FACTORS IMPACT HEALTH

TEAM MEMBERS

One

Two

Three

Slide on who is on your team, and their background

Centers for Disease Control and Prevention. (2017-2020). National Health and Nutrition Examination Survey (NHANES) 2017-2020. Retrieved from https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/P_DEMO.XPT

Hypothesis:

There is a **positive** correlation between heart disease, obesity, and an unhealthy diet, indicating that individuals who are obese and have poor dietary habits are **more likely to develop heart disease.**

As a weight management company, our initial focus was on empowering clients and patients to take control of their health. To support this goal, we formulated a hypothesis suggesting a **correlation between heart disease, obesity, and an unhealthy diet.**

National Library of Medicine: Cardiovascular health market is projected to increase at an annual compound rate of 1.8% per year.

Centers for Disease Control and Prevention. (2017-2020). National Health and Nutrition Examination Survey (NHANES) 2017-2020. Retrieved from https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/P_DEMO.XPT

What data did we mine?

National Health and Nutrition Examination Survey

NHANES

- A program run by the Centers for Disease Control and Prevention (CDC)
- States heart disease is the leading cause of death for men and women across most racial and ethnic groups.
- Evaluates the health and nutritional status of people in the United States through a combination of interviews and physical examinations.

Table 1. Questionnaire components: National Health and Nutrition Examination Survey, 1999–2022—Con.

Component or laboratory test conducted on original sample description | Change from original sample description | Component or laboratory test not conducted

Section	Sample description	1999–2000	2001–2002	2003–2004	2005–2006	2007–2008	2009–2010	2011–2012	2013–2014	2015–2016	2017–2018	2019–2020	2021–2022
Oral health (OHQ)	2 years and over												
Osteoporosis (OSQ)	20 years and over												
Physical activity and physical fitness (PAQ)	2 years and over												
Physical functioning (PFQ)	1 year and over												
Respiratory health and disease (RDQ)	1 year and over												
Sleep disorders (SLQ)	16 years and over												
Smoking and tobacco use (SMQ)	20 years and over												
Social support (SSQ)	60 years and over												
Taste and smell (CSQ)	40 years and over												
Tuberculosis (TBQ)	1 year and over												
Vision (VIQ)	50 years and over												
Weight history (WHQ)	16 years and over												
Family questionnaire													
Consumer behavior (CBQ)	All ages												
Demographic background (HFQ) occupation (DMQ)	All ages												
Food security (FSQ)	All ages												
Health insurance (HIQ)	All ages												
Housing characteristics (HOQ)	All ages												
Income (INQ)	All ages												
Pesticide use (PUQ)	All ages												
Salt use (SUQ)	All ages												
Smoking household (SMQ)	All ages												
Mobile examination center (MEC) questionnaire													
Audio computer-assisted self-interview (ACASI)	8 years and over												
Alcohol use (ALQ)	12–19 years												
Alcohol use (ALQ)	20 years and over												
Current health status (HSQ)	1 year and over												
Drug use (DUQ)	12–69 years												

2017-2020 SURVEY

- The NHANES dataset includes variables such as age, gender, race/ethnicity, socioeconomic status, body measurements, blood pressure, cholesterol levels, dietary intake, physical activity, and prevalence of various health conditions.
- By considering these variables, we can conduct analyses to better comprehend the connections between various factors and health outcomes.

III. Data Output and Metrics

Centers for Disease Control and Prevention. (2017-2020). National Health and Nutrition Examination Survey (NHANES) 2017-2020. Retrieved from https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/P_DEMO.XPT

Libraries Imported: *pandas, numpy, matplotlib, seaborn, sklearn (NearestNeighbors, train_test_split, LogisticRegression, accuracy_score)*

Data Preparation

Dataset Selections

- Demographic
- Body Measurement
- Blood Pressure
- Diet Interview

Feature Selections

- least missing data
- relevant to the topic

Problem:

