INFO 648

Data Science Assignment

Allison Schneider

November 6, 2022

*1. Dataset Retrieval*

This report serves as an overview of my data science assignment. For this assignment, I have chosen a diabetes health indicators dataset that I found on Kaggle.com. Diabetes is a serious chronic disease in which individuals lose the ability to effectively regulate levels of glucose in the blood and can lead to reduced quality of life and life expectancy. The table below outlines the columns in this dataset. This dataset was created by The Behavioral Risk Factor Surveillance System (BRFSS), which is a health-related telephone survey that is collected annually by the Centers for Disease Control (CDC) and was cleaned by a Kaggle.com user. The dataset represents the 2015 BRFSS survey results and contains responses from 253,680 individuals.

This dataset may be used to build a machine learning model that predicts if a patient has diabetes. Additionally, this dataset may be used to understand what other conditions patients have concurrently with diabetes. This dataset may also answer questions around the risk of getting diabetes. For example, does smoking increase the risk of diabetes? Or do highly active people have a lower risk of having diabetes?

|  |  |
| --- | --- |
| ***Column Name*** | ***Column Description*** |
| Diabetes\_012 | 0 = no diabetes  1 = prediabetes  2 = diabetes |
| HighBP | 0 = no high blood pressure  1 = high blood pressure |
| 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  0 = no  1 = yes |
| Stroke | Have you ever been told you have had a stroke?  0 = no  1 = yes |
| HeartDiseaseorAttack | Coronary heart disease (CHD) or myocardial infarction (MI)  0 = no  1 = yes |
| PhysActivity | Physical activity in the past 30 days – no including job  0 = no  1 = yes |
| Fruits | Consume Fruit one or more times per day  0 = no  1 = yes |
| Veggies | Consume vegetables one or more times per day  0 = no  1 = yes |
| HvyAlcoholConsump | Heavy drinkers (adult men having more than 14 drinks per week, and adult women having more than 7 drinks per week)  0 = no  1 = yes |
| AnyHealthcare | Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc.  0 = no  1 = yes |
| NoDocbcCost | Was there a time in the past 12 months when you needed to see a doctor but could not because of cost?  0 = no  1 = yes |
| GenHlth | Would you say that in general your health is:  1 = excellent  2 = very good  3 = good  4 = fair  5 = poor |
| MentHlth | Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?  Scale 1 – 30 days |
| PhysHlth | Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?  Scale 1 – 30 days |
| DiffWalk | Do you have serious difficulty walking or climbing stairs?  0 = no  1 = yes |
| Sex | 0 = female  1 = male |
| Age | 13-level age category  1 = 18-24  9 = 60 – 64  13 = 80 or older |
| Education | Education level  1 = never attended school or only kindergarten  2 = grades 1 through 8 (elementary)  3 = grades 9 through 11 (some high school)  4 = grade 12 or GED (high school graduate)  5 = college 1 year to 3 years (some college or technical school)  6 = college 4 years or more (college graduate) |
| Income | Income scale  1 = less than $10,000  5 = less than $35,000  8 = $75,000 or more |

*2. Data Exploration*

In exploring this dataset, I have learned that each of the columns have a datatype of float64 and are continuous data. A datatype of float64 means that the data is a numerical value inclusive of decimal points. You will note from the column definitions above, that the numbers used in the columns do not include decimal points, which means that this data can be stored as the datatype integer, however it is currently stored as a float64 which does allow for decimal places if needed.

Additionally, I have determined that this is a relatively clean dataset with no null (i.e., missing) values. As discussed above, the shape of this dataset is (253680, 22) which means that this dataset includes 253,680 rows and 22 columns of data. Of the 253,680 rows, 213,703, or 84.2%, are patients that do not have diabetes. This leaves only 39,977 rows, or 15.8% of the records, represent patients with diabetes. This can be an issue of an unbalanced dataset when it comes to training machine learning models. In the example of predicting if a patient has diabetes, we would like to see a more balanced dataset that represents more patients with diabetes for training purposes.

Finally, in exploring this dataset, we have identified the columns that are positively and negatively correlated with diabetes. We have found that GenHlth is the most positively correlated feature of diabetes at 0.3, which is a pretty low correlation factor. Additionally, we have found that income is the most negatively correlated feature of diabetes with an extremely low correlation value of -0.17.

*3. Experience and Queries*

The first question I would like to answer using this dataset is: “*Are patients with high blood pressure AND high cholesterol more likely to also have diabetes?”*. The features that will help me in answering this question include: HighBP, HighChol, and Diabetes\_012. As shown in the screenshot below, I have written a query to determine the percent of patients who have both high blood pressure and high cholesterol that also have diabetes vs those patients with both high blood pressure and high cholesterol who do not have diabetes. I have found that 32.84% of high blood pressure and high cholesterol patients also have diabetes, while 67.16% of these patients do not have diabetes.

