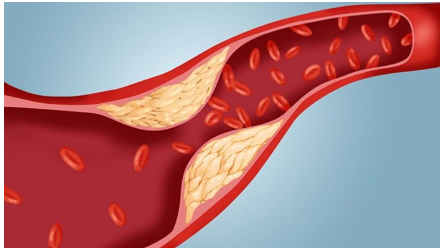
# Predicting Cholesterol Without a Clinical Test

**Zehui Lin, Dylan Mendonca, Shimona Narang, Esmond Tang**

April 6, 2020

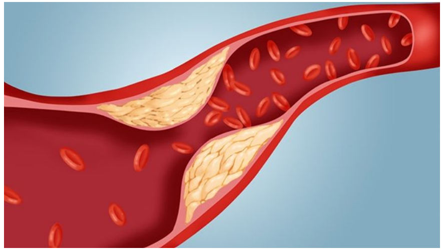
# Why Cholesterol?



- Cholesterol is a ‘waxy’ substance produced by the body and found in foods
- Levels ≥240 mg/dL are considered borderline/high and increases risk of heart disease, stroke, etc.

https://health.costhelper.com/cholesterol-test.html#extres1
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5485409/

# Why Cholesterol?



- Cholesterol is a ‘waxy’ substance produced by the body and found in foods
- Levels ≥240 mg/dL are considered borderline/high and increase risk of heart disease, stroke, etc.

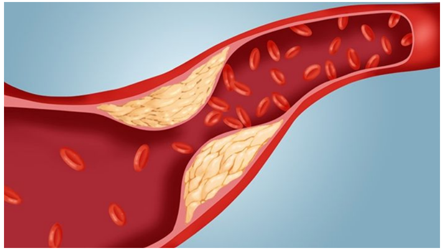
**$25**

Per clinical lab test to check cholesterol level

Expensive for people in developing nations!

https://health.costhelper.com/cholesterol-test.html#extres1
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5485409/

# Why Cholesterol?



- ➢ Cholesterol is a ‘waxy’ substance produced by the body and found in foods
- ➢ Levels ≥240 mg/dL are considered borderline/high and increases risk of heart disease, stroke, etc.

**$25**

Per clinical lab test to check cholesterol level

Expensive for people in developing nations!

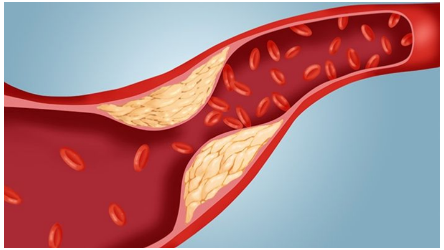
**~27%**

Of urban pop. has high level of cholesterol (>200 mg/dL)

This number is growing!

https://health.costhelper.com/cholesterol-test.html#extres1
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5485409/

# Why Cholesterol?



- Cholesterol is a 'waxy' substance produced by the body and found in foods
- Levels ≥240 mg/dL are considered borderline/high and increases risk of heart disease, stroke, etc.

| | | |
|---|---|---|
| **$25** | **~27%** | **~82%** |
| Per clinical lab test to check cholesterol level | Of urban pop. has high level of cholesterol (>200 mg/dL) | Were **not aware** that they had high cholesterol |
| Expensive for people in developing nations! | This number is growing! | That's not good... |

*Cholesterol Statistics for the Indian population*

https://health.costhelper.com/cholesterol-test.html#extres1
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5485409/

# Supervised ML Algorithms Can Help

**Problem Statement:**

Develop a supervised ML model to predict whether a person has a high total cholesterol level, without having to go to the clinic

# Supervised ML Algorithms Can Help

**Primary Goal:**

Develop a supervised ML model to predict whether a person has a high total cholesterol level, without having to go to the clinic

Inputs

**National Health and Nutrition Survey (NHANES) Data**

- Demographics
- Diet
- Questionnaire
- Examinations

# Supervised ML Algorithms Can Help

**Primary Goal:**

Develop a supervised ML model to predict whether a person has a high total cholesterol level, without having to go to the clinic