The correlation between nutrients and BMI are all negative.

```
data.corr()
```

	Weight_KG	BMI	Protein_Rate	Carbohydrate_Rate	Total_Sugar_Rate	Fiber_Rate	Total_Fat_Rate	Calories_Rate
Weight_KG	1.000000	0.917481	-0.592870	-0.660801	-0.586515	-0.574439	-0.580374	-0.651992
BMI	0.917481	1.000000	-0.530418	-0.573214	-0.499910	-0.498060	-0.518056	-0.577844
Protein_Rate	-0.592870	-0.530418	1.000000	0.792157	0.684131	0.712666	0.849197	0.888471
Carbohydrate_Rate	-0.660801	-0.573214	0.792157	1.000000	0.929391	0.791945	0.831960	0.959779
Total_Sugar_Rate	-0.586515	-0.499910	0.684131	0.929391	1.000000	0.649022	0.733841	0.866879
Fiber_Rate	-0.574439	-0.498060	0.712666	0.791945	0.649022	1.000000	0.689101	0.776706
Total_Fat_Rate	-0.580374	-0.518056	0.849197	0.831960	0.733841	0.689101	1.000000	0.940897
Calories_Rate	-0.651992	-0.577844	0.888471	0.959779	0.866879	0.776706	0.940897	1.000000

III. Data Output and Metrics

Centers for Disease Control and Prevention. (2017-2020). National Health and Nutrition Examination Survey (NHANES) 2017-2020. Retrieved from https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/P_DEMO.XPT

Libraries Imported: *pandas, numpy, matplotlib, seaborn, sklearn (NearestNeighbors, train_test_split, LogisticRegression, accuracy_score)*

Data Preparation

Dataset Selections

- Demographic
- Body Measurement
- Blood Pressure
- Medical_Conditions
- Cardiovascular Health

Analysis Strategy

We created a column based on BMI. When BMI > 30, Obesity is 1

BMI < 30, Obesity is 0.

Problem:

R-Squared is 0.1

In [42]:

```
# Implementing the model
import statsmodels.api as sm
logit_model=sm.Logit(y,x)
result=logit_model.fit()
print(result.summary2())
```

```
Optimization terminated successfully.
         Current function value: 0.617629
         Iterations 8
                            Results: Logit
================================================================================
Model:                  Logit               Pseudo R-squared:    0.101
Dependent Variable:     y                   AIC:                 5619.3090
Date:                   2023-05-18 22:46    BIC:                 5786.0570
No. Observations:       4507                Log-Likelihood:      -2783.7
Df Model:               25                  LL-Null:             -3095.7
Df Residuals:           4481                LLR p-value:         1.1468e-115
Converged:              1.0000              Scale:               1.0000
No. Iterations:         8.0000
--------------------------------------------------------------------------------
                            Coef.  Std.Err.    z    P>|z|   [0.025  0.975]
--------------------------------------------------------------------------------
Age                        -1.2127   0.1345 -9.0176 0.0000 -1.4763 -0.9491
Ratio_of_poverty            0.3531   0.1196  2.9534 0.0031  0.1188  0.5875
Blood_Pressure              0.3282   0.0861  3.8127 0.0001  0.1595  0.4969
Asthma                      0.1763   0.0924  1.9084 0.0563 -0.0048  0.3573
Arthritis                   0.4375   0.0714  6.1317 0.0000  0.2977  0.5774
Stroke                      0.0518   0.1326  0.3907 0.6960 -0.2081  0.3117
Thyroid_Problem             0.2677   0.0933  2.8688 0.0041  0.0848  0.4506
Liver_Disease               0.0705   0.1336  0.5278 0.5976 -0.1913  0.3323
Abdominal_Pain             -0.1860   0.0830 -2.2418 0.0250 -0.3485 -0.0234
gallstones                  0.6315   0.0986  6.4044 0.0000  0.4382  0.8247
Cancer                     -0.0796   0.0968 -0.8219 0.4111 -0.2693  0.1102
Had_Chest_Pain              0.0451   0.0780  0.5778 0.5634 -0.1078  0.1980
Shortness_Breath_On_Stairs  0.6572   0.0723  9.0866 0.0000  0.5154  0.7990
Heart_Disease               0.1414   0.1059  1.3349 0.1819 -0.0662  0.3489
```

III. Data Output and Metrics

Centers for Disease Control and Prevention. (2017-2020). National Health and Nutrition Examination Survey (NHANES) 2017-2020. Retrieved from https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/P _DEMO.XPT

Prepped Dataset

1000 rows of data were analyzed, encompassing various health conditions such as obesity, chest pain, and cancer.

