ensemble learning, the creator was able to meet its stated goal of maximizing the accuracy of a classifier for predicting heart disease.

Outcomes

Fairness Measures

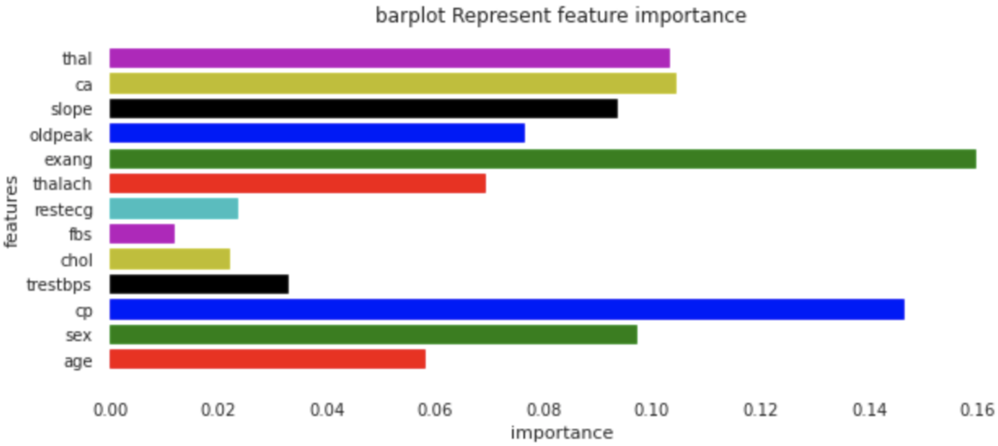
We evaluated the accuracy of the model across different subpopulations (divided by sex). The overall accuracy of the model is 90.2%, but the accuracy of the model on the female subpopulation is 88.2% while the accuracy of the model on the male subpopulation is 90.9%. The difference in accuracy between the genders is 2.6%.

Since the ADS system is a diagnostic tool instead of a benefit rewarding tool, it would not make sense for us to use measures such as the difference in mean outcomes or disparate impact as a measure of fairness because a positive label in this application does not indicate a positive benefit (1 indicates being at risk for heart disease). Therefore, we use the difference in false positive rate (FPR) and false negative rates (FNR) for each subpopulation as our fairness measures. The FPR for male patients is much lower than the FPR for female patients (4.7% versus 33.3% respectively). On the other hand, the FNR for male patients is higher than the FNR for female patients (13% versus 0% respectively).

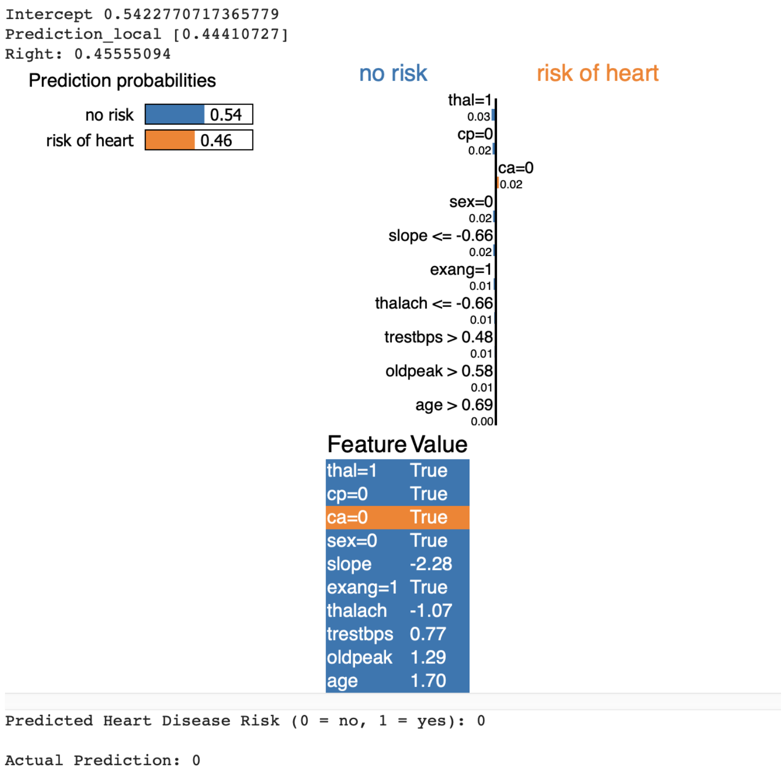
Interpretability

In addition to fairness measures, we use interpretability tools including LIME and SHAP in order to better understand which features are contributing the most to the predictions. For healthcare tools, we think it is important to include an explanation for predictions so that doctors can trust the output of the model and communicate the reasoning behind a decision to their patients. In addition to increasing patient trust, explainable features can also be used to understand the risk factors underlying a disease.

