ALGORITHMIC FAIRNESS

FINANCE AND HEALTHCARE

AGENDA

01 Fairness in Finance

02 Detecting bias in dataset

03 Mitigating age bias on credit decisions

04 Fairness in Health

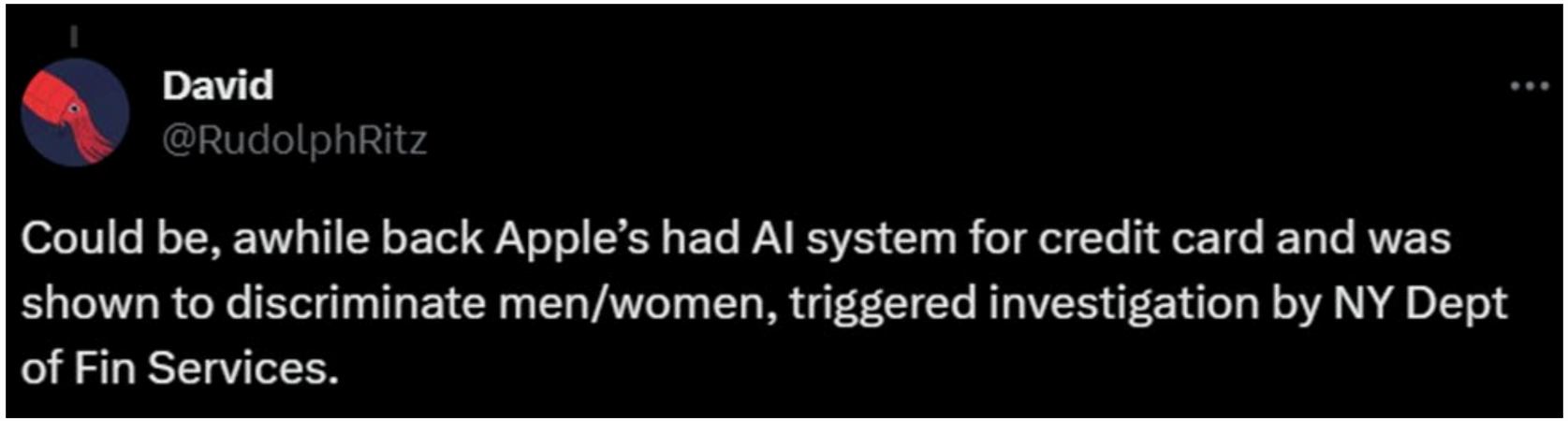
05 Mitigating bias in classifiers

06 Data interpretability using LIME

LOL - GORITHM: DENIED BY THE VOWEL POLICE!

Me and my friend, with identical financial backgrounds, apply for credit cards. The algorithm decides my fate based on an unexpected factor: the number of vowels in your names. my friend, Ram, gets approved, but me, with a longer name, get denied. Turns out, algorithms can be a real "consonant conspirator"!

APPLE CARD CREDIT LIMIT CONTROVERSY



Apple Card faced criticism for alleged gender bias in credit limits

"My wife and I filed joint tax returns, live in a community - property state, and have been married for a long time,"Yet Apple's black box algorithm thinks I deserve 20x the credit limit she does."