0 = No Heart Disease
1 = Yes heart disease

	SEQN	BMI	Age	Gender	Ratio_of_poverty	Obesity	Blood_Pressure	Asthma	Arthritis	Stroke	Thyroid_Problem	Liver_Disease	Abdominal_Pain	gallstones	Cancer	Had_Chest_P
975	112747.0	22.8	66.0	1.0	2.98	0	3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1634	114985.0	24.2	80.0	1.0	1.40	0	3	0.0	1.0	0.0	1.0	0.0	1.0	0.0	0.0	
3467	121290.0	26.1	58.0	2.0	5.00	0	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	
4130	123536.0	37.1	65.0	1.0	3.08	1	3	1.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	
4054	123258.0	29.2	77.0	2.0	5.00	0	2	0.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
4131	123537.0	23.5	45.0	2.0	2.43	0	2	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	
4144	123578.0	30.9	68.0	1.0	1.44	1	2	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	
178	109963.0	24.2	48.0	2.0	0.88	0	3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
2760	118839.0	27.0	58.0	1.0	5.00	0	2	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	
2571	118216.0	22.2	55.0	2.0	3.82	0	2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

1000 rows × 46 columns

Data Modeling

Logistic Regression Model

In [87]:
```
#Creating Model for Logistic Regression

y = Data1.Heart_Disease.values
x_data = Data1.drop(['Heart_Disease',"SEQN"], axis = 1)
```

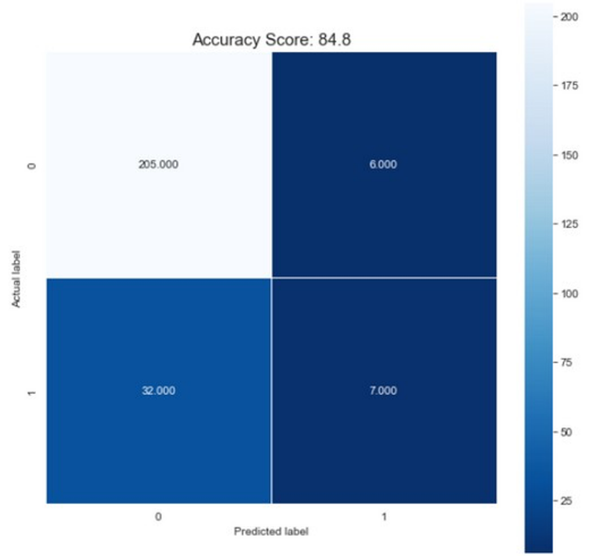
In [88]:
```
# Normalize
x = (x_data - np.min(x_data)) / (np.max(x_data) - np.min(x_data)).values
```

In [89]:
```
#We will split our data. 80% of our data will be train data and 25% of it will be test data.
x_train, x_test, y_train, y_test = train_test_split(x,y,test_size = 0.25,random_state=0)
```

In [90]:
```
accuracies = {}
# all parameters not specified are set to their defaults
lr = LogisticRegression()
# Model is learning the relationship between digits (x_train) and labels (y_train)
lr.fit(x_train,y_train)
acc = lr.score(x_test,y_test)*100

accuracies['Logistic Regression'] = acc
print("Test Accuracy {:.2f}%".format(acc))
```

```
Test Accuracy 84.80%
```

Confusion Matrix

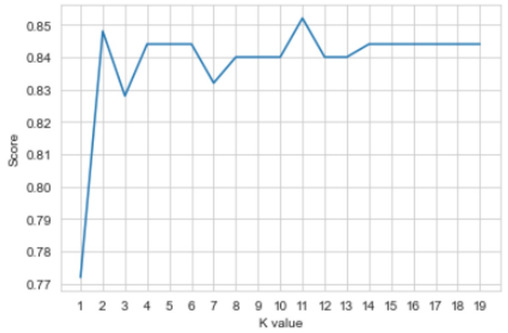


III. Data Output and Metrics

Centers for Disease Control and Prevention. (2017-2020). National Health and Nutrition Examination Survey (NHANES) 2017-2020. Retrieved from https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/P_DEMO.XPT

Maximum KNN Score

```
# try ro find best k value
from sklearn.neighbors import KNeighborsClassifier
scoreList = []
for i in range(1,20):
    knn2 = KNeighborsClassifier(n_neighbors = i)  # n_neighbors means k
    knn2.fit(x_train, y_train)
    scoreList.append(knn2.score(x_test, y_test))

plt.plot(range(1,20), scoreList)
plt.xticks(np.arange(1,20,1))
plt.xlabel("K value")
plt.ylabel("Score")
plt.show()

acc = max(scoreList)*100
accuracies['KNN'] = acc
print("Maximum KNN Score is {:.2f}%".format(acc))
```




```
Maximum KNN Score is 85.20%
```

Data Modeling (cont.)

