Nekeitrius Davis

March 1, 2023

DSC630 - T302 Predictive Analytics

Predicting Diabetes

Introduction:

Diabetes is the 7th leading cause of death currently in the United States and is associated with increase in certain cancers such as liver, pancreas, uterine, colon and breast. More than 37 million people in the US have diabetes while 96 million adults in the US are prediabetic and they have no idea.

For this course project I have chosen a dataset from a telephone questionnaire that the CDC conducted that focuses on questions that may relate to having diabetes, hypertension and/or a stroke. There are 18 different categories that were focused on in this dataset such as “Age”, “Sex”, “BMI”, “Smoker”, “Physical Activity”, etc. I found interest in this dataset because I know that diabetes is common throughout both sides of my families along with some family members experiencing strokes as well. It would be good to know that through this analysis I could help the next person with some preventative measures based on how related some categories are opposed to others.

One of the correlations that I am most interested in finding the correlation is fruits and vegetables relationship to one of the health conditions. I think it would be interesting to possibly -if not closely related – tell someone that the number of fruits and vegetables they consume is total unrelated to their health condition in this particular aspect. Also, just being able to look into different correlation between “simple” things that are so much more related to major health

complication is intriguing to know. Such as someone's self-reflection of their mental and physical health and how that can even help determine a greater thing.

What types of model or models do you plan to use and why?

I plan to use a clustering model to place the data into separate groups that are based on common characteristics. I would use this model in particular because I believe if I consider certain variables as a whole group opposed to them separately the results might show an influence on the chances of having one of the health conditions mentioned. I would also like to try and investigate time series models with this data set because since this data is from the past it will be able to help predict future outcomes depending on how someone answer certain questions. That would be a great early catch preventive measure.

How do you plan to evaluate your results?

When using clustering models such as K Means I will be plotting the data using scatter plot. With trial and error, if using the k means "silhouette" method approach, I will attempt to find the proper number of "Clusters" based on which data is most closely related in distance essentially. Those distances will give me different mean result which will determine how closely related the clusters are with 1 being the most related and -1 is the most unrelated. The silhouette approach to me seems a little more favorable because it is taking in account variable variance.

What do you hope to learn?

Nekeitrius Davis

March 1, 2023

DSC630 - T302 Predictive Analytics

I hope to strengthen my model building interpretation because I feel I grasps the idea of building models in a former class, but I really did not understand the results of the different models I built as good as I should have. I also hope to learn more about different predictive modeling techniques through trial-and-error analysis and seeing what the best fit for this particular data is as well as future datasets. I hope to learn some new python libraries that might be beneficial in my career field one day.

Assess any risks and ethical implications with your proposal.

A major risk in this data set along with any data set that I could run is to is that things could be unrelated overall despite how many models are tried to be used. Another risk could be not having enough data to make the proper conclusion about this topic. With clustering if I do not have enough data then I might not be able to access the data accurately.

Identify a contingency plan if your original project plan does not work out.

If my original plan doesn't work, I will investigate "PyHealth", I might even still look into it because it is a python-based model and algorithm tool that focuses on building models when working with healthcare data. It might be a more intensive way to look into health care models but considering it is meant for actual health care it seems like a pretty good second idea.

Nekeitrius Davis

March 1, 2023

DSC630 - T302 Predictive Analytics

Will I be able to answer the questions I want to answer with the data I have?

I will still be able to answer the questions that I Plan to answer with my data, I do believe I need to tweak my questions because I did a pretty thorough run through of my project's data and I realize that I could just answer some pretty general but import questions with the data that I have already down. I know I need to just go back in and insert the answers to my questions.

What visualizations are especially useful for explaining my data?

The main visualizations that I need to use is clustering charts for sure. I think I am going to keep it real simple instead of using fancy charts that a viewer may have never seen that may cause confusion.

Do I need to adjust the data and/or driving questions?