The creator of the notebook uses the built-in feature importance function to look at the most important features in the Extreme Gradient Boost model. The creator notes that exang (exercise induced angina), cp (type of chest pain), and ca (number of major vessels colored by fluoroscopy) are top predictors for the model, but he does not make note that sex is also in the top five most important features. We extend the exploration further by using LIME and SHAP.



First we used LIME, a package that provides “Locally Interpretable Model-agnostic Explanations” of machine learning models. For each specific prediction, LIME can provide a relevant explanation. We ran LIME on the XGBoost model in order to generate explanations that are comparable to the built-in feature importance. In the first example below, we look at the LIME explanation for a correct prediction (a true negative). For this example, the features that contribute most to the prediction are thal, cp, ca, and sex, which are the top four features produced by the built-in feature importance function of XGBoost, but in a different order.



```
Intercept 0.5422770717365779
Prediction_local [0.44410727]
Right: 0.45555094
```

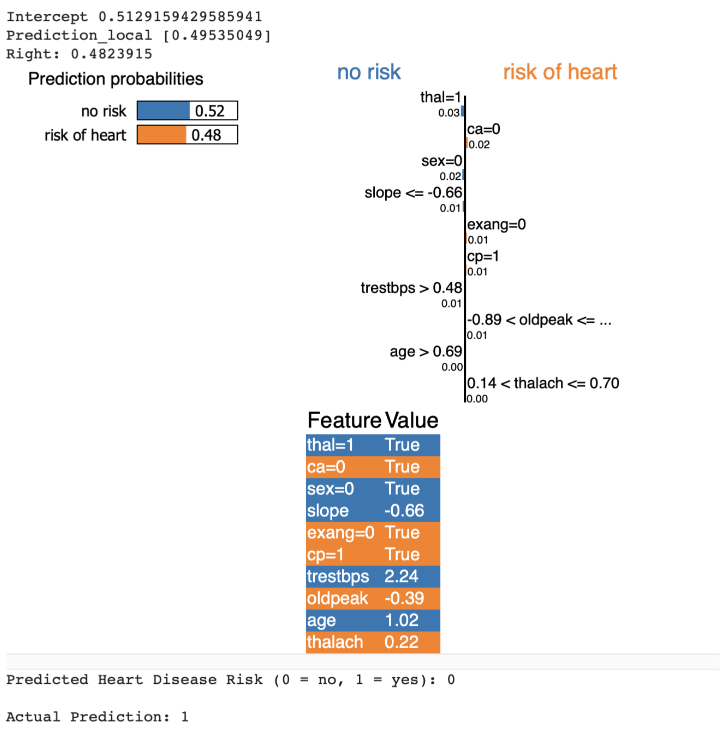
Prediction probabilities

no risk	0.54
risk of heart	0.46

Feature	Value
thal=1	True
cp=0	True
ca=0	True
sex=0	True
slope	-2.28
exang=1	True
thalach	-1.07
trestbps	0.77
oldpeak	1.29
age	1.70

```
Predicted Heart Disease Risk (0 = no, 1 = yes): 0

Actual Prediction: 0
```

In the second example, we look at the explanation for an incorrect example (a false negative). The top features are slightly different than the explanation for a correct example. Exang and cp have dropped out of the top four, and sex has gone up to be the third most important feature. In conclusion, we see that the LIME weights are sometimes consistent with the built-in feature importances, but does vary between individual data points.

