Cardiovascular Diseases Risk Prediction

Kevin Shi & Rithvik Sukumaran

Presentation Roadmap

1. Introduction and Problem
2. Overview of EDA
3. Models and Results
4. Conclusion

Introduction

- Cardiovascular diseases (CVDs) are a significant health issue for millions around the world.
- Existing health conditions and lifestyle factors play a crucial role in determining an individual's vulnerability to CVDs.
- Objective:
  - Develop an understanding of the risk factors associated with CVDs
  - Create predictive models that can assess an individual's vulnerability to CVDs
- Our project highlights the significance of the application of AI in medicine
  - Support understanding, knowledge, & new findings
  - Inform data driven decisions

Our Dataset

- The Behavioral Risk Factor Surveillance System (BRFSS) collects comprehensive health-related data from U.S. residents through telephone surveys conducted by various state health departments in collaboration with the Centers for Disease Control and Prevention (CDC).
- Our dataset contains 308854 rows (records) and 19 columns (features)
  - Of the 19 features, 7 are numerical, 9 are categorical, and 3 are ordinal
- Features in the dataset are related to lifestyle factors that have commonly been associated with an increased risk of various CVDs

```
Index(['General_Health', 'Checkup', 'Exercise', 'Heart_Disease', 'Skin_Cancer',
       'Other_Cancer', 'Depression', 'Diabetes', 'Arthritis', 'Sex',
       'Age_Category', 'Height_(cm)', 'Weight_(kg)', 'BMI', 'Smoking_History',
       'Alcohol_Consumption', 'Fruit_Consumption',
       'Green_Vegetables_Consumption', 'FriedPotato_Consumption'],
      dtype='object')
```

Details of Our Methodology

Task:

Predict CVD from a few select risk factors and identify which risk factors are the most significant

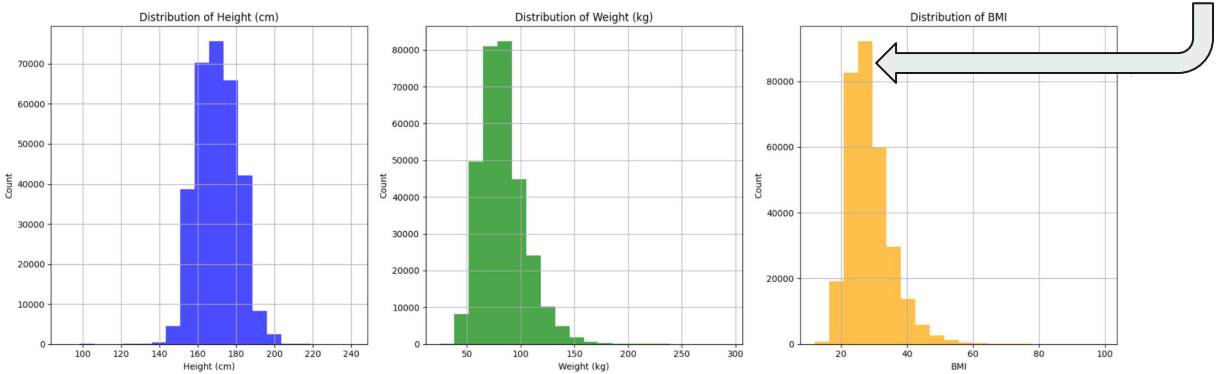
Solution:

Data Analysis and Machine Learning

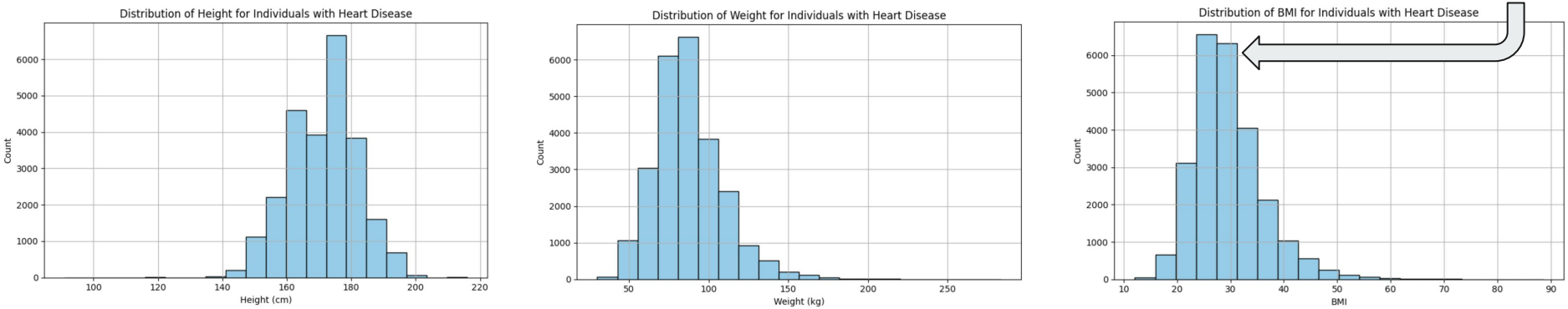
Tools:
Python, sklearn, pandas, matplotlib, Google colab

