Exploratory Data Analysis for Diabetes Dataset Write Up

Introduction

According to Kaggle, this public domain dataset is from the National Institute of Diabetes and Digestive and Kidney Disease. The dataset was created to predict if a patient has diabetes based on their diagnostic measurements. There are 768 women in the sample, with a total of 8 attributes that can be examined for their relation to diabetes. The features were selected from a more extensive database and constrained to females of at least 21 years old of Pima Indian heritage. The features include the number of pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, BMI (body mass index), diabetes pedigree function (influence of genetic inheritance and family history), and age. The outcome variable is 0 (no diabetes) or 1 (yes diabetes).

Research Questions

The overarching research question was what main diagnostic factors predict whether adult women will develop diabetes? Possible relationships between diagnostic factors as potential predictors can aid the prevention of diabetes.

I began with a preliminary research question regarding BMI. BMI is a health indicator. Which diagnostic factors (pregnancies, glucose, blood pressure, skin thickness, insulin, diabetes pedigree function, age) affect BMI in adult women? To what degree do the diagnostic factors affect BMI? Does skin thickness affect BMI? I found mid-way through my data analysis that there is very little correlation between age and diagnostic attributes. Thus I pivoted to a more exploratory data analysis method by looking at relationships between the attributes.

Getting to Know Dataset

I began my study by first familiarizing myself with the dataset. I imported the csv file. Then looked through the columns and characteristics. They were all numerical attributes. There were no string attributes. However, the Outcome is numerical, though, really, it is categorical. I noticed there were 0 values or null values as well. The data is also a bit unbalanced for Outcome. I looked at the min-max values and tried to identify outliers based on how the attribute was measured. For example, blook glucose levels range from 70mg/dL to 126 mg/dl. I kept the column names as they made sense and were self-explanatory. The DiabetesPedigreeFunction was a bit confusing though I learned it was a measurement of the genetic factor for diabetes. Overall, there were no outliers. The range of values was within expected real-world measurements.

Data Cleaning

I updated all the zero values to NaN values so they would not affect the data analysis. I then replaced them with mean values per category. The means changed after the null values were altered. I also checked for duplicate rows. There were none. I used a heatmap to find missing values as well. At first, I dropped the outcome column as I was interested in investigating the numerical values. I changed my mind mid-way through the analysis as I realized I could test the correlation between the Outcome attribute with other attributes, so I added the column back in. Overall, the data was clean. I mainly reformatted the data based on the analysis I was conducting. For example, ANOVA required a categorical variable, which I changed BMI into. I referred back to the data descriptions repeatedly while doing data analysis. They helped me familiarize myself with the data.

Resources

Diabetes Data Set: https://www.kaggle.com/datasets/mathchi/diabetes-data-set
https://my.clevelandclinic.org/health/diagnostics/12363-blood-glucose-test
https://www.healthline.com/health/high-blood-pressure-hypertension/blood-pressure-reading-explained#danger-zone
https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.drop.html
https://seaborn.pydata.org/generated/seaborn.lmplot.html
Hadley, Wickham. "Tidy data." *The Journal of Statistical Software* 59, no. 1 (2014).
https://datagy.io/python-pearson-correlation/
https://www.machinelearningplus.com/plots/python-scatter-plot/
https://pythonexamples.org/pandas-dataframe-delete-column/
https://pythontic.com/pandas/dataframe-manipulations/sample%20selection
https://datatofish.com/dataframe-to-numpy-array/
https://www.statology.org/two-sample-t-test-python/
https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html
https://seaborn.pydata.org/generated/seaborn.kdeplot.html
https://www.aaps.ca/blog/pros-with-a-diploma-in-nutrition-know-bmi-findings-may-be-inaccurate