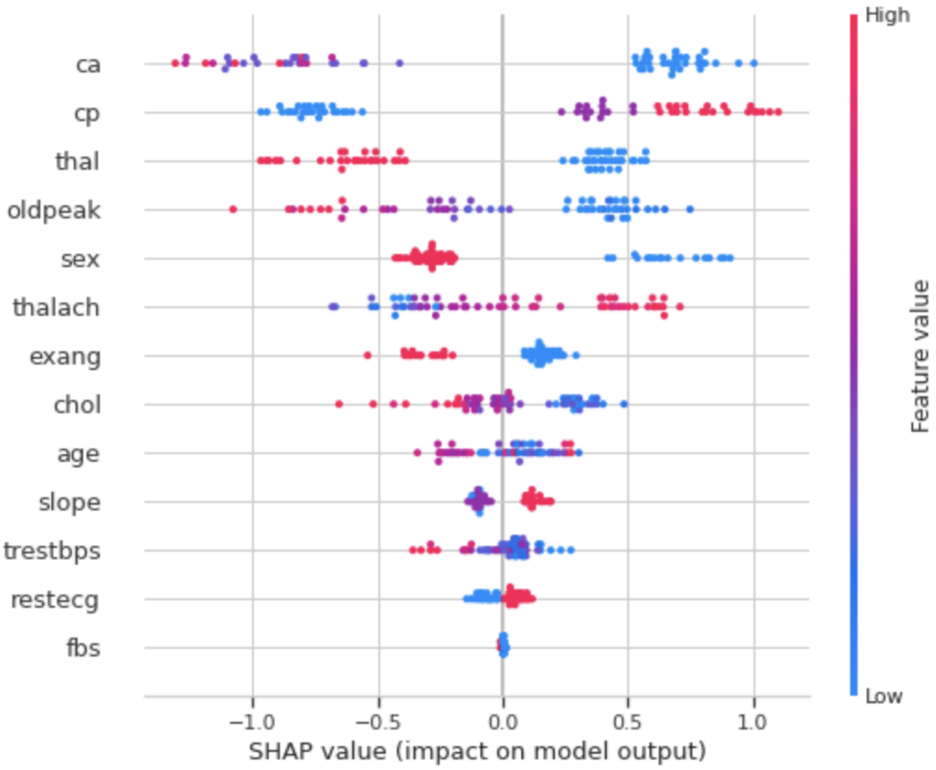


```
Intercept 0.5129159429585941
Prediction_local [0.49535049]
Right: 0.4823915
```

Feature	Value
thal=1	True
ca=0	True
sex=0	True
slope	-0.66
exang=0	True
cp=1	True
trestbps	2.24
oldpeak	-0.39
age	1.02
thalach	0.22

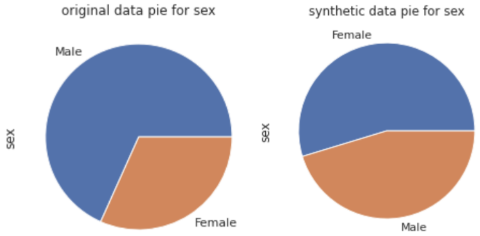
```
Predicted Heart Disease Risk (0 = no, 1 = yes): 0

Actual Prediction: 1
```

Next, we used SHAP (SHapley Additive exPlanations) because it is a unified approach to providing global and local interpretability. We use a SHAP summary plot to visualize which features are most important, and to see the range of effects over the entire dataset. The color indicates the value of the feature and the horizontal location of each datapoint for each row indicates the impact on the model output. For example, higher values for cp (type of chest pain) contribute positively to the model output while lower values for ca (number of vessels) contribute positively to the model output. We also explored the force plots for each individual examples. These plots are included in our Colab notebook but not included in this report to save space and since the LIME plots already include explanations for individual examples.



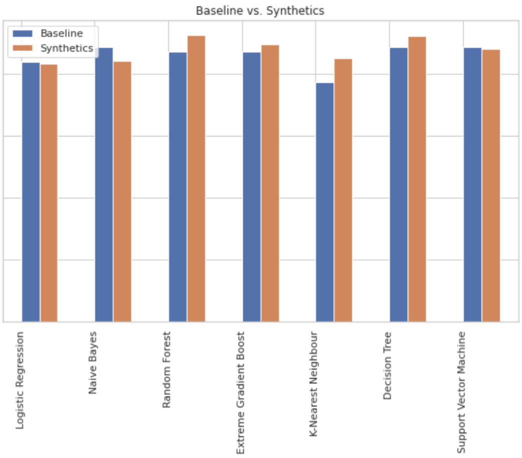
Privacy and Performance via Synthetic Datasets