In order to thoroughly get the most out of my data, yes I will tweak the driving question , just a little bit or I might change my main question and keep those as like filler questions.

Do I need to adjust my model/evaluation choices?

I do not need to adjust my model or evaluation choices.

Are my original expectations still reasonable?

Yes, my original expectations are still reasonable.

Nekeitrius Davis

March 1, 2023

DSC630 - T302 Predictive Analytics

Data Information:

These are the columns of the dataset so that it makes sense about what numbers mean what in the beginning.

- Age: 13-level age category: (1 = 18-24 / 2 = 25-29 / 3 = 30-34 / 4 = 35-39 / 5 = 40-44 / 6 = 45-49 / 7 = 50-54 / 8 = 55-59 / 9 = 60-64 / 10 = 65-69 / 11 = 70-74 / 12 = 75-79 / 13 = 80 or older)
- Sex: patient's gender (1: male; 0: female)
- HighChol: 0 = no high cholesterol 1 = high cholesterol
- CholCheck: 0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years
- BMI: Body Mass Index
- Smoker: Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes
- HeartDiseaseorAttack: coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes
- PhysActivity: physical activity in past 30 days - not including job 0 = no 1 = yes
- Fruits: Consume Fruit 1 or more times per day 0 = no 1 = yes
- Veggies: Consume Vegetables 1 or more times per day 0 = no 1 = yes
- HvyAlcoholConsump: (adult men >=14 drinks per week and adult women>=7 drinks per week) 0 = no 1 = yes
- GenHlth: Would you say that in general your health is: scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor
- MentHlth: days of poor mental health scale 1-30 days
- PhysHlth: physical illness or injury days in past 30 days scale 1-30
- DiffWalk: Do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes
- Stroke: you ever had a stroke. 0 = no, 1 = yes
- HighBP: 0 = no high, BP 1 = high BP
- Hypertension: 0 = no hypertension, 1 = hypertension

Nekeitrius Davis

March 1, 2023

DSC630 - T302 Predictive Analytics

- Stroke: 0 = no, 1 = yes
- Diabetes: 0 = no diabetes, 1 = diabetes

Milestone 4:

Data Preparation:

Data Cleaning

- When originally checking for missing values using "df.isnull().sum()" I found that none of my columns were missing data which was really good. I did however realize that a lot of the columns had zeros instead of leaving them blank:

```
Age                         0
Sex                     38386
HighChol                33529
CholCheck                1749
BMI                         0
Smoker                  37094
HeartDiseaseorAttack    60243
PhysActivity            20993
Fruits                  27443
Veggies                 14932
HvyAlcoholConsump       67672
GenHlth                     0
MentHlth                48091
PhysHlth                39915
DiffWalk                52826
Diabetes                35346
Hypertension            30860
Stroke                  66297
        dtype: int64
```

  So, I had to go back in with the "replace" function and replace those zero values with their means for each individual column such as :

```
df['CholCheck']= df['CholCheck'].replace(0,df['CholCheck'].mean()).
```