EDA: CVD by Height, Weight, & BMI

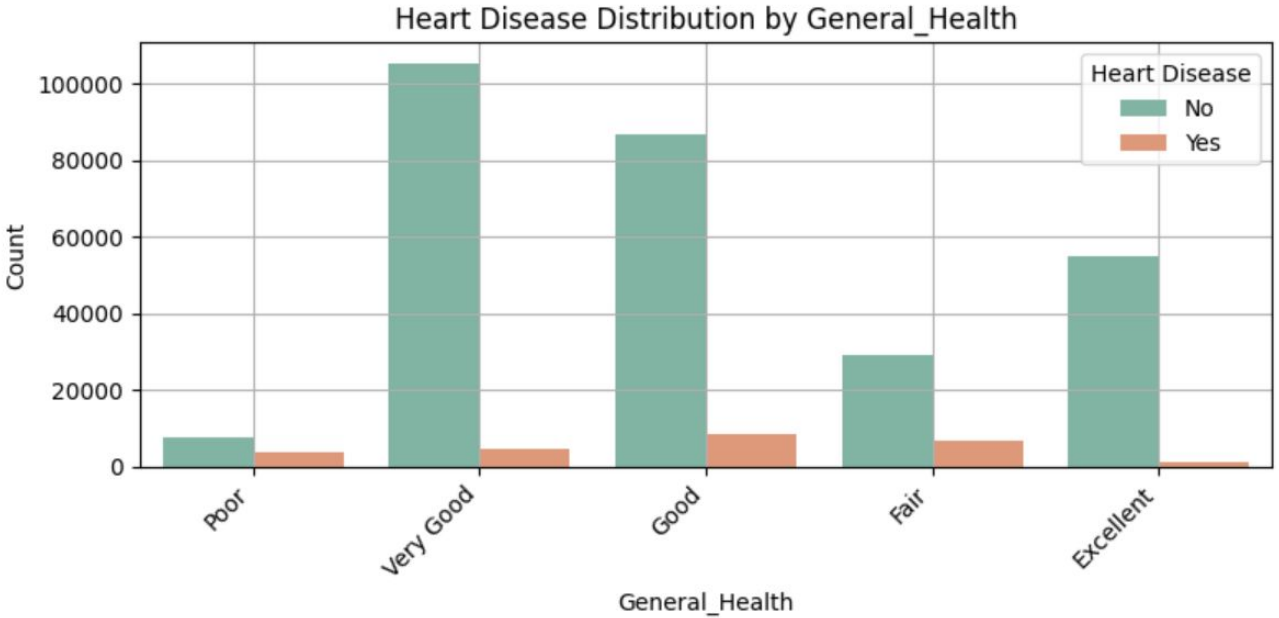
Median BMI: 20-30



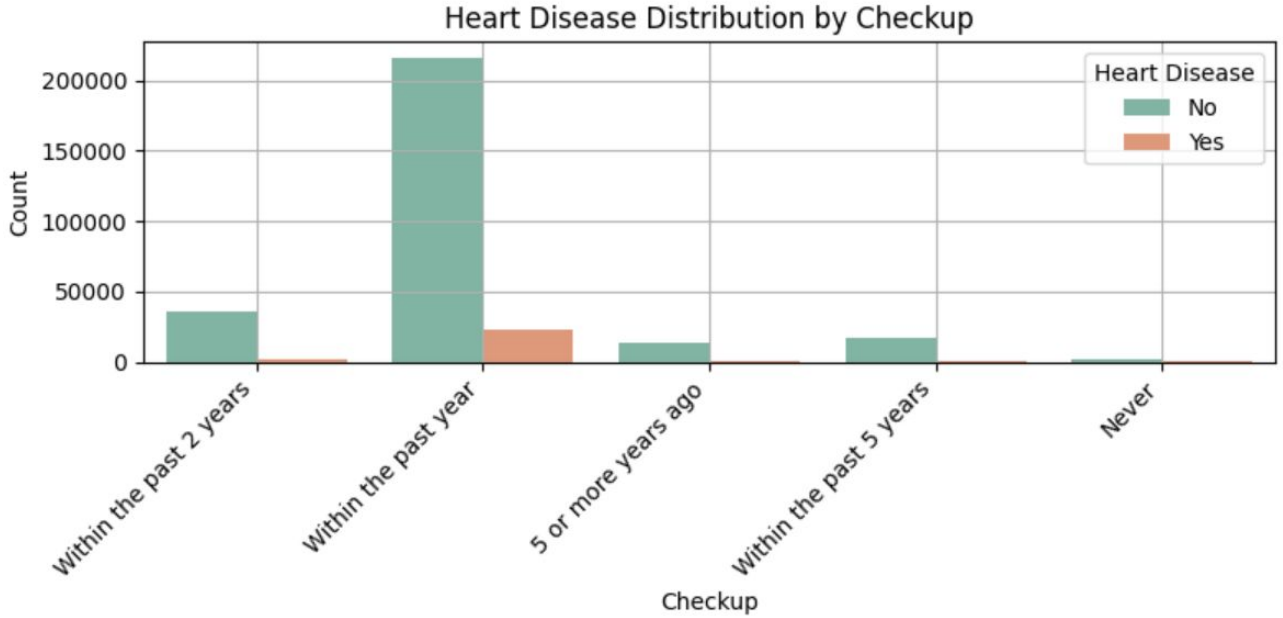
Median BMI: 25-30



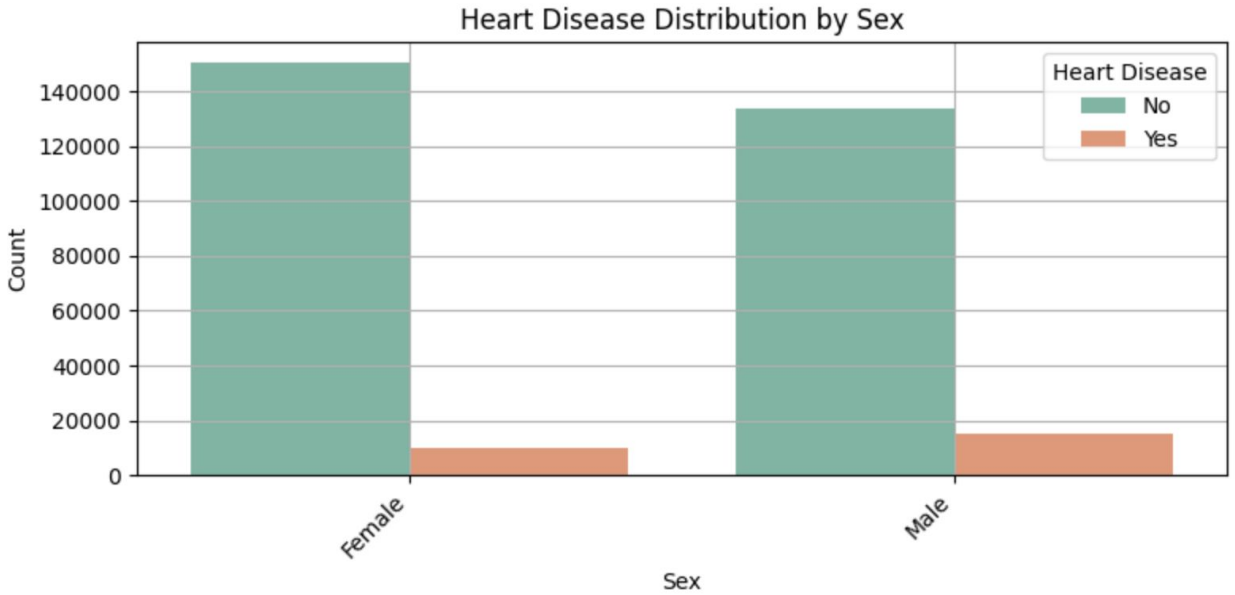
EDA: CVD by General Health



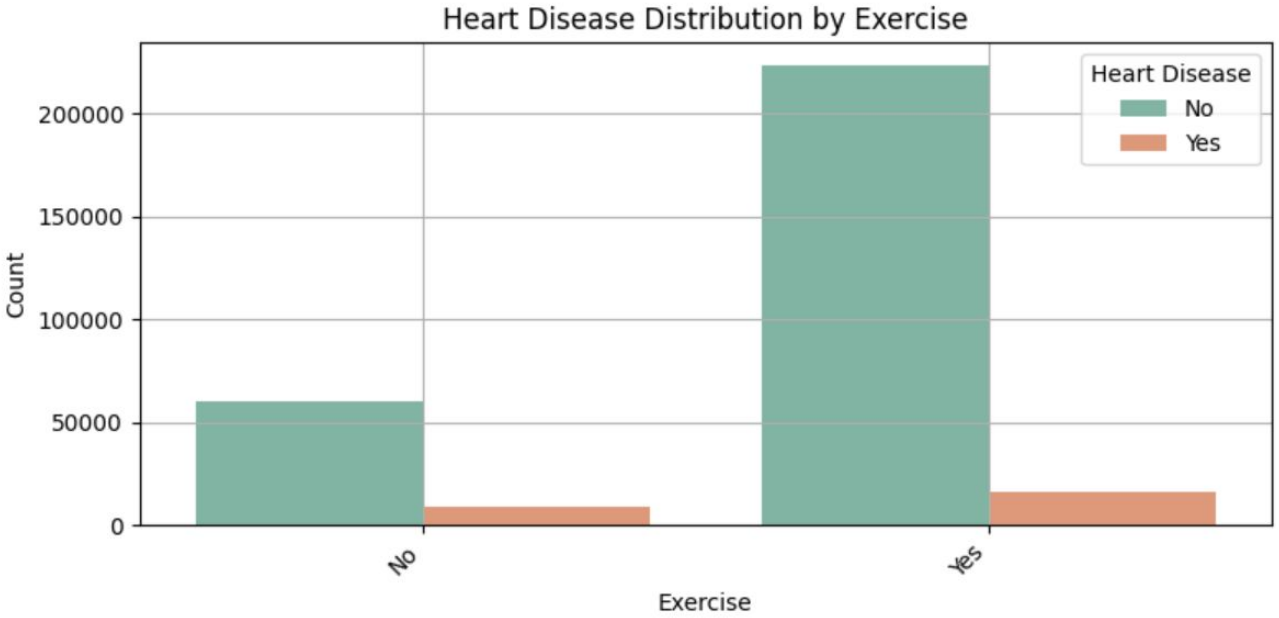
EDA: CVD by Checkup Frequency



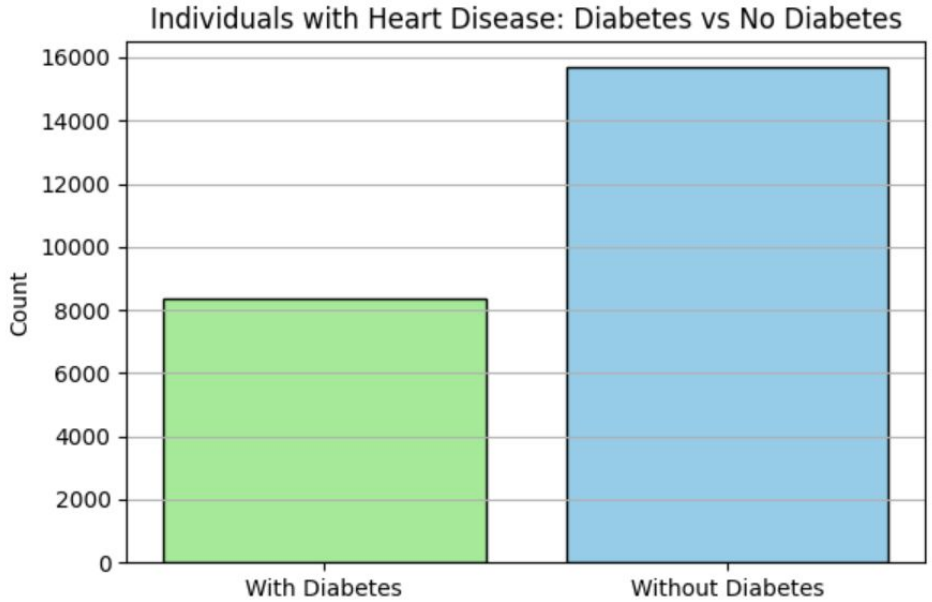
EDA: CVD by Sex



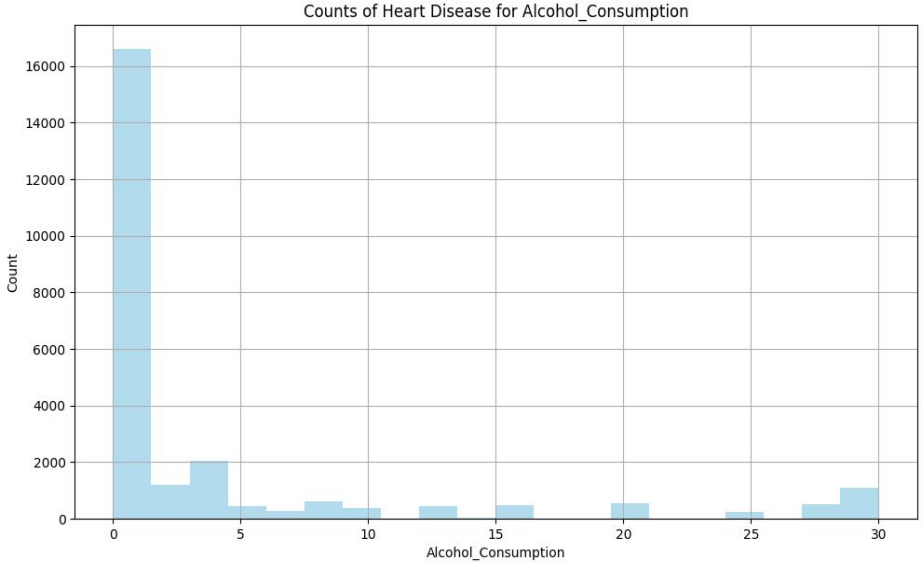
EDA: CVD by Exercise

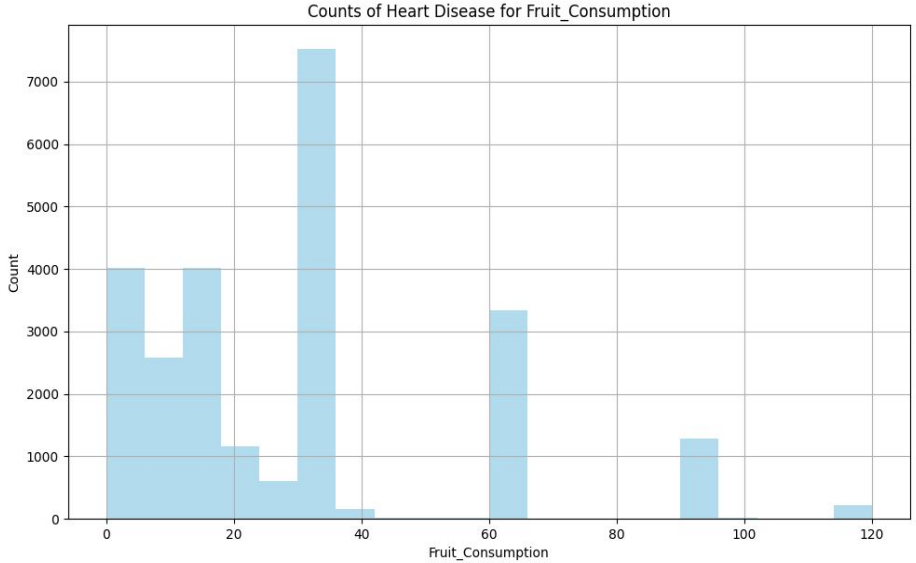


EDA: CVD by Diabetes