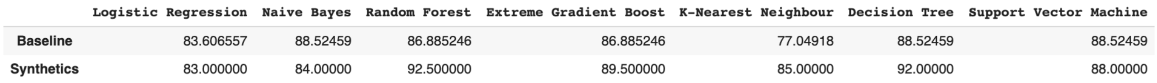
Another area we wanted to explore is whether we could generate synthetic data to train an ADS with similar or better performance compared to an ADS trained on the original dataset. Hospital records of patients are inherently confidential. We plan on using the DataSynthesizer to generate a synthetic dataset, and train a model on the synthetic dataset. We use the DataSynthesizer to generate additional female patients to see if it would help with mitigating the bias in the model. We hope that the mitigated synthetic datasets can improve prediction for both male and female patients equally. We tried the three DataSynthesizer modes and decide to test the synthetic data model based on synthetic correlated mode because it gives a good approximation of the sex distribution.



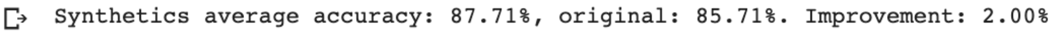
As we can see in the plot below, the synthetic data increased accuracy in some classification models, achieving the highest in the trees model like Random Forest from 86% to 92.5%, and

Decision Tree classifier from 82% to 92%. Also, the XGboost classifiers increase from 86% to 89.5%, and K-nearest Neighbour from 77% to 85%. For Naive Bayesian, the accuracy dropped might due to the algorithms assume that all features are independent, and hence synthetic data model might learned correlations in the training data. The average performance improvement across all the models was 2%.





	Logistic Regression	Naive Bayes	Random Forest	Extreme Gradient Boost	K-Nearest Neighbour	Decision Tree	Support Vector Machine
Baseline	83.606557	88.52459	86.885246	86.885246	77.04918	88.52459	88.52459
Synthetics	83.000000	84.00000	92.500000	89.500000	85.00000	92.00000	88.00000



```
Synthetics average accuracy: 87.71%, original: 85.71%. Improvement: 2.00%
```

Synthetic Data Fairness Measures - Correlated Mode

Using the synthetic dataset, the best performing model is Random Forest which has an accuracy of 92.5%, which is even higher than the accuracy of the ensemble StackingCVClassifier trained on the real data. We recalculated the fairness measures for males and females for the Random Forest Model. The accuracy for females increases from 88% to 93% and the accuracy for males increases from 90% to 91%. The false positive rate for males decreases from 4.7% to 1.3% and for females decreases from 33.3% to 8.3%. However, the false negative rate increases for males from 13% to 46.6% and increases for females from 0% to 0.04%. Even though the accuracy increases, the increase in false negative rate for both males and females is undesirable.

Summary

In conclusion, we believe that this heart disease dataset was appropriate for building an ADS system to predict heart disease but it has some limitations. For using LIME and SHAP, we are able to see that the dataset contains features with strong predictive power for predicting heart disease. However, from calculating the fairness measures, we can also see that the models trained on this dataset is less accurate for females than males. The accuracy for females was 88% and the accuracy for males was 90%, so this ADS system cannot be considered fair. We believe this is due to the fact that the data is very imbalanced. The dataset is not very large (only 303 examples) and less than one third of the data was females. For the main stakeholders of this ADS system, the patients, the accuracy of the system is extremely important because a false negative could become life threatening.

Therefore, we would not be comfortable deploying this ADS tool directly to the public. We think that this tool should only be provided as a diagnostic tool for doctors to proactively treat heart disease. Also, the ADS system cannot be deployed as is - it needs to include an interpretability tool such as LIME and SHAP, so when the doctor receives a prediction for a patient he could look at the main factors contributing to that prediction and determine if it makes sense.

In addition to add interpretability to the predictions, we also would like to recommend collecting more training examples that are female. The imbalanced dataset causes accuracy to be lower for females, and thus the ADS system is not fair. We considered using synthetic data to augment the number of female samples but even though accuracy increased for both genders, the false negative rate also increased. The fact that the synthetic dataset, which was more balanced in gender, was able to train a model with higher accuracy indicates that the ADS system could be improved by collecting more female examples.

We believe that an ADS system, especially in a healthcare setting, should be robust, accurate, and fair. The system that we evaluate in this project shows potential as a preventative diagnostic tool used by doctors, but the system could be augmented to be more interpretable and to improve the fairness of the tool such that it is equally accurate for both men and women.