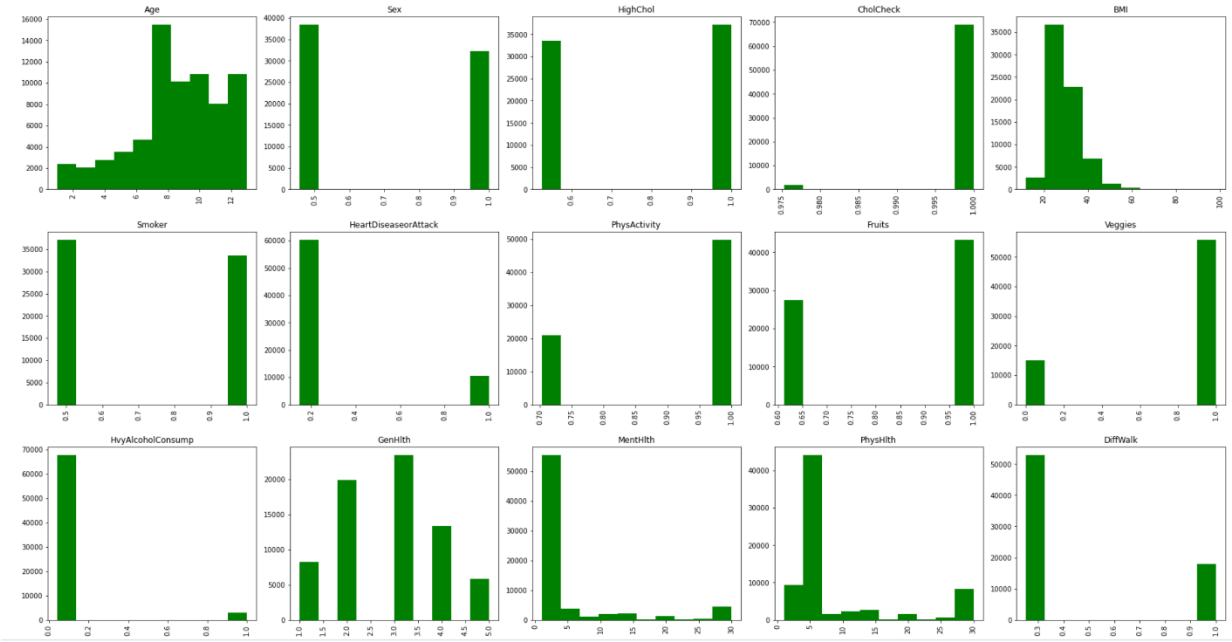
  - Update 8/1 : Originally I thought that was a problem and replaced everything with the mean values but I realized that my actual dateset values for "No" = 0 which is why I had so many values equal to zero so I did do away with this part of the code. I write this because I want to show the transparency and willingness to go back and look at the data for insufficient information.

- Using the "matplotlib.pyplot" library , I was able to create a loop that checked the frequency of all values for each individual column which help me get an idea of what type of balances I was looking at.