Inputs → Supervised ML Algorithms

**National Health and Nutrition Survey (NHANES) Data**

- Demographics
- Diet
- Questionnaire
- Examinations

**Leverage Supervised Models incl.:**

- Linear
- Non-Linear
    - Tree-based
    - Other

# Supervised ML Algorithms Can Help

**Primary Goal:**

Develop a supervised ML model to predict whether a person has a high total cholesterol level, without having to go to the clinic

| Inputs | Supervised ML Algorithms | Outputs |
| --- | --- | --- |

**National Health and Nutrition Survey (NHANES) Data**

- Demographics
- Diet
- Questionnaire
- Examinations

**Leverage Supervised Models incl.:**

- Linear
- Non-Linear
    - Tree-based
    - Other

**Binary Prediction:**

- High (1) ➜ $\geq$ 240 mg/dL
- Low (0) ➜ $<$ 240 mg/dL

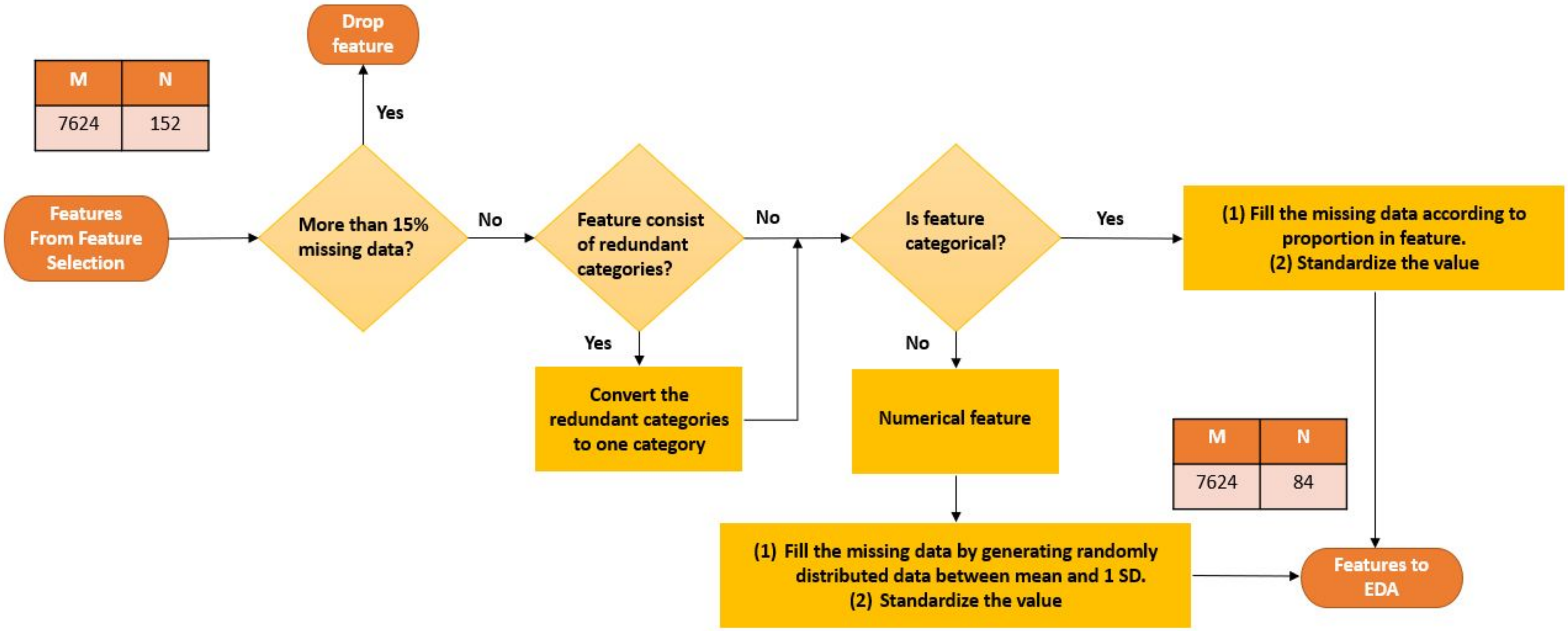
# Explaining the Dataset

## Feature Selection

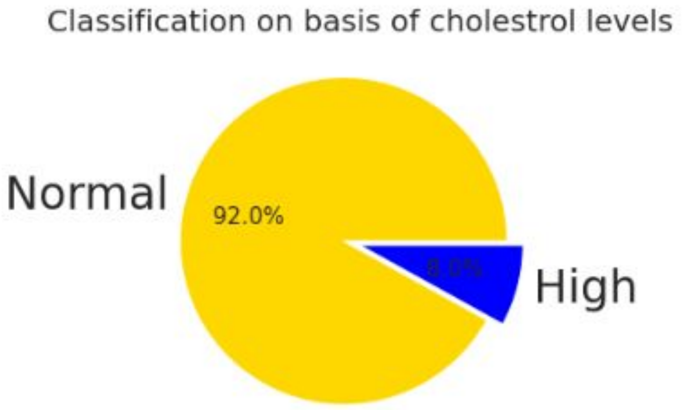
| Steps | Observation (M) | Number of Features (N) |
|---|---|---|
| (1) Original dataset | 9813 | 1390 |
| (2) After feature selection | 9813 | 152 |
| (3) Dropping the observations with no target output | 7624 | 152 |