USE OF ALGORITHMS IN FINANCIAL SERVICES

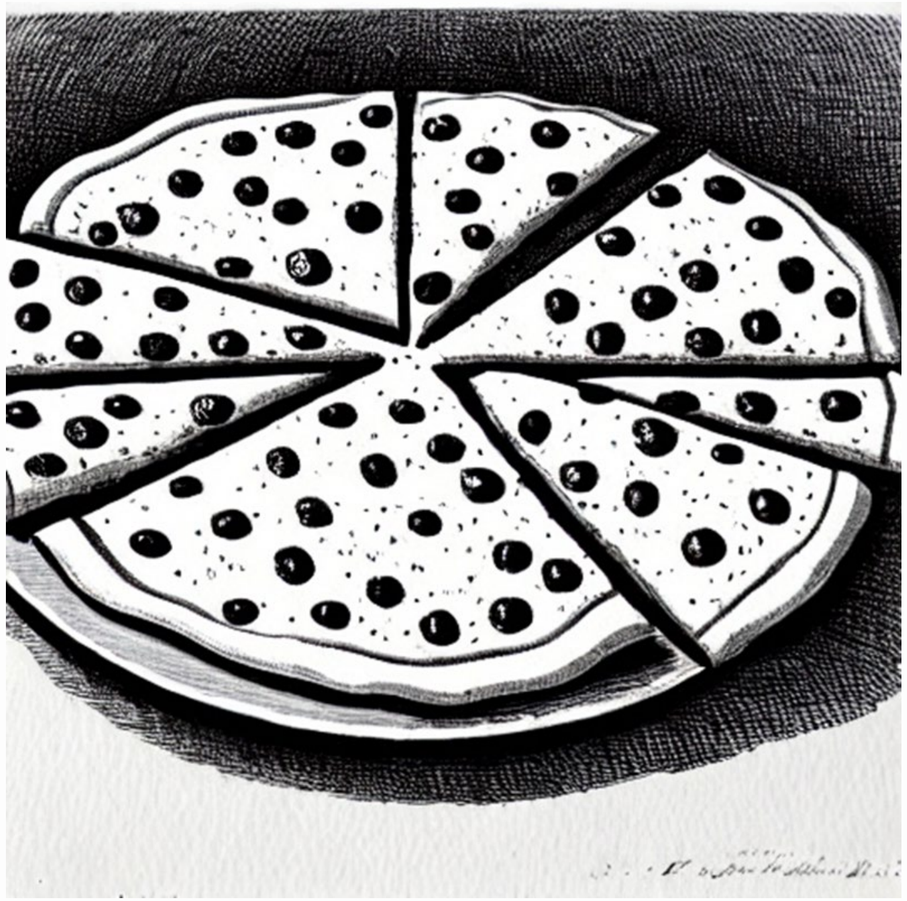
- Automated credit scoring and decisioning
- Automated pricing
- Real-time fraud detection

FACTORS IMPACTING ALGORITHMIC BIAS

There are several factors that can lead to algorithmic bias. The main factors include:

- Biases in data
- Biases in algorithmic design
- Biases in human use.

PIZZA BIAS: WHEN DATA CHEESES OUT!



Think of data like a pizza. Bias is like putting way more cheese on one slice than the others. It's not fair for all the slices, and that's how bias in data works—some parts get more attention than they deserve. Let's aim for a balanced topping distribution!

MITIGATION STRATEGIES

- Data defining
- Data gathering
- Data labelling
- Data pre-processing

DETECTING AND MITIGATING AGE BIAS ON CREDIT DECISIONS

step 1:

Set bias detection options, load dataset, and split between train and test

Step 2:

Compute fairness metric on original training dataset

Step 3:

Mitigate bias by transforming the original dataset

Step 4:

Compute fairness metric on transformed training dataset

DATA PREPROCESSING

```
dataset_orig = GermanDataset(protected_attribute_names=['age'],           # this dataset also contains protected
                                                                          # attribute for "sex" which we do not
                                                                          # consider in this evaluation
                             privileged_classes=[lambda x: x >= 25],      # age >=25 is considered privileged
                             features_to_drop=['personal_status', 'sex']) # ignore sex-related attributes

dataset_orig_train, dataset_orig_test = dataset_orig.split([0.7], shuffle=True)

privileged_groups = [{'age': 1}]
unprivileged_groups = [{'age': 0}]
```

```
Original one hot encoded german dataset shape:  (1000, 57)
Train dataset shape:  (700, 57)
Test dataset shape:  (300, 57)
```