Nekeitrius Davis

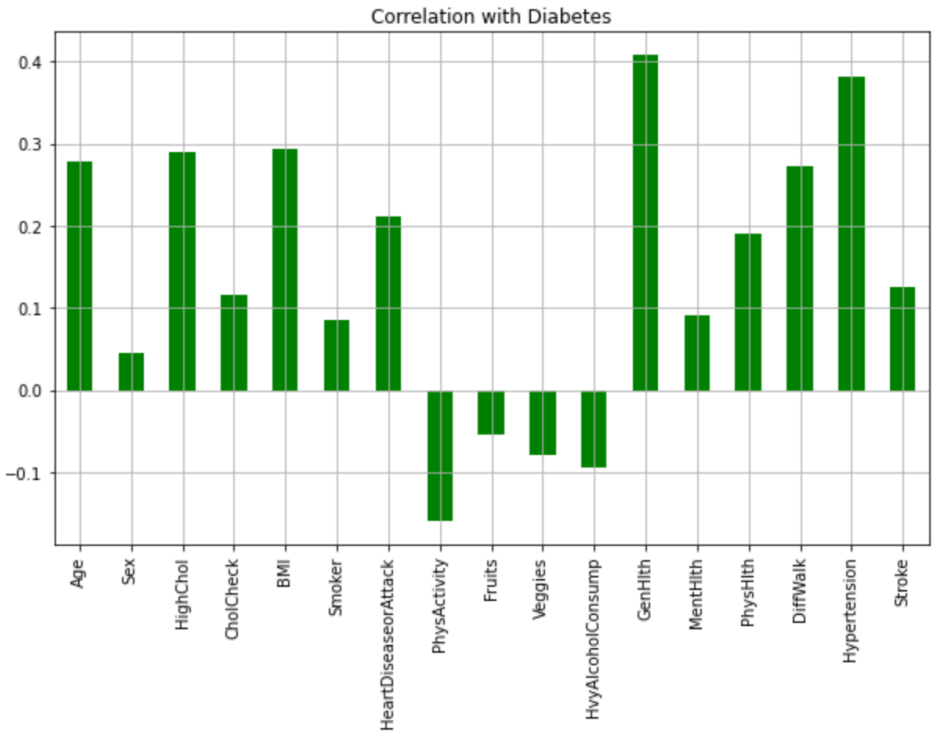
March 1, 2023

DSC630 - T302 Predictive Analytics



- Quick glances observations:
  - Age: The most common age is between 50-54 with the least common age is between 30-34 for this particular study. Also, after 50-54 it is still pretty high in number of participants, I just would assume maybe younger individuals are not really likely to sit through a cell phone questionnaire with this many questions. Also, diabetes are probably not a serious health concern for the younger community as much as it should be.
  - Sex: More females than males but by a couple thousands which isn't that much.
  - High Cholesterol: Most have high cholesterol but its almost as equal to those without.
  - Cholesterol Check: Nearly every has had a cholesterol check within the last five years.
  - BMI: Average is between 24-29 which order to the CDC "If your BMI is **18.5 to 24.9**, it falls within the Healthy Weight range. If your BMI is 25.0 to 29.9, it falls within the overweight range".
  - Smoker: More nonsmokers than smokers but a little.
  - Heart Disease or Attack: Around 75% have never experienced either.
  - Physical Activity: Most have been active in the past 30 days.
  - Fruit: More people have consumed at least one fruit per day but those who haven't are pretty close by. With my own assumption I know that some fruits have natural sugar so it could cause spikes in blood sugars and also that natural sugar might could play a part in diabetes.

  - Veggies: Almost all have consumed at least one vegetable per day.
  - Heavy Alcohol Consumption : Nearly all do not drink over 14 drinks (men) and 7 drinks(women) , that does not mean they do not drink, just that they do not drink a lot.
  - General Health: Visually it is a bell-shaped curve with good being the middle of the curve with the most answers and very good being the next one to follow.
  - Mental Health: Nearly all said they have not experienced any poor mental health days in the last 30 days.
  - Physical Health : Nearly all haven't experienced any physical illness or injuries in the last 30 days.
  - Difficult Walking: Most have no problems with walking or climbing stairs.
  - Stroke: Most never had a stroke.
  - High Blood Pressure: More have high blood pressure.
- I checked the correlation of every column with Diabetes:



  - General Health seems to have the highest correlation alone which is interesting considering this is one of the columns that is based off of their own opinion of themselves. The next to follow is Hypertension which is correlated to the chances of someone having diabetes.

Nekeitrius Davis

March 1, 2023

DSC630 - T302 Predictive Analytics

    - Update: 8/1 - I did drop all the data that was below .1 in correlation in hopes that it would improve accuracy but they all stayed the same so it does show that it is something with the dataset and not my model fitting.

- Splitting data into train and test sets
    - I dropped the “diabetes” column from the data set and set that as my feature matrix while setting “Diabetes” as the target matrix.
    - Using “Label Encoder” I changed the continuous values of Diabetes to categorical.
    - I then split my dataset into training and test sets.
    - The most imperative part to this was creating pipelines and a dictionary for Logistic Regression, KNeighbors, StandardScalers, Decision Trees, Random Forest & Gradient Boosting to see which one would give the best accuracy.