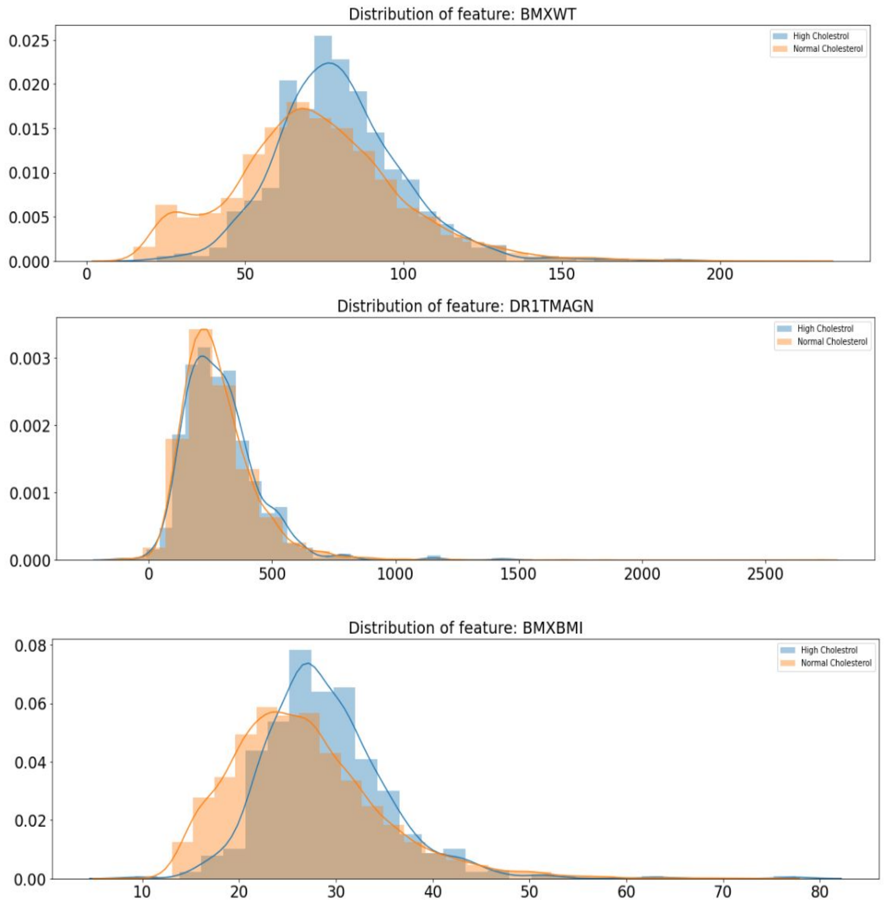
# Explaining the Dataset

## Data Imputation Flow Chart



# Exploratory Data Analysis




<u>Findings from Pie Chart:</u>

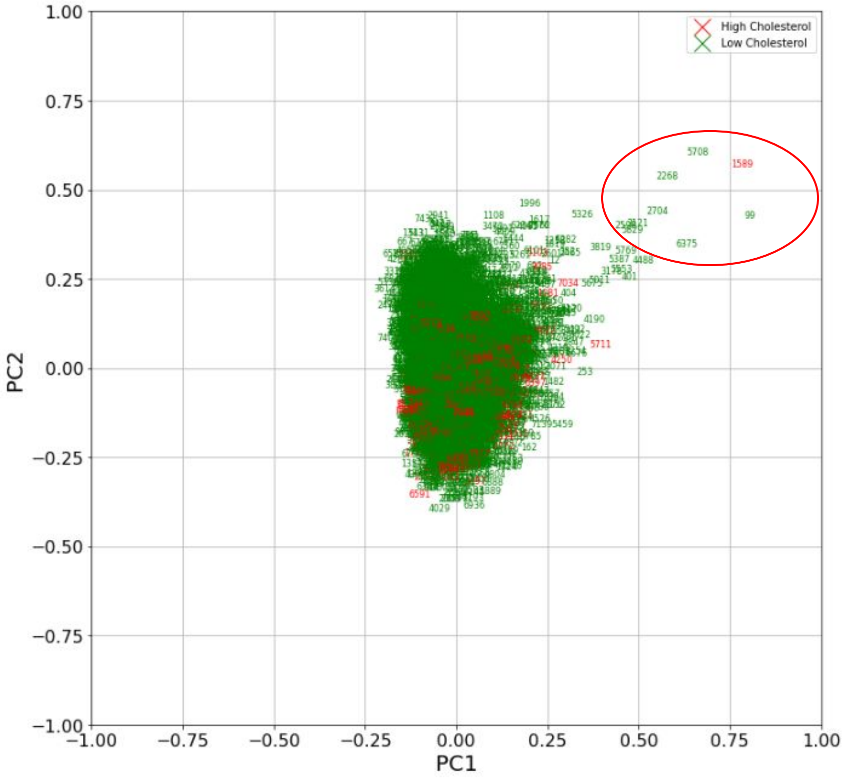
➢ Created a balanced train set and an imbalanced test set

