Heart Disease Prediction:

Nutritional Labels for An Automated Decision System in Healthcare

Team Members:
Ray Chen (yjc464)
Nancy Wen (nw1334)

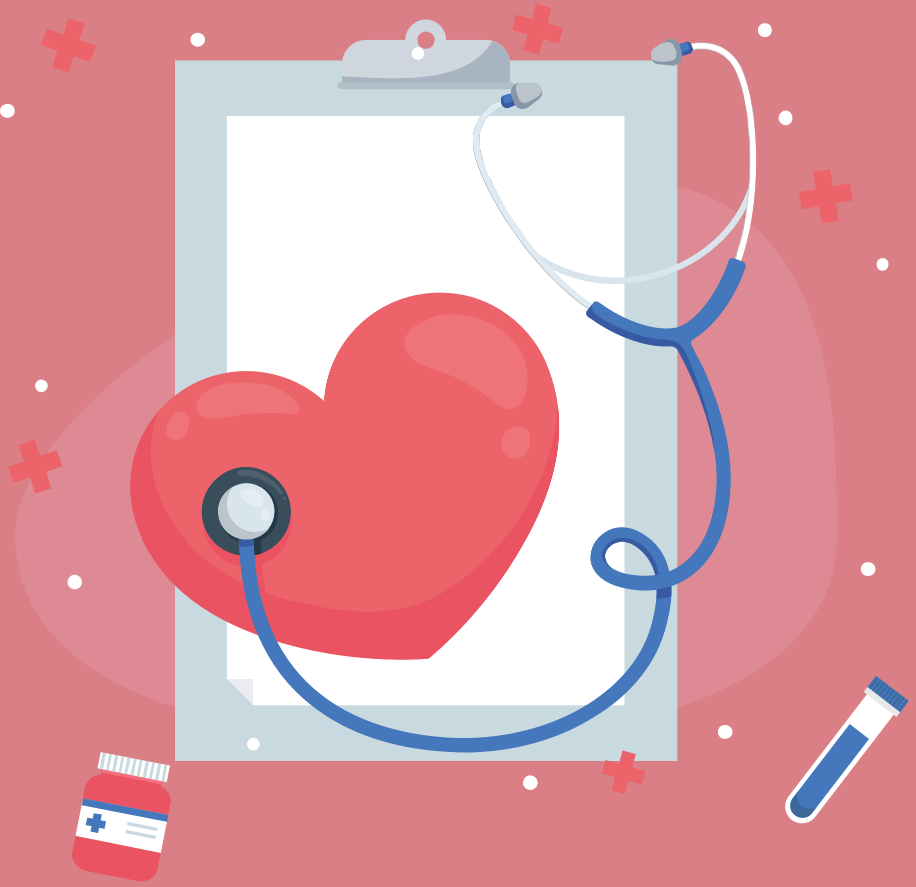


Table of contents

1

Background

Purpose of the heart disease ADS

2

Input and output

Data input features and output of the system

3

Implementation

Data cleaning and pre-processing

4

Validation

ADS validation

5

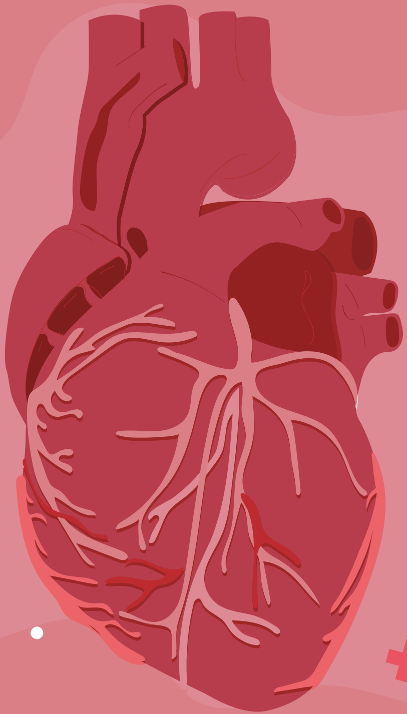
Outcomes

Effectiveness and Performance of ADS

6

Summary

Observations and Suggestions



Background

We propose to build a nutritional label for an ADS system that predicts heart disease using a healthcare dataset from Kaggle.

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.

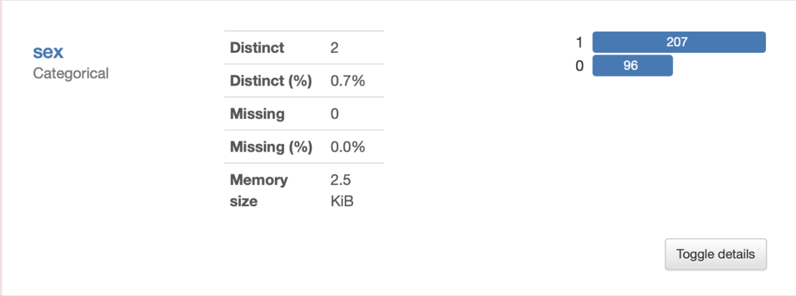
source : https://www.kaggle.com/nareshbhat/eda-classification-ensemble-92-accuracy/notebook#Model-Evaluation