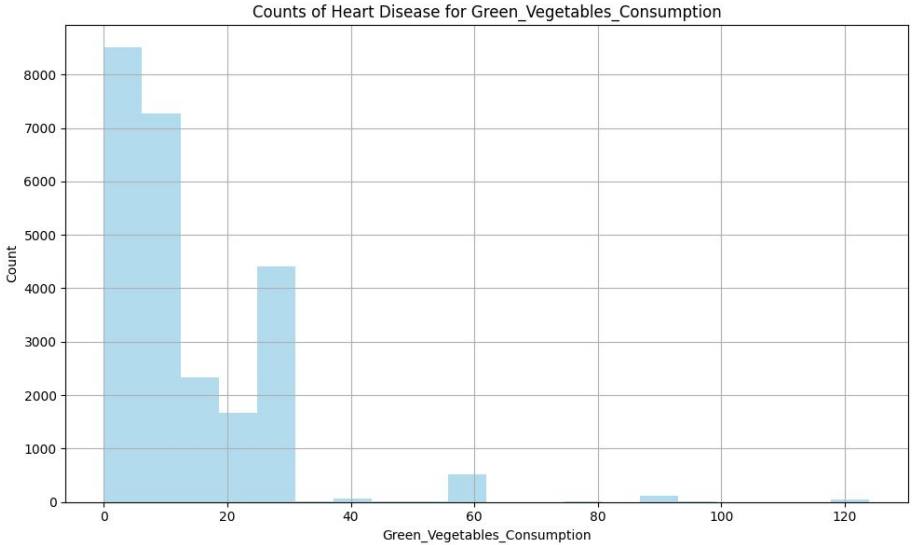


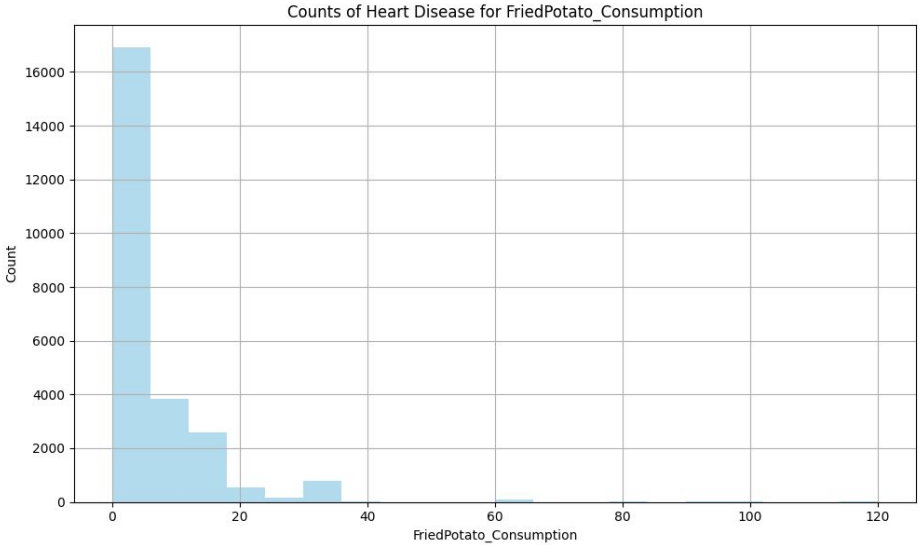
EDA: CVD by Alcohol & Fruit Consumption





EDA: CVD by Vegetable & Fried Potato Consumption





Preprocessing for Modeling

```
ohedf.columns
```

```
Index(['Exercise', 'Heart_Disease', 'Skin_Cancer', 'Other_Cancer',
       'Depression', 'Diabetes', 'Arthritis', 'Sex', 'Height_(cm)',
       'Weight_(kg)', 'BMI', 'Smoking_History', 'Alcohol_Consumption',
       'Fruit_Consumption', 'Green_Vegetables_Consumption',
       'FriedPotato_Consumption', 'General_Health_Excellent',
       'General_Health_Fair', 'General_Health_Good', 'General_Health_Poor',
       'General_Health_Very Good', 'Checkup_5 or more years ago',
       'Checkup_Never', 'Checkup_Within the past 2 years',
       'Checkup_Within the past 5 years', 'Checkup_Within the past year',
       'Age_Category_18-24', 'Age_Category_25-29', 'Age_Category_30-34',
       'Age_Category_35-39', 'Age_Category_40-44', 'Age_Category_45-49',
       'Age_Category_50-54', 'Age_Category_55-59', 'Age_Category_60-64',
       'Age_Category_65-69', 'Age_Category_70-74', 'Age_Category_75-79',
       'Age_Category_80+'],
      dtype='object')
```

Models

- KNN
  - 10-Fold Cross Validation
  - Tested with 5, 10, 15, 20, 25, and 30 Neighbors
- Random Forest
  - Tested with 5-Fold, 10-Fold, 15-Fold, and 20-Fold Cross Validation
- Naive Bayes
  - Tested with 5-Fold, 10-Fold, 15-Fold, and 20-Fold Cross Validation