Maximum KNN Score = NN Score = 85.20%
Test Accuracy of SVM Algorithm = 84.40%
Accuracy of Naive Bayes = 71.20% Accuracy of Decision Trees = 76.80% Accuracy of Random Forest = 84.40%

Accuracy

In [95]:
```
# KNN Model

knn = KNeighborsClassifier(n_neighbors = 11)  # n_neighbors means k
knn.fit(x_train, y_train)
prediction = knn.predict(x_test)

print("{} NN Score: {:.2f}%".format(11, knn.score(x_test, y_test)*100))
```

```
11 NN Score: 85.20%
```

In [96]:
```
from sklearn.svm import SVC
svm = SVC(random_state = 1)
svm.fit(x_train, y_train)

acc = svm.score(x_test,y_test)*100
accuracies['SVM'] = acc
print("Test Accuracy of SVM Algorithm: {:.2f}%".format(acc))
```

```
Test Accuracy of SVM Algorithm: 84.40%
```

In [97]:
```
from sklearn.naive_bayes import GaussianNB
nb = GaussianNB()
nb.fit(x_train, y_train)

acc = nb.score(x_test,y_test)*100
accuracies['Naive Bayes'] = acc
print("Accuracy of Naive Bayes: {:.2f}%".format(acc))
```

```
Accuracy of Naive Bayes: 71.20%
```

In [98]:
```
from sklearn.tree import DecisionTreeClassifier
dtc = DecisionTreeClassifier()
dtc.fit(x_train, y_train)

acc = dtc.score(x_test, y_test)*100
accuracies['Decision Tree'] = acc
print("Decision Tree Test Accuracy {:.2f}%".format(acc))
```

```
Decision Tree Test Accuracy 76.80%
```

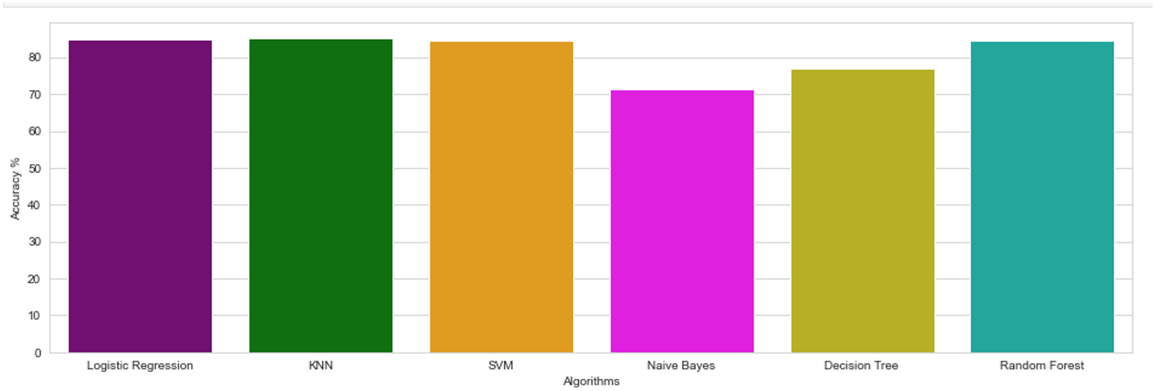
In [99]:
```
# Random Forest Classification
from sklearn.ensemble import RandomForestClassifier
rf = RandomForestClassifier(n_estimators = 1000, random_state = 1)
rf.fit(x_train, y_train)

acc = rf.score(x_test,y_test)*100
accuracies['Random Forest'] = acc
print("Random Forest Algorithm Accuracy Score : {:.2f}%".format(acc))
```

```
Random Forest Algorithm Accuracy Score : 84.40%
```

Centers for Disease Control and Prevention. (2017-2020). National Health and Nutrition Examination Survey (NHANES) 2017-2020. Retrieved from https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/P_DEMO.XPT

Data Evaluation & Modeling

Confusion Matrices

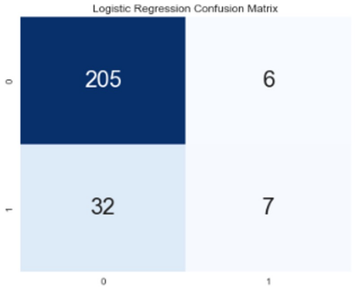


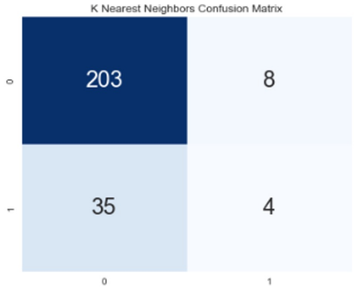
High True Negative (TN) Frequency:

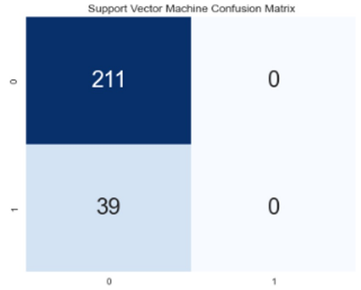
The six models (logical regression, KNN, SVM, Naive Bayes, Decision Tree, Random Forest) correctly predict the absence of heart disease and normal blood pressure in a significant number of cases.