Step 1

Load the German credit card dataset

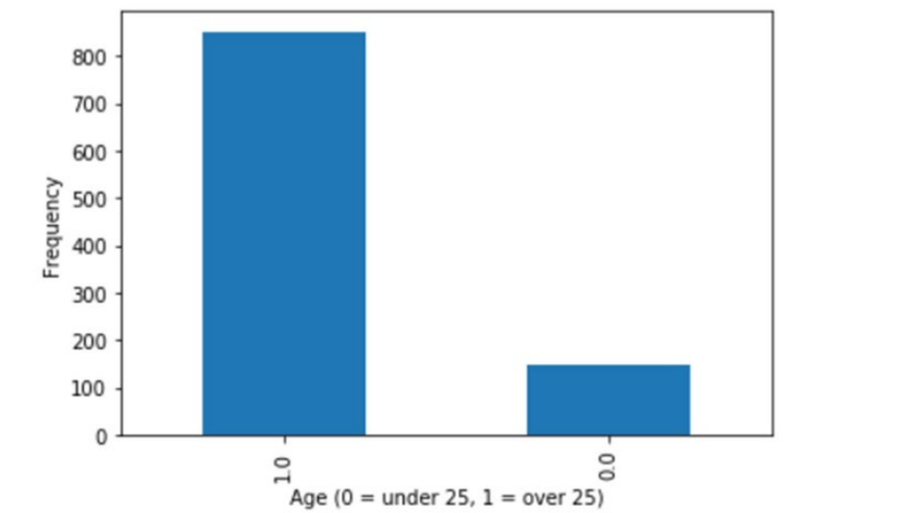
Step 2

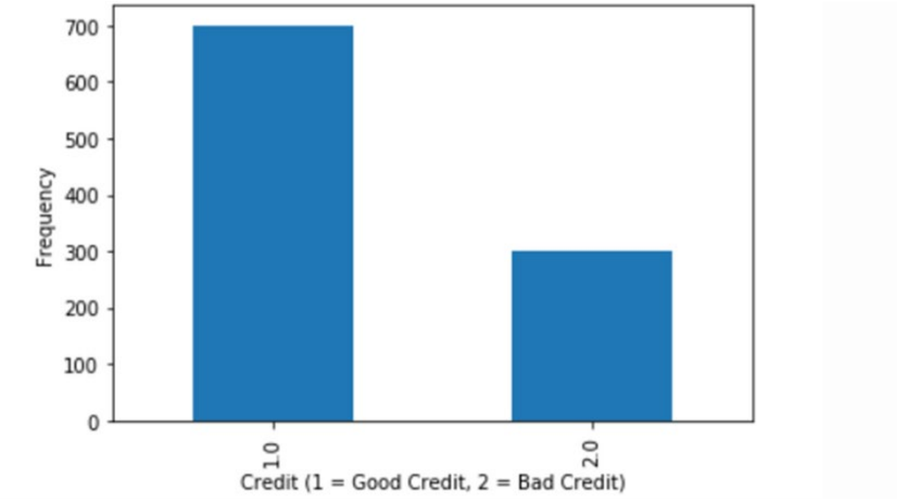
Specify the protected attribute

Step 3

split dataset into train and test datasets

BIAS IN DATASET





COMPUTING FAIRNESS METRIC

```
metric_orig_train = BinaryLabelDatasetMetric(dataset_orig_train,
                                             unprivileged_groups=unprivileged_groups,
                                             privileged_groups=privileged_groups)
```

Original training dataset

```
Disparate Impact = 0.766430
```

Disparate Impact

Computed as the ratio of rate of favorable outcome for the unprivileged group to that of the privileged group. The ideal value of this metric is 1.0.

DATA TRANSFORMATION

```
RW = Reweighing(unprivileged_groups=unprivileged_groups,
                privileged_groups=privileged_groups)
dataset_transf_train = RW.fit_transform(dataset_orig_train)
```

Reweighting algorithm transforms the dataset to have more equity in positive outcomes on the protected attribute for the privileged and unprivileged groups.