Experimental Results

Naive Bayes

	Precision	Recall	f1	Accuracy	num_folds
0	0.197206	0.471347	0.277879	0.801757	5
1	0.197510	0.472109	0.278205	0.801991	10
2	0.198835	0.471428	0.278534	0.801900	15
3	0.198219	0.471789	0.278480	0.802000	20

KNN

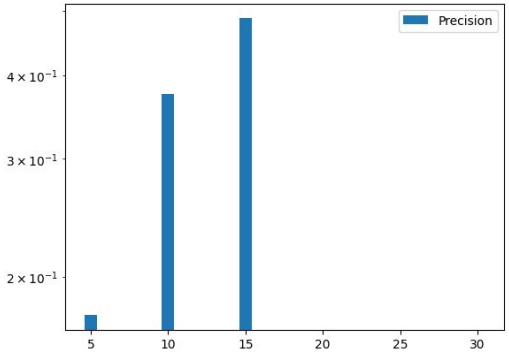
	Precision	Recall	f1	Accuracy	num_neighbors
0	0.174976	0.013135	0.024428	0.915206	5
1	0.375000	0.000761	0.001518	0.919111	10
2	0.486667	0.000561	0.001120	0.919162	15
3	0.000000	0.000000	0.000000	0.919146	20
4	0.000000	0.000000	0.000000	0.919146	25
5	0.000000	0.000000	0.000000	0.919150	30

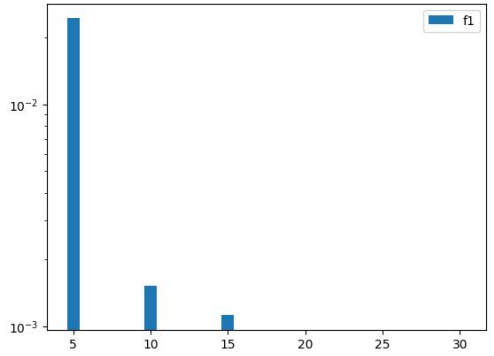
Random Forest

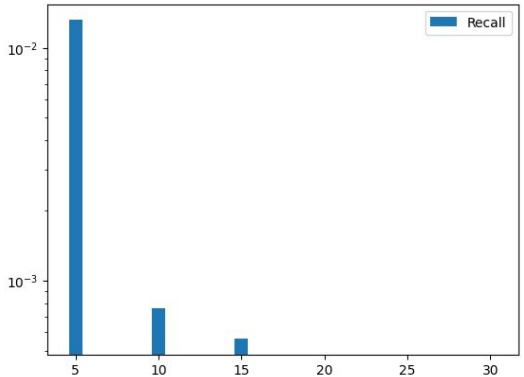
	Precision	Recall	f1	Accuracy	num_folds
0	0.441357	0.044211	0.080358	0.918194	5
1	0.443334	0.045613	0.082675	0.918188	10
2	0.439369	0.044972	0.081411	0.918104	15
3	0.444500	0.045173	0.081898	0.918211	20

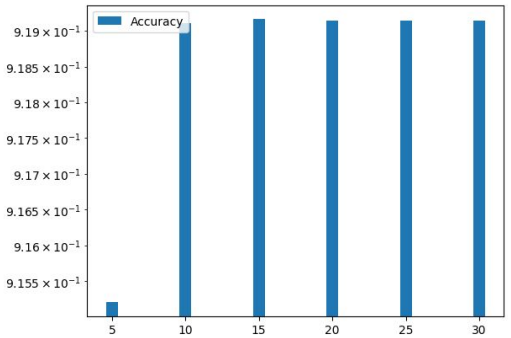
	Model	F1
1	Logistic Regression	0.32564
2	Naive Bayes	0.26982
3	Decision Tree Classifier	0.22237
4	K Neighbors Classifier	0.27350
5	Random Forest Classifier	0.17830

KNN


Precision
4×10^{-1}
3×10^{-1}
2×10^{-1}
5
10
15
20
25
30

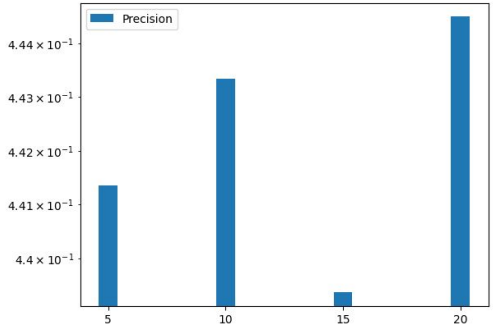

f1
10^{-2}
10^{-3}
5
10
15
20
25
30

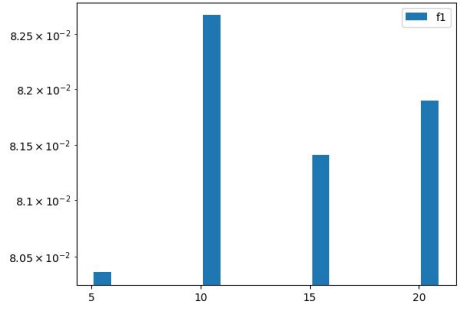

Recall
10^{-2}
10^{-3}
5
10
15
20
25
30

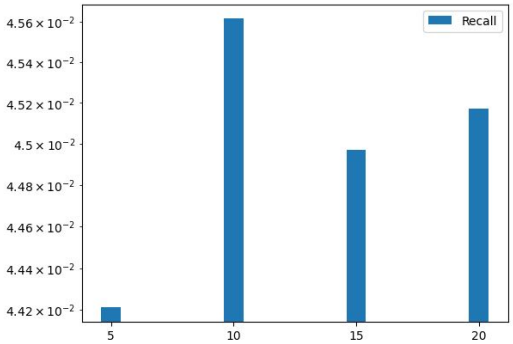

Accuracy
9.19×10^{-1}
9.185×10^{-1}
9.18×10^{-1}
9.175×10^{-1}
9.17×10^{-1}
9.165×10^{-1}
9.16×10^{-1}
9.155×10^{-1}
5
10
15
20
25
30