Nekeitrius Davis

March 1, 2023

DSC630 - T302 Predictive Analytics

```
# Create pipelines
# Standard scaler to measure everything on the same level
pipeline_logreg = Pipeline([('scaler1', StandardScaler()),
                            ('logreg_classifier', LogisticRegression())])
pipeline_knn = Pipeline([('scaler2', StandardScaler()),
                         ('knn_classifier',KNeighborsClassifier())])
pipeline_svc = Pipeline([('scaler3', StandardScaler()),
                        ('svc_classifier',SVC())])
pipeline_dt = Pipeline([('dt_classifier',DecisionTreeClassifier())])
pipeline_rf = Pipeline([('rf_classifier',RandomForestClassifier(max_depth=3))])
pipeline_gbc = Pipeline([('gbc_classifier',GradientBoostingClassifier())])
```

```
pipelines = [pipeline_logreg,
             pipeline_knn,
             pipeline_svc,
             pipeline_dt,
             pipeline_rf,
             pipeline_gbc]
```

```
#Train pipelines
for pipe in pipelines:
    pipe.fit(X_train,y_train)
```

```
#create pipeline dictionary
pipe_dict = {0:'LR',
             1:'KNN',
             2:'SVC',
             3:'DT',
             4:'RF',
             5:'GBC'}
```

- The accuracies were:
  - LR Test Accuracy:74.63045477049296
  - KNN Test Accuracy:71.32046113586533
  - SVC Test Accuracy:74.53851050286443
  - DT Test Accuracy:65.74722399038122
  - RF Test Accuracy:73.63321309852182
  - GBC Test Accuracy:75.23162882806422
- I then created a prediction data with each variable and values that I came up with :

Nekeitrius Davis

March 1, 2023

DSC630 - T302 Predictive Analytics

```
#Testing prediction data
pred_data = pd.DataFrame({
    'Age':3, # 32
    'Sex':0, # Female
    'HighChol': 0 ,# No high cholesterol
    'CholCheck':0, # No cholesterol check in 5 years
    'BMI': 24, # BMI of 24
    'Smoker': 1, # yes
    'HeartDiseaseorAttack': 0, # No heart disease or attack
    'PhysActivity' : 0, # no in last 30 days
    'Fruits': 0, # no did not consume fruit
    'Veggies' : 1 , # yes consumed veggies
    'HvyAlcoholConsump': 0, # no heavy alcohol consumption
    'GenHlth': 4, # general health is "fair"
    'MentHlth': 20, # 20/30 day bad mental '
    'PhysHlth': 1, # 1 illnesss day in last 30 days
    'DiffWalk': 1, # difficulty walking/climbing stairs
    'Hypertension': 0, # no hypertension
    'Stroke': 0 # no
}, index = [0])
```

Results & Conclusion:

I did build a model and used it to predict if diabetes can be determined based on those models and with the use of Random Forest and Standard Scaler I can see that my results are the same for both which was non-diabetic so that makes me believe that my model did in fact work. From the beginning I did in fact plan to use KNN but when I tested my models accuracy it was on the lower end out of the six I tested. It wasn't a major tweaked considering that I accomplished my goal. I did not run into many problems during this project, and I feel like it added to my strengths as a data scientist because this is one step further than what I have been using these past couple of semesters.

References:

1. Frengzkermova, K. (2022, December 3). *Diabetes prediction*. Kaggle. Retrieved January 26, 2023, from https://www.kaggle.com/code/frengzkermova/diabetes-prediction
2. Center for Disease Control . (2016, August 23). *Behavioral Risk Factor Surveillance System*. Retrieved January 27, 2023, from https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf

Nekeitrius Davis

March 1, 2023

DSC630 - T302 Predictive Analytics

3. Centers for Disease Control and Prevention. (2022, June 3). *Assessing your weight*. Centers for Disease Control and Prevention. Retrieved January 26, 2023, from https://www.cdc.gov/healthyweight/assessing/index.html
4. Centers for Disease Control and Prevention. (2022, September 30). *Diabetes Quick Facts*. Centers for Disease Control and Prevention. Retrieved February 1, 2023, from https://www.cdc.gov/diabetes/basics/quick-facts.html
5. Centers for Disease Control and Prevention. (2022, September 30). *Risk factors for diabetes-related complications*. Centers for Disease Control and Prevention. Retrieved February 1, 2023, from https://www.cdc.gov/diabetes/data/statistics-report/risks-complications.html