<u>Findings from Histograms:</u>

➢ Most of the numerical features have overlapping classes

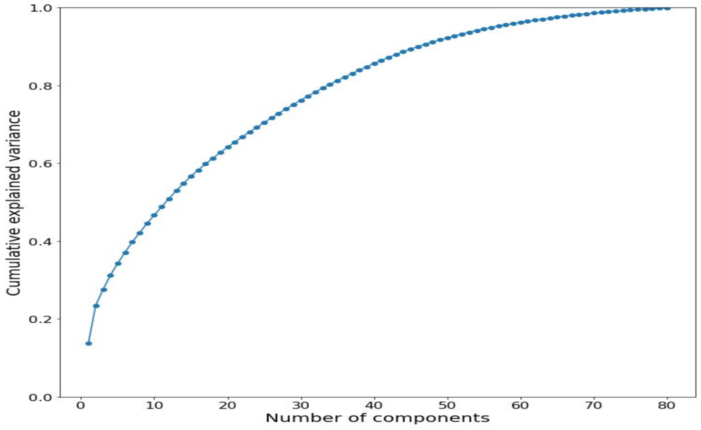
# Exploratory Data Analysis, contd.

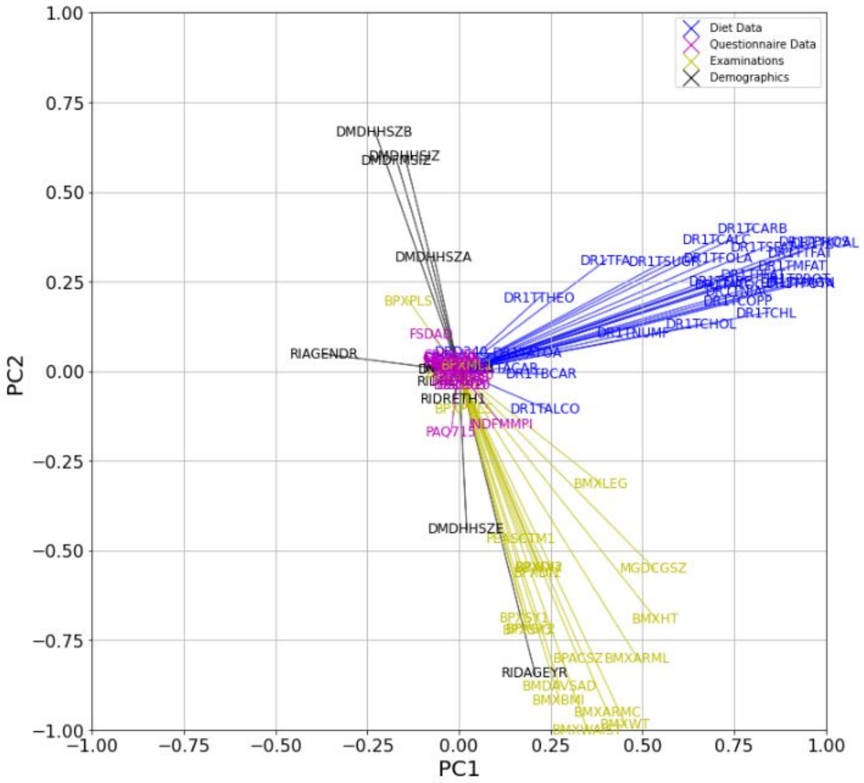
- ➢ Our data doesn't seem to be linearly separable
- ➢ Outliers are mostly low cholesterol and have really high values for diet-related variables



# Exploratory Data Analysis, contd.

- ➢ Loadings show that features coming from the same dataset match
- ➢ Explained variance makes dimensionality reduction less “appealing”