- Package used: aif360.algorithms
- Algorithm : Reweighting
- Bias mitigation technique : Data preprocessing

FAIRNESS METRIC

```
metric_transf_train = BinaryLabelDatasetMetric(dataset_transf_train,
                                               unprivileged_groups=unprivileged_groups,
                                               privileged_groups=privileged_groups)
```

Transformed training dataset

Disparate Impact = 1.000000

THE AI NURSE

The AI Nurse, an Avocado Advocate, Keeps Recommending Guacamole for Every Ailment, Ignoring Other Food Preferences . Patients Wonder if Avocado Toast is the Universal Cure!

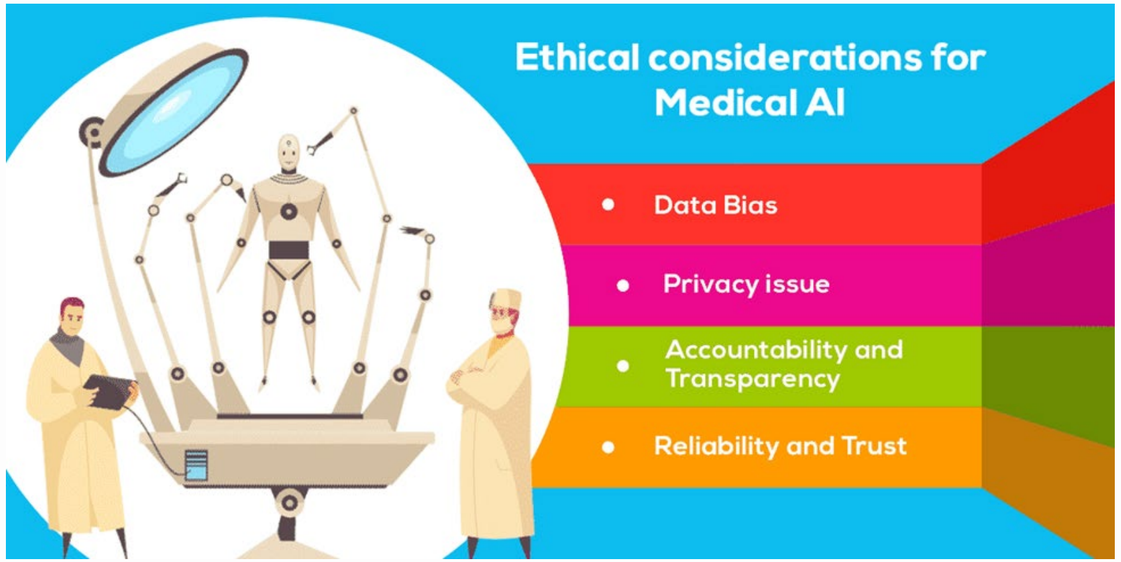
NAVIGATING FAIRNESS IN DIAGNOSES

How machine-learning models can amplify inequities in medical diagnosis and treatment

MIT researchers investigate the causes of health care disparities among underrepresented groups.

