CMSE 830 Midterm Project: PCOS Diagnosis EDA

Angelica Louise Gacis

2022-10-29

Goal of the project

I chose this topic because besides being one of the main causes of women's infertility, women with PCOS also have increased risks for type 2 diabetes and heart diseases which are some of the most common causes of medical-related death. The exact parameters that cause PCOS are not yet identified, but since it is a hormonal disorder, there are some features that are suspected to help distinguish the presence or absence of the disorder which can cause several/daily inconveniences to anyone might have it.

What you can learn from this app

From this app, you can learn the features of the data set that I was able to retrieve from Kaggle. You can see the head and summary of the data at the topmost part below the title. Next, you will find the univariate analysis where I split it into two: categorical variables and numerical variables. For each, I used the 'PCOS diagnosis' variable for comparing. Below that is the bivariate analysis between two numerical variables with a hue which you can select from the categorical variables. At the end is the correlation matrix where you can check which variables are correlated to one another. Then, you can go back to the univariate and bivariate plots to check the more detailed trend between the correlated pairs. I did this in preparation of a possible classification model for the next project.

Visualizations

I used plotly for all of my plots because they are interactive unlike seaborn plots yet simpler unlike altair plots. For the univariate categorical, I used a dn unstacked or dodged histogram plot. I used this so that the user can see the imbalance in the data set among different variables especially with our main variable which is 'PCOS diagnosis'. For the univariate numerical, I used plotly violin plots with box plots inside to combine the power of the two plots. This is also the plot which will show you the outliers and whether the data should be transformed or not. I removed the background of the plots for a cleaner view and mainly used standard primary colors with blues and reds, so that the colors will not be distracting.

Preprocessing

I had two or three missing values overall from three different variables and used the median to replace them instead of deleting them to retain other information. I used the median so that they won't be affected by the outliers. I also combined a two variables. For example, the number of follicles originally had two variables, one for the right ovary, and another for the left. However, since these two are highly correlated and are equal most of the time, I created a new variable which is represents the average follicle number per ovary. I also did this for follicle size. I also removed some redundant variables such as BMI which is redundant with the height and weight variables. Lastly, I changed the binary variables from 0 and 1, for example to, no and yes respectively for the plots except for the correlation matrix.

Additional features

I made versions of the data set with removed outliers, standardized values, and both. I made a multiselectbox on the sidebar where the user can choose how to transform the data based on what they think is appropriate. Users can also filter the data using different variables and values in the sidebar. Lastly, I added a slider for the correlation matrix wherein the user can choose the range of coefficient magnitude they want to see regardless of positive or negative signs. This will result in a resized heatmap with only the variables producing a correlation coefficient in the specified range remaining. The purpose of this is because I have around 35 variables and bigger heatmaps are harder to read, when in reality we mostly just want to extract the variables with stronger relationships especially if this project will lead to a classification model. All these additional features in the web app were made to give more freedom and tools for the user to explore the data considering that different people may discover different things when exploring.