# Approach for Model Building

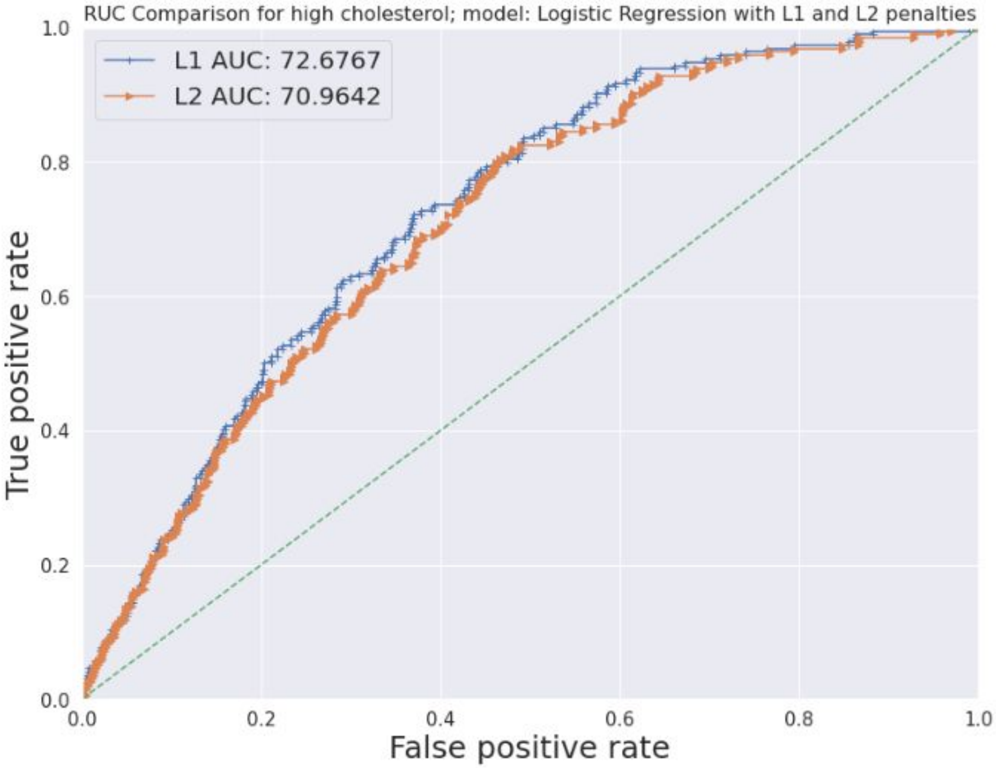
1. Training a model on balanced dataset and tested on imbalanced dataset
2. Fine-tuning hyperparameters using GridSearchCV
    - Score: Recall (exception: XGBoost with Area under precision and recall curve)
    - CV folds: 5
3. Comparing the models based on recall.

Recall measures the proportion of actual high cholesterol cases that are correctly identified as ‘high’ by the model.

Therefore, higher the recall, better the model

# Linear methods for classification

Logistic Regression with L1 and L2 regularization

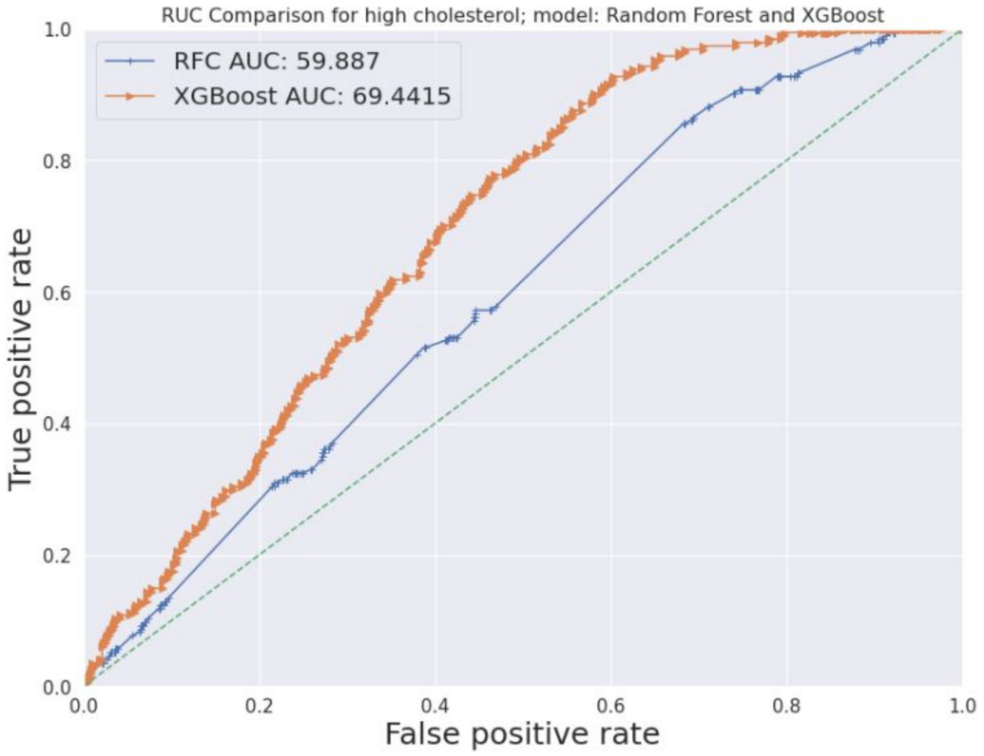



- Parameters
  - Solver: ‘saga’
  - Regularizer strength C (Fine Tuned)