Steve Nadis | MIT CSAIL
August 17, 2023

ETHICAL CONCERNS IN AI HEALTHCARE



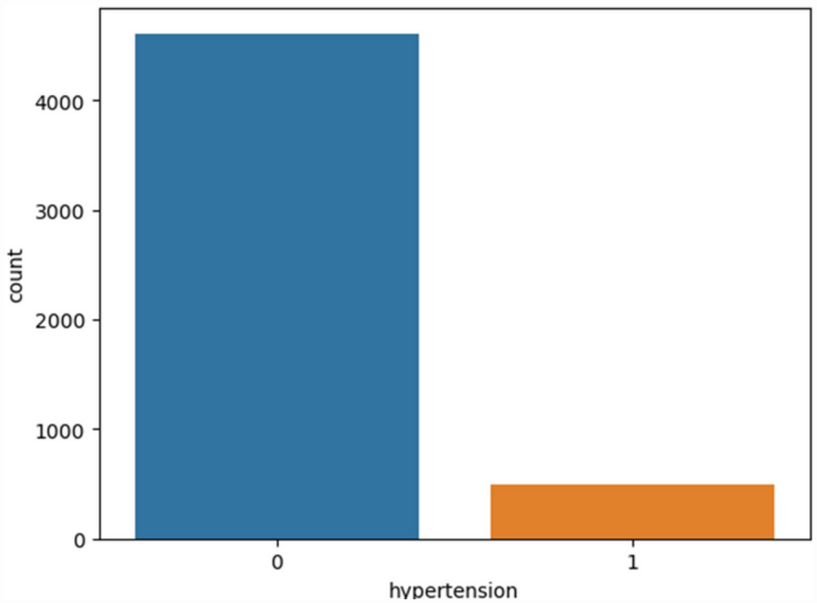
- Artificial Intelligence (AI) holds great promise in transforming healthcare, but it also raises ethical concerns that demand careful consideration.
- We are focusing on mitigating bias in our heart attack prediction model, recognizing the importance of fair and unbiased predictions for patient well-being.
- By implementing techniques such as SMOTE and post-analysis with LIME, we aim to enhance the fairness and interpretability of our AI model.

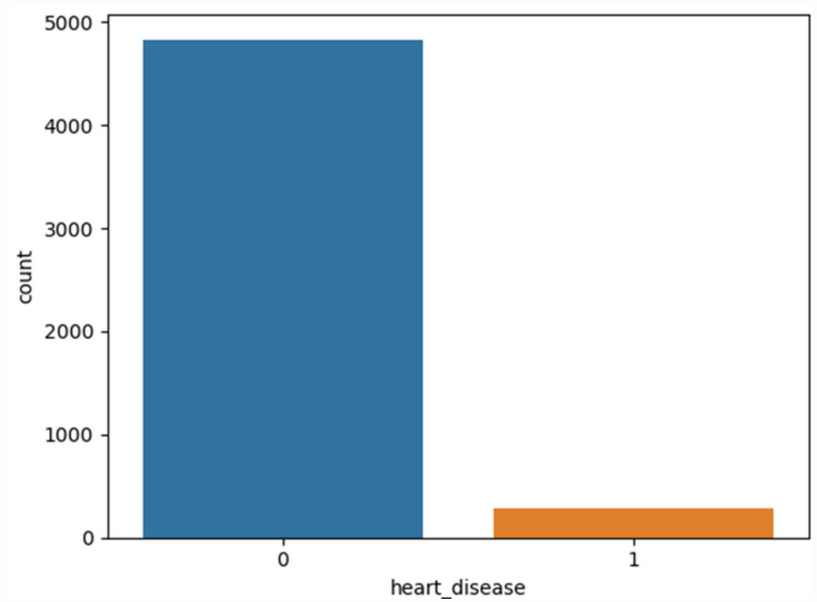
DATA OVERVIEW

Dataset: Healthcare -dataset-stroke-data

Key Parameters:

- id
- gender
- age
- hypertension
- heart_disease
- ever_married
- work_type
- Residence_type
- avg_glucose_level
- bmi
- smoking_status
- stroke





SMOTE - SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE

SMOTE is a resampling technique designed to balance class distribution by generating synthetic instances of the minority class.

```
[52]
    print('Before OverSampling, counts of label 1: {}'.format(sum(y_train==1)))
    print('Before OverSampling, counts of label 0: {} \n'.format(sum(y_train==0)))

...
Before OverSampling, counts of label 1: 195
Before OverSampling, counts of label 0: 3893
```

```
[53]
    sm = SMOTE(random_state=2)
    x_train_res, y_train_res = sm.fit_resample(x_train,y_train.ravel())

    print('After OverSampling, the shape of train_x: {}'.format(x_train_res.shape))
    print('After OverSampling, the shape of train_y: {}'.format(y_train_res.shape))

    print('After OverSampling, counts of label 1: {}'.format(sum(y_train_res == 1)))
    print('After OverSampling, counts of label 0: {}'.format(sum(y_train_res == 0)))

...
After OverSampling, the shape of train_x: (7786, 10)
After OverSampling, the shape of train_y: (7786,)
After OverSampling, counts of label 1: 3893
After OverSampling, counts of label 0: 3893
```