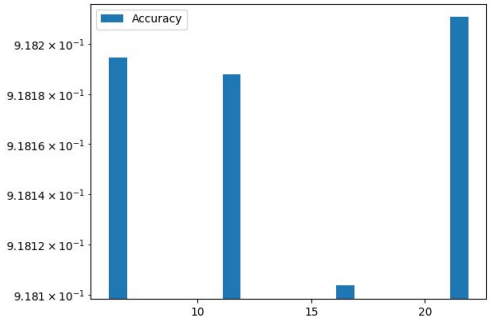
Random Forest

```
array([0.01960809, 0.01427426, 0.01579944, 0.0187046 , 0.02144972,
       0.01837596, 0.01548138, 0.09530568, 0.1258011 , 0.14161415,
       0.01630808, 0.06008439, 0.09162513, 0.09413556, 0.08859432,
       0.00640482, 0.01449094, 0.00771149, 0.02193178, 0.0079308 ,
       0.00238815, 0.00068243, 0.00562183, 0.00299067, 0.00818475,
       0.00140625, 0.0014288 , 0.00206393, 0.00270284, 0.00331855,
       0.00410394, 0.00517215, 0.00681821, 0.00837412, 0.00927102,
       0.01075815, 0.01135102, 0.01773149])
```









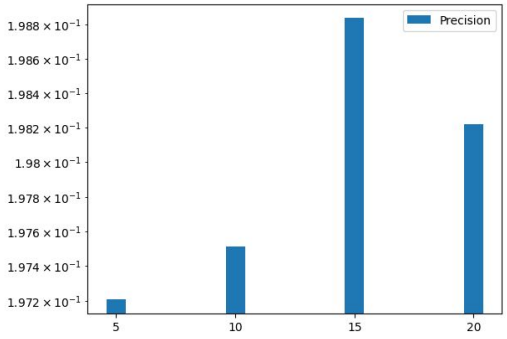
Random Forest Feature Importance

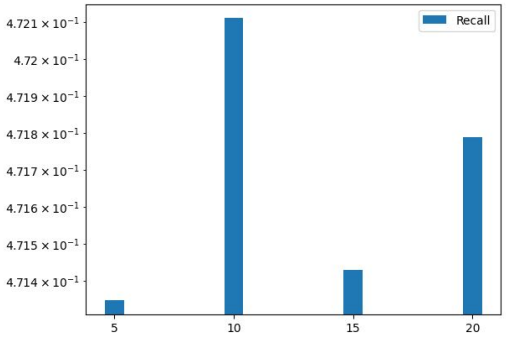
index	feature	importance
9	BMI	0.1416141488530385
8	Weight_(kg)	0.12580110206974682
7	Height_(cm)	0.095305677169544
13	Green_Vegetables_Consumption	0.09413555643070755
12	Fruit_Consumption	0.09162513491299387
14	FriedPotato_Consumption	0.08859432127399156
11	Alcohol_Consumption	0.060084390002987556
18	General_Health_Poor	0.02193178433355509
4	Diabetes	0.021449720928456668
0	Exercise	0.019608088628981078