| | Model Type | |
|---|---|---|
| **Metrics** | **Penalty L1** | **Penalty L2** |
| **Recall** | **0.74** | **0.70** |
| **Accuracy** | **0.61** | **0.61** |

# Non-Linear methods for classification

Random Forest Classifier and XGBoost



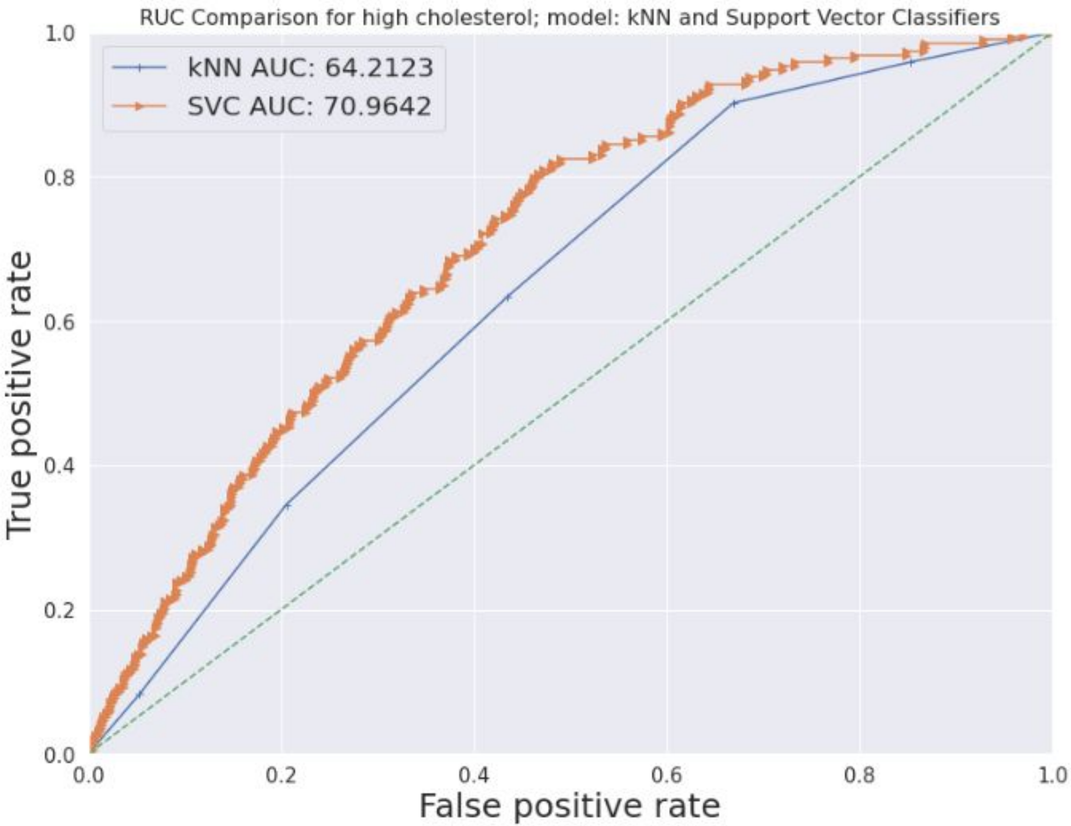

- Parameters to train & fine tune Random Forest
  - n_estimators
  - Max_depth

- Parameters to train & fine tune XGBoost
  - Eta (Learning rate)
  - Max_depth
  - Gamma
  - subsample

| | Model Type | |
|---|---|---|
| **Metrics** | **Random Forest** | **XGBoost** |
| **Recall** | **0.80** | **0.70** |
| **Accuracy** | **0.56** | **0.60** |