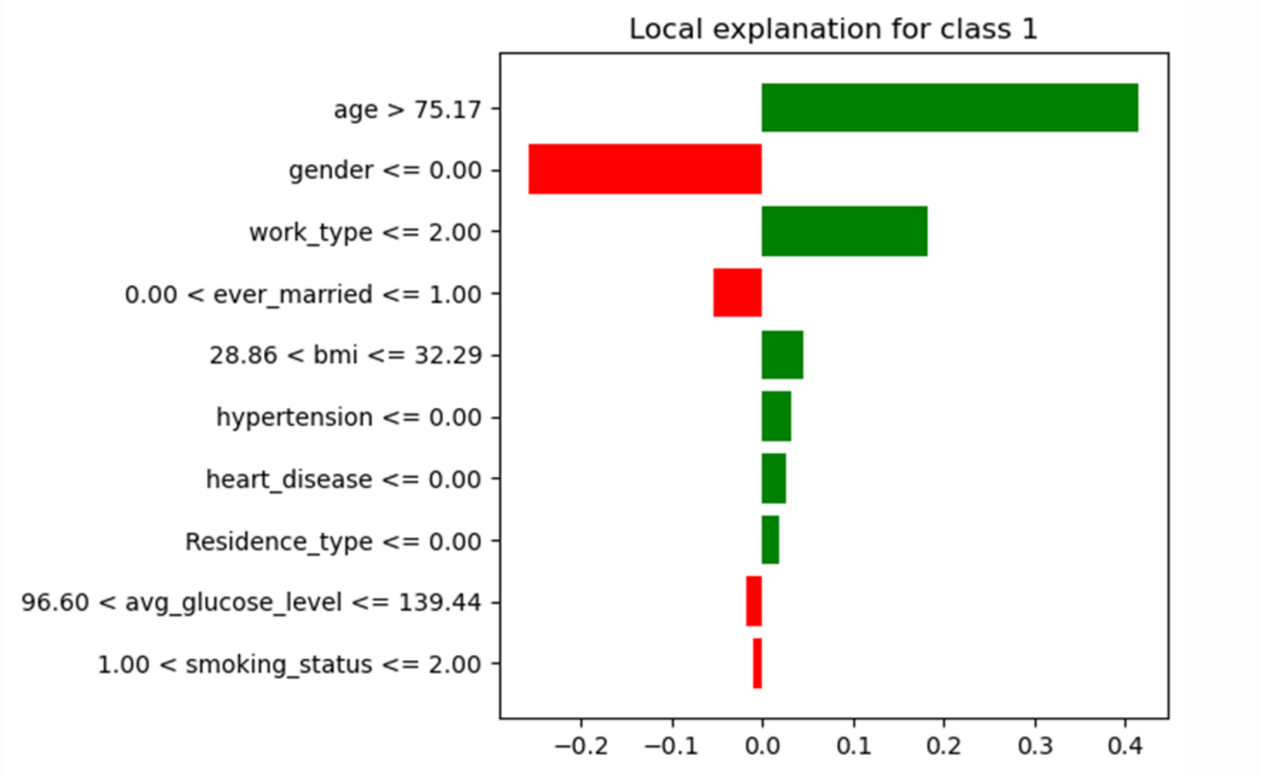
RESULTS

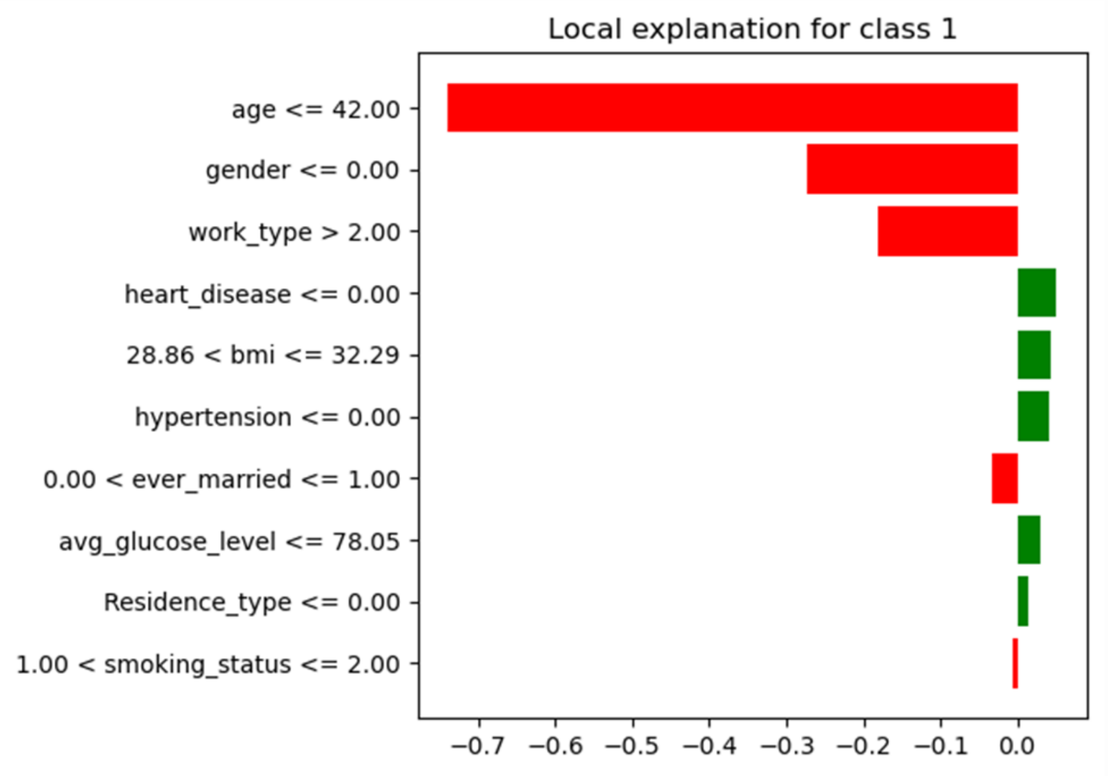
```
RandomForest: Average Fold Accuracy - 0.9571
GradientBoosting: Average Fold Accuracy - 0.9023
LogisticRegression: Average Fold Accuracy - 0.7962
SVM: Average Fold Accuracy - 0.7860
KNeighbors: Average Fold Accuracy - 0.8889
NaiveBayes: Average Fold Accuracy - 0.7972
DecisionTree: Average Fold Accuracy - 0.9362
AdaBoost: Average Fold Accuracy - 0.8639
XGBoost: Average Fold Accuracy - 0.9612
Overall OOF Accuracy: 0.3264
```

```
              precision    recall  f1-score   support

           0       0.97      0.96      0.96       777
           1       0.96      0.97      0.96       781

    accuracy                           0.96      1558
   macro avg       0.96      0.96      0.96      1558
weighted avg       0.96      0.96      0.96      1558

[[745  32]
 [ 26 755]]
```

RandomForestClassifier

LIME - LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS





CONCLUSION

- As AI technologies advance, the increasing ethical concerns directly impact human trust.
- We emphasize the importance of concurrently advancing ethical considerations to ensure the development of safe and transparent AI models in the future.

Thank You