Input of data

Variable name	Variable description	Datatype	Type	Num of distinct vlaues	Num of missing vlaues
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 defect	int	categorical	4	0
target	1 or 0	int	categorical	2	0

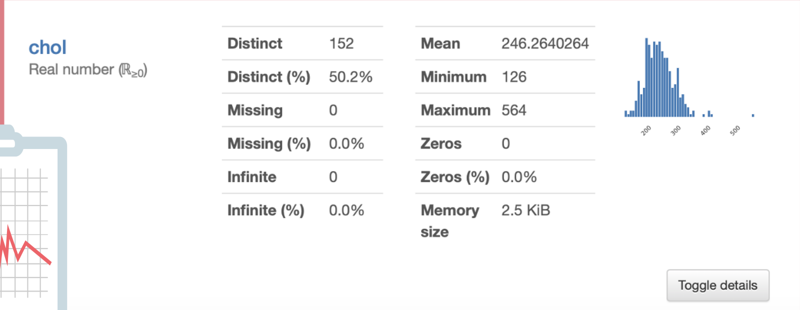
source : https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Output of Data



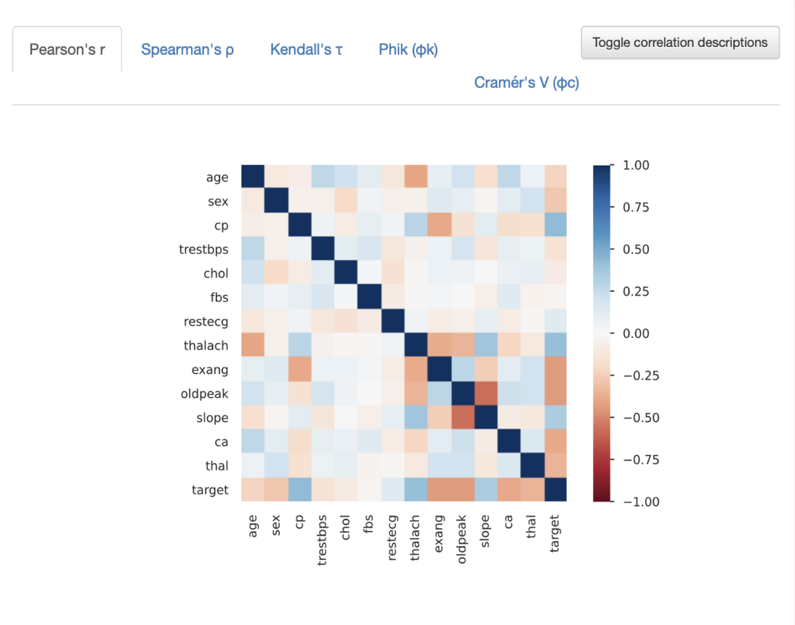
sex
Categorical

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



chol
Real number ($\mathbb{R}_{\geq 0}$)

Distinct	152	**Mean**	246.2640264
Distinct (%)	50.2%	**Minimum**	126
Missing	0	**Maximum**	564
Missing (%)	0.0%	**Zeros**	0
Infinite	0	**Zeros (%)**	0.0%
Infinite (%)	0.0%	**Memory size**	2.5 KiB



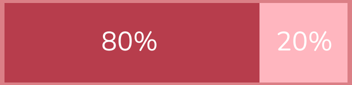
Observation:

1. The values for the ‘sex’ feature is imbalanced: fewer women then men in the dataset (96 versus 207).
2. Positive correlation between chest pain (cp) and target (our predictor).
3. Negative correlation between exercise induced angina (exang) and our predictor.

source : https://www.kaggle.com/nareshbhat/health-care-data-set-on-heart-attack-possibility

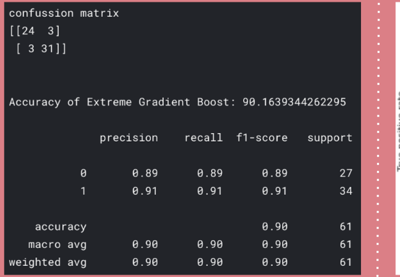
Implementation and Validation

Model Preparation



Train Set Test Set

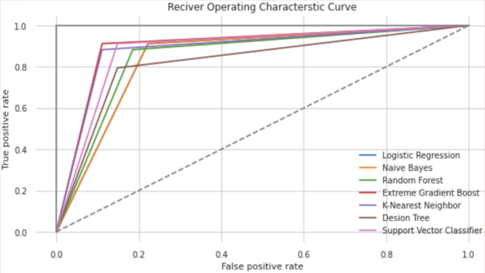
Modeling / Training



```
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
```

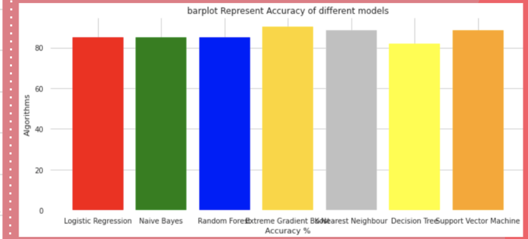
E.g. Extreme Gradient Boost model

Model Evaluation



E.g. ROC curve of all of the models

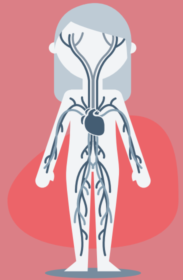
Model Output



E.g. Accuracy of different models

source : https://www.kaggle.com/nareshbhat/health-care-data-set-on-heart-attack-possibility