Text

Description automatically generated with low confidence

The second question I would like to answer using this dataset is: “*Does diabetes affect more men or women?”*. The features that will help me in answering this question include Sex and Diabetes\_012. As shown in the screenshot below, I have written a query to understand the percentage of diabetes patients that are female vs male. I have found that diabetes is slightly more common in females with 52.57% of diabetes patients being female.

Text

Description automatically generated

The third question I would like to answer using this dataset is: “*Does diabetes affect more patients with poor health vs good health?”*. The features that will help me in answering this question include: PhysActivity, HvyAlcoholConsump, GenHlth, PhysHlth, DiffWalk, and Diabetes\_012. As shown in the screenshot below, I have written a query to determine the percent of patients with poor health that high diabetes, vs the percent of patients with good health that have diabetes. I have found that 23.53% of patients with poor health also have diabetes, while only 9.4% of patients with good health have diabetes.

Text

Description automatically generated

The fourth and final question I would like to answer using this dataset is: “*Do patients with lower incomes have a higher chance of diabetes?”*. The features that will help me in answering this question include: AnyHealthcare, NoDocbcCost, Income, Education, and Diabetes\_012. As shown in the screenshot below, I have written a query to determine the percent of low-income patients that have diabetes, vs the percent of high-income patients that have diabetes. I have found that diabetes is more common in low-income patients than in high income patients, with 19.27% of low-income patients having diabetes, while 8.78% of high-income patients have diabetes.

Text

Description automatically generated

*4. Software Program to Analyze Data*

In a real-world scenario, I would use Python or R to analyze this diabetes health indicators dataset. You can see from the screenshots above, that I have written my queries to better understand the data in Python. In analyzing this dataset, I would use Python to implement a machine learning regression model to predict if a patient has diabetes or not.

I would first use Python to clean the dataset, however, as we noted above not much cleaning is necessary for this dataset. Our features are already identified in a numeric or continuous datatypes, and there are no missing or null values. However, the dataset is unbalanced with only 16% of the records in this dataset representing patients with diabetes. To build a strong training model, I would want to balance the dataset so that there are more records of diabetes cases. This will allow for a better training model which will in turn produce more accurate prediction of diabetes.

Once I have the dataset cleaned and balanced, I would split the dataset into a training set and a testing set. I would then implement the regression model of my choosing onto the training set and test the accuracy of the results using the testing set.

The regression model will also give me insights into the variables that are most significant in predicting diabetes. As we saw in our exploratory data analysis, general health, high blood pressure and BMI have the highest correlation with diabetes. Does this mean that these variables are significant in predicting diabetes? Iterations of a regression model will be able to answer this question for us.

*5. Summary*

In my experience with this diabetes health indicators dataset, I have learned a lot about diabetes patients. I have learned that patients with high blood pressure and high cholesterol have a higher chance of also having diabetes then patients identified as having poor health based on physical activity, alcohol consumption, general health, physical health, and difficulty walking. I have also learned that diabetes occurs in more women than men, slightly. Additionally, I have learned that even patients with good health (identified based on physical activity, alcohol consumption, general health, physical health, and difficulty walking) have a risk of having diabetes, however the risk is much lower than those with poor health. Finally, I have learned that diabetes is more prevalent in low-income patients than in high-income patients which are identified by healthcare coverage, the need to see a doctor but didn’t because of cost, income greater the $75,000, and a college education.

The insights that I have been able to quickly learn about diabetes patients from this dataset could be extremely insightful to a clinician. Having the ability to understand how demographics and other health indicators affects a patient’s risk of a specific illness or disease could help a clinician to narrow down possible diagnoses for a patient with unclear symptoms. Additionally, data science can tremendously help public health. If clinicians are able to run machine learning models on health datasets, they may be able to more accurately predict if a patient may get a specific illness or disease. Machine learning models will also provide clinicians with significant variables in predicting specific illnesses or diseases which will be beneficial to preventative care, because it will show the importance of other health indicators such as high blood pressure, high cholesterol, physical activity, etc.

*6. Data Science Course*

I suggest that health informaticians/informaticists take the School of Data course that is referenced in the Hoyt & Muenchen Data Science Resources. I suggest this course because it covers multiple topics in data exploration including data fundamentals, data cleaning, exploring data, extracting data, and mapping data. I think that having these skills as a health informatician/informaticist is important because it provides an understanding of the information and knowledge we can gain from data. It also shows how data is stored and why standardization in health informatics is so important. If health informaticians are able to at least explore data and understand how it is stored and accessed then interoperability in health informatics may be easier to achieve, and the importance of data in a healthcare setting will be better understood.