# Non-Linear methods for classification

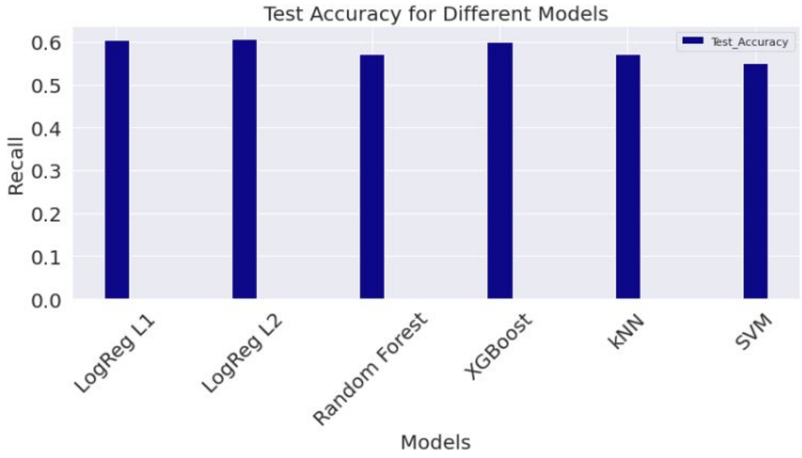
K Nearest Neighbors (kNN) and Support Vector Classifier (SVC)

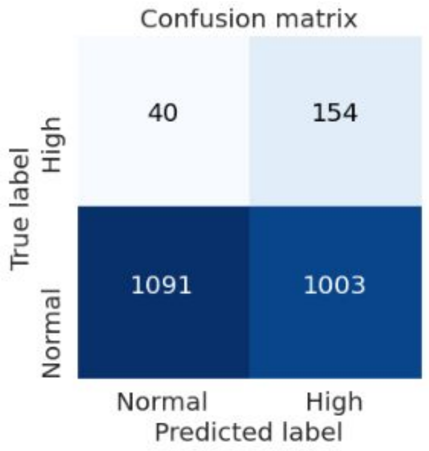



- Parameters to train & fine tune kNN
  - n_neighbors
  - Distance metric: euclidean
- Parameters to train & fine tune SVC
  - Kernel: RBF kernel
  - Regularization parameter C

| | **Model Type** | |
|---|---|---|
| **Metrics** | **kNN** | **SVC** |
| **Recall** | **0.63** | **0.79** |
| **Accuracy** | **0.57** | **0.55** |

# Model Comparison


**Random Forest**

- ➢ Random Forest predicts high cholesterol more efficiently
- ➢ Random Forest > SVM > XGBoost > Logistic Regression L1 > Logistic Regression L2 > kNN
- ➢ Normal cholesterol cases are not classified accurately; this may be due to overlapping of classes

# Building for Deployment

**Challenge:**

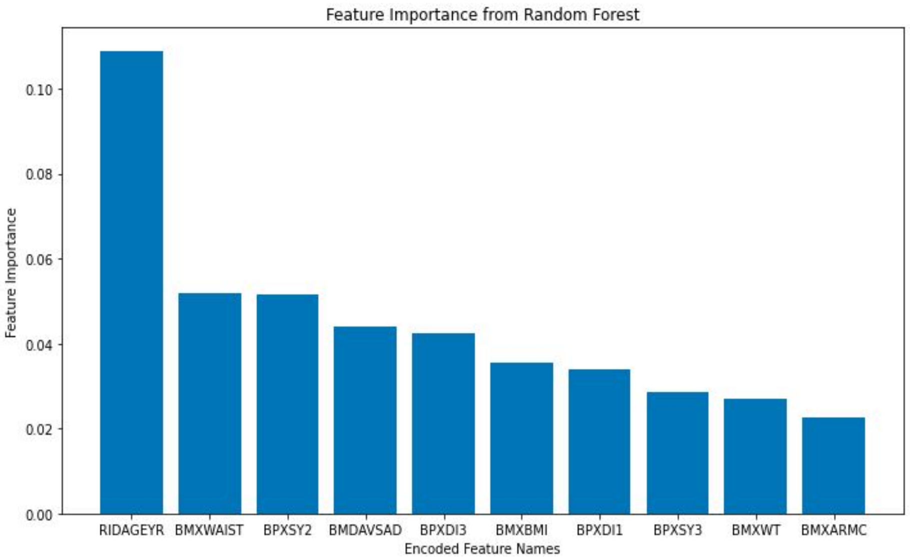
- ➢ Determine cholesterol level by filling out online surveys
- ➢ 84 questions are too overwhelming for an online survey