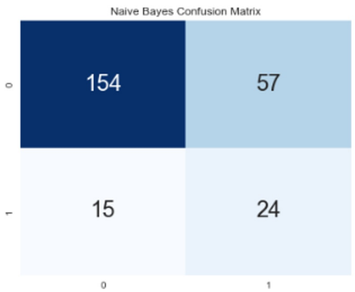
>> Indicates a strong, positive outcome for reliability and accuracy of the models.

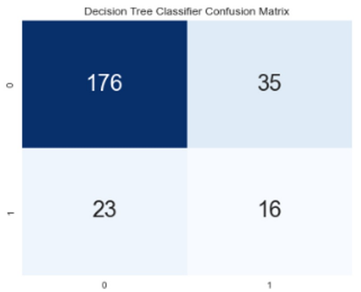
>> Suggests that the model is reliable in ruling out the presence of heart disease and identifying individuals who are at a lower risk

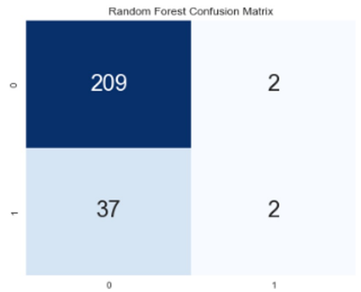












IV. Application of Findings

Centers for Disease Control and Prevention. (2017-2020). National Health and Nutrition Examination Survey (NHANES) 2017-2020. Retrieved from https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/P_DEMO.XPT

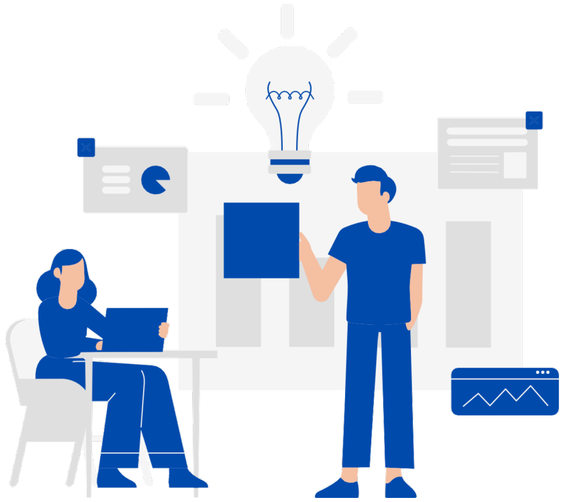
Hypothesis:

There is a positive correlation between heart disease, obesity, and an unhealthy diet, indicating that individuals who are obese and have poor dietary habits are more likely to develop heart disease.

Findings:

The strongest indicator for heart disease are **Had_Chest_Pain** and **Shortness_Breath_On_Stairs**

Business Concept



1. **Accurate Diagnosis and Treatment Planning:** Provide models and findings to healthcare professionals to help them confidently rule out individuals with normal blood pressure, and symptoms such as chest pain and shortness of breath from having heart disease.
2. **Lower Risk Identification:** By identifying individuals at lower risk, our model assists in targeting interventions and resources more effectively, minimizing unnecessary medical procedures and reducing costs.

Goal: Facilitate accurate diagnoses for patients and healthcare professions.

As of 2020, the cardiovascular disease market is at $49.8 billion. We're hoping for a reasonable estimate of 20% of the market share for our team's goal to assist with the accurate identification and diagnosis of heart disease.

Citations

Centers for Disease Control and Prevention. (2017-2020). National Health and Nutrition Examination Survey (NHANES) 2017-2020. Retrieved from https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/P_DEMO.XPT

Precedence Research. (n.d.). Cardiovascular Devices Market. Retrieved from https://www.precedenceresearch.com/cardiovascular-devices-market

Centers for Disease Control and Prevention. (n.d.). Heart disease facts. Retrieved from https://www.cdc.gov/heartdisease/facts.htm

Cardiovascular disease market set to grow very slowly to $146.4 billion by 2022, says GBI Research. (2016). Cardiovascular Journal of Africa, 27(5), 293