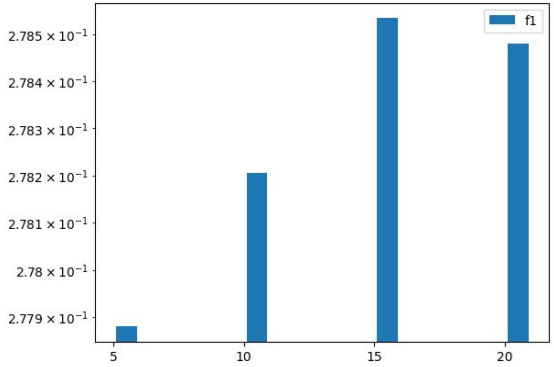
Naive Bayes

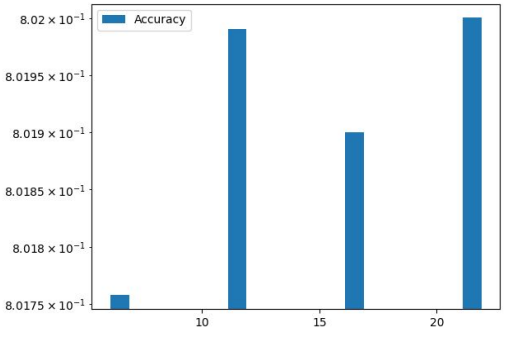
```
nb.feature_log_prob_
```

```
array([[ -6.08105346,  -8.26327604,  -8.25710757,  -7.46889393,
         -7.79044706,  -7.02852677,  -6.59639476,  -0.7033811 ,
         -1.41998473,  -2.49091597,  -6.78556852,  -4.19645163,
         -2.44098568,  -3.11874165,  -3.999295  ,  -7.48790398,
         -8.12422024,  -7.02536255,  -9.45128236,  -6.83266903,
         -8.93068506, -11.18605692,  -7.91817557,  -8.65699069,
         -6.11133183,  -8.56378528,  -8.76798576,  -8.59306486,
         -8.47787562,  -8.43051283,  -8.48194761,  -8.3249523 ,
         -8.22672143,  -8.10991895,  -8.10656749,  -8.20688181,
         -8.65276599,  -8.63781085],
        [ -6.29246387,  -7.5243825 ,  -7.5212606 ,  -7.25970316,
         -6.838367  ,  -6.41156422,  -6.34918824,  -0.70440368,
         -1.3818661 ,  -2.4600255 ,  -6.38747557,  -4.43902832,
         -2.51125458,  -3.21842546,  -4.03897913,  -8.9625987 ,
         -7.1495114 ,  -6.91247497,  -7.77358756,  -7.48864447,
        -10.09531904, -11.83988282,  -8.67881643,  -9.80115246,
         -5.94659426, -11.39957098, -11.27118981, -10.6595925 ,
        -10.36702044,  -9.89701681,  -9.46944025,  -8.88652516,
         -8.36224988,  -7.95328674,  -7.73110393,  -7.53898548,
         -7.75337027,  -7.48727404]])
```









Naive Bayes Probabilities

index	feature	0	1
7	Height_(cm)	-0.7033811003828951	-0.7044036810446599
8	Weight_(kg)	-1.4199847296416799	-1.3818661026698766
9	BMI	-2.4909159741193534	-2.460025500146795
12	Fruit_Consumption	-2.440985680117972	-2.511254575254359
13	Green_Vegetables_Consumption	-3.1187416486812864	-3.218425458348131
14	FriedPotato_Consumption	-3.999294997234893	-4.038979134662814
11	Alcohol_Consumption	-4.196451627305674	-4.439028318393484
24	Checkup_Within the past year	-6.111331828951377	-5.9465942572964
0	Exercise	-6.081053458858651	-6.29246387358071
6	Sex	-6.59639476163049	-6.349188241643962

Bonus: Neural Network

```
Epoch 1/10
7239/7239 [==============================] - 19s 2ms/step - loss: 0.2456 - accuracy: 0.9156 - precision_3: 0.3977 - recall_3: 0.0749
Epoch 2/10
7239/7239 [==============================] - 23s 3ms/step - loss: 0.2327 - accuracy: 0.9174 - precision_3: 0.4513 - recall_3: 0.0783
Epoch 3/10
7239/7239 [==============================] - 14s 2ms/step - loss: 0.2296 - accuracy: 0.9180 - precision_3: 0.4696 - recall_3: 0.0693
Epoch 4/10
7239/7239 [==============================] - 15s 2ms/step - loss: 0.2282 - accuracy: 0.9183 - precision_3: 0.4782 - recall_3: 0.0653
Epoch 5/10
7239/7239 [==============================] - 14s 2ms/step - loss: 0.2272 - accuracy: 0.9184 - precision_3: 0.4846 - recall_3: 0.0593
Epoch 6/10
7239/7239 [==============================] - 14s 2ms/step - loss: 0.2263 - accuracy: 0.9185 - precision_3: 0.4899 - recall_3: 0.0616
Epoch 7/10
7239/7239 [==============================] - 14s 2ms/step - loss: 0.2256 - accuracy: 0.9188 - precision_3: 0.5019 - recall_3: 0.0566
Epoch 8/10
7239/7239 [==============================] - 14s 2ms/step - loss: 0.2253 - accuracy: 0.9187 - precision_3: 0.5003 - recall_3: 0.0486
Epoch 9/10
7239/7239 [==============================] - 14s 2ms/step - loss: 0.2249 - accuracy: 0.9189 - precision_3: 0.5136 - recall_3: 0.0431
Epoch 10/10
7239/7239 [==============================] - 14s 2ms/step - loss: 0.2245 - accuracy: 0.9191 - precision_3: 0.5238 - recall_3: 0.0438
<keras.callbacks.History at 0x7e99774aeda0>
```

```
2413/2413 [==============================] - 8s 3ms/step - loss: 0.2252 - accuracy: 0.9172 - precision_3: 0.4319 - recall_3: 0.1264
[0.22517704963684082,
 0.9172041416168213,
 0.4319066107273102,
 0.12638255953788757]
```

Conclusion Main Idea - Contributing Factors to CVDs

- Our Feature Importance
  - Height, weight, BMI
  - Fruit, Green Vegetable, Fried Potato, Alcohol Consumption
  - General health
  - Checkup Frequency
  - Diabetes
  - Sex
  - Exercise
- Research Paper's Feature Importance
  - Sex
  - Diabetes
  - General Health

Summary

- We have achieved high accuracy with a variety of different models
- Contrary to one of our initial hypotheses, the neural network proved surprisingly effective for this simple task
- Challenges we faced:
  - Training the models on large volumes of data
  - Learning the tensorflow APi
- Moving forward:
  - Create a more specialized neural network
  - Tune more hyperparameters
  - Address the severe overfitting issues (possibly with data augmentation)