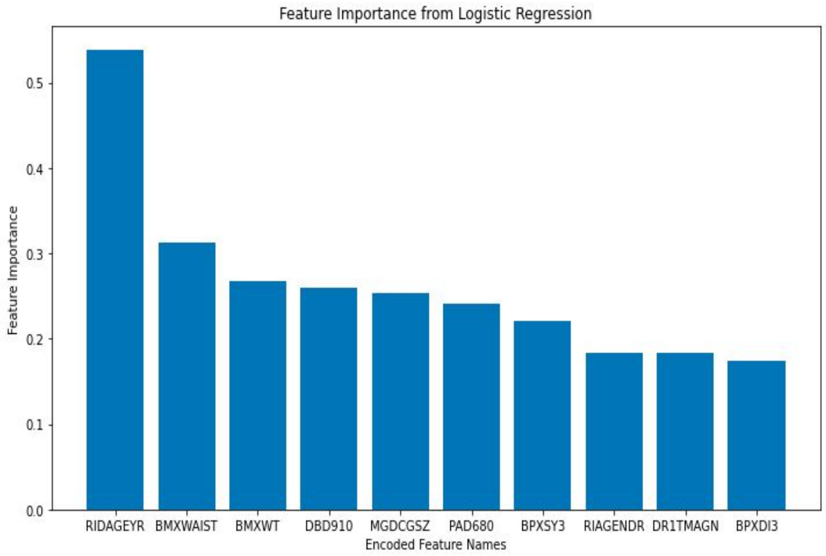
**Solution:**

- ➢ Select the 10 most important features from logistic regression model
- ➢ Design a survey based on these 10 features
- ➢ Non-linear models such as Random Forest shows promising results

# Building for Deployment Cont'd.

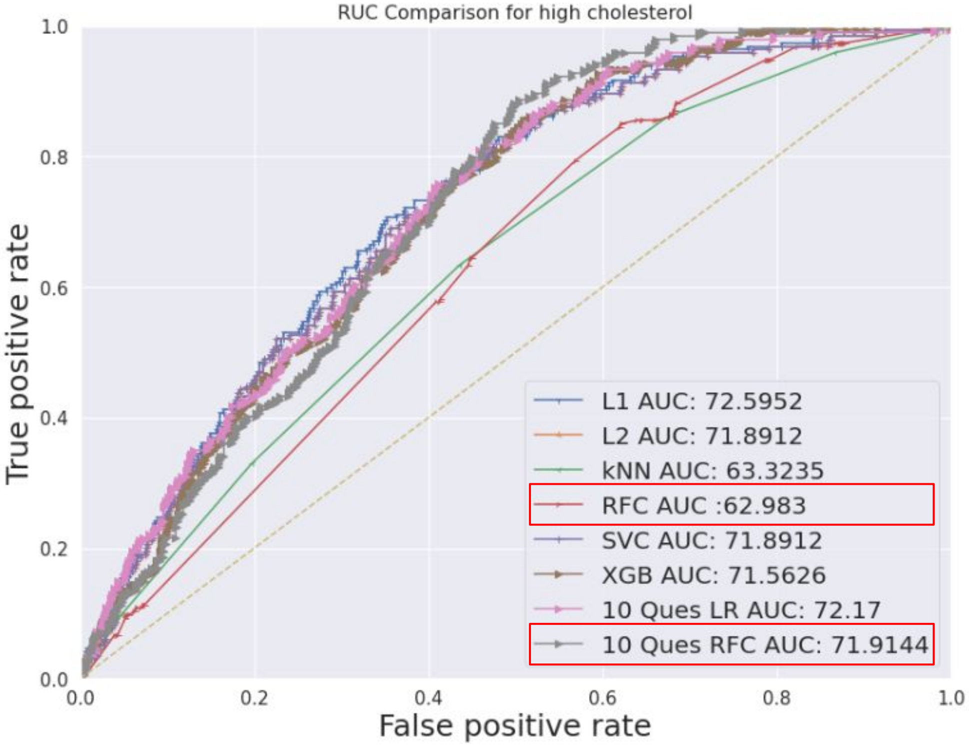
**Ten Most Important Features for Random Forest (Left) and Logistic Regression (Right):**




- ➢ Age and waist circumference are ranked as the 2 most important features by both methods
- ➢ Features related to blood pressure are important as well
- ➢ Majority of the features reflect the participants' diet habit and lifestyle

# Building for Deployment Cont'd.




| | Model Type | |
|---|---|---|
| **Metrics** | **Logistic Regression** | **Random Forest** |
| **Recall** | **0.77** | **0.84** |
| **Accuracy** | **0.61** | **0.56** |

# Conclusion

- ➢ Random forest algorithm can build the best performance model
- ➢ Questionnaire with only 10 questions can be used to classify the patient's cholesterol level
- ➢ Dataset Limitations: Only represent a population from USA



https://www.dreamstime.com/online-survey-internet-surveying-hands-holding-tablet-test-form-mobile-questionnaire-customers-voting-vector-concept-image134503931
https://www.bhf.org.uk/informationsupport/heart-matters-magazine/medical/tests/blood-cholesterol

# Thank You!

Questions?