Fairness Measures: evaluate different subpopulations (divided by sex)

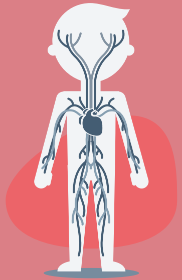


Female

Accuracy on female subpopulation is 88.2%

confussion matrix

11 (TP)	0 (FN)
2 (FP)	4 (TN)



Male

Accuracy on male subpopulation is 90.9%.

confussion matrix

20 (TP)	3 (FN)
1 (FP)	20 (TN)

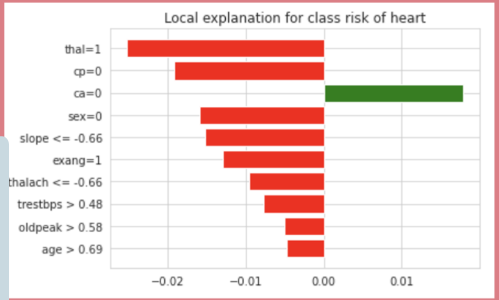
Observation:

1. The difference in accuracy between the genders is **2.7%**.
2. The False positive rate (FPR) for male patients is **much lower** than the FPR for female patients (4.7% versus 33.3% respectively).
3. The False negative rate (FNR) for male patients is **higher** than the FNR for female patients (13% versus 0% respectively).

source : RDS Lab 1: ProPublica's Analysis of the COMPAS Tool, RDS Lab 4 & 5: Predict early readmission for diabetes patients based on their laboratory and medications record

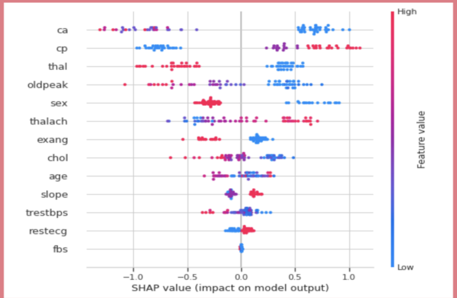
Interpretability

LIME



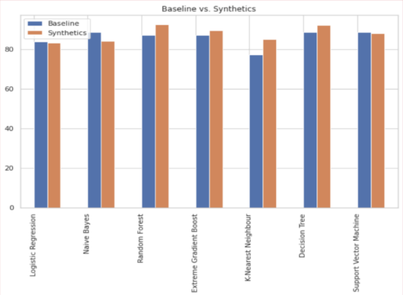
Result: Share some of the top features comparing with XGboost feature importance, but sometimes the prediction is completely opposite.

SHAP



Result: Summary plot is explainable and replaces the typical bar chart of feature importance.

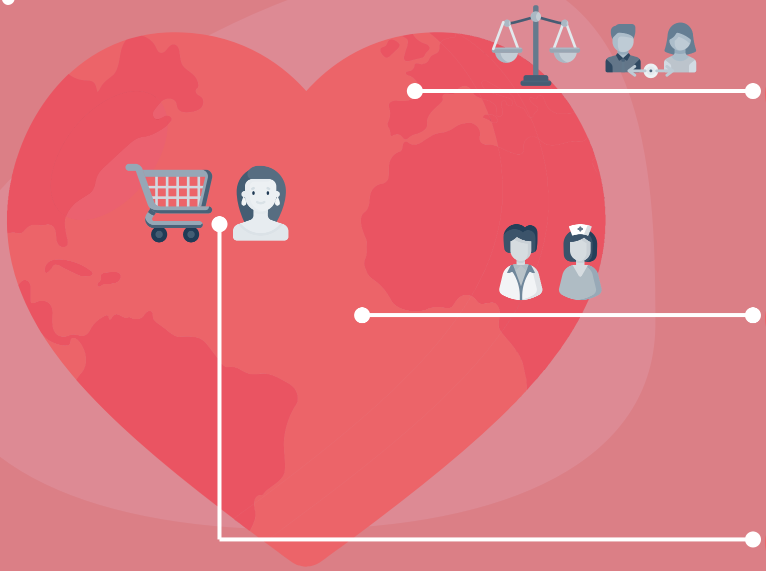
Synthetic Data



Result: The synthetic data model shows improvements in 4 out of 6 classification.

source : RDS Lab 1: ProPublica's Analysis of the COMPAS Tool, RDS Lab 4 & 5: Predict early readmission for diabetes patients based on their laboratory and medications record

Summary



1. Fairness is crucial because the ADS should be equally accurate for both men and women.

2. The ADS tool should only be used by medical professions in conjunction with in-person health checkups.

3. Collect more data from female patients. Deal with imbalanced dataset issue using synthetic data to reduce algorithmic biases.

Thank you !